A Theory-Based Interpretable Deep Learning Architecture for Music Emotion

Hortense Fong, Vineet Kumar, K. Sudhir
Yale School of Management
hortense.fong@yale.edu, vineet.kumar@yale.edu, k.sudhir@yale.edu

September 1, 2021

Music is used extensively to evoke emotion throughout the customer journey. This paper develops a theory-based, interpretable deep learning convolutional neural network (CNN) classifier—MusicEmoCNN—to predict the dynamically varying emotional response to music. To develop a theory-based CNN, we first transform the raw music data into a format—mel spectrogram—that accounts for human auditory response as the input into a CNN. Next, we design and construct novel CNN filters for higher-order music features that are based on the physics of sound waves and associated with perceptual features of music, like consonance and dissonance, which are known to impact emotion. The key advantage of our theory-based filters is that we can connect how the predicted emotional response (valence and arousal) are related to human interpretable features of the music. Our model outperforms traditional machine learning models and performs comparably with state-of-the-art black-box deep learning CNN models. Our approach of incorporating theory into the design of convolution filters can be taken to settings beyond music. Finally, we use our model in an application involving digital advertising. Motivated by YouTube's mid-roll advertising, we use the model's predictions to identify *optimal emotion-based ad insertion positions* in videos. We exogenously place ads at different times within content videos and find that ads placed in *emotionally similar contexts are more memorable* in terms of higher brand recall rates.

Key words: emotion, deep learning, interpretable AI, music theory, digital advertising

---

Click here for most recent version.

1. Introduction

Emotions play a central role in many elements of marketing and consumer behavior, such as consumer choice, advertising response, customer satisfaction, and word of mouth (e.g., Holbrook and Hirschman 1982, Bagozzi et al. 1999, Huang 2001, Laros and Steenkamp 2005). Though emotions research has primarily been in lab settings with self-report data (Cohen et al. 2018), the recent availability of novel data (e.g., biometrics, real-time consumption and social media content) and advances in machine learning that allow for the inference of emotion induced by speech, music, video, and text now offer novel possibilities to measure and study consumer emotions in field settings (Du et al. 2021). In fact, "affective computing" is now an important sub-field of computer science.[1]

Music is widely regarded as among the most effective and efficient of channels to influence emotion; it is often called the *language of emotion* (Corrigall and Schellenberg 2013). As such, it is used extensively to evoke emotion all along the customer journey, from need recognition to purchase, in advertising, content marketing, and physical stores (Gorn 1982, Vermeulen and Beukeboom 2016, Krishna et al. 2016). The use of music in advertising is almost universal and advertising creators spend considerable effort crafting it to elicit the desired emotion.[2] Given the strong links between music and emotion, mapping music into a dynamic sequence of emotions can be extremely valuable in many applications. For example, marketers can use it to design ads and match content with music on the basis of emotion. With online programmatic advertising, the ability to automate emotion-based contextual matching of ads and content at scale over hundreds of millions of ads and content can be extremely valuable (Shukla et al. 2017). Further, on music and video platforms, it can be used to improve the automation of developing mood-based playlists and "next song" recommendations for users from their large catalogs.[3]

Given the above background, we develop an *interpretable, theory-based deep learning convolutional neural network (CNN) model* for music emotion, MusicEmoCNN (pronounced

[1] The term, introduced by Picard (2000), involves systems that can recognize, interpret, process, and simulate human affect (emotion).

[2] A content analysis of over 3,000 ads showed that 94% use music (Allan 2008). Further, over 75% of advertising hours in broadcast media uses music in some form (Huron 1989). Ad creators also spend considerable effort in crafting advertising music to generate a desired emotion and marketing outcomes. As Huron (1989) states: "on a second-for-second basis, advertising music is the most meticulously crafted and most fretted about music…"

[3] As of 2021, YouTube has over 2 billion users, spending over 2 billion hours per day. Spotify has 356 million users whose playlists span over 70 million tracks.

MusicEmotion), that maps a music clip to the sequence of emotions that it evokes.[4] Our modeling contribution is in two distinguishing features of our CNN model. The first is the use of music theory to construct the model; we use the theory to *engineer appropriate inputs* to the model and then *design filters to capture theoretically known constructs* that link music and emotion. Such filter construction turns out to be challenging, because current CNN implementations for music are based on adaptations of models developed for computer vision where *spatial contiguity* is meaningful. Since this is not true for musical constructs, filters inspired from computer vision and adapted to model music emotion are often inadequate.

The second distinguishing feature is the interpretability of the CNN and its predictions. Deep learning models are typically black-box predictors, but for managers to develop trust in the model and adopt it at scale, the predictions need to be interpretable, such that they can see a clear connection between conceptually well understood music characteristics and predicted emotion. We build in interpretability in the model construction stage and obtain interpretability in the post-estimation visualization stage. In the model construction stage, we design filters for higher-order constructs like consonance and dissonance that have theoretical meaning based on acoustic physics and human auditory response and have well-established empirical links to human emotion. Consonance refers to a combination of notes that sound pleasant when played simultaneously and dissonance refers to a combination of notes that sound harsh or jarring when played simultaneously (Müller 2015). In the post-estimation stage, we construct visualizations adapted from recent advances in the visualization of deep learning models for computer vision to show the relationship between our theory-motivated filters and emotion. Together, this allows us to go beyond prediction, and provide interpretable output that links music characteristics with predicted emotion.

Finally, we illustrate the practical value of our model using an application motivated by YouTube's mid-roll video ads. YouTube serves tens of millions of video content pieces a day; within each of these content pieces, multiple ads can be placed. Inserting ads in effective locations of the video requires automation. We conjecture that the appropriate matching of the ad's emotion with the emotion in video content reduces ad skipping and

[4] We characterize emotion using the well-established and widely used valence-arousal framework developed by Russell (1980), but our model can easily be adapted for other emotion frameworks. It may be also useful to draw a parallel to sentiment analysis in text, where each sentence generates a particular sentiment around an attribute. See for example Büschken and Allenby (2016), Chakraborty et al. (2021).

increases recall. We insert ads at various locations that vary in the emotional similarity between the ad and video content to assess the effectiveness of ad-content insertion based on emotional similarity.

We provide an outline of the modeling details of MusicEmoCNN to help build intuition for what the critical challenges are in (i) constructing a theory-based CNN model for predicting music emotion; and (ii) constructing a model that is also interpretable. Our methodological approach uses the raw sound wave (music clip) as the starting input. The audio at each time point is the composite of the amplitudes (sound volume) of several underlying frequencies. Sound waves can be represented in acoustics by a two-dimensional image with frequencies along the y-axis and time along the x-axis, and the square of the amplitude corresponding to the frequencies represented in color. This 2D representation is called a *spectrogram* and is useful because it allows us to "read" some of the clip's musical features, such as pitch range and tempo. A mathematical operation called the short-time Fourier transform (STFT) allows us to convert the raw sound wave into its STFT spectrogram. However, even though the STFT spectrogram faithfully represents the sound, it does not account for the complexities of human hearing. The human ear does not perceive differences within small bands of frequencies, and differentiates much more between lower frequencies than between higher frequencies, as in a log scale. We therefore transform the STFT spectrogram to a mel spectrogram to reflect the fact that humans perceive frequency on a log scale. Recent research developing CNN for music uses mel spectrograms as input (e.g., Pons et al. 2016, Chowdhury et al. 2019). We also use the mel spectrogram as input into a CNN.

CNNs are built for image processing, where objects and shapes are contiguous across both x- and y-dimensions, which have spatial meaning based on physical reality. Convolution filters play a critical role in determining the performance of CNNs and filters designed for image processing take advantage of contiguity to perform effectively. However, spectrograms generated from music audio are not like regular images in that the x-axis represents time and the y-axis represents frequency. In spectrograms, non-contiguous regions in the frequency space impact the perception of music, in particular consonance and dissonance, which are associated with emotion evoked by music. For example, the simultaneous playing of an octave (e.g., A4 (440 Hz) and A5 (880 Hz)) produces a consonant sound while the simultaneous playing of a tritone (e.g., A4 (440 Hz) and D5 (587 Hz)) produces a dissonant

sound. We can only capture such constructs using non-contiguous filters, highlighting the importance of incorporating domain knowledge into the design of deep learning models. We develop novel non-contiguous filters that specifically highlight the frequencies of interest and integrate them into the CNN. We use mel filter banks[5] to transform the non-contiguous filters, which capture specific mathematical relationships, so that they can be applied to the mel spectrogram.

While deep learning models have been successful in prediction and automation, a key concern with such models is interpretability. To build trust as AI is integrated into automating decisions, models need to go beyond prediction and become "transparent" by explaining why they predict what they predict. For example, Grad-CAM (Selvaraju et al. 2017) is a tool that uses a heatmap to visualize which areas of an image contribute most to the classification of the image into a specific class. If the Grad-CAM heatmap highlights the areas of an image that correspond to large ears and a trunk while classifying an elephant, we would characterize the model as transparent and interpretable because we can understand why the model made the classification. However, if the prediction is based on spurious correlations, its generalizability would be questionable. For example, suppose a CNN trained to distinguish dogs from wolves in pictures uses the presence of snow as a primary predictor, since pictures of wolves disproportionately have snow whereas those of dogs have grass. If this spurious correlation is the basis for prediction, a new example that has a dog playing in snow could well be classified incorrectly as a wolf (Ribeiro et al. 2016). A model that is able to provide transparent and trustworthy explanations for its predictions is more likely to be useful and widely adopted.

In the context of music, for a model to be interpretable, it should relate the top-level label of interest (e.g., emotion) to a mid-level set of features (e.g., harmony, rhythm, pitch) related to music.[6] While low-level features (e.g., frequency, time) provide some degree of transparency, Fu et al. (2010) argues they do not have a clear interpretable link to top-level labels. To overcome this challenge, we identify mid-level music features closely connected to emotion to motivate the design of our convolution filters. Combining these mid-level features with a model visualization technique enables interpretability of the CNN by connecting a feature with clear musical meaning to the emotion classification. We design

[5] Mel filter banks map linear frequencies to the mel scale. See Section B in the Appendix for details.

[6] Table C1 in the Appendix organizes features used in music classification by level of interpretability.

the convolution filters to clearly show which parts of a song relate to the mid-level feature of consonance and the final emotion classification. Our paper highlights that a strategy for interpretability is to think about constructs whose links with the classification have clear theoretical motivation and/or are managerially actionable.

Summarizing, our key contributions are as follows. First, we develop a theoretically-motivated interpretable deep learning framework that allows us to model and predict dynamically varying emotional responses throughout the duration of a music clip. Existing music emotion classifiers can be grouped into two types based on interpretability and accuracy. Classifiers that use hand-crafted features and more traditional machine learning methods are typically more interpretable but less accurate while classifiers that use data-driven features and deep learning methods are typically less interpretable but more accurate. Our proposed classifier is not only accurate, but also interpretable through the incorporation of theory into the design of the model. Second, our approach integrates theoretically motivated features of music, like consonance and dissonance, that are known to impact emotion. The deep learning literature has not examined these features obtained from the physics of sound waves. We design CNN filters for consonance and dissonance that capture not just how they are mathematically represented in harmonics, but also how they are perceived by human auditory processing. Third, we provide an interpretation of the specific musical features learned and their connection with emotion. Finally, we demonstrate an application that examines the impact of matching the emotional content (valence and arousal) of a video advertisement with that of the content video. More generally, we note that our conceptualization of filter design for deep learning in terms of theoretically or managerially relevant constructs aids with more robust prediction as well as greater interpretability of deep learning models in domains outside of music.

The rest of the paper is structured as follows. Section 2 overviews the relevant literature. Section 3 discusses the model and the deep learning pipeline, which takes music audio waves as input and outputs a measure of emotion based on valence and arousal. Section 4 describes the details of the model training. Section 5 describes an application of MusicEmoCNN for ad insertion in online videos. Finally, Section 6 concludes with a discussion of limitations and avenues for future research.

2. Related Literature

Our paper builds upon several streams of literature across different academic fields. We organize our discussion in three sections: (1) Listener Response to Music; (2) Machine Learning with Audio Data; and (3) Explainable and Interpretable AI.

2.1. Listener Response to Music

Music creates feelings and induces emotion, as shown by a wide literature using a range of methods from surveys to brain scans (Johnson-Laird and Oatley 2016, Juslin and Laukka 2003). To measure emotion, researchers have often used Russell's circumplex model (e.g., Yang and Chen 2011b). Certain musical structure settings, such as fast tempo and loud music, are related to certain valence and arousal settings. For example, fast tempos are associated with high arousal (energetic) music while slow tempos are associated with low arousal music. Similarly, other features like loudness, timbre, pitch and harmony are also associated with valence and arousal (Gabrielsson and Lindström 2010, Gabrielsson 2016).[7] While our paper focuses on emotion, music also impacts listeners through other mechanisms, such as through music associations (e.g., classical music associated with an upscale setting) and perceptions of time (North and Hargreaves 2010).

These listener responses have marketing implications, and there is a substream of literature that focuses on marketing outcomes like ad impact and even sales, based on music characteristics. For example,Yang et al. (2021a) use low-level acoustic features to predict ad audio energy levels and find that energetic commercials are more likely to be watched for longer. Boughanmi and Ansari (2021) use a Bayesian nonparametric approach to predict album sales using multi-modal data that includes high-level audio features of music in the songs of the album. They use structured music data produced by Spotify like danceability, energy, etc. and therefore do not use or require the raw high-dimensional audio data.

Our research can be broadly viewed as connected to the literature on sensory marketing, and music is a crucial way to achieve impact through the auditory sense (Krishna 2012, Krishna et al. 2016).[8] Music elicits different moods, impacting ad outcomes (Bruner 1990, Alpert and Alpert 1990). Certain types of music can also have an attentional impact, contributing a different mechanism for ad effectiveness (Bruner 1990). Others have found

[7] See Table A1 in the Appendix for definitions of the musical structures.

[8] To quote Krishna et al. (2016): "Sensory marketing can be defined as marketing that engages the consumers' senses and affects their perception, judgment and behavior."

that moving music can also have a persuasive impact (Strick et al. 2015), and that repetition can be especially helpful (Ward et al. 2014). Consumer behavior in commercial settings has been found to be impacted by music type (North and Hargreaves 2010, North et al. 1997). Musical properties associated with high emotional arousal have been found to increase the speed with which customers shop (Smith and Curnow 1966), eat (Milliman 1986), and drink (McElrea and Standing 1992). Overall, music has a variety of implications for marketing outcomes, motivating our digital advertising application studied in Section 5.

