Aug 18, 2023

Prof. Tat Chan
Senior Editor
Marketing Science

Dear Professor Chan

Thank you very much for offering us the opportunity to revise our paper (**MKSC-22-0323**). We greatly appreciate the positive evaluations and constructive feedback from all members of the review team. We have used their feedback to extensively revise and improve the paper.

We first provide an overall summary of the major changes in this new submission. We then provide specific point-by-point responses to each member of the review team. We hope you will retain the current review team,

We trust and expect that this new submission addresses all of the concerns in our previous submission. We thank you, the AE, and the reviewers for your detailed and constructive feedback, which has helped us produce a paper with a much clearer and stronger contribution.

Best Regards,
The Authors

Contents

1 Summary of Major Changes

We provide below an overview of the major changes made to the paper along with pointers to specific sections and results in the revised paper.

1. **Model Performance and Benchmarking:** All members of the review team raised concerns about the performance of the model and provided us various suggestions to improve the accuracy of the emotion prediction. We appreciate that various members of the review team also noted that while they want us to improve the predictive accuracy, the goal was not to simply outperform benchmarks, as our goal was to create a theory-driven explainable deep learning model for predicting music emotion. In this revision, we seek to balance these goals; we incorporate almost all of the review team's thoughtful suggestions to improve accuracy by estimating a large number of new models, while preserving our primary goal of incorporating theory-guided structure in designing the deep learning filters.

Table 1 details the list of model changes we have made or assessed to improve accuracy. We have examined a wide range of models suggested by the review team. We find some of the new model specifications help with performance. In particular, we find that combining our proposed deep learning model with theory-based harmonics filters and handcrafted features improves emotion classification performance. The results for these different model specifications are presented in Table A.1 of this response letter and Appendix B provides details of the models. In the paper, the results can be found in Tables X, X, and X.

Overall, the F1 of the revised models still remain in the 53-55% range. Here is what we say in the paper on p. X for how to think about these absolute levels of predictive performance.

> In interpreting and assessing the results, it is important to consider the specific prediction task involved. Prediction tasks involving *subjective human response* (e.g., emotion recognition, humor detection) typically have lower prediction accuracy than more *objective* recognition tasks (e.g., object identification, instrument identification). This is because for subjective tasks, in the absence of objective ground truth, humans have heterogeneous responses implying less agreement on the "ground-truth." In such cases, even our best approximation is bounded above by this subjectivity, as with music emotion tagging.

Hence we should compare our model's performance only against other state-of-the-art approaches for emotion classification. We also note that even within music emotion classification, we predict the emotion for short music clips rather than full music clips (given our business application), which results in less information for classification, making ours a more challenging emotion classification task.

2. **Data:** In the revised paper, we train and test our model on a new dataset that has more observations with more precise labeling and that is also more representative of the valence-arousal space. We combine the Soundtracks dataset, which is representative of the music in our business application, with the DEAM dataset, which is labeled over time and therefore more precise. Details on the dataset can be found on p. XX of the paper.

(i) *More accurately labeled dataset:* The Soundtracks and 4Q datasets that we had previously used were labeled for the entire song and not at the six-second clip level. While the Soundtracks music clips were selected to evoke a single emotion over the length of the clips, this was not true of the 4Q music clips, resulting in measurement error for the emotion labels. To address this issue, we now use a dataset (DEAM) where the labels are made at a more granular level and so

Table 1: Models – Existing and New

No.	Model Element	Rationale
Retained from Original Submission		
O1	MFCC	
O2	CNN with Square Filters	
O3	CNN with Time Filters	
O4	CNN with Frequency Filters	
O5	CNN with Freq-Time Filters	
O6	CNN with Harmonics Filters	
New Models in the Revision		
N1	STFT	Using STFT spectrograms as input instead of mel to evaluate whether mel is required and/or helpful
N2	CNN - Rectangular Filters	Rectangular filter treats time and frequency dimensions independently (spectrograms are different from regular images)
N3	CNN - Harmonics + Tempo Feature	Music and emotion literature suggests tempo helps predict emotion
N4	CNN - Harmonics + Tempogram	Tempo can be alternatively represented as a tempogram, similar to a spectrogram; CNN can include both tempogram and spectrogram as input
N5	Top Handcrafted Features	Handcrafted features improved accuracy in predicting music emotion using the 4Q dataset in Panda et al. (2018); the features include tone color/timbre, suggested by the review team
N6	CNN - Harmonics + Handcrafted Features	We evaluate CNN with Harmonics Filters + Top Handcrafted Features and with MFCCs to improve accuracy but retain explainability
N7	CNN with RNN	RNN architecture in the final layer could capture positional or temporal information better than CNN alone

the emotion labels are more precise (Aljanaki et al. 2017). We combine the Soundtracks and DEAM datasets since they are more precisely labeled.

(ii) *More observations:* Our training data includes 2,176 six-second music clips. This is intended to address the review team's requests to increase the number of observations in the dataset. We recognize that this is still not a huge dataset. But large scale labeling of datasets at the six-second level by multiple human coders is quite expensive—and is usually done by computer scientists with dedicated grants for data creation. Developing these datasets is often viewed as a separate contribution in its own right. Even with the current dataset, we note that *our accuracy levels are comparable or better* with respect to state-of-the-art benchmarks for music emotion classification.

(iii) *Dataset is more representative of full span of emotion space:* Finally, the Soundtracks and 4Q datasets we used in our previous submission had songs/clips designed to be *exemplars* for the various emotion quadrants, i.e., the songs in 4Q were chosen based on high emotional content across the valence and arousal quadrants. This selection may leave gaps in the valence-arousal space, which could limit the broader generalizability of the model to music less high in emotional content. The DEAM dataset collection process instead begins with a pool of royalty-free songs and obtains emotion tags for these songs, making the data more representative of the overall emotion space. The DEAM dataset therefore nicely complements the Soundtracks dataset in terms of covering the valence-arousal emotion space.

3. **The Ad Insertion Application:** Our method focuses on human emotional response to music, and in the application we examine how emotional congruence versus contrast of video ads and content impacts ad skip and brand recall. We have now focused on making the empirical application more generalizable by (i) including more content videos and ads, and (ii) by demonstrating how to incorporate emotion information from multiple modalities. We provide more details in Appendix E of this response letter. In summary, we increase the scale and scope of the application as follows:

(i) *Use more content videos and ads in the experiment:* The review team had asked us to increase the number of ads and content videos in order to improve generalizability. We have expanded the number of ads (from 2 to 4) and the number of content videos (from 1 to 4), increasing the number of experimental cells from 12 to 96. While this helps to demonstrate generalizability, we recognize this is still a limited set. We explain the cost constraints that requires us to constrain the set of ad and content videos in Appendix E of this response letter.

(ii) *Include emotion features from other modalities*: We agree that demonstrating how to include emotion information from multiple modalities can increase the flexibility of our approach and hence its value to practitioners. The review team suggested that other modalities (e.g., video images, speech) might convey emotion information and that practitioners might want to understand how to account for such information. Please see Appendix E.2 of the paper for details.

We have extensively rewritten the application portion of the paper to provide more detail and clarify the steps we took to determine the optimal emotion-based ad insertion points using the various models.

4. **Exposition:** We have clarified and simplified the paper's exposition along several dimensions to make the positioning and contribution clearer to the journal's readers.

(i) *Contribution and Positioning around Theory-based Deep Learning Models:* Beyond the predictive accuracy of the model and the managerial application toward emotion-based ad insertion, we have now further highlighted an important methodological goal which was perhaps not sufficiently emphasized in the last submission. Our goal was to illustrate and demonstrate how using theory to inform filter construction in deep learning models can lead to more efficient and explainable models. From the review team's feedback, we recognize that this aspect of the contribution had not been sufficiently salient in our prior submission. We believe this is an important goal, and we hope that this approach to thinking about deep learning can have broader impact in developing applied and explainable deep learning models.

(ii) *Incorporating Additional Features while Retaining Explainability:* Per the review team's suggestion, we incorporate handcrafted features into the model and show how doing so can be achieved without sacrificing explainability, a primary goal of our method. We find an incremental improvement in model performance by including handcrafted features.

(iii) *Theory-based Filters Capture Features Beyond Consonance:* The filters impose structure that enables Grad-CAM visualization of consonance but the filters do not restrict the model to learn only consonance. There is enough flexibility in the model to learn other musical features that impact emotion, like whether high harmonics or low harmonics are present, loudness, and pitch. It is this flexibility that enables the model to classify music emotion accurately.

While these form the major changes in the revision, we expand on these points (and others) in our separate responses to the SE, AE, Reviewer 1, and Reviewer 2 below.

2 Response to the Senior Editor

The review team likes the general idea and the methodology you have presented. However, there are major revisions that would need to be done if the paper is going to be accepted. Given the major extent of these needed revisions, there is considerable uncertainty as to the final outcome. I invite you to respond to the review team's comments and revise your manuscript.

After reading your paper and the reports from the review team, I think your revision effort has to focus on addressing the contribution to the marketing literature, modeling, performance comparison, explainability, and the empirical application. As all of these issues have been discussed in detail in reviewers' reports, and have been nicely summarized by the AE's report, I will not further elaborate here. You should use the AE report as the main guidance for the revision. Below, I will offer a few general comments which may help your revision effort:

Reply: Thank you very much for giving us the opportunity to revise the paper. We appreciate the detailed and constructive feedback of the review team.

Overall, as we outlined in our summary of the major changes, we have addressed each of the issues raised. We provide detailed responses to each member of the review team. As appropriate, we have revised the paper and included additional appendices with details either in the main paper or in the reviewer response. We hope the detailed responses address all of the concerns of the review team.

SE 1:

1. Model: There are many good suggestions made by reviewers, but addressing all of them (e.g. including tone color, tremolo, and melody as R2 mentioned) could be challenging. I am willing to accept your work as the first step to study how music evokes emotions. However, you should better justify why you focus on the music features such as consonance and dissonance in this paper. Based on my (very limited) knowledge of music, consonance and dissonance are important but there are other important features (e.g. sequence as suggested by R1 and tempo as suggested by R2) that represent music as well. Why are consonance and dissonance so unique in terms of affecting emotions?

Reply:

We appreciate the various suggestions by the reviewers to add additional musical features that may potentially improve accuracy. We also appreciate your advice (in the next point) that we should keep the focus of our paper's contribution around explainability. As suggested, we estimated many new models and explored the inclusion of additional features while trying to retain explainability. Please see Table 1 of this response letter for details.

Why focus on consonance and dissonance? We should start by saying that the filters are designed with consonance in mind but do not restrict the model to only learn consonance as detailed on p. XX of the paper. In short, the filters enable the visualization of consonance for explainability while providing enough flexibility to learn other music features related to emotion. We design the convolution filters with consonance in mind for three main reasons. First, consonance is one of the few features that has a strong relationship with valence. Models that predict music emotion have historically been more successful at classifying arousal and less successful at classifying valence. The creators of the DEAM dataset ran a competition for music emotion prediction and after analyzing the submissions wrote, "The algorithms show very good performance on arousal and completely unsatisfactory performance on valence. It is a known issue, that valence is much more difficult to model than arousal, but not to the extent that we observe" (Aljanaki et al. 2017). We therefore focused on a musical feature known to have a strong relationship with valence (Gabrielsson and Lindström 2010, Gabrielsson 2016).[1]

Second, using the language of Fu et al. (2010), consonance is a mid-level feature that has a clear relationship with the high-level construct of emotion but also has a well-studied relationship with low-level frequencies that is based on the physics of sound.[2] It is this connection that enables explainability. To draw a parallel to image recognition, as humans we cannot understand the relationship between individual pixels to the image label but we can understand the relationship between clusters of pixels (e.g., sharp teeth, yellow eyes) to the image label (e.g., shark, wolf).