2.2. Machine Learning with Unstructured Audio Data

The second stream of literature is machine learning from unstructured audio data. For a detailed survey, see Fu et al. (2010). While audio broadly includes both speech and music, our focus in this paper is on the latter. Traditional machine learning methods like SVM and Gaussian mixture model-hidden Markov model (GMM-HMM) previously produced high classification performance in many settings by using hand-crafted features (also known as "feature engineering"). Typical examples of such features include mel frequency cepstral coefficients (MFCCs), which were originally developed for speech recognition but have also demonstrated success in other tasks (Tiwari 2010). However, the performance of deep learning algorithms have overtaken almost all other methods in audio applications, similar to vision applications (Hinton et al. 2012). The crucial advantage that deep learning methods have is that features are automatically learned from data, rather than pre-specified (Choi et al. 2017). With deep learning methods, audio is typically converted to a visual representation called a spectrogram and used as the input to the learning algorithm (see Section 3 for details).

A few researchers have attempted to build music-specific classifiers. To predict the ballroom music genre, Pons et al. (2016) suggest using musically-motivated CNN filters to capture low-level timbral and temporal elements of music. This translates to using various rectangular convolutional filters—tall and skinny filters for timbral elements and short and wide filters for temporal elements. Others have designed deep learning models to predict mid-level features in a data-driven fashion, replacing human-designed transformations that have been proposed. For example, Elowsson and Friberg (2019) design a CNN with pitch chroma input to predict music modality and achieve state-of-the-art prediction performance. Dubois et al. (2019) propose a deep learning model to predict sensory dissonance in piano chords. Chowdhury et al. (2019) build a deep learning model that includes an

interpretable mid-level feature layer, including features such as melodiousness and articulation, which are used to predict emotion. We contribute to the machine learning from audio literature by designing and developing CNN filters that characterize consonance and dissonance, and demonstrate how they impact emotion and are critical to prediction as well as interpretation.

2.3. Explainable and Interpretable AI

Machine learning and AI methods, in particular deep learning, have often been regarded as black boxes that provide excellent predictive performance but are uninterpretable. Humans often cannot understand why algorithms make the predictions they make (LeCun et al. 2015, Castelvecchi 2016). The challenge with deep learning is that the models often feature millions of parameters, much more than the number of data points, which makes them particularly opaque (Doshi-Velez and Kim 2017).

There are a number of reasons why it is important to have explainable or transparent learning methods. First, without transparency it is challenging to trust AI systems, since we do not understand when and how they may break; the example of the dog in grass and wolf in snow in the introduction highlighted the potential pitfalls of not having transparency. A second concern is that deep learning may often just memorize data, some of which could be correlated across training and test data (Arpit et al. 2017, Dhar et al. 2019). A third is reproducibility, since these algorithms are known to be quite sensitive to hyperparameter tuning (Bergstra and Bengio 2012, Bergstra et al. 2011). Researchers need to ensure that the model's predictions and explanations are robust to variations in these parameters. Fourth, the ability to perform "what if" kind of analyses of deep learning requires a strong degree of generalization so that when we move outside the scope of the data the methods still perform well. Finally, the issue of bias in AI and ML is a very active line of inquiry, examining issues of fairness all the way from axiomatic principles (Honegger 2018) to challenges with current state-of-the-art implementations (Buolamwini and Gebru 2018, Raji and Buolamwini 2019).

We take a different approach to interpretability by combining music theory domain knowledge with a gradient visualization technique, gradient-weighted class activation mapping or Grad-CAM (Selvaraju et al. 2017). Grad-CAM increases the model transparency of CNNs by producing visual explanations. Our proposed consonance and dissonance filters place an interpretable structure on what the CNN sees, allowing us to make sense of the

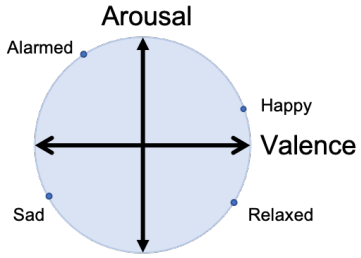
model. The key implication that arises is that filter design must be done thoughtfully in advance in order to ensure interpretability, and not just ex post after training the model.

3. Model

We develop a deep learning model based on convolutional neural networks for emotion classification that includes several theoretically motivated components relating to the physics of sound waves and the perception of music by listeners. For a detailed definition of the music terms used, please see Table A1 in Section A of the Appendix.

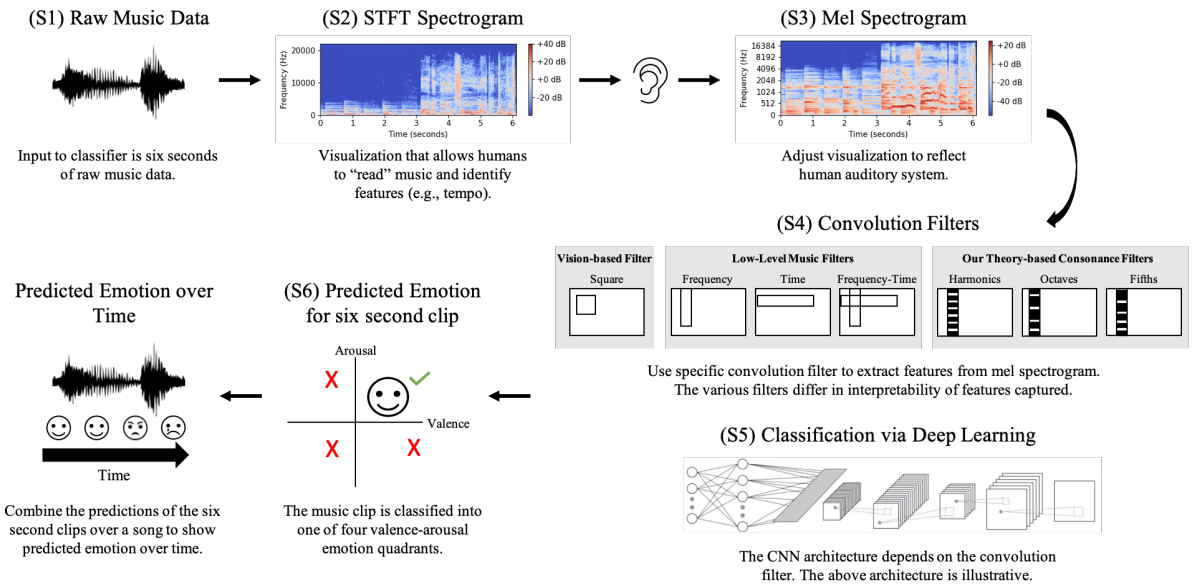
We characterize emotion using the valence-arousal circumplex model represented in Figure 1, based on Russell (1980). Valence measures how positive or negative a listener feels and higher valence maps to a more positive feeling. Arousal measures how energetic a listener feels and higher arousal maps to greater excitement and energy. Discrete emotions such as "happy," "sad," "relaxed," and "alarmed" can be mapped onto the valence and arousal dimensions. Both the discrete and dimensional models have been used in the literature. The main drawbacks of the discrete model are that researchers have yet to reach a consensus on the appropriate level of emotional granularity for music and there is ambiguity in language (Yang and Chen 2011b). In contrast, the valence-arousal model implicitly offers an infinite number of emotion descriptions and as a result, many researchers have adopted the valence-arousal framework for emotion classification (e.g., Panda et al. 2018, Yang and Chen 2011a, MacDorman 2007, Korhonen et al. 2006).

Figure 1 Russell's Circumplex Model of Emotion



We begin with an overview of the steps of our deep learning model that maps music to emotion in Figure 2. Step **S1** takes six seconds of raw audio sound wave data as input. In Step **S2**, the music clip is converted to a short-time Fourier transform (STFT) spectrogram, which is a visual representation of the frequencies present in the sound wave. In Step **S3**, we transform the STFT spectrogram to a mel spectrogram, which characterizes how the

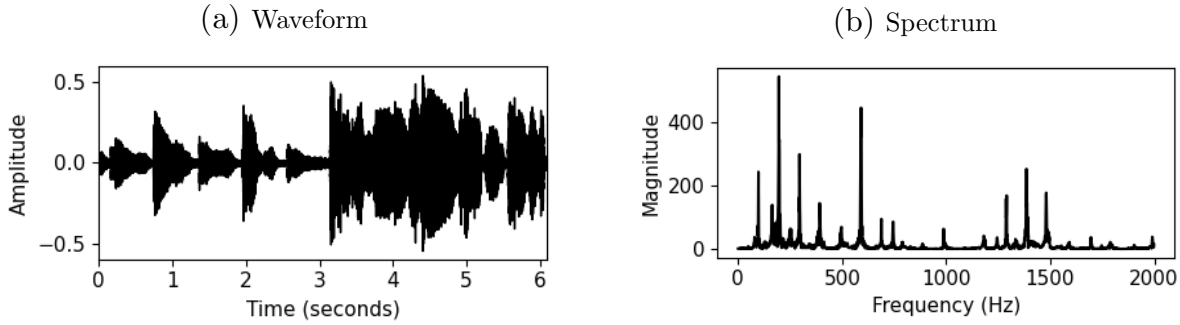
Figure 2 Music Emotion Classification Schematic



sound is perceived by the human ear. In Step **S4**, the mel spectrogram is used as a visual input to the deep learning convolutional neural network (CNN) with one of the convolution filter types. We use a number of theoretically motivated filters to reflect aspects of music that we expect to impact listener emotion, as well as an atheoretical square filter that is commonly used in image processing. In Step **S5**, the input from **S4** is put through the remaining layers of the CNN. In Step **S6**, the CNN generates a classification prediction for the six second sound clip, indicating the quadrant of the dimensional model for which the sound is closest. Finally, the model combines the predictions of the six second clips and shows the emotion predictions of the music over time. Below we describe each of these steps in detail.

(S1) Physical Properties of Sound Waves: The reader may recall that music (or any sound) is a pressure wave that travels through the air until it reaches the listener's ear. The waveform illustrated in (S1) of Figure 2 graphs the change in air pressure at a certain location over six seconds (Müller 2015). Audio data can be represented in a number of ways. While a waveform is one way to visually represent sound, it does not model how humans hear. Two different waveforms can sound the same wherein it is difficult to discern from it musically relevant features which relate to emotion, such as pitch and rhythm.

Figure 3 Example of Waveform and Spectrum



Notes: (a) Waveform of six seconds of New Soul by Yael Naim. A waveform represents audio data by graphing the change in air pressure at a certain location over time. Amplitude measures the deviation of the air pressure from the average air pressure. (b) Spectrum of the first second of the waveform. A spectrum graphs the magnitude of the frequencies that compose the waveform, showing what the sound is made of.

To get to musically relevant features, we need a representation of the different frequencies that the sound wave is composed of in terms of fundamental sine waves. This is because sine waves determine what humans hear and are at the foundation of musical concepts like pitch and harmony. The mathematical representation of this process is the Fourier transform, which decomposes a sound wave into its constituent sine waves.[9] Any sound wave can be represented as a combination of sine waves of different frequencies, amplitudes, and phases, known as the *partials* of the sound wave. The complete set of partials makes up the *spectrum*. Figure 3b shows the spectrum of the first second of Figure 3a. From the spectrum, we can identify the main frequencies that make up the sound and the amplitude of each frequency.

(S2) Short-Time Fourier Transform Spectrogram: A spectrogram visualizes frequency and time features of audio data (Müller 2015). The fundamental spectrogram is the short-time Fourier transform (STFT) spectrogram, which is produced by taking the Fourier transform of short overlapping time windows of the waveform, decomposing a sound wave into its individual frequencies and their respective magnitudes. The STFT maps the squared magnitude of each frequency over time.[10]

[9] Sound waves cause the eardrum to vibrate. The basilar membrane in the cochlea wiggles in response and different frequencies cause different parts of the basilar membrane to respond, determining what we hear. The Fourier transform captures this process.

[10] To operationalize this procedure, we first digitize the analog audio signal by sampling from the signal since we are working with digital technology. The sampling rate represents the number of samples taken per second and is measured in Hertz. The optimal sampling rate depends on the context. We will use a sampling rate of 44,100 Hz, which is also used for CD recordings, since it generates an STFT spectrogram that covers the range of human hearing, which spans from roughly 20 Hz to 20,000 Hz (Müller 2015). Since time has been discretized, we now measure time

The parameters that go into generating an STFT spectrogram are the sampling rate, window type and size, and hop length.[11] Let x represent the discrete-time signal of the audio signal, w the window function, which takes in N samples, and H the hop size. The window function specifies how we weight the audio signal within each window of time and the hop size specifies how many samples we jump between each window. The discrete STFT X of signal x is:

$$X(m,k) \coloneqq \sum_{n=0}^{N-1} x(n+mH)w(n)exp(-2\pi ikn/N), \tag{1}$$

where m is the time index, $k \in [0:K]$ is the frequency index, and $i \coloneqq \sqrt{-1}$. The largest frequency index is $K = \frac{N}{2}$ because higher frequency indices are redundant, which is why a sampling rate of 44,100 Hz generates a spectrogram that extends up to 22,050 Hz.

The STFT spectrogram can then be written as:

$$S(m,k) \coloneqq |X(m,k)|^2. \tag{2}$$

The magnitude of the complex number $X(m,k)$ captures the presence of each frequency at each time sample. Squaring the magnitude yields the power of each frequency at each time sample. We generate an STFT spectrogram for each six second clip of music. The resulting frequency $\times$ time dimensions of the STFT spectrograms are $1,025 \times 517$.

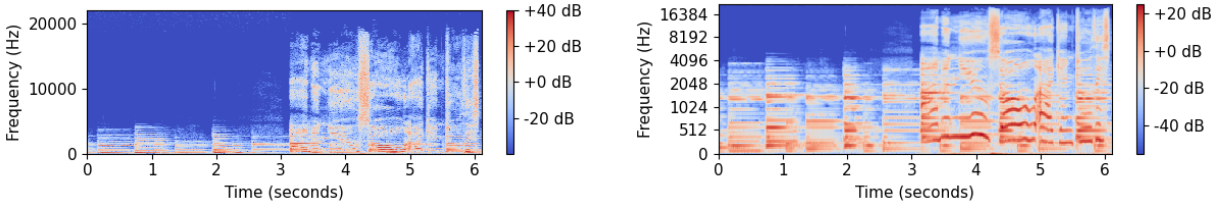
The STFT spectrogram is represented in Figure 4a, with the x-dimension representing time, the y-dimension representing frequency, and color representing the power of each frequency bin at each time sample. Note that the frequency and time dimensions are discretized since we are working with a digital signal. In the STFT spectrogram, frequency and time are shown on linear scales while power is shown on a log scale and measured in decibels (dB) since humans perceive volume on a log scale. By using a log scale, small intensity values of relevance are visible to a human reader. Higher intensity values are shown in red while lower intensity values are shown in blue. In Figure 4a, the large patch of blue before the third second indicates a lack of high frequencies early in the music clip.

in terms of sample number rather than in seconds. Each second contains 44,100 samples. The sampling theorem states that if an analog signal contains no frequencies greater than half the sampling rate, then it can be perfectly reconstructed from its sampled version. Frequencies greater than half the sampling rate suffer from aliasing, where some frequency components become indistinguishable from each other. As a result, we limit the frequencies to half the sampling rate.

[11] We set the sampling rate to 44,100 Hz, the window type to Hann, the window size to 2,048 samples, and the hop length to 512 samples, which are standard choices in the literature (Müller 2015). A Hann window is a bell-shaped window that places more weight on the center of the window and less weight on the edges of the window.

Figure 4 Spectrograms

(a) STFT Spectrogram (b) Mel Spectrogram



Notes: (a) Short-time Fourier transform (STFT) spectrogram of six seconds of New Soul shown in 3a. The STFT spectrogram visualizes the time and frequency features of audio data. The x-axis represents discretized time, the y-axis represents discretized frequency, and color represents the squared magnitude of each frequency bin over time. It enables one to "read" musical features, such as the range of frequencies played. (b) Mel spectrogram of six seconds of New Soul. The mel spectrogram transforms the linear frequency scale of the STFT to a log-frequency scale that reflects human auditory perception.

(S3) Mel Spectrogram based on Auditory Perception: Humans are better at perceiving frequency differences at low pitches than at high pitches (Müller 2015).[12] Thus, the STFT spectrogram does not represent human hearing of sound with equal sensitivity across the frequency spectrum. The mel spectrogram transforms the STFT spectrogram by mapping the frequencies onto the mel scale, a log-frequency scale created to reflect human hearing. Equal distances on the mel scale have the same perceptual distance in pitch.

The additional parameter that goes into generating a mel spectrogram from an STFT spectrogram is the number of mel bands. The number of mel bands specifies the mel filter banks, which are the weights that map the STFT frequencies to the mel frequency scale. We use 128 mel bands, a commonly used number in the literature (e.g., Huzaifah 2017). Figure 4b shows the mel spectrogram of six seconds of New Soul. Compared to Figure 4a, the differences among the lower frequencies are much clearer.

(S4 Theory) Consonance and Dissonance from the Physics of Sound Waves: In order to explain the convolution filters shown in (S4), we must first provide some background information about our mid-level features of interest—consonance and dissonance—and their relationship with the physics of sound waves. The STFT of (S2) decomposes the sound wave into its constituent sine waves, known as partials. The harmonics of the sound wave are the partials that are integer multiples of its fundamental frequency (or lowest

[12] Pitch is a subjective measure of frequency and is defined as the attribute of sound that allows it to be ordered on a scale from low to high. For a pure tone sine wave, the pitch and frequency are the same, and are determined by its fundamental or lowest frequency. However, they can differ for more complex and realistic sounds. In all cases, the higher the frequency, the higher the perceived pitch.

partial).[13] For example, the harmonics of note A4 are $f_0 = 440$ Hz, $f_1 = 2f_0 = 880$ Hz, ..., $f_n = (n+1) \cdot f_0$. If all of the partials in a spectrum are integer multiples of the fundamental frequency, the spectrum is considered harmonic. This property forms the basis for understanding consonance, and the inspiration for our convolution filters.

Harmony captures the perception of simultaneous pitches and is characterized as being consonant or dissonant. Sethares (2005) characterizes it thus: "a musical interval is consonant if it sounds pleasant or restful." In general, consonant sounds, such as the *octave* and the *fifth* in Western music, are considered pleasing to the ear, while dissonant sounds are considered rough and jarring.[14]

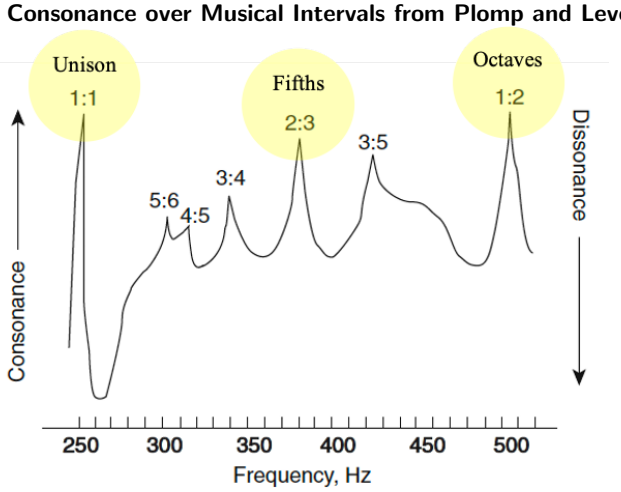
Consonance and dissonance have a strong connection with the physics of sound waves. It is this connection that allows us to make precise use of the theory in our deep learning CNN model. Experimental studies have revealed that consonance and dissonance are not binary categories, but rather represent the opposite ends of a continuum. When two notes have identical frequencies, or unison, they are judged as consonant (Plomp and Levelt 1965, Rasch and Plomp 1999). As shown in Figure 5, borrowed from Plomp and Levelt (1965), as the frequencies diverge from unison (ratio $1:1$), the degree of consonance initially decreases and then increases, producing a U-shaped relationship. The unison is the point of global maximum of consonance, whereas specific other frequencies form local maxima. The precise frequencies where we get local maxima of consonance correspond to specific frequency intervals, including octaves (ratio $2:1$) and fifths (ratio $3:2$). In general, consonance is associated with small integer ratios, also known as simple ratios, of pitch frequencies. Music theorists have suggested that the physics underlying consonance is the occurrence of overlapping harmonics (Sethares 2005), which occurs with small integer ratios.

We use these mathematically represented properties of consonance and dissonance (i.e., overlapping harmonics and simple ratios) to design filters for use in a CNN to predict the listener emotion of a music clip. In the next two steps, we discuss the convolution filters and CNN design associated with the consonance filters. The atheoretical square filters and low-level frequency and time filters are discussed in Section 4.

[13] The fundamental frequency of a violin string corresponds to the entire length of string, whereas higher frequencies are obtained by fixing another point. For example placing a finger at the middle of the string to fix the vibrations to half the length of the string doubles its fundamental frequency. The concept is similar for other instruments like the flute.

[14] A classic example of a dissonant sound is the tritone, which refers to two notes that are three whole steps apart being played simultaneously. The tritone was often used in music from medieval times and has been used in contemporary movies and music to provide a negative connotation or of something foreboding or fear-inducing like the Devil for instance (Peretz and Zatorre 2003, Lerner 2009).

Figure 5 Consonance over Musical Intervals from Plomp and Levelt (1965)



Notes: The graph, borrowed from Plomp and Levelt (1965), shows the relationship between consonance and changing musical interval frequency ratios. The points of greatest consonance occur at small integer ratios of the frequencies of a musical interval. We specifically include octaves and fifths in our model and the filters trivially capture unison.

(S4) Filter Design and Construction: CNNs have typically been used for problems involving visual image inputs, beginning with early problems of digital recognition (LeCun et al. 1989). Convolution filters in CNNs are matrix operations operating on a part of the image and are used for a variety of image processing tasks, including basic ones like edge detection or increasing the sharpness of an image.

We propose specific convolution filter designs to capture music features associated with the concepts of consonance and dissonance. These could in theory impact both valence and arousal in terms of listener emotion. Consonance is associated with positive valence emotions (e.g., joy, tenderness) while dissonance is associated with negative valence emotions (e.g., sadness, fear) (Gabrielsson 2016). Consonance is also associated with low arousal emotions (e.g., contentment) while dissonance is associated with high arousal emotions (e.g., fear) (Gabrielsson 2016).

Since sounds with overlapping harmonics, and in particular octaves and fifths, produce more consonant music in Western music, we propose consonance filters based on harmonics, octaves, and fifths. Music theorists have proposed different theories to explain the perception of consonance (Sethares 2005). One relates to tonal fusion, or the perception of multiple tones as a single tone. The idea behind fusion is that the ear cannot tell which specific tone a partial comes from. When many partials coincide, the sound produced is perceived as a single tone, the most consonant of sounds, rather than multiple tones. In

other words, music with notes that have overlapping harmonics are perceived to be more consonant. Another theory relates to beating, which is the phenomenon that occurs when two sine waves of slightly different frequencies are played together and create an interference pattern, resulting in dissonance. Notes with overlapping harmonics produce less of the interference pattern. These two explanations are the motivation for designing harmonics filters. A third explanation is that the ear simply likes simple ratios (Sethares 2005), motivating the design of octaves and fifths filters.

The consonance filters use "blinders" to select which frequencies are input to the CNN and a convolution filter that considers a large range of frequencies at each time frame. We use the term "blinders" to refer to the matrix operation that selects and weights the mel bands in the mel spectrogram prior to convolution. We use the term "consonance filter" to refer to the combination of the blinders and convolution filters.

Harmonics. As previously discussed, the harmonics of a pitch are the frequencies that are integer multiples of the fundamental frequency. Mathematically, the set of harmonics $\mathcal{H} \coloneqq \{\omega_n\}$ of a tone with fundamental frequency f_0 where $n \in \mathbb{Z}^+$ contains frequencies:

$$\omega_n = n f_0. \tag{3}$$

Octaves. Besides unison, octaves and fifths are typically the most associated with consonance. An octave is defined as a musical interval corresponding to a pair of pitches with a 2:1 frequency ratio. For example, A3 (220 Hz) and A4 (440 Hz) are an octave apart. Given the frequency ratio of 2:1, octaves generate overlapping harmonics. The set of frequencies $\mathcal{O} \coloneqq \{\omega_n\}$ of a given pitch class with a lowest pitch of fundamental frequency f_0 contains frequencies:

$$\omega_n = 2^n f_0 \tag{4}$$

where $n \in \mathbb{Z}^+$. A pitch class captures the set of all pitches that are an integer number of octaves apart. Thus, the lowest A pitch has a fundamental frequency of $f_0 = 27.5$ Hz so the A pitch class includes all the powers of 2 detailed in eq. (4). There are 12 pitch classes in Western music: C, C♯, D, D♯, E, F, F♯, G, G♯, A, A♯, B.

Fifths. A fifth is defined as a musical interval corresponding to a pair of pitches with a 3:2 frequency ratio. For example, C4 (262 Hz) and G4 (392 Hz) are a fifth apart. The set of perfect fifths $\mathcal{F} \coloneqq \{\omega_n\}$ with starting fundamental frequency f_0 where $n \in \mathbb{Z}^+$ contains frequencies:

$$\omega_n = \left(\frac{3}{2}\right)^n f_0. \tag{5}$$

Pitch Class Blinders. We use the set of frequencies defined in eq. (3), eq. (4), and eq. (5) to design blinders that retain only the frequencies of interest. We build the blinders using the fundamental frequency of the lowest pitch within hearing range in each pitch class (shown in Table 1). In total, 12 blinders are constructed for each filter type.

Table 1 Fundamental Frequency of Lowest Pitch in each Pitch Class

Pitch	C	C♯	D	D♯	E	F	F♯	G	G♯	A	A♯	B
Frequency (Hz)	16.35	17.32	18.35	19.45	20.60	21.83	23.12	24.50	25.96	27.50	29.14	30.87

Consonance Filters. Below we overview the steps to create the convolution filter for a given pitch class i and filter type j.

i *Calculate pitch class filter frequencies:* Beginning with the fundamental frequency f_0 of pitch class i, calculate the set of frequencies relevant to filter type j. For example, the lowest C has a fundamental frequency of 16.35 Hz, so the C harmonics frequencies are 1×16.35 Hz, 2×16.35 Hz, 3×16.35 Hz, etc. The number n of harmonics, octaves, and fifths is set to cover the entire range of the spectrogram. For harmonics $n = 1400$, for octaves $n = 11$, and for fifths $n = 20$.

ii *Calculate frequency bands:* For each frequency ω_n where $n \in \mathbb{Z}^+$, calculate a band of frequencies such that frequencies that fall within this band are perceived as being the same pitch. Frequency bands reflect human hearing. The exact size of the band depends on a number of factors, including duration, intensity, and frequency, so there is not a hard rule for calculating the width of the bands. In general, the width of a band is constant for lower frequencies and increases with increasing frequency for higher frequencies. We therefore set the band size to be based on a constant for frequencies less than 1,000 Hz and proportional to frequency for frequencies greater than 1,000 Hz.

iii *Construct STFT indicator column:* Since the STFT spectrogram discretizes frequency, extract the center frequencies of the spectrogram's y-axis. Next, construct an indicator variable that retains only center frequencies that fall within one of the frequency bands.

iv *Construct mel blinders:* To convert the STFT blinders to mel blinders, multiply the STFT indicator column by the mel filter bank to generate a mel weight column that has $N_{mel} = 128$ dimensions. Extending the mel weight column over the time dimension generates the mel blinders. Multiply the mel spectrogram by the mel blinders to generate the input to the convolution layer of the CNN. Section B in the Appendix details the relevant steps.

v *Apply convolution filter:* Apply the convolution filter to the transformed mel spectrogram. We discuss in (S5) below the design of the convolution filter.

(S5) Convolutional Neural Network Architecture: A CNN is typically comprised of multiple types of layers, including convolutional layers, pooling layers, and fully connected layers. Designing a neural network involves a number of architectural decisions and hyperparameter choices, such as the dimensions of the convolution filters and the choice of activation function. Often times, these modeling choices are empirically driven in image processing because the basic model elements have already been optimized for images. However, many model elements have not been optimized for music and require developing. We design the convolutional and pooling layers to incorporate the physics of sound waves and their relationship with consonance.