Third, the structure underlying consonance (i.e., overlapping harmonics) still leaves a lot of flexibility for other musical features to be learned. For each pitch class filter, we restrict the frequencies seen by the model but the model can still learn from these frequencies features related to tone color, loudness, pitch, etc.

Inclusion of other features: Although the model can learn many different music features, it is possible that including handcrafted features can still improve classification performance given the limited training data. We include several other handcrafted music features across new model specifications. First, we include the Top Handcrafted Features based on Panda et al. (2018), which largely connect to tone color/timbre. Second, we include tempo in a few different ways. These model specifications are detailed in Appendix B of this response letter. As shown in Table A.1 of this letter, we find that inclusion of handcrafted features slightly improves classification performance but not dramatically and we believe this is because the model already captures several important musical features that influence human emotional response.

SE 2: *2. Model performance is important, but the real contribution of your paper is about explainability. R2 has a series of questions about how to interpret the diagrams. Please address them carefully.*

Reply: We very much appreciate that you want us to keep our paper's contribution around explainability. We have tried to find a good balance in considering the tradeoff between reviewer's suggestions on model's performance and the need for explainability. We have also added the features in a manner that preserves the explainability of the theory-based deep learning model.

Regarding R2's questions, we have responded directly to the reviewer as recommended.

SE 3: *3. Application: R1's comment on randomizing multiple music clips and ads insertion times is very important for establishing the validity of your findings. Please follow the suggestion in the revision.*

[1] There is a body of evidence that finds that humans respond favorably to consonance, even as infants (Trainor and Heinmiller 1998).

[2] In an interesting test to tease apart the mechanisms by which consonance has impact, research has found that individuals with congenital amusia, a neurogenetic disorder, did not display a preference for consonance, highlighting a primary role for pitch perception, related to harmonics.

Reply: We appreciate R1's concern regarding multiple music clips. We therefore randomize ad insertion across multiple content videos (4) and ads (4) at various positions within the content in a fully factorial design. Overall, our conclusions remain robust to the use of multiple content videos and ads. We detail the cost considerations and constraints that impact the scaling of the application in Appendix E of this response letter.

I consider addressing the above issues as necessary but not sufficient for being accepted for the publication. You should use the AE report as the contract for the revision.

Reply: We greatly appreciate the opportunity to revise the paper. We believe addressing these comments have significantly improved the paper. We trust that our extensive and detailed revision addresses all of the concerns raised by the review team.

3 Response to the Associate Editor

Thank you very much for your positive evaluation of our paper and for the opportunity to revise the paper. We provide our point-by-point responses to your comments below. For ease of reference, we copy your original comments here.

AE 1: *1. Motivation/Contribution*

I agree with both referees that there is a lot to like about this manuscript. The idea of using music theory to develop MusciEcoCNN is novel. The ad insertion exercise also shows some promise for business application of this proposed CNN classifier. In order to strengthen the motivation/contribution to the ad insertion piece, I agree with R2.4 that a literature review on ad insertion in videos is needed in the front end of the manuscript. I would also like to see a discussion on why the authors believe that the proposed method might be superior to prior emotion- or non-emotion-based ad insertion methods that have been used in marketing.

Reply: Thank you for the kind comment about the promise and novelty of the manuscript. We respond to your multiple comments in turn below:

1. Literature Review on Ad Insertion: Thank you for this suggestion. We agree with the value of adding the literature review on ad insertion to the front end of the manuscript, and have added this on p. X.

2. Superiority of Proposed Model/Method for Ad Insertion: We clarify that we do not want to claim that our proposed theory-based explainable model of emotion will be superior to prior emotion and non-emotion based methods for ad insertion in marketing. Emotional targeting using music emotion can also be used in combination with other methods of targeting. We have added a brief discussion as to why we believe emotion-based targeting is valuable to p. XX of the paper. Regarding comparing to other emotion-based targeting methods, we do a comparison between using only music and using music along with images, details of which can be found in Appendix E of this letter. We provide additional discussion in the response to **AE11**.

AE 2: *2. Model: Given that the manuscript is positioned as a method paper, both referees have raised a number of valuable questions on the proposed MusicEmoCNN. In what follows, I will provide a summary of their comments/suggestions.*
2.1. Modifications to MusicEmoCNN: a. Given that the current CNN classifier cannot distinguish between Q1 (exuberance) and Q3 (sadness) in terms of consonance, R2.1a recommends adding tempo to CNN architecture so that these two quadrants can be differentiated. I agree that being able to distinguish Q1 and Q3 would be an important next step in the revision process. This perhaps also relates to the low predictive accuracy issue, which was raised by both referees.

Reply: We agree that it is critical for the model to be able to differentiate between Q1 and Q3. We have provided a more detailed response in our reply to **R2.1** and provide a summary below.

1. Our exposition of this issue in the previous round should have been more clear. Our model can and does indeed distinguish between Q1 and Q3, as can be seen in the confusion matrix for

the CNN with harmonics filters (see Appendix X of the paper). Our statement about Q1/Q3 disambiguation only referred to the Grad-CAM heatmaps and we have extensively revised this section to be more clear.

2. We provide evidence that in addition to learning consonance (as shown by the Grad-CAM heatmaps), the model learns to pick up the presence of high harmonics, which helps to differentiate high arousal emotion (Q1/Q2) from low arousal emotion (Q3/Q4). Thus, in the simplest case we can think that consonance helps to differentiate positive valence from negative valence and the presence of high harmonics helps to differentiate high arousal from low arousal.

3. Following the guidance provided by you and R2, we train a few models that incorporate tempo, details of which can be found in Appendix B of this response letter. We capture tempo using a tempogram and using handcrafted tempo features. We find that when we include handcrafted tempo features, classification performance improves (see Appendix A Table A.1 of this letter).

AE 3: *b. Inclusion of the other possibly relevant features such as tone color, tremolo, mode (major or minor), melody (range and direction) as suggested by R2.1b. I agree with R2 that the additions of these features may potentially improve accuracy and enhance interpretability.*

Reply:

After addressing the issue of tone color etc in R2 response, it would be good to revisit this a little.

We appreciate the review team's suggestion to include additional musically relevant variables, along with the guidance by you and the SE that inclusion of all such possible variables can be challenging and beyond the scope of an initial study.

We provide a detailed reply in our response to **R2.2** and summarize our main points below. We were able to collect 11 out of the 17 Top Handcrafted Features as identified in Panda et al. (2018) Table 6.[3] Most of the features correspond to tone color and one of the features to rhythm. See Table **??** in Appendix C of this response letter for a description of the Top Handcrafted Features. There are potentially hundreds of variables, and we wanted to be thoughtful about what variables are included in the model. We chose the 11 Top Handcrafted Features because they were shown in prior research to have predictive power specifically for music emotion, as compared to other less relevant music prediction problems like genre.

We incorporate these 11 features into the model following the approach used in the revised paper to combine our deep learning model and handcrafted features as detailed in Section XX of the paper, with the goal of maintaining explainability. We observe that the inclusion of the Top Handcrafted Features did improve the predictive performance of the model, increasing the F1 from 0.5380 to 0.5450 relative to the mel - harmonics model (see Table A.1).

[3] The features can be divided into "base" features that exist in the literature and "novel" features proposed in Panda et al. (2018). We were able to obtain base features from Matlab's MIR Toolbox but not the proposed novel features because the code has not yet been released and there are several design decisions that need to be made in implementation that we are unsure of, e.g. hyperparameters for feature construction. When we reached out to the authors, they offered to compute features for a few audio clips but could not compute them for the number requested because of high computation costs associated with the features.

Regarding enhancing interpretability, it is not quite obvious that including handcrafted features always leads to greater interpretability. It depends on exactly what features are included. While we know mathematically how the features are defined, it's not obvious how to interpret them all. For instance, MFCC1 represents "cosine fits of log energies" and does not have clear human interpretability.[4] For others, like Spectral Entropy, it is perhaps easier to see, since it represents the spread of frequencies across the spectrum.

We thank the review team for the suggestions to include more relevant features in order to not only improve accuracy but also enhance interpretability.

AE 4: *c. Explore the use of inception network with multiple convolutional filters concatenated in parallel and possibly incorporating RNN and/or some position encoding in the classifier as suggested by R1 (under some other suggestions).*

Reply:

We thank the review team for the various model modification suggestions for the goal of improving predictive performance. We have added the RNN layer after the CNN. As noted in the Summary of Major Changes and in our response to **R1.1**, this did not improve the accuracy—and this is consistent with the fact that in Choi et al. (2017) (Figure 3), they also find that RNN marginally reduces the classification accuracy for the two emotions they classify (happy, sad).

Regarding R1's suggestion to consider the use of the inception network, we have replied in detail in our response to **R1.4**. To summarize: Inception is designed for vision since objects (e.g., animal faces) are comprised of smaller elements (e.g., eyes) of different sizes. In music, unlike images, dependencies span the entire frequency spectrum, and are non-contiguous, so there is no conceptual way to capture this information using inception. It is unclear how to design an inception net for music that is theory-based and explainable. Additionally, inception net was motivated by gradient propagation problems for very deep nets (the authors discuss this focus quite a bit). In contrast, our nets do not suffer from these issues. Our sense is that the above reasons make the inception net architecture less relevant for our practical problem.

AE 5: *2.2. Provide additional details/motivations of the proposed algorithm*

a. Additional explanations/motivations as to why mel transformation is needed (R2.1c). Without mel transformation, how much model accuracy/explainablity do we lose?

Reply:

To address this question, we investigated how the performance would change if we directly used STFT spectrograms without the mel transformation, which accounts for knowledge about how humans hear. Appendix Table A.1 summarizes the results. Although not greatly different, we find that the average performance measures are slightly higher with the mel spectrogram than with the STFT spectrogram. In addition to the slightly better performance, the mel model is more efficient in terms of number of

[4] See `https://dsp.stackexchange.com/questions/38830/whats-the-correct-graphical-interpretation-of-a-series-of-mfcc-vectors/38877` for instance.

parameters (since the input image is smaller) and mel spectrograms are more frequently used in the literature (e.g., Choi et al. (2017), Pons et al. (2016), Chowdhury et al. (2019)). We therefore retain the mel spectrogram as the primary model input and include a footnote about the STFT results on p. X of the paper.

AE 6: *b. Additional justification on why avg pooling is used to combine feature maps across time and max pooling is used to summarize information over pitch classes.*

Reply: We have now provided more detail on p. XX of the paper. In short, typically either max pooling or average pooling is used in deep learning models to summarize the feature space. We empirically test the four possible pooling combinations by training models according to the four architectures. As shown in Table B.1 in Appendix B of this letter, we find that average pooling across time and max pooling over pitch classes results in the best classification performance. We believe that this is because first, consonance over a time period is based on the average consonance during that period (and not the max). A clip that is perceived to be consonant throughout is more likely to be perceived to be consonant overall than a clip that is consonant at one point but otherwise dissonant. And second, we believe overall consonance is dominated by the most consonant pitch class. A clip with one highly consonant pitch class (i.e., many overlapping harmonics) is perceived to be more consonant than many pitch classes that are slightly consonant. We therefore use average pooling over time and max pooling over pitch classes in our main model specification.

AE 7: *c. Provide additional details to answer R2.1.d-f. On a related point, R1 asks for more visualization on the training and testing process. All of these will help us better understand important details about the proposed algorithm.*

Reply: We have added visualizations on the training and testing process to Appendix D of this response letter (and the paper too). We have answered R2.1.d-f in detail in our response to the reviewer.

AE 8: *2.3. Stronger benchmarks: R1 raises the question of whether the current benchmark models are strong enough, particularly because the authors did not include some of the cited different CNN architectures as benchmarks. I agree with R1 that the manuscript should at least include additional music CNNs that use m-by-n rather than simple square filter as additional benchmarks. Some comparisons with existing RNN music classification would also be helpful.*

Reply: We have added a number of benchmark models that include CNNs with $m \times n$ filters, and test $m = 2n$ and $m = \frac{n}{2}$, as well as a CNN with an RNN layer. More specifically, we have added filters that are tall and skinny rectangles, which may better capture features that span a larger portion of the frequency space, and filters that are short and wide, which may better capture features that span a larger portion of the time space. As shown in Table X in the paper, we find that the rectangular filters do outperform the square filters but do not outperform the harmonics filters. We also train a CNN + RNN model. We find that this model underperforms our baseline CNN mel - harmonics model (Table A.1), which is consistent with what Choi et al. (2017) also found for music emotion classification. Additional model details can be found in Appendix B of this letter.

AE 9: *3. Model Performance and Explainability*

3.1. Low classification accuracies As noted above, both referees commented on how the classification accuracies in Table 2 are very low (R1.2 and R2.2a). I noticed the same. The referees have been very constructive and proposed several ways to improve predictive accuracy. I recommend that the authors explore all of their suggestions. More specifically, please try the following: 1) adding additional features; 2) modifying the CNN architecture; and 3) reporting binary classification results as suggested by the referees.

Reply:

Thank you and the review team for the various questions about model performance and suggestions to improve them. We address this issue in detail in Point 1 of the Summary of Major Changes at the beginning of this response letter. The critical point is that performance accuracy is very much a function of the classification/prediction task. Objective recognition tasks (e.g., object identification in computer vision, instrument/music genre identification) tend to have much higher accuracy rates than subjective human response tasks (e.g., emotion recognition, humor detection, art evaluation), like the task in our current application.

That said, we also estimated a large number of other models suggested by you and other members of the review team to improve accuracy. In brief, as shown in Table A.1 we find that our proposed approach achieves similar or better performance than benchmarks for the emotion classification of short music clips.

We hope you find our clarification about the relationship between the nature of the classification task and absolute performance helpful in framing the overall accuracy of our results and how we choose benchmarks to compare. We very much appreciate your overall encouragement to increase the accuracy of our model and trust that you will find our efforts on this dimension satisfactory.

AE 10: *3.2. Improve explainability arguments*

R2.2b suggests using the same music clip to compare model in Figures 7 and 9. This referee also has a few other great suggestions/comments regarding explainability, which I will not repeat.

Reply:

We thank the referee for these suggestions and comments. We now use the same music clips "to get an apples-to-apples comparison" as the reviewer suggested. We have answered the reviewer's other points directly in the reviewer response.

AE 11: *4. Ads Insertion Application*

As both referees, I find the ad insertion lab experiment interesting but not sufficiently convincing. I will provide some additional details below. Even if the manuscript only aims to provide a proof of concept (which is totally fine in my opinion), the ad insertion exercise still need to be strengthened.

1. *The use of only 1 video: As R1 correctly pointed out, when only 1 video is used in the lab experiment, various confounding issues arise. I agree with R1 that multiple videos and multiple ads need to be used in the ads insertion exercises. The review team is quite reasonable by not*

asking a high number of videos and ads. I think this will be an important step in the revision process for us to gain more trust in the business application.

2. *Only consider music emotions in focal video and ad: I fully understand the empirical challenge of accounting for emotions evoked by things like the story lines, scripts, visuals, etc. in the focal video and the ad. At the same time, if industry practitioners were to use the proposed algorithm for automated ad insertion, I feel that the authors need to make some effort in controlling for emotions evoked by non-music-related content in the focal video and the ad. With many off-the-shelf text-mining and image-processing machine learning algorithms, I don't think it would be extremely challenging to control for some of these additional factors. I would be more sympathetic if such controls are not perfect. But at the minimum, I felt that the authors need to make some attempts in this regard. Otherwise, I can see a lot of resistance from industry practitioners in terms of embracing and adopting this algorithm for automatic ad insertions.*

Reply: We appreciate your guidance that the intention of the application is to provide a proof-of-concept. We strengthen the ad insertion exercise by increasing its *scale* (both number of content videos and number of ads) and *scope* (by including emotion from other modalities), and discuss them in turn.

1. Increasing Scale of Application: One concern raised by the review team was whether our results might have been an artifact of the particular content video (and only two ads) used in the empirical application. We assess the robustness of the results across multiple ads and content videos, and find the result from the prior version that "higher congruence between ad and [content] videos improves brand recall" to be robust. We further find that higher congruence also decreases ad skipping. We did face practical cost constraints in scaling even further—please see Appendix E for details.

2. Increasing Scope of Application: First, note that content-level controls would already be captured in our analysis, since we use fixed effects for each content video and for each ad. The question then is whether we have data from other modalities (especially facial emotion, voice emotion, and speech content emotion) that might help to improve the emotion matching between content video and ad. We demonstrate how this can be done, and use a publicly available tool, Microsoft Azure, to extract emotion from facial expressions and speech. Appendix E.2 of this letter details our implementation procedure and findings. We summarize the procedure and findings in the revised paper (p. XX) as follows:

> Our primary analysis has focused on using emotion evoked from music, and characterizing that into four quadrants based on valence and arousal. However, with videos, emotional content may be present across multiple modalities (e.g., facial expressions, the text of speech). Multimodal emotional content can also be used to predict emotional distance. When the videos have human faces, we can use publicly available tools to estimate emotion from facial expressions. Similarly, emotional content can also be obtained from voice tonality and speech text.
>
> In our application, we observe that not every content video has speech or human faces and the same to be true for the first six seconds of the ads, implying this approach is not always feasible. Including face emotion improves the recall rate when used in combination with the mel - harmonics model but hurts the skip rate when used in combination with the mel - harmonics + Top Handcrafted Features model. Appendix E.2 details the analysis that includes face emotion.
>
> Overall, there is potential in incorporating emotion information from images and text but the existing tools are limited in their ability to extract emotion information from short clips

(i.e., first six seconds of ads) and animated videos. However, this is a moving target and as these methods steadily improve these findings could well change.

AE 12: *4.3. Provide additional details on the ads, human tagging, etc.: R2 asked the authors to provide a lot of additional details on the ads, human-tagged distribution of emotions, JS distance for Ad2, insertion times, etc. I am also particularly puzzled by how humans were instructed to insert the ads. It seems that the MusicEmoCNN ad insertion performed slightly worse (or about the same as) than the human insertions in terms of brand recall (Table 4). I am curious to learn how humans decide where to insert the ads. Some visualization on how ads were inserted at different moments of the focal video under different methods would also be interesting. Additionally, I assume that the manuscript uses probabilities of classification (rather than discrete classification outcome) to determine optimal ad insertions, am I right? Otherwise, there will be ties in terms of where to insert the ads. If there were ties, how did the authors break the tie?*

Reply: We have extensively rewritten the application portion of the paper to include more detail on the ads, content videos, human tagging, etc. and have included more graphs of the JS distances. In the previous version, we did not ask humans to select ad insertion times. What we had meant was that we used the human emotion labels to calculate the ground truth emotion distributions to determine the points of greatest emotional similarity in terms of JS distance. We have now clarified our procedure. You are correct that we are using classification probabilities and not the discrete classification outcomes to do the emotion matching, and thus the issue of ties is not critical.

In the revised version, we have not included ad matching based on human emotion labels ("human insertion") as a relevant benchmark, since asking humans to label emotion for content videos and ads over time is a non-scalable task. Even if humans performed much better than any algorithm, the approach would not be practical, and hence not useful in practice.

AE 13: *5. Limitations: In addition to the limitations listed by the authors, I think another limitation is that the MusicEmoCNN is trained on very short music clips. It's hard to tell whether we learn here generalizes to longer videos. Given the novelty of the manuscript, I will not push further on this dimension. But I think the authors should point this out as a potential limitation of their proposed algorithm.*

Reply: We now directly discuss this point in the conclusion. We mention that for other applications that may require training on longer clips, one should assess the generalizability of the claims on longer clips. While the method itself is not limited to a specific duration, the details (architecture, hyperparameters) of appropriate models could vary when the duration is different. For example, related to your point, we might want to use a modified architecture (e.g., CNN + RNN) with longer clips since positional information might become more important. The motivation for training the model using six-second clips is that duration is what is appropriate for the application. Specifically, users are able to first make the ad skip decision at the six second mark. Of course, this is an institutional setting and under the control of the platform (e.g., YouTube).

We thank you and the review team for seeing potential in the research. We trust you will find that the current version addresses the concerns that arose during the prior review. We are also very appreciative that the overall quality and contributions of the paper have been enhanced significantly in this process.

4 Response to Reviewer 1

The paper proposes a new set of CNN filters to process music spectrum diagrams in order to classify the emotion of the music based on valence-arousal quadrant. The paper then uses the emotion classifier to position ads within a music to demonstrate that such emotion classification could be useful in marketing applications. In essence, the paper makes two claims:

- *The proposed method can be as good as the state-of-art algorithm in music classification and it is explainable.*
- *The emotion classification from the proposed method could be helpful for marketers.*

In general, I like the idea of the paper and think the paper is publishable if the authors could convincingly prove both of their claims (we need both claims since the paper will not have enough contribution if the proposed method does not perform well and the paper will not be a fit to Marketing Science if the emotion classification does not help in marketing contexts). However, there are several issues that prevent me from believing the authors claims and I list them in the following:

Reply: Thank you very much for your helpful comments and feedback/suggestions on the paper. We provide our point-by-point responses to your comments.

R 1.1: *Does the algorithm really outperform traditional methods in music emotion classification?*

First, it is unclear if the benchmark is strong enough. The authors have cited various papers that propose different CNN architectures to process music in Page 8 (such as Pons et al. (2016), Dubois et al. (2019), etc.). As most of these papers noted, while most image processing tasks use square diagrams, most music CNN uses m-by-n diagrams (assuming the spectrogram has dimension M-by-N) because music spectrograms correspond to time and frequency. If the authors want to do a more representative comparison, they should compare their methods with various architectures proposed in these papers rather than a simple square filter. Moreover, the literature has proposed various RNN architecture for music classification due to the fact that music is sequential. I am surprised to see that the authors have not utilized sequential information beyond the window size of the filter and the authors, when doing performance benchmarking, should also compare with these RNN models (for example, Choi et al. (2017))

Reply:

In our original manuscript, we benchmarked against models that use square filters as well as time, frequency, and frequency-time filters as proposed in Pons et al. (2016). Per your request, we have now included two additional sets of benchmarks.

1. *m-by-n rectangular filters*: We have now included *m-by-n rectangular filters* as additional benchmarks. We trained one model with filters that are tall and skinny rectangles, which may better capture features that span a larger portion of the frequency space, and another model with filters that are short and wide, which may better capture features that span a larger portion of the time space. As shown

in Table XX in the paper, we find that the rectangular filters do in fact outperform the square filters. However, they do no better than the proposed harmonics filters. Table A.1 in the appendix to this response letter summarizes the additional benchmark models we compare against.

2. *RNN Architecture:* In response to your request to account for sequential information, we now consider a CNN+RNN model. For this, we modify our CNN model (with theory-based consonance filters) by adding an RNN layer post-convolution to capture sequential information. We find that the RNN layer did not improve performance; in fact, the classification accuracy for emotion is worse (Table A.1). The reduced emotion classification accuracy of CNN+RNN is also consistent with the results reported in Choi et al. (2017), who found their CRNN (CNN combined with RNN) did worse for emotions like happiness and sadness. Please refer to Figure 3 in their paper for the "sad" and "happy" classification. Given the results, we did not pursue the RNN route to improve emotion classification accuracy.

Additional model details can be found in Appendix B of this letter.

R 1.2: *Second, and probably more importantly, the basic classification accuracies in Table 2 are too low. The average Precision and Recall are around 50-55% while the accuracy rate of random guess is 25%. This is very low compared to image processing (where base rate now is around 90%+ on image net with random guess rate to be 0.1%) and music genre prediction (70%+ with random guess rate to be 10%, see Dong 2018). I am not sure if this is because the model is ill-structured, the data is too limited or the training is premature (my hunch is that the data is too limited). However, I think the authors should improve their model performance and I will provide some suggestions later.*

Last, let me be clear, I do not think that the authors need to propose a method that achieves the state-of-art performance in prediction because this is not a CS journal. What the authors should do is to propose a method that is reasonably good and useful in a generalizable marketing context. Therefore, the true constraint here is whether the model will be useful in a practical marketing context, which we will discuss later.