In addition to these layers, deep neural networks should guard against overfitting and facilitate learning. We determine these modeling choices empirically and use standard modeling elements from the deep learning literature (Table D1 in the Appendix describes these standard modeling choices in detail). Dropout (alongside early stopping) is used to prevent overfitting. Batch normalization and a rectified linear activation unit (ReLU) are known to facilitate model learning.

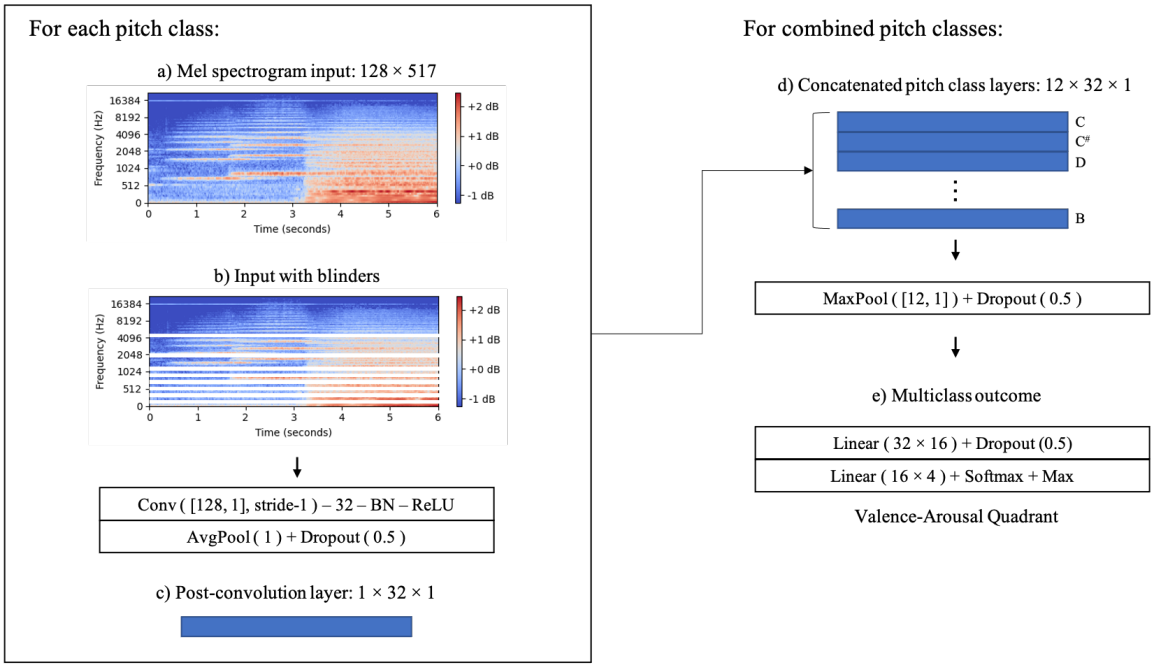
The objective of the model is to learn parameters to minimize the loss between the predicted outputs and the actual target outputs. Since we seek to predict which of four valence-arousal quadrants a music clip falls into, our problem is a multiclass classification problem. The standard loss function for such problems is cross-entropy loss. Let k represent one of the four quadrants, y_{ik} a binary indicator for whether k is the correct class label for music clip i, $p(\hat{y}_{ik})$ the predicted probability that i is of class k, and N the total number of music clips. The cross-entropy loss L_{CE} over the set of music clips is:

$$L_{CE} = -\sum_{i=1}^{N}\sum_{k=1}^{4} y_{ik} log(p(\hat{y}_{ik})). \tag{6}$$

The operation of MusicEmoCNN for any of the consonance filters (i.e., harmonics, octaves, fifths) is as follows:

1. For each pitch class:
   (a) Apply its blinders by multiplying the input mel spectrogram by the blinders.
   (b) Apply convolution and then batch normalize and take ReLU of the feature map.

Figure 6 MusicEmoCNN Architecture



Notes: Overview of our proposed CNN architecture with consonance filters. The input to the CNN is the mel spectrogram. For each pitch class, mel blinders are applied to place structure on what the CNN sees and then convolution and average pooling are applied. The outputs are concatenated together and then max pooled before going through two fully connected layers. The final output is the valence-arousal quadrant prediction.

   (c) Average pool over time frames and apply dropout.

2. Concatenate the hidden layers generated by each of the pitch classes.
3. Max pool over the pitch classes and apply dropout.
4. Use two fully connected layers and apply softmax to output a probability distribution over the four valence-arousal quadrants.
5. Output the quadrant with the max probability.

Figure 6 summarizes the overall architecture of MusicEmoCNN. Since our innovation is in the design of the consonance filters, we discuss the related model ingredients (i.e., convolution and pooling layers) below. We summarize the standard CNN architectural decisions in Table D1 in the Appendix.

Convolution Filter. For a spectrogram, the convolution filter (i.e., kernel) height determines the number of frequency bins included and the width determines the number of

time frames included in the convolution. Mathematically, discrete convolution over two dimensions is defined by:

$$F(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$$

where I is the input image, K is the two-dimensional convolution filter, and m and n index the two dimensions (Goodfellow et al. 2016). The output $F(i,j)$ denotes a feature map. The stride specifies how much the filter slides over the image before performing another operation. A common practice is to learn convolution filters over multiple channels. The idea is that different channels learn different features of the input.

We design the convolution filter of the consonance filters with two objectives: 1) to capture the set of frequencies relevant to consonance and 2) to aid model interpretability. A tool that has been used to provide some visibility into what a CNN learns is gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al. 2017). Grad-CAM produces a heatmap that highlights the regions of the input that predict the target by plotting the gradients of a target class (e.g., Q1—exuberance) that flow into the final convolutional layer. The design of the convolution filter should therefore produce a feature map after the final convolution that captures the information in the mel spectrogram in an interpretable way. The first objective of capturing the relevant set of frequencies can be met by using a tall convolution filter since consonance depends on non-contiguous frequency bins. The second objective of interpretability can be met by setting the filter height to the height of the mel spectrogram, the filter width to one time sample, and the stride to one time sample[15] since it produces a one-dimensional Grad-CAM heatmap over time. We argue that the resulting heatmap captures the concept of consonance over time since the CNN is shown only specific frequencies.

Pooling. Average pooling and max pooling are used at different stages of the CNN to summarize the hidden layers. Pooling applies a function over all units within a specified shape and is typically used for computational efficiency and to make the model invariant to small translations of the input (Goodfellow et al. 2016). To combine the information in the feature maps over the time frames we use average pooling. We use average pooling because we believe the average level of consonance (versus a maximum or minimum level)

[15] This convolution transforms the 128×517 mel spectrogram to a 1×517 vector as the filter slides across the image over time. We learn filters over 32, 64, 128, and 256 channels and find that 32 channels performs the best.

over the music clip is the best predictor of the clip's emotion. After concatenating the resulting vectors for the 12 pitch classes, we use max pooling to summarize the information over the pitch classes. The motivation for using max pooling is that so long as any pitch class has high consonance, the associated time frame should be considered consonant.

(S6) Predicted Emotion: The model maps each six second music clip into one of the valence-arousal quadrants. Combining the predictions over time allows us to observe the dynamics of music emotion.

4. Empirical Analysis

In this section, we begin by describing the datasets used to train the model and comparison models. We then report the performance of our proposed architecture, which uses consonance filters, and compare it against standard benchmark models. Finally, we show how our model is interpretable using gradient-based model visualizations and compare it to visualizations generated by other CNN models which use low-level feature filters.

4.1. Datasets

We use two public datasets compiled by music emotion researchers to demonstrate our classification strategy, namely the Soundtracks and 4Q datasets. The Soundtracks dataset is tagged using continuous valence-arousal labels and contains music similar to what is used in placement and ad videos. The 4Q dataset is a balanced dataset over the four valence-arousal quadrants and more diverse in terms of types of music included.[16] Below we describe these two music datasets in greater detail.

Soundtracks Dataset. The Soundtracks dataset (Eerola and Vuoskoski 2011) is composed of 360 excerpts from movie soundtracks that range in duration from 10 to 30 seconds. One benefit of using movie soundtracks is that they are composed to elicit emotion. The music clips are all instrumental and do not contain any lyrics, dialogue, or sound effects. The clips were also selected to be unfamiliar to prevent song familiarity from impacting the emotion tagging.

[16] Music emotion researchers group emotion into expressed emotion, perceived emotion, and evoked emotion. Expressed emotion refers to the emotion the performer tries to communicate, perceived emotion refers to the emotion a listener perceives from a song (cognitive), and evoked emotion refers to the emotion a listener actually feels in response to a song (emotive) (Jaquet et al. 2014, Yang and Chen 2011b). Most often, the emotion of interest is evoked emotion but because of its subjectivity researchers typically build music datasets that use perceived emotion labels, as is the case for Soundtracks and 4Q. Perceived and evoked emotion are typically positively related (Evans and Schubert 2008, Kallinen and Ravaja 2006, Juslin and Vastfjall 2008). We therefore do not distinguish between perceived and evoked emotion.

University students and staff with musical expertise annotated the song emotions, and six annotators tagged each music excerpt. Perceived valence and arousal were separately annotated on a scale of 1 to 9. To convert the continuous valence-arousal labels to the four discrete quadrants (Q1 to Q4), we discretize the valence-arousal space around the midpoint $(4, 4)$. Q1 captures positive valence, positive arousal, Q2 captures negative valence, positive arousal, Q3 captures negative valence, negative arousal, and Q4 captures positive valence, negative arousal. To provide a reference emotion for each quadrant, we borrow the language from Panda et al. (2018) and label the four quadrants as follows: Q1—exuberance, Q2—anxiety, Q3—depression, Q4—contentment.

Our focal dataset is the Soundtracks dataset because the perceived emotion comes entirely from the audio without the influence of lyrics or song familiarity.

4Q Dataset. In addition to Soundtracks, we train a set of classifiers using the 4Q dataset to test the robustness of our model. The 4Q dataset (Panda et al. 2018) is composed of 900 30-second excerpts equally distributed over the four valence-arousal quadrants. The excerpts cover a range of genres, including pop, country, and jazz, and are from AllMusic. Some clips are entirely instrumental, while others also include lyrics.

Panda et al. (2018) follow a two-stage procedure to obtain the emotion tags. First, they collect the tags provided by AllMusic, which claims to tag the emotion of a song using professional editors. Then they ask survey participants to validate the quadrant tags from AllMusic to ensure the validity of the perceived emotion tags.

4.1.1. Benchmark Models for Comparison In addition to CNNs that use the mid-level consonance filters, we train a number of additional models for comparison. The benchmark models can broadly be characterized as either atheoretical from a music and emotion perspective or musically-motivated but focused on low-level features. The atheoretical models include support vector machine (SVM) with mel frequency cepstral coefficient (MFCC) features, a standard model in music information retrieval, and CNN with square filters, which is borrowed from image recognition. The musically-motivated but focused on low-level features models include CNN models with filters designed to extract either frequency or time features, first proposed by Pons et al. (2016). We detail below the benchmark models used for comparison.

SVM with MFCCs. Prior to the introduction of deep learning to music emotion recognition, using MFCCs as features to SVM classification proved to be a successful strategy. MFCCs

were originally developed for automatic speech recognition and describe the overall spectral envelope shape of audio, capturing timbre. Although MFCCs have proven successful in music classification, there is no theoretical motivation for why they relate to music emotion.

Following this literature (Bhardwaj et al. 2015, Dahake et al. 2016), we extract the first 13 MFCC coefficients of each time window as well as the first and second derivatives of the MFCC coefficients. This procedure generates 39 features per time window. To summarize the features over time, we average each feature over the time windows. We use Python's librosa package to extract the MFCC coefficients.

SVM is a machine learning technique based on separating class by characterizing a hyperplane boundary and using a well-known kernel transformation can also achieve separation in the case of nonlinear and non-monotonic effects (Cortes and Vapnik 1995).

CNN with Square Filters. Image recognition CNN models typically use square filters that capture associations across two orthogonal spatial dimensions. Although mel spectrograms visualize music, the vertical and horizontal dimensions represent frequency and time rather than spatial dimensions. Thus, in music, the dimensions have very different meanings and resulting properties. Image recognition models have been fine-tuned to reflect how we see and recognize images but these models do not represent how we hear and process audio. Therefore, square filters are atheoretical from the perspective of sound wave physics and acoustics. Square filters capture features but it is unclear what these are and how they relate to music emotions.

To operationalize the CNN with square filters, we borrow the architecture used by Chowdhury et al. (2019) with the Soundtracks dataset to classify emotion, which is based on the VGG image classification model. The model includes nine convolutional layers that primarily use 3×3 square filters alongside batch normalization, ReLU, and dropout. After the ninth convolutional layer, the model uses average pooling to summarize the information over different channels and then a fully connected layer to produce the valence and arousal predictions. See Table D2 in the Appendix for the overview of the CNN with square filters architecture.

CNN with Time and Frequency Filters. We compare our proposed mid-level consonance filters against low-level time and frequency filters proposed by Pons et al. (2016). The authors experiment with rectangular filters of different shapes and sizes to capture low-level timbral and temporal features. Tall and skinny filters are designed to capture timbral

features across the frequency spectrum, e.g., a specific combination of notes, while short and wide filters are designed to capture temporal features, e.g., tempo. Pons et al. (2016) apply these ideas to ballroom genre classification and find that, individually, these filters do not perform as well as a CNN which uses "black-box" square filters, but that combining the two types of filters along with an additional fully connected layer results in comparable performance.

We design a model that uses frequency filters, a model that uses time filters, and a model that combines the two types of filters, much like Pons et al. (2016). However, we allow the models additional flexibility by including an additional fully connected layer after pooling and before the final classification.

4.2. Model Performance

To evaluate our multiclass model, we measure precision, recall, and F_1-score. These metrics are calculated for each class (quadrant) and a weighted average by the number of samples in each class determines each overall measure. Table 2 summarizes the performance of the various models for the Soundtracks and 4Q datasets. The performance metrics are averaged over each fold of the held-out test data from 10-fold cross-validation. The standard deviations of the performance measures calculated over the ten folds are shown in parentheses.