Reply:

Thank you for your questions about model performance. We address this issue in detail in Point 1 of the Summary of Major Changes at the beginning of this response letter. The critical point is that performance accuracy is very much a function of the classification/prediction task. Objective recognition tasks (e.g., object identification in computer vision, instrument/music genre identification) tend to have much higher accuracy rates than subjective human response tasks (e.g., emotion recognition, humor detection, art evaluation) and the task in our current application is very subjective.

That said, as mentioned in point R1.1, we also estimated a large number of other models suggested by you and other members of the review team to improve classification performance. In the spirit of your suggestions, we now use as the main specification a combination of deep learning and handcrafted features that has the highest classification performance. In brief, we find that our proposed approach achieves similar or better performance than *benchmarks focusing on music emotion classification* for short music clips. We also respond to your suggestion about revisiting the data in our reply to your point **R1.5** and have added graphs showing the train and validation loss over epochs in Figure D.1 in Appendix D of this response letter.

We hope you find our clarification about the relationship between the nature of the classification task and absolute performance helpful in framing the overall performance of our results and how we choose benchmarks to compare against. We appreciate your overall encouragement to increase the accuracy of our model and trust that you will find our efforts on this dimension satisfactory.

R 1.3: *Is the algorithm useful in a practical marketing context? The major problem with the lab experiment is that it is not valid. The authors only utilize one video which creates serious confounding issues. For example, suppose the MusicEmoCNN inserts the ad at 45 seconds while the human inserts it at 30 seconds of the video. How do we know whether the difference in recall is caused by the difference in JS distance or is it caused by some other systematic difference between 30 and 45 seconds of the video? A more proper experiment will involve multiple videos/musics (such as 30) and multiple ads (such as 6), and for each random pairing, each of the insertion method's results are reported. The authors should then conduct proper inference based on their lab experiment results. Properly conducting the lab experiment is very important since this is the only way for us to make sure that the proposed method works in practice which makes us less worried about whether the performance of the algorithm is state-of-art or not.*

Reply: We agree that it is important to generalize beyond a single video in assessing the robustness/validity of our results. We have now expanded the number of content videos (from 1 to 4) and the number of ads (from 2 to 4), increasing the number of experimental cells from 12 to 96 (4 content videos x 6 ad insertion points x 4 ads). We pre-tested ads and content videos, and chose 4 ads that span the 4 emotion quadrants and 4 content videos that span the emotion dimensions over time. Finally, we used a full factorial design of the 6 insertion points for ad and content, which allows us to compare any set of ad insertion algorithms. Overall, with this new expanded scale of the experiment, our conclusion that congruence is helpful remains consistent with the prior version.

While having more content videos/ads/insertion points would be even better from the point of view of generalizability of the application, there are practical cost constraints that require us to limit the set of ad and content videos. The experimental design in the revision, with 96 cells, 15 respondents per cell, and $2.40 per respondent, cost ∼$3,500. Using 30 content videos and 6 ads with 6 insertion points (as suggested) would cost ∼$39,000. Please see Appendix E for details.

We hope you agree that our current design balances the tradeoff between validity/robustness of our conclusions and costs in a reasonable manner.

R 1.4: *As aforementioned, the authors may benefit from the design from inception network (Szegedy et al. 2017) where multiple convolutional filters are concatenated in parallel so that they can utilize different filters' power at the same time. Noted that the parameter setting in these inception networks are pretty empirical so this requires the authors to conduct more empirical experiments and probably find more data.*

Reply: We appreciate the suggestion in the spirit of helping us to improve accuracy. But after spending a fair amount of time understanding the details of the inception network, which was developed and fine-tuned to address vision problems at scale, we believe that this model may not be a good fit for our music application.

As you know, the inception network is commonly used for image classification (Szegedy et al. 2014). First, the key idea of the inception network is to approximate and cover an optimal local sparse structure of a convolutional vision network using readily available dense components–a technique that is particularly helpful computationally for object identification when objects vary in their scale. This issue of object scale is not directly relevant to music spectrograms. Second, in vision, contiguous areas of images have meaning, and so larger filter dimensions are meaningful in capturing information in contiguous areas. But our key insight in filter design to capture consonance and dissonance in music is that we should design it to be non-contiguous. This would not be possible or quite challenging to design with the inception network. Finally, we note the inception network was motivated to solve the problems that arise when using deep nets with high depth, including gradient propagation, but our model does not seem to suffer from these issues.

Overall, aspects of the vision problem that motivate the need for inception networks seem less relevant for our problem setting. As such, while we have explored all of the other modeling suggestions by the review team, we did not explore the estimation of the inception model. We trust that you would agree with this assessment.

R 1.5: *The authors could also benefit from more data. There are several programs that allow researchers to automatically generate music data. The authors should utilize these programs along with Prolific/MTurk tagging to construct a larger dataset.*

Reply:

We appreciate the suggestion to increase the size of the dataset as an approach to improve accuracy and robustness. We thought hard about how to address the key limitations of the 4Q and Soundtracks datasets for our application. As you note, there are only a relatively small number of clips in the 4Q (900) and Soundtracks (360) datasets that we had used earlier. In addition, the data was also limited in that they had only a single label for each full-length clip (30 seconds for 4Q and 10-30 seconds for Soundtracks). While each Soundtracks clip was selected on the basis of evoking a single emotion throughout, this was not the case for 4Q, which could introduce noise in labeling the clips every six seconds.

We first considered your suggestion to construct a new labeled dataset, but given the challenges involved, we searched for another dataset that can address the limitations of the 4Q and Soundtracks datasets. Fortunately, we found a different dataset (DEAM) that addressed most of the concerns. We elaborate on the considerations below.

1. Construct a New Labeled Dataset: Constructing our own custom dataset would have been (i) technically challenging since it involves expertise and effort on many dimensions and (ii) financially costly. This is usually why constructing such datasets is typically its own separate and valued contribution in computer science, and often funded with specialized grants. To construct a new dataset, we would need to generate thousands of emotionally-varying songs with programs like you suggest. After generating the data, we would need to gather the emotion labels for each short clip (six seconds) from several human respondents using mTurk, Prolific, or another similar resource. This would require creating an interface that can play music and capture emotion every six seconds. We would then need to evaluate intercoder reliability and potentially recruit more human evaluators to achieve consistent evaluation. The cost of effort and resources to build such an interface and gather labels would be a non-trivial amount. Next, even after publishing such a dataset, it takes time for such datasets to become accepted and trusted in the community. Given these costs, we felt that if we could find a dataset that addressed most of the issues, that would make the most sense for the current paper's goals.

2. Dataset on Music Clips with Dynamically Varying Emotion (DEAM): Fortunately, the **DEAM dataset** collected by computer science scholars (Aljanaki et al. 2017) addresses the concerns we mentioned with the previous data. DEAM is particularly valuable in our application since it includes 1,802 songs with emotion labeled at frequent intervals within a music clip. This allows us to have emotion labels for each six-second clip. The dataset also covers the span of the valence-arousal space—making the training robust in all quadrants and also in ambiguous cases that are closer to quadrant boundaries. We combine the DEAM data with the Soundtracks data to increase the size of the training data. We retain Soundtracks since it is representative of the music in our application and is less noisy in that it was selected to evoke a single emotion throughout, unlike 4Q. We detail how we combine the datasets on p. XX of the paper. After subsampling the data so that no quadrant is too dominant, we obtain a dataset comprised of 2,176 six-second music clips.

Overall, our paper's goal is to illustrate how to estimate theory-based explainable models for music emotion and then illustrate its practical utility through the ad application. We view the combination of the DEAM and Soundtracks datasets to be an excellent compromise in enabling the paper's goals, while addressing the concerns regarding limited data. We hope you also find this approach reasonable.

R 1.6: *The authors may benefit from some ideas of coding sequential information in their model. They can use traditional RNN with CNN as the citation suggestions or they can use some position encoding.*

Reply: Thank you for this suggestion. We too expected that some form of positional information would improve prediction performance. As we had noted in **R1.1**, we tried CNN + RNN, but we found that the performance did not improve, and was actually worse than our main model. Since this is consistent with the results in Choi et al. (2017) (Figure 3), we decided not to pursue this route further in the revision.

We now note that it would be useful to pursue in more detail why sequential information does not add to the prediction accuracy in future research. Is the above a typo?

R 1.7: *The authors should publish their code and the training/testing datasets (which I assume is public) so that the results and paper are replicable.*

Reply: We agree. We plan to share all of the code on Github, consistent with the replication policy of the journal. The datasets we have used are publicly available, so the availability of both code and data will ensure replicability.

R 1.8: *The authors should provide more visualization on their training and testing process so that we can understand more how losses are changing over training epochs.*

Reply: We have added graphs of the training process over epochs for the training and validation data sets in the online appendix to the paper and the appendix to this response note (Appendix D).

In summary, I like the general idea of the paper and see a path forward if the two claims from the authors are convincingly supported. However, the current execution fails to do so. I think the major next step is to improve the execution, which, in my opinion, is reasonably focused and clear. Therefore, I recommend a major revision and hope the suggestions are helpful.

Reply: We trust you will find this revision to address the following main points. First, the proposed method is explainable, and is as good in terms of performance as benchmark algorithms that only focus on accuracy and not explainability. Second, our method can use additional features in conjunction with the theory-based deep learning model to improve accuracy without impacting explainability. Third, given the increase in scale and scope of the experiment, we find that using the proposed method in an ad insertion application obtains reasonable performance (in terms of skip rate and brand recall rate), making it useful for marketers. We appreciate your framing of the contribution in these terms.

Thank you again for your many helpful suggestions and constructive comments to make the paper better along multiple dimensions. Thinking through your comments/suggestions and addressing them has helped us both improve the paper and also clarify our contributions. We hope the various analysis/estimation and clarifications that we have provided have collectively addressed your concerns.

5 Response to Reviewer 2

The paper builds a new deep learning model to predict emotion (both valence and arousal) in music audio files. The model architecture is theory-driven and interpretable. The idea behind the model comes from music theories implying that consonance leads to positive valence and low arousal, whereas dissonance leads to negative valence and high arousal. So, the model's core is to create a consonance filter to find harmonics, octaves, and fifths that can produce a consonant sound. In an application, the paper shows that inserting an ad at a point where the ad and the background video have emotional similarity can increase ad recall. There are many things to like about this paper. The authors have rich knowledge of music and audio signal processing. The proposed consonance filter is innovative. The paper is well written and easy to follow. I offer the following suggestions to strengthen the contributions of this paper.

Reply:

Thank you for your excellent summary and the overall positive reaction to our paper. We greatly appreciate it. We provide our point-by-point responses to your comments below.

R 2.1: *1. Model: a. One drawback of the proposed model is that "with the single dimension of consonance, Q1-exuberance and Q3-sadness cannot be disambiguated (p25)." This drawback could limit the prediction accuracy, explainability, and application of the model. As the authors indicated, tempo, which can be seen in the spectrogram, can help differentiate these two quadrants. Could you change the architecture of the deep learning model, detect tempo via the width of the notes, and add the tempo feature between step d) and step e) in Figure 6?*

Reply:

We thank you for raising this question. Our unclear writing must have contributed to the confusion. We would like to clarify that the CNN classifier using music theory-based consonance filters *can and does distinguish* between Q1 (exuberance) and Q3 (sadness). As can be seen in the confusion matrix for the CNN with harmonics filters in Appendix X of the paper, correct Q1 and Q3 classification rates are fairly high and the misclassification rates between Q1 and Q3 are fairly low. The statement on the previous p. 25 referred only to the Grad-CAM heatmaps, which we discuss below. Since not all features can be easily visualized using the Grad-CAM heatmaps, we conduct additional analysis on the features learned by the theory-based deep learning model to understand what features the model learns besides consonance. We find evidence that the model learns to differentiate high harmonics from low harmonics, which helps to separate out high arousal from low arousal since higher harmonics are associated with higher arousal. Therefore, Q1/Q3 classification is not a concern for prediction accuracy or application of the model. Nevertheless, we thought it was a good idea to explore incorporating tempo into the model as it is a feature associated with arousal and discuss the results below.

a. Clarification of Grad-CAM Explainability: While our model can separate out Q1 and Q3 in its classification, it is challenging to visually differentiate the Grad-CAM heatmaps for Q1 and Q3. High consonance correlates with high valence and low arousal, and low consonance correlates with low valence and high arousal. Therefore it is easy to distinguish quadrants Q4 (high valence-low arousal) and Q2 (low valence-high arousal) from the Grad-CAM heatmaps as they nicely separate between high and low consonance. However, Q1 (high valence-high arousal) and Q3 (low valence-low arousal) do not map directly with high and low consonance. We can therefore only form two sets of hypotheses based on the theory as discussed on p. XX of the paper: 1) for a consonant clip, we expect the Q4 heatmap to be the brightest, the Q2 heatmap to be the darkest, and the Q1 and Q3 heatmaps to be in between; 2) across clips predicted to be in the different quadrants, we expect the following average heatmap brightness patterns: Q1 $>$ Q2, Q4 $>$ Q3, Q4 $>$ Q1, Q3 $>$ Q2. The Grad-CAM heatmaps are consistent with these sets of hypotheses, suggesting that the model learns to capture consonance from the data in order to predict emotion. Since we know that consonance is a strong predictor of emotion from past literature (Gabrielsson 2016, Luck et al. 2008), we can be more confident that our model can generalize outside of the training data.

b. Model Learns More Than Consonance

As noted on p. X of the paper, our proposed harmonics filters impose structure on the frequencies seen by the model, but the model is still free to learn features useful in classification beyond consonance. For example, within the set of harmonic frequencies, the model can learn features like whether more high or low frequencies are present. Indeed, one feature that helps to distinguish high arousal (Q1/Q2) from low arousal (Q3/Q4) is the presence of high harmonics. More high harmonics is associated with higher arousal (Gabrielsson 2016).

While it is challenging to visualize these other features within the Grad-CAM framework, we can assess the correlations between the features learned by the model and the emotion labels and handcrafted features. We find that when we extract the 32 features learned by the theory-based deep learning model for each music clip, the features correlate not only with valence and arousal, but also several handcrafted features. Spectral skewness is a handcrafted feature that measures the presence of high harmonics and positive skewness indicates more energy at higher frequencies (Kazazis et al. 2022). We find that the most negative correlation between the features and spectral skewness is -0.54. The corresponding channel also has a correlation of -0.07 with arousal. Meanwhile, the most positive correlation between the features and spectral skewness is 0.45 and the corresponding channel has a correlation of 0.07 with arousal. These patterns suggest that the model learns features that capture the presence of high harmonics, which can separate out low arousal music from high arousal music.

Therefore, we can think that in the simplest case consonance helps to separate out high valence from low valence and the presence of high harmonics helps to separate out high arousal from low arousal.

c. Including Tempo in Model:

Your suggestion to include tempo was helpful and we tested multiple ways to implement this using tempo inputs standard in the literature: a) incorporating a tempogram as an additional input into the deep learning model beyond the mel spectrogram (Hsu et al. 2021), and b) using tempo features as additional explanatory variables beyond the features learned by the deep learning model. Appendix B of this response letter provides more detail for both these methods and Table A.1 shows the classification performance. We find that incorporating a tempogram did not improve performance but incorporating the tempo features marginally did. We are grateful for this suggestion to include an interpretable feature that also improved classification performance.

R 2.2: *b. Other features: Panda et al. 2018 The paper focuses on consonance as a prediction of* *emotion, but ignores many other related features that the literature has found. For instance, in the cited paper, Panda et al. 2018, Tone color and Tremolo are found to be highly associated with the four emotion quadrants. Other features, including Mode (major or minor) and Melody (range and direction), seem relevant too. Would it be possible to incorporate some of these features? Adding them can make the model more accurate. And because these are interpretable features, maybe the model can become more explainable too.*

Reply: Thank you for this suggestion. We have included some of these features, and do so in a principled manner. The Panda et al. (2018) paper represents a heavily feature-engineered approach, trying to identify the features that matter most for music emotion classification. We have tried our best to translate the ideas and musical constructs from Panda et al. (2018) into our model. We were able to obtain 11 out of 17 features listed in Table 6 of Panda et al. (2018), which specifies the Top 5 features that most impact classification performance within each quadrant for the 4Q dataset. A description of these features can be found in Appendix C of this response note. Most of the features relate to tone color and one of the features relates to rhythm. We found that including the features slightly improved classification performance. Specifically, when using the Top Handcrafted Features in conjunction with the mel spectrogram and our harmonics filters, we found that F1 increased from 0.5380 to 0.5450 relative to the mel - harmonics model (see Table A.1).

However, there are several practical challenges associated with using all the features in our empirical setting, which we detail below.

1. We have had a detailed correspondence with the lead author Dr. Renato Panda, and we requested their code to replicate and compute the other features. Dr. Panda commented that their software is not the most stable and hence they have not yet made their code publicly available. In trying to extract features based on the paper, we learned that the code depends on a number of different softwares (specifically Marsyas, Psysound3, and Matlab), of which we are only able to replicate the Matlab features. We found that the other programs had significant compatibility issues and could not even be installed on currently available hardware/operating systems.[5]

2. We also learned that obtaining these features is highly computationally intensive for any new dataset, especially for the "novel" features from Panda et al. (2018). Panda also expressed that his team is now looking into deep learning solutions because of the computational burden of constructing handcrafted features.[6] This computational cost significantly limits the usability of such an approach in our empirical application, where the goal would be to at scale compare different insertion points across a wide set of content videos and a wide set of ads to identify which combination might result in higher ad engagement.

[5] To quote Dr. Panda (private e-mail correspondence): *"Regarding your request, the feature extraction setup is quite heavy and complex since I ended up extracting features with several academic frameworks that are not the most easy and stable thing to use (marsyas - c++, pysound3 - matlab and very bug ridden, mirtoolbox - matlab), and in addition to that also developed some novel features that depend on other things (the process is more or less audio → MELODIA from Salomon or Dressler* → Notes segmentation by Paiva → several scripts to extract info (this was repeated with an extra step of source segmentation to get only the voice, and the Dressler part got me into an NDA with Fraunhofer)."*

[6] To quote Dr. Panda (private e-mail correspondence): *" [T]his is the main issue with handcrafted features, and one of the reasons why we now are also exploring deep learning solutions in parallel. Both proposing and extracting new audio features (DSP) is heavy/complex. Especially in this case where we tested several sets of different natures, so it implies running multiple stacks of technology that are quite slow (i.e., MATLAB)."*

3. Your point regarding interpretability is an important one to understand carefully. It is not quite obvious that including handcrafted features always leads to greater interpretability. It depends on exactly what features are included. For instance, Appendix C of this letter details the Top Handcrafted Features from Panda et al. (2018). While we know mathematically how the features are defined, it's not obvious how to interpret them all. For instance, MFCC1 represents "cosine fits of log energies" and does not have clear human interpretability.[7] For others, like Spectral Entropy, it is perhaps easier to see, since it represents the spread of frequencies across the spectrum.

Again, while we could have included many other features, we are mindful that Panda et al. (2018) tested hundreds of features and found the above ones to have the biggest impact for music emotion.

R 2.3: *c. Mel transformation Could the authors explain why it is necessary to transform the STFT to mel spectrogram? It seems that the transformation to mel spectrogram makes the design of the consonance blinder extra complex, involving the mel filter bank. Without the transformation to mel spectrogram, would the model accuracy and interpretability be compromised?*

Reply: The mel spectrogram is typically useful since it is a transformation of the STFT spectrogram designed to correspond to human hearing sensitivity across the frequency spectrum, and is almost always used as a pre-processing step in the literature (Müller 2015, Choi et al. 2017, Pons et al. 2016, Chowdhury et al. 2019). At low frequencies, the mel spectrogram and STFT spectrogram are very similar. It is only at high frequencies that the two start to differ because human hearing cannot differentiate close frequencies at high frequencies.

To address this question, we investigated how the performance would change if we directly used STFT spectrograms without the mel transformation. Table A.1 provides the results from using STFT spectrograms with square filters and harmonics filters. We find that the average mel spectrogram performance measures are slightly higher than those from the STFT spectrograms. In addition to the slightly better performance, the mel model is more efficient in terms of number of parameters since the input image is lower resolution (256 rows in mel versus 2049 rows in STFT). Given these points and to be more consistent with the literature, we retain the mel spectrogram as the primary model input and include a note about the STFT results in the paper (p. XX footnote XX).

R 2.4: *d. Dimensions: 1025, 127*

*On page 12, the paper says the dimension of the STFT is 1,025 * 517. Could you explain where is the number 1,025 from?*

Reply: In the revised paper, we have changed the STFT spectrogram to be 2,049 x 517 and the mel spectrogram to be 256 x 517 to increase the granularity of the frequency bins.

The frequency dimension of the STFT is a result of the window size choice. We use a window size of 4,096 samples and results from signal processing lead to the 2,049 frequency bins (window size / 2 + 1). We add more detail to the paper (p. XX footnote XX) and include it here as well for easy reference.

[7] See https://dsp.stackexchange.com/questions/38830/whats-the-correct-graphical-interpretation-of-a-series-of-mfcc-vectors/38877 for instance.

> The dimensions of the spectrogram are obtained in a standard way as per Müller (2015) and come from the length of music, sampling rate, window size, and hop length. The time dimension is equal to length of music clip × sampling rate / hop length = (6 seconds)(44,100 Hz)/512 samples = 517. The frequency dimension is equal to the window sample size / 2 + 1 = 4,096 / 2 + 1 = 2,049. A discrete Fourier transform (DFT), which underlies the STFT, generates redundant information when it transforms audio information from the time domain to the frequency domain and only the first half of the frequencies are meaningful. The DFT transforms the 4,096 samples to 4,096 frequency bins but we retain only half of them because the other half is redundant. The larger the window sample size the greater the frequency resolution but the lower the time resolution.

A lot of this footnote still sounds mechanical–and that may be ok, if we refer to it as a formula but with some intuition—let us discuss and refine

R 2.5: *e. Step c) and d) why all blue?*

Are the colors in Figure 6 meaningful? If I understand correctly, the color in Mel Spectrogram (steps a) and b) in Figure 6) represents the squared magnitude of each frequency bin. Then why, after the convolution filter, do the post-convolution layer c) and the concatenated pitch class layers d) become all blue?

Reply:

Thank you for pointing this detail out. The colors in Figure 6 in the prior version did not have a specific meaning. We have changed the blue used in Figure 6 c) and d) to gray as shown in Figure **??** on p. XX to not confuse readers. The color itself does not have a meaning here; our objective is to show the rectangle that represents the vector output of the post-convolution layer.

R 2.6: *f. Threshold Page 27 says, "We quantify the relationship between heatmap brightness and emotion quadrant by summing up the heatmap values that are greater than a threshold. We impose a threshold." What is the threshold value and how did you choose it?*

Reply: We tested various thresholds and found that the rank ordering of brightness values did not depend on the threshold. We have therefore removed the threshold to simplify exposition and instead report the raw brightness measures averaged over the four emotion quadrants (p. XX).