We begin by discussing the performance results of the various classifiers on the Soundtracks dataset. First, consistent with the trend in the MIR literature, the CNN with square filters greatly outperforms SVM with MFCC features. Although not musically-motivated, the CNN with square filters is nevertheless able to learn emotion-discriminating features. Second, consistent with Pons et al. (2016), the CNNs with frequency or time filters underperform the CNN with square filters. The combination of the two filter types does better than either on its own but is still less accurate than the square filters. Third, our proposed harmonics and octaves filters perform comparably to the square filters despite being part of a shallower network. The consonance filters impose structure on the input to the CNN that has empirically been observed to relate to emotion. In doing so, the model is able to learn features relevant to emotion recognition with fewer model parameters. The CNNs with consonance filters have 50,900 trainable parameters while the CNN with square filters has nearly five million, the CNN with frequency filters over one million, the CNN with time

Table 2 Classification Performance

		Soundtracks Data			4Q Data		
Features	Model	Precision	Accuracy/Recall	F_1	Precision	Accuracy/Recall	F_1
Current Atheoretical Filters							
MFCC	SVM	0.3720	0.4555	0.3926	0.4965	0.4922	0.4830
		(0.0879)	(0.0557)	(0.0604)	(0.0724)	(0.0550)	(0.0604)
Square	CNN	0.4381	0.5888	0.4975	0.5657	0.5633	0.5438
		(0.1100)	(0.0895)	(0.1031)	(0.0503)	(0.0619)	(0.0565)
Current Theoretic Low-Level Filters							
Frequency	CNN	0.3080	0.4277	0.3227	0.4976	0.4955	0.4605
		(0.1255)	(0.0765)	(0.0824)	(0.0964)	(0.0699)	(0.0848)
Time	CNN	0.3316	0.4694	0.3429	0.4572	0.4644	0.4365
		(0.1598)	(0.0757)	(0.1090)	(0.0974)	(0.0678)	(0.0807)
Frequency-Time	CNN	0.4169	0.4888	0.4095	0.5562	0.5611	0.5366
		(0.1292)	(0.0809)	(0.1041)	(0.0867)	(0.0739)	(0.0758)
Proposed Mid-Level Consonance Filters							
Harmonics	CNN	0.4937	0.5777	0.5092	0.5589	0.5377	0.5338
		(0.1404)	(0.0933)	(0.1043)	(0.0852)	(0.0697)	(0.0731)
Octaves	CNN	0.4988	0.5861	0.5066	0.5564	0.5244	0.5201
		(0.1199)	(0.0592)	(0.0859)	(0.0759)	(0.0548)	(0.0625)
Fifths	CNN	0.4379	0.5111	0.4430	0.5725	0.5366	0.5329
		(0.1266)	(0.0860)	(0.0928)	(0.0821)	(0.0723)	(0.0776)

Notes: Precision $= \frac{\text{True Positive}}{\text{Predicted Positive}}$. Recall $= \frac{\text{True Positive}}{\text{Actual Positive}}$ $F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$. Precision is particularly useful when false positives are costly (e.g., spam detection). Recall is particularly useful when false negatives are costly (e.g., disease detection). Accuracy is equivalent to weighted average recall and captures the proportion of correct predictions out of the entire set of data. F_1 is useful when we want a balance between precision and recall.

filters over six million, and the CNN with frequency-time filters over seven million. Interestingly, the octaves filters perform comparably with the harmonics filters even though the octaves blinders more greatly restrict the set of frequencies seen by the CNN.

Next, we discuss the model performance metrics on the 4Q dataset, which is noisier than the Soundtracks dataset because of the inclusion of music with lyrics and more familiar music. Once again, the CNN with square filters does better than the SVM with MFCC features. The model with the combined frequency and time filters shows a dramatic improvement over the individual filters and performs comparably to the CNN with square filters. Finally, the consonance filters have similar or slightly worse performance than the square filters, but outperform the SVM with MFCCs and individual low-level frequency and time filters. The confusion matrices shown in Figures E1c and E1d in the Appendix suggest that both the square and harmonics models have difficulty differentiating between the valence levels for a given arousal level.

Overall, the CNNs that use the mid-level consonance filters perform well despite having fewer parameters than the other networks. In the Soundtracks data, the harmonics and octaves filters perform comparably with the square filters. In the 4Q data, the consonance

filters perform slightly worse than the square filters but not greatly so. However, as we see below, our proposed consonance filters enable interpretability in contrast to the square filters.

4.3. Model Interpretability

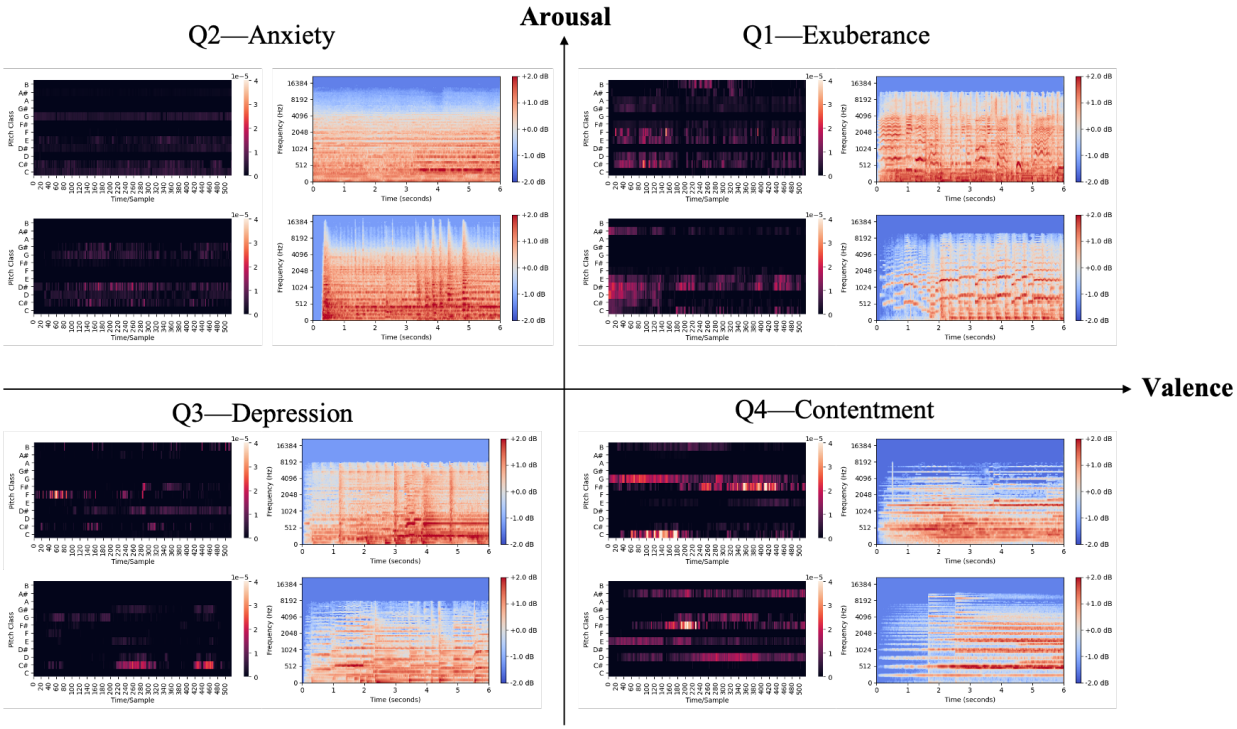
A common concern about deep neural networks is that they lack interpretability, decreasing trust in such models and making them difficult to fix when broken. A key benefit of our proposed music theory-based consonance filters is that they enable post-hoc local interpretability when combined with gradient-based visualizations.

From a top-down perspective, the music and emotion literature tells us that positive valence (vs. negative valence) music and low arousal (vs. high arousal) music are typically perceived as being more consonant. From a bottom-up perspective, music theory tells us that harmonies with overlapping harmonics produce consonance. Combining these two perspectives generates hypotheses around what we should expect to see in the gradient-based visualizations.

We use gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al. 2017) to visualize which parts of the mel spectrogram contribute to a model's classification decision. Grad-CAM uses the gradients of a target class (e.g., Q1—exuberance) that flow into the final convolutional layer to produce a heatmap that highlights the input regions that predict the target (Selvaraju et al. 2017). We know what set of frequencies the CNN sees by placing structure on what is input to the CNN (i.e., the harmonics, octaves, and fifths blinders). We expect areas of the spectrogram with overlapping harmonics to "light up" in the heatmap of a consonant song. In contrast, we expect areas with overlapping harmonics to be dark in the heatmap of a dissonant song.

Therefore, positive valence, low arousal music (i.e., Q4—contentment) should have a bright heatmap with consonance filters, whereas negative valence, high arousal music (i.e., Q2—anxiety) should have a dark heatmap. With the single dimension of consonance, Q1—exuberance and Q3—depression cannot be disambiguated by the Grad-CAM heatmaps because of the opposing valence-arousal and consonance predictions. However, inspection of Grad-CAM in conjunction with the original mel spectrograms helps differentiate the quadrants. Music and emotion research associates tempo with different arousal levels. A fast tempo is associated with high arousal, while a slow tempo is associated with low arousal. Tempo can be seen in a mel spectrogram based on the width of the notes. Being

Figure 7 Harmonics Grad-CAM Heatmaps by Quadrant



Notes: Within each emotion quadrant, the figures on the right are mel spectrograms of music clips and the figures on the left are their associated Grad-CAM heatmaps produced by harmonics filters. The heatmaps are organized by the pitch class of the starting frequencies used to construct the harmonics filters (Table 1). Brightness in the heatmaps captures consonance over time.

high arousal, Q1—exuberance should, on average, have a faster tempo and have shorter notes. Being low arousal, Q3—depression should have a slower tempo and longer notes.

The architecture of MusicEmoCNN generates Grad-CAM heatmaps that are 1×517 (517 representing the number of time frames) for each pitch class. Since the consonance filters are based on pitch class, the Grad-CAM visualizations allow us to visualize the heatmaps over the 12 pitch classes. Figure 7 shows a few prototypical heatmaps generated by the harmonics filter and their associated mel spectrograms for each of the four quadrants. The y-axis of a heatmap represents the 12 sets of frequencies associated with the harmonics of each pitch class. The x-axis represents time. Color represents the importance of the frequencies within each pitch class towards their respective classification, with brighter colors representing greater importance.

In general, positive valence, low arousal songs have the brightest heatmaps and negative valence, high arousal songs have the darkest heatmaps. Appendix Figure G1 displays more

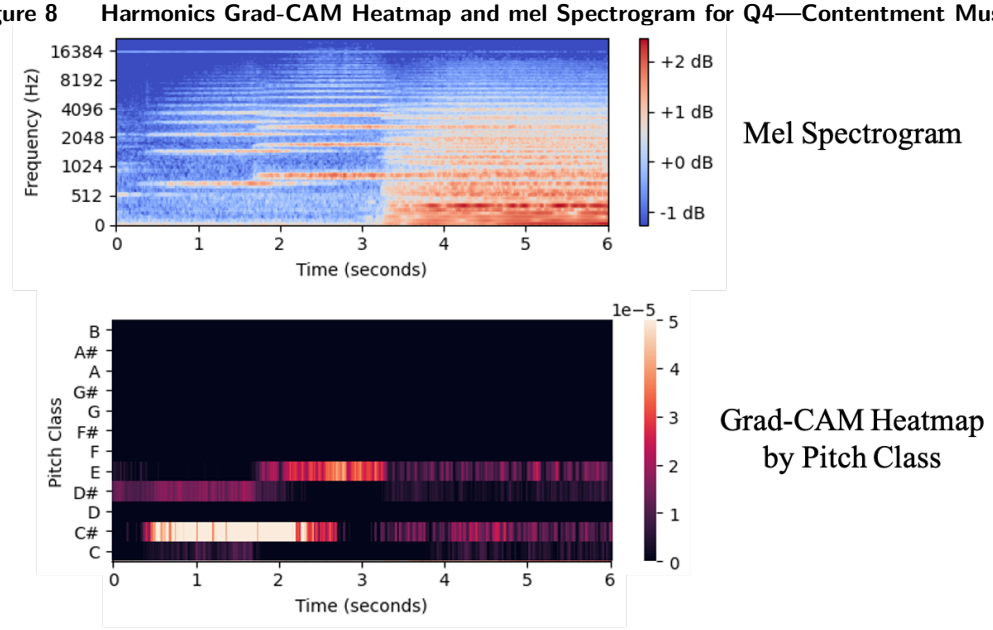
Grad-CAM heatmaps for each of the emotion quadrants, which are generally consistent with this pattern. The heatmaps in Q1 and Q3 lie somewhere in between in terms of brightness. A change in sound can be seen in the mel spectrogram by what looks like a vertical line, capturing a rapid change in frequencies. More vertical lines represents a faster tempo, as can be seen in the high arousal quadrants. The mel spectrograms show that the notes in Q3 are generally played for longer than the notes in Q1, helping to separate the two quadrants. Overall, the results suggest that in low arousal states, people become even more emotionally sensitive to consonance. Thus, consonance has a greater impact on valence in the contentment (say calm music), than in the exuberant case.

It is important to note that music emotion classifiers have generally had greater success differentiating arousal. This fact is evident by looking at the mel spectrograms. Our proposed filters instead focus on the concept of consonance, which measures the pleasantness of music to the human ear and is closely tied to emotional valence. The filters are particularly helpful in classifying valence. Furthermore, while we can get a sense of concepts like tempo, loudness, and timbre by reading a mel spectrogram, we cannot visually extract consonance (except in rare cases such as a single violin string being played). Figure 8 shows the Grad-CAM heatmap and mel spectrogram for a song that begins with a violin, which is then joined by additional instruments shortly after three seconds. As can be seen in the frequencies of approximately 700 Hz, 1,400 Hz, 2,100 Hz from 0.25 seconds to 1.75 seconds, violins produce harmonic sounds. The overlap of the lit-up portion of the heatmap with the harmonic frequencies provides further validation that the deep learning model captures overlapping harmonics.

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 because low brightness levels spread out over different pitch classes are not equivalent to a high level of brightness for a single pitch class in terms of consonance. The average brightness levels for the 4Q and Soundtracks datasets are 8.7 for Q1, 2.3 for Q2, 3.7 for Q3, and 28.2 for Q4, further confirming our consonance interpretation of the Grad-CAM heatmaps. The octaves and fifths Grad-CAM heatmaps follow the same pattern as was observed for the harmonics filters.