R 2.7: *2. Model performance and explainability a. Industry benchmarks, classification by quadrant Could you make Table 2 more informative by adding benchmarks in the previous literature? For instance, Panda et al. 2018 attained a 76.4% F1 score with SVM. Right now, Table 2 reports metrics of a 4-class classification, which is hard to compare against metrics of a binary classification. Could you also report binary classification results, for each quadrant?*

Reply: We have added all comparison benchmarks outlined in the Summary of Major Changes to Table A.1 of this response letter to make them more informative like you suggested. We incorporate models/features suggested in the previous literature and report the results from the six-second Soundtracks and DEAM music clips.

In particular, we have tried using the Top Handcrafted Features suggested by Panda et al. (2018) and reported the results in the paper. Overall, we find that: a) the model using the Top Handcrafted Features does not have better predictive performance for our dataset and problem than our proposed theory-based deep learning model, and b) the Top Handcrafted Features used in conjunction with our theory-based deep-learning model does improve performance.

We report the classification results for each quadrant in Table A.2 in Appendix A of this letter, just as Panda et al. (2018) reported in Table 4 of their paper. Panda et al. (2018) does not do binary classification, and if we report that, we would not be able to directly compare the models. The Panda et al. (2018) results with the 76.4% F1 score are quite different in our empirical setting due to the following reasons:

1. Using a Subset of Top Handcrafted Features: As explained in our response to your point **R2.2** above, we are able to include 11 of the 17 Top Handcrafted Features, which Panda et al. (2018) (Table 6) found to be the most impactful features for classification performance.
2. Shorter Duration Clips: We use clips of six seconds (with the Soundtracks and DEAM data), whereas the clips used in Panda et al. (2018) are 30 seconds long (with the 4Q dataset). When we use the 11 Top Handcrafted Features on six-second segments from the 4Q dataset, we find precision is 0.577, recall is 0.561, and F1 is 0.539. The F1 is lower than the 0.764 found by Panda et al. (2018) because the clip durations are shorter (6 seconds vs. 30 seconds)[8] and because we use 11 features rather than 100.
3. Use of Exemplars in 4Q: The DEAM data, which represent 63% of our training data, include songs that have varying emotions across all parts of the valence-arousal circumplex while the 4Q dataset was more selective, with the authors choosing *exemplars* of each quadrant. Thus, a song might be strongly in Q1, indicating a highly positive valence accompanied by a high degree of arousal. Songs with weakly high or low valence/arousal were not included in the construction of this dataset. In other words, exemplar songs are selected to have a "large distance" between songs belonging to different quadrants, and more uncertain songs are not included in the 4Q dataset. In contrast, in the DEAM dataset, no such selection is used. We note that it is easy to see why the accuracy would be higher when we filter out the more ambiguous or more uncertain songs from the dataset.
4. Generalizability: Identifying Top Handcrafted Features implies that results within the 4Q dataset may be good, since the features connect well with the training data. However, generalizability across datasets (for example to Soundtracks and DEAM) could be adversely impacted. We find that the Top Handcrafted Features perform worse on the Soundtracks and DEAM dataset; precision is 0.400, recall is 0.402, and F1 is 0.394.

R 2.8: *b. Use the same music clips to compare models The paper compares the explainability of the proposed model against the square filter CNN using Figure 7 and Figure 9. However, the music clips in the two figures seem to be different (in all four quadrants) because the mel spectrograms look very different. Could you use the same music clips to get an apples-to-apples comparison? The same problem applies to Figures H3 and H4. Moreover, how are the music clips*

[8] Predicting listener emotion is more challenging with "less" data (specifically $\frac{1}{5}$ of the data in 4Q).

selected? Could you choose the clips that are the most representative, with the highest contentment/sadness/anxiety/exuberance rating?

Reply: We thank you for this helpful suggestion as it allows for easier comparison over models. Where possible, we have updated the figures to show the Grad-CAM heatmaps associated with the same music clips over different models. The same eight music clips are shown for the harmonics filters, the square filters, and the frequency filters. However, since the goal of these figures is to show how Grad-CAM heatmaps give us a sense of why a classifier makes the predictions it makes and not all of the models always make the same emotion classification, there are a few cases where we could not use the same eight clips. In particular, for the time filters we had to show different clips for two of the quadrants.

We selected clips that were representative in terms of brightness for the harmonics clips (i.e., close to the average brightness value for each quadrant) and clips that were predicted to be in the same emotion quadrant by the CNN models using harmonics filters and square filters.

Your suggestion also inspired us to generate a new figure (Figure **??** on p. XX) that shows Grad-CAM heatmaps for the same music clip over the four quadrants for several of the models. Grad-CAM tells us which parts of a music clip contributed to the probability of a given class even if the probability is very low. This figure shows that the model using harmonics filters is more explainable relative to the other deep learning models using atheoretic and low-level theory-based filters.

R 2.9: *c. Plot axes*

What are the x and y-axes in the Grad-CAM heatmaps in Figures 9, H3, and H4?

Reply: These axes correspond to the dimensions of the final feature map post-convolution. Therefore, the x-axis captures time and the y-axis captures frequency. We have added how to interpret the axes to the notes below the figures.

R 2.10: *d. Bright dark patterns for square CNN also hold The authors think Figure 7 is explainable because it is consistent with the theoretical prediction that "positive valence, low arousal songs have the brightest heatmaps (p25)." However, it seems that the same pattern holds for square CNN. In Figure 9, the heatmaps in Q4 are also the brightest. In this sense, square CNN is also explainable.*

Reply: For the consonance filters, music and emotion theory provides us predictions for the Grad-CAM heatmaps based on the structure we have placed on the model. We can then assess empirically whether the predictions hold in the data. However, for square filters we are unable to form predictions and so it is unclear what the heatmap brightness relates to in terms of musical features. In the case of square filters, we can only say that some subset of the spectrogram image, i.e., some subset of frequencies and time segments, contributes relatively more to the classification into a particular quadrant. This is similar to specific pixels or regions within images contributing to image classification. However, frequency and time are low-level sound concepts, unlike consonance, a mid-level concept, which is more meaningful to humans (Fu et al. 2010). It is therefore difficult to understand what the square filters learn.

R 2.11: *e. The Time Grad-CAM heatmaps are also explainable To some degree, I find the Time Grad-CAM heatmaps in Figure H4 also explainable. High arousal is associated with long-lasting, steady bright colors, whereas low arousal is associated with intermittent bright colors. Am I reading it correctly? Overall, the authors need to justify better why the proposed model is more explainable than existing ones.*

Reply: Yes, you were reading the previous Grad-CAM heatmaps for time filters correctly. The high-arousal quadrants had long-lasting, steady bright colors, while the low-arousal quadrants had intermittent bright colors. While it is good that the Grad-CAM heatmaps could be clustered in this way, it is still not clear what features the heatmaps highlight.

We have extensively revised the explainability portion of the paper (p. XX - XX) to better articulate why the proposed model is more explainable and hope you agree that it is now more clear.

Thank you again for your many helpful suggestions and constructive comments to make the paper better along multiple dimensions. Thinking through your comments/suggestions and addressing them has helped us both improve the paper and also clarify our contributions. We hope the various analysis/estimation and clarifications that we have provided have collectively addressed your concerns.

References

Aljanaki A, Yang YH, Soleymani M (2017) Developing a benchmark for emotional analysis of music. *PloS one* 12(3):e0173392.

Choi K, Fazekas G, Sandler M, Cho K (2017) Convolutional recurrent neural networks for music classification. *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 2392–2396 (IEEE).

Chowdhury S, Vall A, Haunschmid V, Widmer G (2019) Towards explainable music emotion recognition: The route via mid-level features. *arXiv preprint arXiv:1907.03572* .

Fu Z, Lu G, Ting KM, Zhang D (2010) A survey of audio-based music classification and annotation. *IEEE transactions on multimedia* 13(2):303–319.

Gabrielsson A (2016) The relationship between musical structure and perceived expression. *The Oxford Handbook of Music Psychology* .

Gabrielsson A, Lindström E (2010) The role of structure in the musical expression of emotions. *Handbook of music and emotion: Theory, research, applications* .

Hsu WH, Chen BY, Yang YH (2021) Deep learning based edm subgenre classification using mel-spectrogram and tempogram features. *arXiv preprint arXiv:2110.08862* .

Kazazis S, Depalle P, McAdams S (2022) Interval and ratio scaling of spectral audio descriptors. *Frontiers in Psychology* 13:835401.

Luck G, Toiviainen P, Erkkilä J, Lartillot O, Riikkilä K, Mäkelä A, Pyhäluoto K, Raine H, Varkila L, Värri J (2008) Modelling the relationships between emotional responses to, and musical content of, music therapy improvisations. *Psychology of music* 36(1):25–45.

Müller M (2015) *Fundamentals of music processing: Audio, analysis, algorithms, applications* (Springer).

Panda R, Malheiro R, Paiva RP (2018) Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing* 11(4):614–626.

Pons J, Lidy T, Serra X (2016) Experimenting with musically motivated convolutional neural networks. *2016 14th international workshop on content-based multimedia indexing (CBMI)*, 1–6 (IEEE).

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2014) Going deeper with convolutions.

Trainor LJ, Heinmiller BM (1998) The development of evaluative responses to music:: Infants prefer to listen to consonance over dissonance. *Infant Behavior and Development* 21(1):77–88.

A Appendix: Model Results

Here, we detail the results from the additional models that we have evaluated.

Table A.1: Classification Performance - Various Models

No.	Model Element	Model	Precision	Accuracy/Recall	F_1
Proposed Theory-based Mid-level Filters					
O6	Mel - Harmonics	CNN	0.5570	0.5434	0.5380
			(0.0873)	(0.0780)	(0.0791)
STFT Spectrogram Input					
N1	STFT - Square	CNN	0.5348	0.5054	0.4905
			(0.0841)	(0.0722)	(0.0802)
N1	STFT - Harmonics	CNN	0.5503	0.5310	0.5209
			(0.0741)	(0.0773)	(0.0795)
Rectangular Filters					
N2	Mel - Tall Rectangle	CNN	0.5500	0.5376	0.5216
			(0.0647)	(0.0618)	(0.0670)
N2	Mel - Wide Rectangle	CNN	0.5488	0.5376	0.5299
			(0.0737)	(0.0734)	(0.0747)
Incorporating Tempo					
N3	Mel - Harmonics + Tempo Feature	CNN + RF	0.5599	0.5572	0.5506
			(0.0765)	(0.0719)	(0.0717)
N4	Mel - Harmonics + Tempogram	CNN	0.5605	0.5471	0.5357
			(0.0846)	(0.0803)	(0.0785)
Hand-crafted Features					
N5	Top Handcrafted Features	RF	0.4005	0.4023	0.3947
			(0.0523)	(0.0424)	(0.0495)
Combined Theory-based Mid-Level Filters + Hand-crafted Features					
N6	Mel - Harmonics + Top Handcrafted Features	CNN + RF	0.5557	0.5532	0.5469
			(0.0782)	(0.0756)	(0.0746)
N6	Mel - Harmonics + MFCCs	CNN + RF	0.5583	0.5547	0.5490
			(0.0785)	(0.0764)	(0.0750)
Incorporating Positional Information					
N7	Mel - Harmonics with RNN	CNN + RNN	0.4744	0.4627	0.4192
			(0.0678)	(0.0420)	(0.0504)

Table A.2: Classification Performance by Quadrant

Features	Quadrant	Precision	Recall	F_1
Benchmark Models				
Atheoretic Filters				
Mel - Square	Q1	0.5709	0.6098	0.5800
	Q2	0.4952	0.4939	0.4789
	Q3	0.5601	0.3582	0.3879
	Q4	0.5339	0.6346	0.5509
Mel - Tall Rectangle	Q1	0.5708	0.6708	0.6029
	Q2	0.5870	0.4924	0.4912
	Q3	0.5047	0.3449	0.3998
	Q4	0.5408	0.6386	0.5839
Mel - Wide Rectangle	Q1	0.6297	0.5865	0.5960
	Q2	0.5410	0.5295	0.5201
	Q3	0.5003	0.3952	0.4342
	Q4	0.5125	0.6430	0.5662
Theory-based Low-level Filters				
Mel - Time	Q1	0.4702	0.5457	0.4771
	Q2	0.2286	0.1894	0.1769
	Q3	0.4054	0.2470	0.2720
	Q4	0.4277	0.5871	0.4838
Mel - Frequency	Q1	0.6020	0.6102	0.5890
	Q2	0.5063	0.5659	0.5151
	Q3	0.5343	0.3687	0.4103
	Q4	0.5233	0.5679	0.5363
Mel - Time-Frequency	Q1	0.5704	0.6326	0.5873
	Q2	0.5612	0.4750	0.4893
	Q3	0.5165	0.3604	0.4124
	Q4	0.5177	0.6001	0.5409
Proposed Theory-based Mid-level Filters				
Mel - Harmonics	Q1	0.6047	0.6567	0.6199
	Q2	0.5302	0.4656	0.4767
	Q3	0.5263	0.4520	0.4804
	Q4	0.5420	0.5829	0.5520

B Appendix: Details of Alternative Models

We consider a range of alternative model specifications to check for robustness of the main effects identified in the paper, and also to determine whether specific modeling features are differentially important to the task of emotion classification.