The interpretability of the consonance filters builds trust in the model by providing some transparency into what the model is learning. It highlights areas of consonance, a mid-level feature not observable by eye. Music theorists have proposed a number of formulas to

Figure 8 Harmonics Grad-CAM Heatmap and mel Spectrogram for Q4—Contentment Music clip



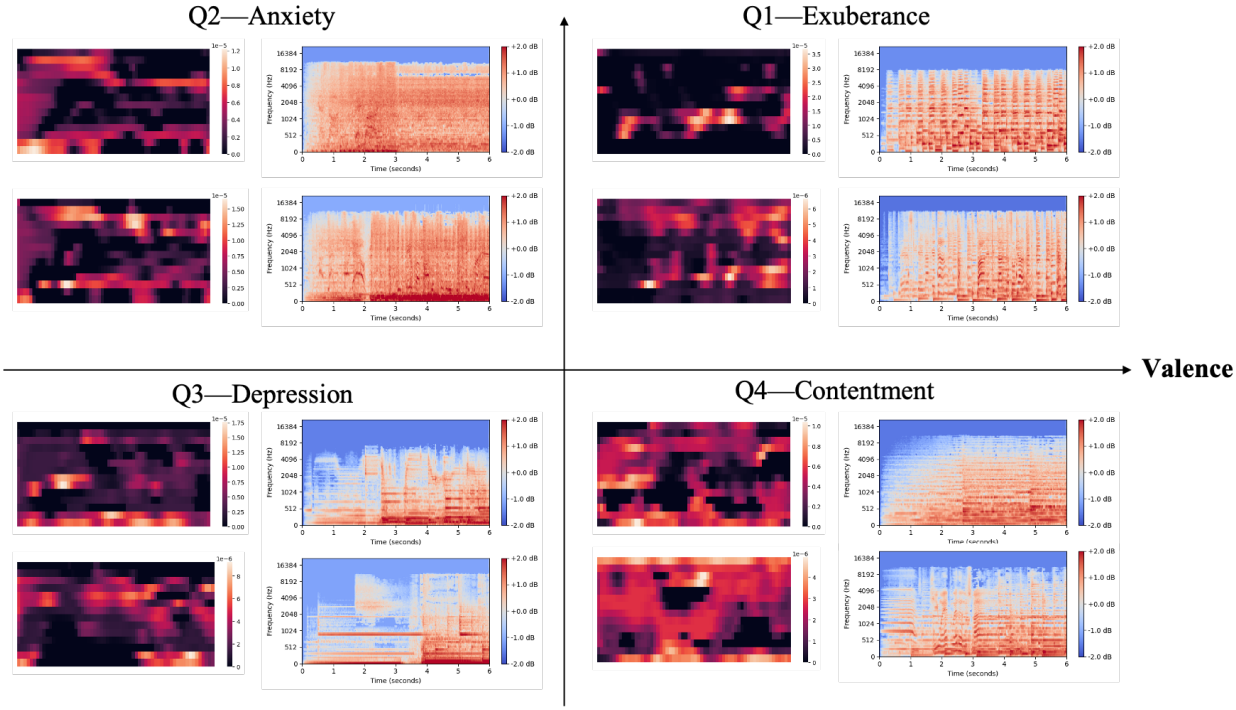
Notes: The Grad-CAM heatmap on the left shows the music clip is very consonant from time sample 40 to 220. The image on the right zooms in on the C♯ heatmap. The music clip starts off with a violin, producing a harmonic sound that can be seen in the mel spectrogram through the red parallel horizontal bars at frequencies that are integer multiples of approximately 700 Hz. The heatmap also captures this period of consonance.

quantify sensory dissonance based on our understanding of the human ear and the physics of sound. Our filters enable the deep learning algorithm to learn this relationship as it relates to listener emotion.

4.3.1. Interpretability of Low-Level Filters. Gradient-based visualizations can also be produced for the other filter types; however, given their low-level focus, it is more difficult to interpret what they are capturing and how the captured features contribute to the emotion classification of a particular class.

The Grad-CAM heatmaps for the square filter CNN are equivalent to heatmaps produced for an image recognition model. The heatmaps show spatially which parts of the input image contribute to the classification of a particular target class. While the heatmaps, shown in Figure 9, provide an idea of the range of frequencies and times contributing to the classification of the target class, it is otherwise difficult to interpret the heatmaps. Figure G3 in the Appendix shows the Grad-CAM heatmaps associated with the frequency filters. The heatmaps indicate that high frequencies contribute to high arousal classifications while low frequencies contribute to low arousal classifications, consistent with the music and emotion literature. Figure G4 in the Appendix shows the Grad-CAM heatmaps

Figure 9 Square Grad-CAM Heatmaps by Quadrant



Notes: Within each emotion quadrant, the figures on the right are mel spectrograms of music clips and the figures on the left are their associated Grad-CAM heatmaps produced by square convolution filters. The bright portions of a heatmap capture the parts of the mel spectrogram that most greatly contribute to the classification of the specified emotion. It is challenging to interpret the Grad-CAM heatmaps.

associated with the time filters. The high arousal quadrants show consistent brightness across time, while the low arousal quadrants light up in columns. The heatmaps are not straightforward to interpret.

5. Application: Emotion-based Ad Insertion in Videos

MusicEmoCNN is a tool that can be used in a number of real-time music matching applications by quantifying the valence and arousal of a music clip. In this paper, we demonstrate its value in determining the optimal emotion-based ad insertion point within a video that varies in emotion over time. Given our focus on music and its effects, we treat the emotion evoked by the background music of a video as a sufficient statistic for the emotion evoked by the overall video.

This question is of significant managerial importance because of the rapid proliferation of user-generated content (UGC) on video platforms and because of the limitations of the data

available to advertisers due to privacy concerns.[17] First, the vast amount of UGC available to advertise within makes non-algorithmic approaches challenging, if not impossible to implement at scale. Past studies have found that emotion impacts ad effectiveness, therefore it is important to incorporate emotion as a variable in determining ad insertion. Second, large tech firms are increasingly placing restrictions on person-specific data they collect, thereby limiting data available to advertisers. As a result, contextual targeting and in particular content targeting will increasingly play a role in ad placement. Our model can be deployed at scale and provide emotion-based ad position suggestions to television networks, video platforms, and content creators.

Past marketing studies have found that matching valence increases ad effectiveness (Coulter 1998, Kamins et al. 1991) and that placing high arousal ads in low arousal contexts decreases ad effectiveness (Puccinelli et al. 2015). In contrast to past work, which focuses on either valence or arousal, we study the impact of both simultaneously. In addition, while past work considered the emotion of videos that differ in overall emotion, we consider optimal emotion-based ad insertion within a video that features variation in emotion over time.

It is an empirical question as to whether ads that are similar to the emotional context increase or decrease ad attention and memorability. On the one hand, many behavioral studies have found that emotion congruence is more effective (Lee et al. 2013, Kamins et al. 1991, Gibson et al. 2000), including studies of matching in persuasion (Teeny et al. 2021) and fluency (Hertwig et al. 2008). On the other hand, other studies have found that consumers have a preference for positive stimuli when feeling negative emotions (Biswas et al. 1994, Andrade 2005, Tamir 2016) and that perceptual contrast draws attention, suggesting emotion contrast may be more effective.

To answer this empirical question, we exogenously insert ads into a video at different emotion distances between the ad and the placement video, as measured by the Jensen–Shannon distance, and obtain outcome measures around ad attention and memorability. The application consists of three parts. First, we take human-tagged emotion as the ground truth and select two points in the placement video with the largest and

[17] Many web browsers have already eliminated third-party cookies (source: https://www.mediapost.com/publications/article/346034/baking-up-new-strategies-for-a-post-cookie-world.html) and in March 2021 Google announced that it would stop tracking the web browsing behavior of individuals (source: https://www.businessinsider.com/google-to-stop-tracking-individuals-web-browsing-precision-ad-targeting-2021-3).

smallest emotional distances between the ad and the placement. This allows us to answer the empirical question about the effect of emotional distance on ad effectiveness. Next, to determine whether there is a more general relationship between emotional distance and ad effectiveness, we select additional ad insertion points with varying emotional distances to test. Finally, we use our trained SVM and CNN classifiers to select ad insertion points and compare them against the human-tagged points.

5.1. Experiment: Response by Emotional Distance

Below we describe the measure we use for emotional distance, the content video and ads used in the experiment, the experimental setup, and the results.

5.1.1. Emotional Distance Measure. To determine where to place the ad in the video based on emotional distance, we must define a measure of distance. We define the emotional distance between an ad and a placement as the Jensen-Shannon (JS) distance between their probability distributions over the four valence-arousal quadrants. Let P_t represent the emotion probability distribution of the placement at time t and Q the emotion probability distribution of the ad. JS distance is defined as:

$$JSD(P_t||Q) = \sqrt{\frac{1}{2}D(P_t||M) + \frac{1}{2}D(Q||M)} \tag{7}$$

where $M = \frac{1}{2}(P_t + Q)$ and D is the Kullback–Leibler (KL) divergence. KL divergence is in turn defined as:

$$D(P_t||Q) = \sum_{x \in X} P_t(x) log\left(\frac{P_t(x)}{Q(x)}\right) \tag{8}$$

where X represents the probability space over which P_t and Q are defined (i.e., the four quadrants). The benefit of JS distance over KL divergence is that it is symmetric between P_t and Q and always finite. The larger the JS distance, the more dissimilar the ad emotion is from the placement emotion at time t. Using human-tagged emotion, we determine the placement times associated with the largest and smallest emotional distances, which serve as the ad insertion points in the experiment.

5.1.2. Video Ad Insertion Survey. To control the ad insertion point, we develop a survey that inserts an ad partway through a placement video, mimicking the concept of YouTube's mid-roll ads. Six seconds into the ad, similar to YouTube, a "Skip Ad" button appears, allowing viewers to skip the remainder of the ad. Upon watching the ad

to completion or skipping it, the placement video picks up where it left off. Each viewer sees only one ad in the video. After viewers finish watching the video, they are asked a number of questions about the video, such as their enjoyment level, as well as the ad.

From the survey, we can see if and when each viewer skips an ad. We capture whether the viewer viewed the ad to measure attention and whether a viewer could remember the advertising brand to capture memorability. A view occurs if the viewer does not skip the ad within the first nine seconds.[18] This definition allows us to capture the emotion interaction of the placement video at the time of ad insertion and of the ad. The survey asks the participant to recall the advertising brand after finishing the video.

5.1.3. Video Placement and Ad Selection. For the placement video, we use Spring,[19] a seven-minute animated short created by Blender Institute and released under a Creative Commons License. It contains background music and varies in emotion over time. To obtain the human emotion tags, we show survey participants on Prolific, an online survey platform, the video in 30-second clips and ask them about their valence and arousal levels.[20] We convert these values to valence-arousal quadrants to obtain a probability distribution of evoked emotion for each 30-second clip.

Figure 10 shows the emotion distribution of the responses over time. At the beginning of the video, viewers largely feel positive (Q1—exuberance and Q4—contentment). In the middle, viewers feel increasingly anxious (Q2—anxiety) with a short period of relief in the middle. At the end, viewers return to a positive state.

The video ads come from a local branch of a large U.S. nonprofit. Prolific workers tag the first six seconds of each ad for their valence and arousal levels. We focus on the first six seconds because this is the length of time viewers on YouTube see an ad for before having the option to skip. Two ads with differing emotion distributions are used in the experiment. Both ads include background music and are 30 seconds long.

Using the JS distance, we determine the times of minimum and maximum emotional distance. We exclude the first minute of video to allow participants to settle into the study and become familiar with the video. Figure 11 plots the JS distance between the placement

[18] This is different from YouTube's definition of view rate. YouTube counts a view as having watched at least 30 seconds of an ad or its duration if it is less than 30 seconds.

[19] https://www.youtube.com/watch?v=WhWc3b3KhnY

[20] The video emotion over time was tagged by 49 Prolific workers. Table H1 in the Appendix summarizes the number of survey participants that participated in each part of this study.

Figure 10 **Spring Human-Tagged Distribution of Emotion over Time**

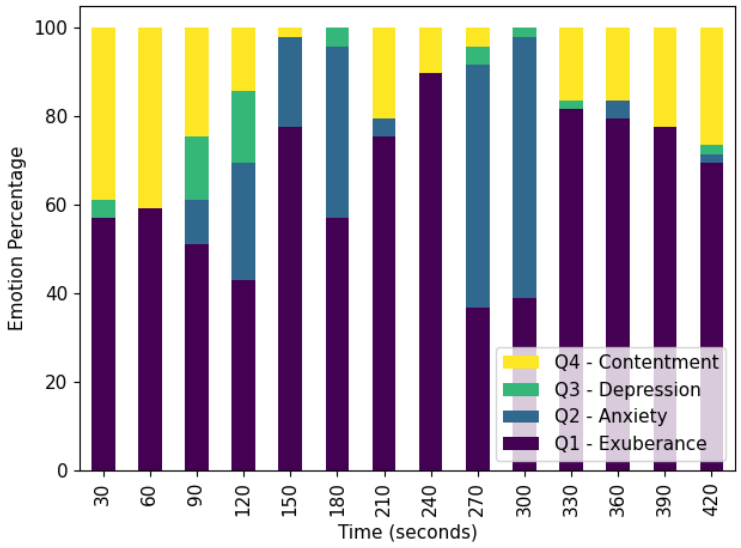
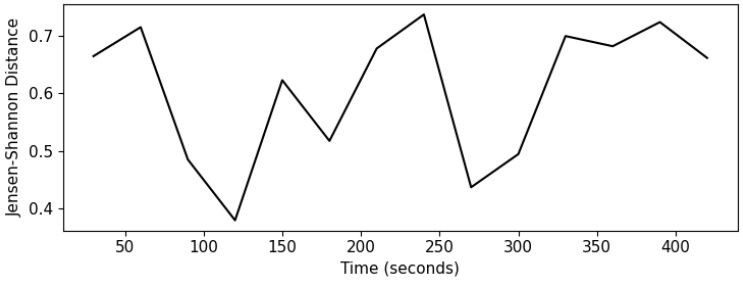


Figure 11 **Jensen-Shannon Distance of Ad 1 at different placement positions**



and Ad 1 over time. The JS distance varies over time. The minimum distance occurs at 120 seconds and the maximum distance occurs at 240 seconds.