B.1 STFT Spectrograms

In our main model specification, we use the mel spectrogram, which is a transformation of the STFT spectrogram that is obtained from the raw audio data. Mel is used here because the human hearing system perceives low frequency at a higher sensitivity than high frequency audio. It is an empirical question how the use of the original STFT instead of the mel impacts performance metrics like accuracy.

To test this, we leave out the mel transformation in the main model, and we find that just the STFT in place of mel has a lower performance (Table A.1).

B.2 Temporal Features

We test two methods of capturing tempo information to see whether model performance can be further improved. We make the following design changes to the model:

1. Tempogram: A tempogram is a 2D representation of the extent to which a tempo, the rate of musical beat which is measured in beats per minute,[9] is present at each time point.[10] We incorporate the tempogram as an additional input into the deep learning model and use typical square CNN convolutional filters since theory is not required to enable explainability in the case of tempograms.[11] The deep learning model combines features learned from the tempogram and features learned from the mel spectrogram to make the emotion classification.

2. Tempo Feature: Given our interest in short six-second clips, it might be sufficient to use a few numbers to capture tempo information rather than a 2D representation. Essentially, we use the tempo with the strongest signal in the tempogram. We use Matlab's MIR Toolbox to extract two measures of tempo, one from the Fourier tempogram and one from the autocorrelation tempogram. To incorporate the tempo features into the model, we follow the approach used in the revised paper to combine our deep learning model and handcrafted features as detailed in on p. XX.

[9] Müller (2015) in his textbook writes: "Intuitively, the beat corresponds to the pulse a human taps along when listening to music." While humans might intuitively understand beats, an algorithmic approach is surprisingly challenging. As the authors says, *"simulating this cognitive process with an automated beat tracking system is much harder than one may think."*

[10] The tempogram is conceptually similar to the spectrogram, which characterizes the extent to which a periodic component corresponding to a specific frequency is present at each time point. We use an autocorrelation tempogram (rather than a Fourier tempogram) since it is known to be more sensitive to changes in signals in the onset (the beginning of a musical note) envelope, and more accurate for music with variable tempo.

[11] A high tempo is associated with high arousal and a low tempo is associated with low arousal and a typical Grad-CAM heatmap would reveal whether high or low tempos contribute to a particular emotion classification.

In both cases, the incorporation of tempo information is theory-based. Table A.1 summarizes the performance of the models. Relative to the baseline model of using the theory-based harmonics filters, we do not find that the addition of a tempogram improves performance. However, we do find that incorporating the tempo handcrafted features improves predictive performance.

B.3 CNN + RNN

Since music is inherently temporal, it is natural to think about using a deep learning architecture that deals with positional data, specifically capturing dependencies over time. Note, however, that our interest is in evaluating emotion classification for relatively short audio clips (six seconds), so it is not clear that long-range dependencies will help significantly with predictive performance.

One issue is that traditionally CNNs have been used with spatial image data and so they are not thought to be as capable of capturing temporal aspects. However, note that when audio is converted to spectrograms, the horizontal axis is time, so temporal data is already captured in the image.

We modify our mel - harmonics model to incorporate positional information by including an RNN layer. Instead of average pooling over time to summarize the consonance information over time, we use an RNN to learn temporal dependencies, as suggested by Choi et al. (2017). However, we see from the results in Table A.1 that this model results in lower classification performance than our proposed mel - harmonics model without positional encoding. This is consistent with the fact that in Choi et al. (2017) (Figure 3), they also find that RNN marginally reduces the classification accuracy for the two emotions they classify (happy, sad). Thus, it might be that the positional encoding captured by the RNN is not as good at characterizing basic emotions as CNN, according to Choi et al. (2017).

B.4 Rectangular Filters

The mel - square convolutional filters, including one 5×5 and 8 3×3 convolutional filters, are based on Chowdhury et al. (2019), who developed a music emotion classifier for the Soundtracks dataset. Square filters are designed for image data, where the x and y axes in an image are both treated the same since they represent spatial dimensions. However, in the case of music or audio more generally, the x axis represents time and the y axis denotes frequency, so they are conceptually distinct. Therefore, it is possible that square filters might not be ideal to capture the features of audio data. Thus, we generalize this by using rectangular filters that are relatively more: (a) tall and narrow (specifically, replace each $k \times k$ convolution filter by a $2k \times k$ filter), and (b) short and wide (specifically, replace each $k \times k$ convolution filter by a $k \times 2k$ filter). We make these replacements for both the 5×5 and 3×3 filters that are used in our baseline square filter model.

The results from these rectangular filters are detailed in Table **??** of the paper. We find that both the tall and narrow as well as short and wide convolutional filters outperform the square filters.

B.5 Top Handcrafted Features

We use a combination of 11 handcrafted features highlighted by Panda et al. (2018) for their ability to predict the four emotion quadrants[12] Most of the identified features are low-level features that capture tone color or timbre. Please refer to Table **??** for the categorization of low-, mid-, and high-level features. Table **??** summarizes the Top Handcrafted Features features.

We would like to combine the handcrafted features with our deep learning model that uses consonance filters while retaining explainability. There are many potential ways to incorporate handcrafted features. One strategy is to concatenate the handcrafted features with the features learned by the model before the classification step (e.g., the fully-connected layer). However, if some of the features are correlated with consonance then the model will try to learn other features besides consonance, reducing our ability to understand the model using visualizations post-training. Given our aim to develop an explainable model, we instead train the deep learning model using only the mel spectrogram as before and instead extract the features learned by the model before the final classification step. We then concatenate these learned features with the handcrafted features, and input this combination to a random forest model to make the final classification.

We find that combining the features learned using the consonance filters with the handcrafted features slightly improves predictive performance, relative to using only a deep learning model. Overall, our results suggest that there is value in combining deep learning with handcrafted features.

B.6 Different Pooling Decisions

Table B.1: Pooling Decisions for Mel - Harmonics Model

Over Time	Over Pitch Classes	Precision	Accuracy/Recall	F_1
Average Pool	Max Pool	0.5570	0.5434	0.5380
		(0.0873)	(0.0780)	(0.0791)
Average Pool	Average Pool	0.5341	0.5259	0.5118
		(0.0854)	(0.0857)	(0.0825)
Max Pool	Average Pool	0.5612	0.5300	0.5248
		(0.0702)	(0.0833)	(0.0828)
Max Pool	Max Pool	0.5392	0.5277	0.5250
		(0.0759)	(0.0784)	(0.0772)

[12] Panda et al. (2018) comprehensively considers a wide set of handcrafted features, and identifies the Top Handcrafted Featureswhich most impact emotion classification into quadrants based on valence and arousal.

C Appendix: Handcrafted Features

Table C.1: Top Music Emotion Base Features from Panda et al. (2018)

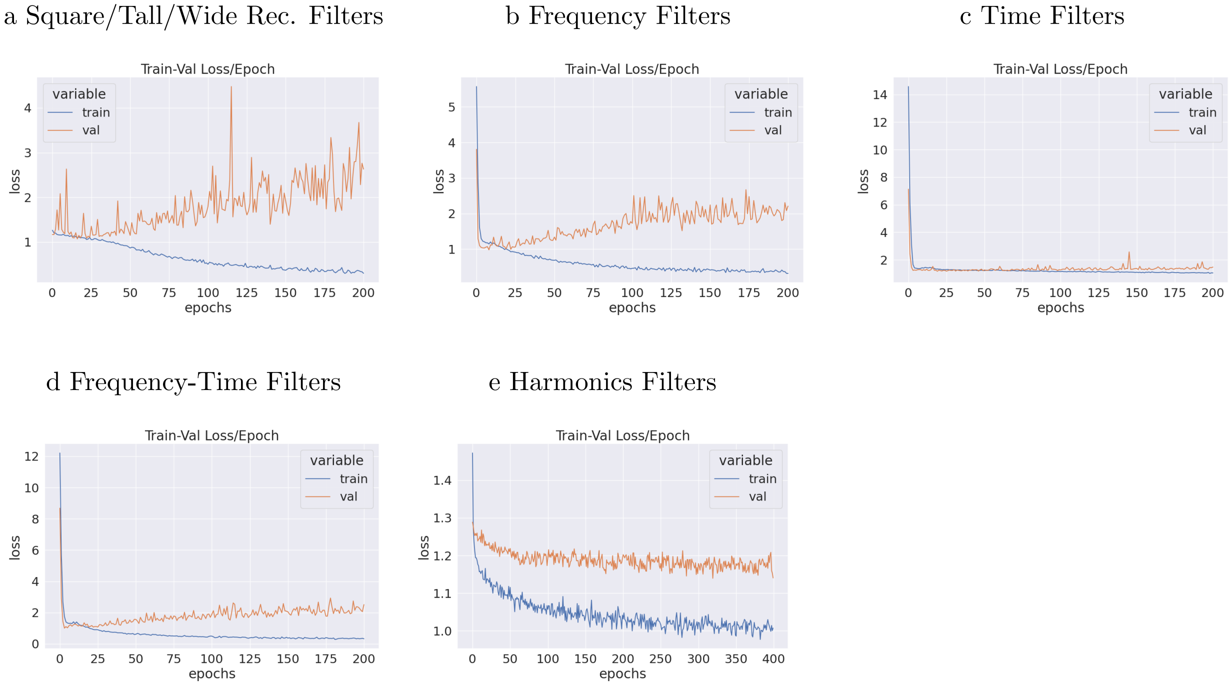
Feature in Panda et al. (2018)	Feature in MIR Toolbox	Musical Concept	Definition
FFT Spectrum - Spectral 2nd Moment (median)	Spectral Spread (median)	Tone Color	The standard deviation of the spectrum; gives a measure of the dispersion or spread of the distribution
FFT Spectrum - Average Power Spectrum (median)	Spectral Centroid (median)	Tone Color	The geometric center of the spectrum distribution can be an indicator of the “brightness” or “sharpness” of the sound
FFT Spectrum - Skewness (median)	Spectral Skewness (median)	Tone Color	The third moment of the spectrum; a measure of the symmetry of the distribution
Spectral Skewness (std)	Spectral Skewness (std)	Tone Color	The third moment of the spectrum; a measure of the symmetry of the distribution
Spectral Skewness (max)	Spectral Skewness (max)	Tone Color	The third moment of the spectrum; a measure of the symmetry of the distribution
MFCC1 (mean)	MFCC1 (mean)	Tone Color	MFCC offers a description of the spectral shape of the sound
MFCC1 (std)	MFCC1 (std)	Tone Color	MFCC offers a description of the spectral shape of the sound
Roughness (std)	Roughness (std)	Tone Color	An estimation of the sensory dissonance
Rolloff (mean)	Rolloff (mean)	Tone Color	The amount of high frequency in the signal; consists in finding the frequency such that a certain fraction of the total energy is contained below that frequency
Spectral Entropy (std)	Spectral Entropy (std)	Tone Color	Shannon entropy offers a general description of the spectral power distribution
Fluctuation (std)	Fluctuation (std)	Rhythm	Estimates the rhythm content based on spectrogram computation transformed by auditory modelling and then a spectrum estimation in each band

Note: We use Matlab’s MIR Toolbox to extract these features. Some of the features in Panda et al. (2018) Table 6 were initially extracted using Marsyas and PsySound3, but these softwares do not appear to have been maintained and are not usable with currently available versions of software. These software compatibility issues therefore prevented us from using them. We use the features from Matlab’s MIR Toolbox that we believe are the most similar to those directly used in Panda et al. (2018) and currently unavailable.