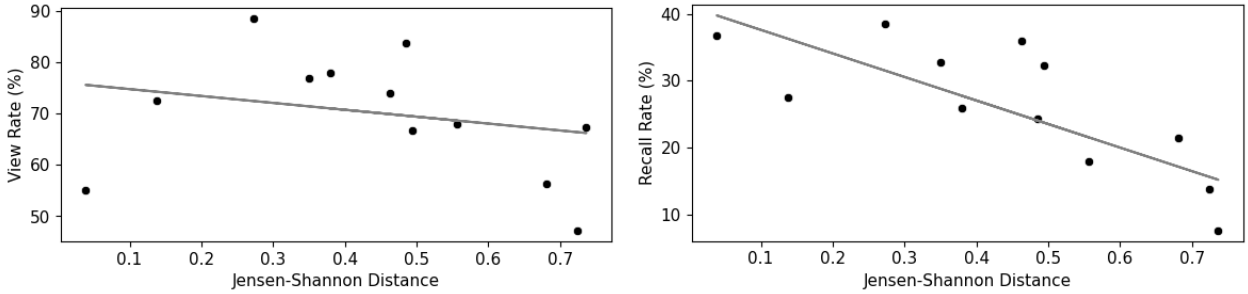
5.1.4. Experimental Results. 201 survey participants recruited through Prolific participated in the first part of this study.[21] Table 3 displays the ad engagement measures of the experiment. The view rates do not differ across the maximum and minimum emotional distances ($p = 0.88$ using a two-tailed z-test).[22] For recall rates however, the difference is significant ($p < 0.01$) with ads placed at the minimum emotional distance generating higher recall rates. These results suggest that reducing the emotional distance between an ad and

[21] Participants were told they would watch a seven-minute animated short and then be asked to answer some questions about the video. The survey was limited to Prolific workers who have U.S. citizenship, are fluent in English, have an approval rating greater than 95%, and have completed at least 25 previous Prolific tasks. The survey took roughly ten minutes to complete and each participant was paid $1.60 for their time.

[22] An alternative definition of a view is to define it as watching the full 30-second ad. In this case, the minimum distance insertion point obtains 40 views and the maximum distance insertion point obtains 32 views; $p = 0.18$.

Figure 12 Ad Engagement vs. JS Distance

(a) View Rate vs. JS Distance (b) Recall Rate vs. JS Distance



its insertion point within content has little impact on attention, but greatly improves ad memorability.

Table 3 Ad Insertion Results

	Counts			Rates	
Emotional Distance	# Impr.	# Views	Correct Recalls	View Rate	Recall Rate
Least	99	66	31	67%	31%
Most	102	69	13	68%	13%

5.2. Additional Ad Insertion Distances

To further assess the relationship between emotional distance and ad engagement, we test additional ad insertion points at various JS distances. Four additional ad insertion distances are selected for each ad and an additional 346 Prolific workers are recruited for the study. Figures 12a and 12b plot view rate and recall rate, respectively, against JS distance. There does not appear to be a monotonic relationship between view rate and JS distance but there does appear to be a monotonic relationship between recall rate and JS distance.

To determine the impact of emotional distance on views and brand recall, we regress both the binary indicator for viewing an ad and the binary indicator for recalling the brand on JS distance, controlling for covariates. We estimate the following regression equation:

$$y_{it} = \alpha + \beta JSD_{it} + \gamma X + \epsilon_{it} \tag{9}$$

where the outcome y_{it} represents the binary indicator for a view or correct recall for ad i at insertion time t, JSD_{it} represents the JS distance between the emotion of the content video at time t and the ad video, X represents covariates depending on the model specification,

and ϵ_{it} the error term. Table 4 shows the regression results of different specifications for the covariates. Columns (1) and (4) do not include any covariates and show that there is no significant effect of emotional distance on views but there is a significant effect on recall.[23] Next, we also include ad fixed effects to control for ad differences and a linear time trend to control for time effects. As shown in Columns (2) and (5), the results remain robust to including these covariates. As a robustness check, we control for time using insertion time fixed effects rather than a linear time trend. The results, shown in Columns (3) and (6), remain consistent. The results confirm that emotional distance significantly impacts brand recall but not ad views. Instead, views appear to decrease as the viewer has spent more time watching the video.

Table 4 Effect of Emotional Distance on Ad Engagement

	Outcome: I(View)			*Outcome: I(Recall)*		
	(1)	(2)	(3)	(4)	(5)	(6)
JS Distance	−0.1343	0.0442	−0.0547	−0.3514***	−0.2546**	−0.4172**
	(0.099)	(0.127)	(0.205)	(0.084)	(0.119)	(0.176)
Time		−0.4250***			−0.0953	
		(0.078)			(0.079)	
Ad FE	N	Y	Y	N	Y	Y
Time Block FE	N	N	Y	N	N	Y
R^2	0.004	0.06	0.06	0.03	0.03	0.04

Note: *p<0.1; **p<0.05; ***p<0.01, 547 observations for all regressions

5.3. Ad Insertion Automation

Through the experiment, we have established the importance of emotional distance on brand recall. Testing this required 640 individuals for a single placement video and two ads (see Table H1 in the Appendix for a summary of the number of participants involved at each step). Human tagging of emotion for all video content is not a viable strategy so we demonstrate the use of MusicEmoCNN in determining the optimal (lowest emotional distance) emotion-based ad insertion point by comparing outcomes from the MusicEmoCNN selected ad-insertion time against those based on human-selection and SVM.

The CNN and SVM classifiers are provided the same information given to the human taggers. Since MusicEmoCNN is trained on six-second clips of content, we feed six-second clips into the two models and average the predicted probabilities over 30-second chunks

[23] Robust standard errors are used since the outcome variables are binary.

to predict the placement valence-arousal quadrant probabilities. We also feed the first six seconds of each ad into the two classifiers.

The JS distances are calculated between the ads and the placement at the six time points included in the experiment for each ad. For each classifier, we determine the least emotionally distant point. Table 5 compares the average JS distance (based on human tagging) and brand recall rate at the insertion points determined by human taggers, SVM, atheoretical CNN, and MusicEmoCNN. The deep learning models outperform SVM and its performance, in terms of recall, is essentially comparable to human taggers. Compared to the atheoretical CNN, MusicEmoCNN is also interpretable, enabling trust.

Table 5 Ad Insertion Automation Results

Tag Source	Avg. JS Distance	Avg. Recall Rate
Human	0.21	31%
SVM	0.64	16%
CNN	0.36	29%
MusicEmoCNN	0.38	30%

5.4. Managerial Implications

Past studies have provided evidence that emotional ads impact attention and memory (Cohen et al. 2018, Petty et al. 1988, Holbrook and Batra 1987). The results of this study support the theory that emotional similarity increases ad memorability. We find that ads which have background music similar in emotion distribution to the background music of placement videos improve brand recall. We develop MusicEmoCNN as a tool to facilitate the determination of emotion based on the background music of videos. We show that it outperforms the baseline SVM classifier in terms of approximating human-tagged emotion and achieves ad recall comparable to that based on human tagging.

We demonstrate the value of MusicEmoCNN in a video advertising setting, but it could be useful in a number of other applications as well. For example, existing Spotify playlists built around a unifying emotion are based on the overall emotion of a song. However, The Echo Nest found that one quarter of songs are skipped in the first five seconds[24] so the interaction of the ending of one song and the beginning of the next is a critical point for a listener's decision to continue with a playlist. MusicEmoCNN can quantify the match

[24] https://www.theguardian.com/music/2014/may/07/one-quarter-of-spotify-tracks-are-skipped-in-first-five-seconds-study-reveals

between the end of one song and the beginning of the next. The classifier can also be used in contexts that do not match on music but match on emotion. For example, advertisers who show video ads in news articles should consider the emotion elicited by the article. A text classifier can be used for the news article while MusicEmoCNN can be used for the video ad. The same applies to radio talk shows and radio ads.

More broadly, any setting that involves emotion and requires music choice could benefit from a music emotion classifier. For example, call center music could be selected based on the incoming emotional tone of the caller. MusicEmoCNN could help select a set of emotionally diverse songs suitable in different emotional contexts to be A/B tested without relying on one's gut or involve labor-intensive human tagging.

6. Conclusion

Our research contributes to the literature that studies consumer response to unstructured data, such as text, audio, images, and video. Music is pervasive in customer interactions with firms. From music in ads to hold music for call centers, from Peloton playlists to background music in retail stores, customers engage with music in a variety of ways. The exponential growth of user-generated content on platforms like YouTube and Tik Tok has created a huge quantity of high-dimensional data, where automated prediction of music-evoked emotion at scale has become critical for many marketing problems.

We develop a deep learning CNN model, MusicEmoCNN (pronounced "music emotion"), to classify the emotion evoked by music in a listener. Our framework integrates a number of theoretically motivated elements from the physics of music, human hearing of sound, as well as human perception of musical notes. We develop novel filters to capture musical features associated with consonance and dissonance using structures that have a foundation in the physics of sound waves (harmonics, octaves, fifths) and integrate them into CNN models. The filters are not only important in predicting emotional outcomes but are also interpretable and help us understand how musical features impact emotion. Our approach achieves similar classification performance (in terms of accuracy and F_1-score) as that of atheoretical models which are not easily interpretable.

We use our emotion classification method in an application where we match the dynamically varying emotion in a content video with the emotion of a short ad. We find that for ad recall, matching the ad emotion to the content emotion improves ad memorability.

With the decline in user-specific tracking due to privacy concerns, contextual targeting is becoming increasingly important. Digital ad targeting will need to move away from strategies that rely on demographic and behavioral targeting to strategies that are based on contextual targeting, such as emotion. We note that besides advertising, our framework can be used in a number of other settings, such as playlist formation and music therapy.

We mention a few limitations and suggestions for future research. First, given the current paper's focus on developing an emotion predictor for music, our empirical application for ad insertion only considers music for ad insertion. Most ads feature a strong musical component, and music has been widely recognized to affect ad performance. However, ads typically have audio features other than music that can impact evoked emotion, like human voice or background sounds. As such, it would be useful to extend our model to incorporate other audio elements in emotion prediction. Further, ads also contain video and text data in addition to music. Text has been found to express emotional content on Twitter (Fong and Kumar 2020) and even in loan applications (Netzer et al. 2019); video features are found to impact sales (Yang et al. 2021b). Integrating emotion recognition with multimodal data (audio, video, and text) would be an important avenue for future research.

In terms of interpretability, we exploit specific elements of music theory to construct filters to capture musical features associated with consonance and dissonance. Deep learning using convolutional neural networks uses highly nonlinear transformations to learn from the data. Such models are often viewed as black-box models, although some recent approaches aim to provide transparency into the visual factors that strongly impact the outcomes of interest. We visualize the model using Grad-CAM (Selvaraju et al. 2017), which applies to any CNN architecture, and focus on the final convolutional layer to identify how gradient information flows into it. This approach provides a visual representation of the areas in an image (spectrogram for sound) of the same size as the feature maps (e.g., our consonance filters). While this provides a degree of transparency, we note that making deep learning models more explainable is an active area of research in machine learning (Ribeiro et al. 2016, Choo and Liu 2018, Angelov and Soares 2020, Singh et al. 2020).

Our research suggests a few aspects that deserve further study. First, although the valence and arousal circumplex model captures the underlying factors for emotions, we could examine discrete emotions like happiness and sadness to understand the acoustic

features that drive them. Second, extending our work to emotions present in human voice and understanding how that impacts listener emotion would be valuable. Third, while we focus on the audio features of ads, other research has shown that video features could also be important (Yang et al. 2021b). Similarly, extending the emotional model to multimodal data could capture differential elements of expression that are uniquely expressed in text or other media. In sum, we believe that the growing presence of multimodal high-dimensional data offers a rich set of opportunities to understand consumer behavior and choices.

References

Allan D (2008) A content analysis of music placement in prime-time television advertising. *Journal of Advertising Research* 48(3):404–417.

Alpert JI, Alpert MI (1990) Music influences on mood and purchase intentions. *Psychology & Marketing* 7(2):109–133.

Andrade EB (2005) Behavioral consequences of affect: Combining evaluative and regulatory mechanisms. *Journal of Consumer Research* 32(3):355–362.

Angelov P, Soares E (2020) Towards explainable deep neural networks (xdnn). *Neural Networks* 130:185–194.

ANSI (1960) Acoustical terminology si. *1-1960, American Standards Association* .

Arpit D, Jastrzebski S, Ballas N, Krueger D, Bengio E, Kanwal MS, Maharaj T, Fischer A, Courville A, Bengio Y, Lacoste-Julien S (2017) A closer look at memorization in deep networks. Precup D, Teh YW, eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 233–242 (PMLR), URL `http://proceedings.mlr.press/v70/arpit17a.html`.

Bagozzi RP, Gopinath M, Nyer PU (1999) The role of emotions in marketing. *Journal of the academy of marketing science* 27(2):184–206.

Bergstra J, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization. *25th annual conference on neural information processing systems (NIPS 2011)*, volume 24 (Neural Information Processing Systems Foundation).

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. *Journal of machine learning research* 13(2).

Bhardwaj A, Gupta A, Jain P, Rani A, Yadav J (2015) Classification of human emotions from eeg signals using svm and lda classifiers. *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, 180–185 (IEEE).

Biswas R, Riffe D, Zillmann D (1994) Mood influence on the appeal of bad news. *Journalism Quarterly* 71(3):689–696.

Boughanmi K, Ansari A (2021) Express: Dynamics of musical success: A machine learning approach for multimedia data fusion. *Journal of Marketing Research* 00222437211016495.

Bruner GC (1990) Music, mood, and marketing. *Journal of marketing* 54(4):94–104.

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, 77–91 (PMLR).

Büschken J, Allenby GM (2016) Sentence-based text analysis for customer reviews. *Marketing Science* 35(6):953–975.

Castelvecchi D (2016) Can we open the black box of ai? *Nature News* 538(7623):20.

Chakraborty I, Kim M, Sudhir K (2021) Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes .

Choi K, Fazekas G, Cho K, Sandler M (2017) A tutorial on deep learning for music information retrieval. *arXiv preprint arXiv:1709.04396* .

Choo J, Liu S (2018) Visual analytics for explainable deep learning. *IEEE computer graphics and applications* 38(4):84–92.