D Appendix: Deep Learning Training Loss

The figures below plot the training and validation loss of one cross-validation fold over training epochs. In general, the plots for the remaining folds look similar to the ones shown below. The square and rectangular images follow the same pattern and so we show the loss from the wide rectangle filters. The CNN models that use square, rectangular, frequency, time, and frequency-time filters learn the "best" model fairly quickly (within 25-50 epochs) while the CNN model that uses harmonics filters learns more gradually.

Figure D.1: Model Training Loss Over Training Epochs



a Square/Tall/Wide Rec. Filters

b Frequency Filters

c Time Filters

d Frequency-Time Filters

e Harmonics Filters

E Appendix: Application Details

E.1 Application Scaling Feasibility

Using 6 ad insertion slots (at intervals of approximately every minute within the content video), we obtain $N_{cells} = 4 \times 4 \times 6 = 96$ cells. In each cell, we recruit $N_{users-per-cell} = 15$ users to obtain performance metrics (of skip rate and brand recall rate), so we need $N_{users} = N_{cells} \times N_{users-per-cell} = 1,440$ users in total. The experiment costs \$2.40 per respondent (user), implying a total cost of $\$3,500$. Using multiple insertion points is essential for our model, since the value of our method is that it can be used with time-varying emotion over the content video. Scaling our approach to 30 videos and 6 ads with 6 insertion points would increase the cost to roughly \$39,000. We are positioning our application more as a proof-of-concept, rather than a separate application paper. The application is intended to show that: a) music emotion has the potential to be useful in identifying both the timing and selection of ad to insert within a content video and b) using our explainable method obtains a relatively good performance (in terms of skip rate and brand recall rate) relative to benchmark methods.

E.2 Incorporating Emotion Data from Images and Text

Although music is a key driver of emotion in video, other features like text (what is said), voice tonality (how it is said), and images can also influence emotion. We use Microsoft Azure to try to extract the emotion related to these other features to explore the impact of incorporating emotion data from other modalities of the video beyond the background audio. Azure's Video Indexer emotion detection algorithm predicts emotion from speech text and voice tonality.[13] The list of possible emotions includes joy, fear, anger, and sadness. Azure's Face API[14] detects emotion based on facial expressions from images.[15] The Face API treats the emotion prediction task as a multiclass problem so the probabilities over the eight possible classes (anger, contempt, disgust, fear, happiness, neutral, sadness, surprise) sum to one.

We find that Azure's emotion predictions obtained a different emotion classification from our collected human-tagged emotion. For all ads, the Video Indexer predicted no emotion for the first six seconds of each ad. This is likely due to the fact that Azure's emotion detection algorithm relies on speech text and voice tonality, and there is very little speech in the first six seconds of each ad. Given the high prevalence of music in video ads, this observation highlights an opportunity for audio emotion detection models to go beyond speech and include music. In terms of the images, the Face API generally found the faces to be neutral. Figure E.1 shares the first frame extracted from each ad. In all of these frames, the highest predicted emotion was always neutral. Figure E.2 summarizes the facial emotions predicted for the first six seconds of each ad.

For the content videos, the Video Indexer predicted no speech emotion for two of the content videos since they do not contain speech. The first row in Figure E.3c shows the distribution of speech emotion for the remaining two videos. Each distribution is defined as the average over

[13] https://learn.microsoft.com/en-us/azure/azure-video-indexer/video-indexer-overview

[14] https://learn.microsoft.com/en-us/xamarin/xamarin-forms/data-cloud/azure-cognitive-services/emotion-recognition

[15] Starting June 21, 2022, Azure customers will need to apply for access to use the Face API due to concerns around accuracy for specific demographic groups.

Figure E.1: Ad First Frames

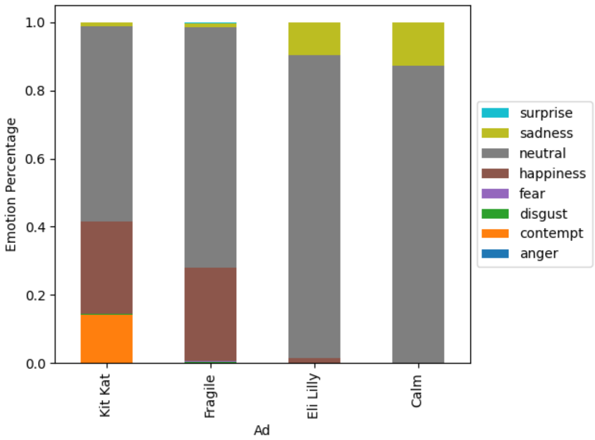
a Kit Kat

b Fragile Childhood

c Eli Lilly

d Calm App

Figure E.2: Ad Facial Emotion Distribution



the 30 seconds prior to the ad insertion time. The third row shows the human-tagged emotion distributions for comparison. For Run With Me, Segment 3 aligns with the human-tagged emotion but Segments 1 and 2 less so. For Unspoken, Segments 4 and 5 are identified to be negative by the Video Indexer as well as by humans.

The Face API predicted no emotion for two of the content videos since they do not contain human faces but instead contain animated animal faces. The second row in Figure E.3c shows the distribution of facial emotion for the remaining two videos. Each facial emotion distribution is defined as the average of the facial emotion distributions over the 30 seconds prior to the ad insertion time. For the most part, the faces are predicted to be either neutral or happy. For Run With Me, we see that there are some similarities with human-tagged emotion in that Segments 3, 4, and 6 are higher in facial happiness and higher in Q1 - High Valence - High Arousal.

Given the little speech present in the first six seconds of the four ads, we cannot use the Video Indexer for speech emotion in our setting. We can, however, use the Face API in combination with the music emotion classifiers to determine the optimal emotion-based ad insertion point for the two content videos Run With Me and Unspoken. We calculate the JS distance based on face emotion and the JS distance based on music emotion for each ad insertion and ad combination and sum the two distances to determine which ad insertion point is the most emotionally similar for each ad and content video combination.[16] Then following the same procedure used with music emotion we

[16] The correlation between the music emotion-based JS distance and the face emotion-based JS distance is 0.153. This low correlation suggests that the face emotion model captures information not present in the music emotion model.

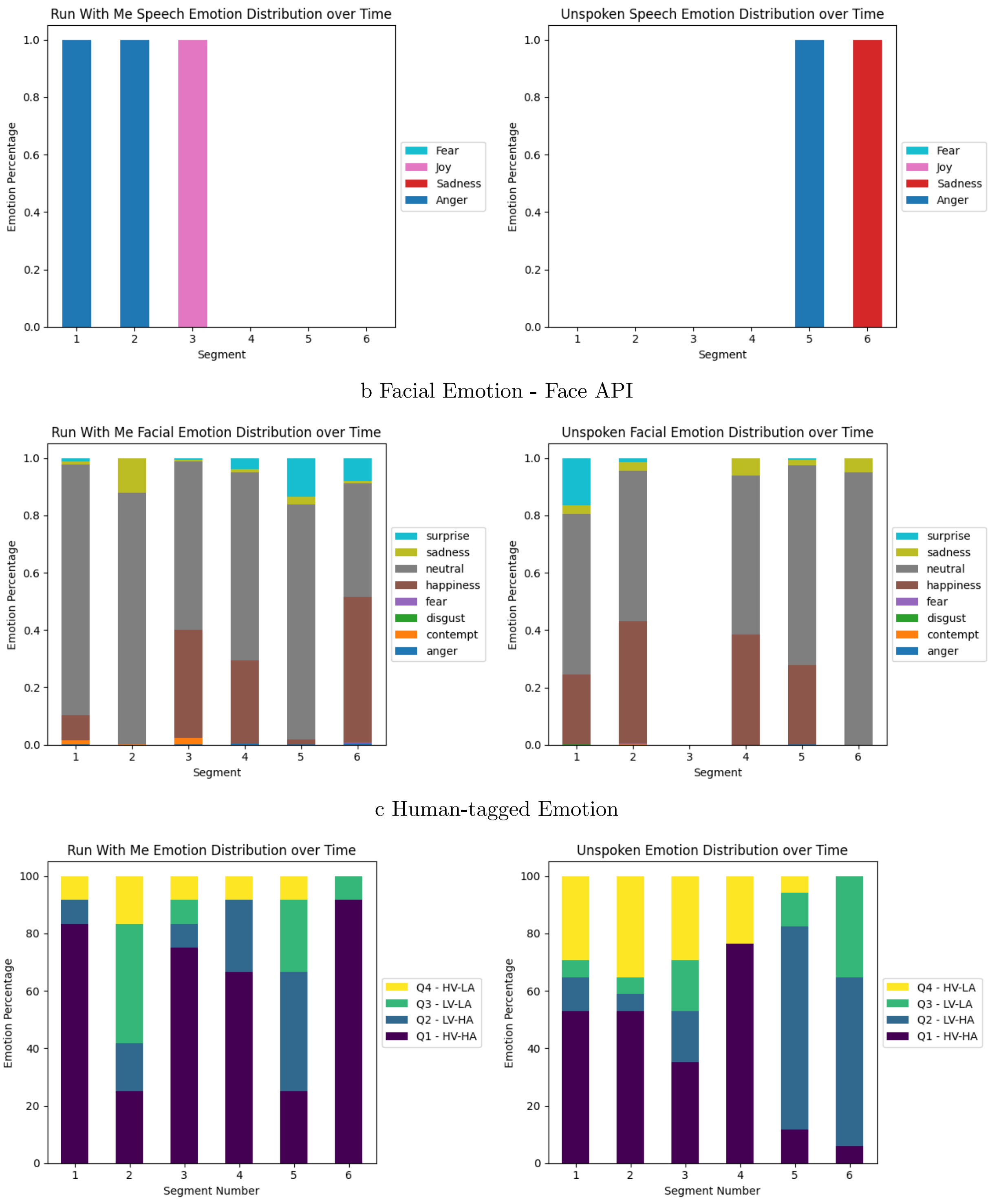
calculate the average human-tagged JS distance, skip rate, and recall rate for each model. Table E.1 compares these measures from using face and music emotion versus using only music emotion for the two content videos Unspoken and Run With Me.

Table E.1: Face and Music Emotion vs. Only Music Emotion - Unspoken and Run With Me

Feature	Model	JS Distance	Skip Rate	Recall Rate
Face and Music Emotion				
Mel - Harmonics	CNN	0.459	46.1%	50.0%
Mel - Harmonics + Top Handcrafted Features	CNN + RF	0.417	43.3%	50.3%
Music Emotion				
Mel - Harmonics	CNN	0.484	44.5%	43.6%
Mel - Harmonics + Top Handcrafted Features	CNN + RF	0.401	39.0%	50.1%

Including face emotion improves the recall rate when used in combination with the mel - harmonics model but hurts the skip rate when used in combination with the mel - harmonics + Top Handcrafted Features model. Overall, there is potential in incorporating emotion information from images and text but the existing tools are limited in their ability to extract emotion information from short clips (i.e., first six seconds of ads) and animated videos. The results suggest that audio models like the one included in Azure's Video Indexer service could benefit from incorporating music emotion classification.

Figure E.3: Content Speech and Facial Emotion Distribution



a Speech Emotion - Video Indexer

b Facial Emotion - Face API

c Human-tagged Emotion