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* .

Cohen JB, Pham MT, Andrade EB (2018) The nature and role of affect in consumer behavior. *Handbook of consumer psychology*, 306–357 (Routledge).

Corrigall KA, Schellenberg EG (2013) Music: The language of emotion. .

Cortes C, Vapnik V (1995) Support-vector networks. *Machine learning* 20(3):273–297.

Coulter KS (1998) The effects of affective responses to media context on advertising evaluations. *Journal of Advertising* 27(4):41–51.

Dahake PP, Shaw K, Malathi P (2016) Speaker dependent speech emotion recognition using mfcc and support vector machine. *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, 1080–1084 (IEEE).

Dhar P, Singh RV, Peng KC, Wu Z, Chellappa R (2019) Learning without memorizing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5138–5146.

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* .

Du RY, Netzer O, Schweidel DA, Mitra D (2021) Capturing marketing information to fuel growth. *Journal of Marketing* 85(1):163–183.

Dubois J, Elovsson A, Friberg A (2019) Predicting perceived dissonance of piano chords using a chord-class invariant cnn and deep layered learning. *16th Sound & Music Computing Conference SMC2019, Malaga, Spain*, 530–536.

Eerola T, Vuoskoski JK (2011) A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music* 39(1):18–49.

Elowsson A, Friberg A (2019) Modeling music modality with a key-class invariant pitch chroma cnn. *arXiv preprint arXiv:1906.07145* .

Evans P, Schubert E (2008) Relationships between expressed and felt emotions in music. *Musicae Scientiae* 12(1):75–99.

Fong H, Kumar V (2020) Using domain knowledge to enhance deep learning for emotional intelligence—extended abstract (Proceedings of the 3rd Workshop on Affective Content Analysis co-located with Thirty-Fourth AAAI Conference on Artificial Intelligence).

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* .

Gibson R, Aust CF, Zillmann D (2000) Loneliness of adolescents and their choice and enjoyment of love-celebrating versus love-lamenting popular music. *Empirical Studies of the Arts* 18(1):43–48.

Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) *Deep learning*, volume 1 (MIT press Cambridge).

Gorn GJ (1982) The effects of music in advertising on choice behavior: A classical conditioning approach. *Journal of marketing* 46(1):94–101.

Hertwig R, Herzog SM, Schooler LJ, Reimer T (2008) Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, memory, and cognition* 34(5):1191.

Hinton G, Deng L, Yu D, Dahl GE, Mohamed Ar, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, et al. (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29(6):82–97.

Holbrook MB, Batra R (1987) Assessing the role of emotions as mediators of consumer responses to advertising. *Journal of consumer research* 14(3):404–420.

Holbrook MB, Hirschman EC (1982) The experiential aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of consumer research* 9(2):132–140.

Honegger M (2018) Shedding light on black box machine learning algorithms: development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv:1808.05054* .

Huang MH (2001) The theory of emotions in marketing. *Journal of Business and Psychology* 16(2):239–247.

Huron D (1989) Music in advertising: An analytic paradigm. *The musical quarterly* 73(4):557–574.

Huzaifah M (2017) Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. *arXiv preprint arXiv:1706.07156* .

Jaquet L, Danuser B, Gomez P (2014) Music and felt emotions: How systematic pitch level variations affect the experience of pleasantness and arousal. *Psychology of Music* 42(1):51–70.

Johnson-Laird PN, Oatley K (2016) Emotions in music, literature, and film. *Handbook of emotions* 82–97.

Juslin PN, Laukka P (2003) Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin* 129(5):770.

Juslin PN, Vastfjall D (2008) Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences* 31(5):559.

Kallinen K, Ravaja N (2006) Emotion perceived and emotion felt: Same and different. *Musicae Scientiae* 10(2):191–213.

Kamins MA, Marks LJ, Skinner D (1991) Television commercial evaluation in the context of program induced mood: Congruency versus consistency effects. *Journal of Advertising* 20(2):1–14.

Korhonen MD, Clausi DA, Jernigan ME (2006) Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36(3):588–599.

Krishna A (2012) An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior. *Journal of consumer psychology* 22(3):332–351.

Krishna A, Cian L, Sokolova T (2016) The power of sensory marketing in advertising. *Current Opinion in Psychology* 10:142–147.

Laros FJ, Steenkamp JBE (2005) Emotions in consumer behavior: a hierarchical approach. *Journal of business Research* 58(10):1437–1445.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *nature* 521(7553):436–444.

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. *Neural computation* 1(4):541–551.

Lee CJ, Andrade EB, Palmer SE (2013) Interpersonal relationships and preferences for mood-congruency in aesthetic experiences. *Journal of Consumer Research* 40(2):382–391.

Lerner N (2009) *Music in the horror film: Listening to fear* (Routledge).

MacDorman KF Stuart Ough Chin-Chang Ho (2007) Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research* 36(4):281–299.

McElrea H, Standing L (1992) Fast music causes fast drinking. *Perceptual and Motor skills* .

Milliman RE (1986) The influence of background music on the behavior of restaurant patrons. *Journal of consumer research* 13(2):286–289.

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

Netzer O, Lemaire A, Herzenstein M (2019) When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research* 56(6):960–980.

Nielsen MA (2015) *Neural networks and deep learning*, volume 25 (Determination press San Francisco, CA).

North AC, Hargreaves DJ (2010) Music and marketing. *Handbook of music and emotion: Theory, research, applications* 909–930.

North AC, Hargreaves DJ, McKendrick J (1997) In-store music affects product choice. *Nature* 390(6656):132–132.

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

Peretz I, Zatorre RJ (2003) *The cognitive neuroscience of music* (Oxford university Press).

Petty RE, Cacioppo JT, Sedikides C, Strathman AJ (1988) Affect and persuasion: A contemporary perspective. *American Behavioral Scientist* 31(3):355–371.

Picard RW (2000) *Affective computing* (MIT press).

Plomp R, Levelt WJM (1965) Tonal consonance and critical bandwidth. *The journal of the Acoustical Society of America* 38(4):548–560.

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).

Puccinelli NM, Wilcox K, Grewal D (2015) Consumers' response to commercials: when the energy level in the commercial conflicts with the media context. *Journal of Marketing* 79(2):1–18.

Raji ID, Buolamwini J (2019) Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Rasch R, Plomp R (1999) The perception of musical tones. *The psychology of music*, 89–112 (Elsevier).

Ribeiro MT, Singh S, Guestrin C (2016) " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Russell JA (1980) A circumplex model of affect. *Journal of personality and social psychology* 39(6):1161.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sethares WA (2005) *Tuning, timbre, spectrum, scale* (Springer Science & Business Media).

Shukla A, Gullapuram SS, Katti H, Yadati K, Kankanhalli M, Subramanian R (2017) Evaluating content-centric vs. user-centric ad affect recognition. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 402–410.

Singh A, Sengupta S, Lakshminarayanan V (2020) Explainable deep learning models in medical image analysis. *Journal of Imaging* 6(6):52.

Smith PC, Curnow R (1966) " arousal hypothesis" and the effects of music on purchasing behavior. *Journal of applied psychology* 50(3):255.

Strick M, de Bruin HL, de Ruiter LC, Jonkers W (2015) Striking the right chord: Moving music increases psychological transportation and behavioral intentions. *Journal of Experimental Psychology: Applied* 21(1):57.

Tamir M (2016) Why do people regulate their emotions? a taxonomy of motives in emotion regulation. *Personality and social psychology review* 20(3):199–222.

Teeny JD, Siev JJ, Briñol P, Petty RE (2021) A review and conceptual framework for understanding personalized matching effects in persuasion. *Journal of Consumer Psychology* 31(2):382–414.

Tiwari V (2010) Mfcc and its applications in speaker recognition. *International journal on emerging technologies* 1(1):19–22.

Vermeulen I, Beukeboom CJ (2016) Effects of music in advertising: Three experiments replicating single-exposure musical conditioning of consumer choice (gorn 1982) in an individual setting. *Journal of Advertising* 45(1):53–61.

Ward MK, Goodman JK, Irwin JR (2014) The same old song: The power of familiarity in music choice. *Marketing Letters* 25(1):1–11.

Yang J, Xie Y, Krishnamurthi L, Papatla P (2021a) High-energy ad content: A large-scale investigation of tv commercials. *Available at SSRN* .

Yang J, Zhang J, Zhang Y (2021b) First law of motion: Influencer video advertising on tiktok. *Available at SSRN 3815124* .

Yang Y, Chen H (2011a) Predicting the distribution of perceived emotions of a music signal for content retrieval. *IEEE Trans. Audio, Speech and Lang. Proc*, volume 19, 2184–2196.

Yang YH, Chen HH (2011b) *Music emotion recognition* (CRC Press).

Appendix

A. Music Definitions

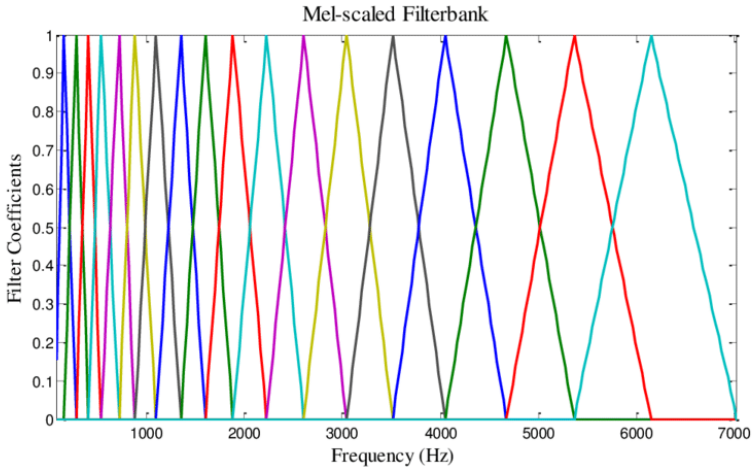
Table A1 Definitions of Musical Concepts

Feature	Type	Definition
Frequency	Physical	The number of cycles a sine wave completes in a second, measured in Hertz (Hz)
Fundamental Frequency	Physical	Lowest natural frequency of a sine wave
Partial	Physical	Any of the sine waves that comprise sound
Harmonic	Physical	A frequency that is an integer multiple of the fundamental frequency
Spectrum	Physical	The range of frequencies contained in a signal
Musical Interval	Physical	Spacing between two sounds in frequency
Octave	Physical	Interval between two sounds with a frequency ratio of 2:1
Fifth	Physical	Interval between two sounds with a frequency ratio of 3:2
Tritone	Physical	Interval composed of three adjacent whole tones (frequency ratio of 45:32 or 64:45)
Pitch	Perceptual	The attribute of sound that allows it to be ordered on a scale from low to high
Note / Tone	Perceptual	A pitched sound
Scale	Perceptual	A finite set of pitches that subdivide an octave into twelve notes
Key	Perceptual	The scale on which music is based
Pitch Class	Perceptual	Set of all pitches that are an integer number of octaves apart
Harmony	Perceptual	Set of pitches played simultaneously
Tonalness		Music that has a specific note on which it is the most stable and at rest
Consonance	Perceptual	A combination of notes that sound pleasant when played simultaneously
Dissonance	Perceptual	A combination of notes that sound harsh or jarring when played simultaneously
Beating	Perceptual	The phenomenon that occurs when two sounds of slightly different frequencies are played together and create an interference pattern, resulting in a periodic variation in volume whose rate is the difference of the two frequencies
Tempo	Perceptual	Perceived beat (sense of equally spaced temporal units) per minute
Loudness	Perceptual	The intensive attribute of an auditory sensation, in terms of which sounds may be ordered on a scale extending from soft to loud (ANSI 1960)
Timbre	Perceptual	The attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar (ANSI 1960)

B. Consonance Blinders Transformation

We describe the process to transform the STFT blinders to mel blinders. We will use octaves blinders as the running example. To simplify the problem, let us assume that the frequency dimension of the STFT is continuous for now. For octaves and fundamental frequency f_0 we retain the frequencies f_0, $f_1 = 2f_0$, $f_2 = 2^2 f_0$,..., $f_n = 2^n f_0$. To account for human auditory perception, we allow for a band of frequencies centered around each frequency. For low frequencies, the bandwidths are based on a constant bandwidth so our retained frequencies with bandwidth δ are of the form: $[f_n - \delta, f_n + \delta]$. For high frequencies, the

Figure B1 Mel Filter Bank



bandwidths are proportional to frequency so our retained frequencies with proportion α are of the form: $[f_n(1-\alpha), f_n(1+\alpha)]$.

We allocate the power associated with the frequencies in each band to the mel bands. The mel filter bank maps frequencies to the mel bands. Each frequency maps on to a maximum of two mel bands. Figure B1 shows the mapping of frequencies to 20 mel bands (we use 128 but it is more challenging to visualize). The top of each triangle represents the center of each mel band. Each triangle represents the weight each frequency contributes to a particular mel band. For example, the right most triangle maps frequencies ranging from roughly 5,400 Hz to 7,000 Hz to the 20th mel band. All frequencies below 5,400 Hz receive zero weight. The triangles grow wider as we move to higher frequencies because human hearing resolution is worse at higher frequencies.

Let β_j represent the function that maps frequencies to mel band j (i.e., the triangles). Let b^- and b^+ represent the lowest and highest frequencies, respectively, which map to mel band j and b the midpoint of the two numbers $(\frac{b^- + b^+}{2})$. β_j is defined as:

$$\beta_j(x) = \begin{cases} 0 & \text{if } x < b^- \\ \frac{x - b^-}{b - b^-} & \text{if } b^- \leq x \leq b \\ \frac{b^+ - x}{b^+ - b} & \text{if } b \leq x \leq b^+ \\ 0 & \text{if } x > b^+ \end{cases} \tag{10}$$

Then the contribution of frequency band $[f_n - \delta, f_n + \delta]$ (or $[f_n(1-\alpha), f_n(1+\alpha)]$) to mel band j, M_{jn}, is:

$$M_{jn} = \int_{f_n - \delta}^{f_n + \delta} \beta_j(x) P(x) \phi(x) dx \tag{11}$$

where x represents frequency, $P(x)$ the associated power of x, and $\phi(x)$ the distribution over frequencies. We assume $\phi(x) \sim U[f_n - \delta, f_n + \delta]$. Since multiple frequency bands could contribute to a single mel band, we sum the contributions so the final power of mel band j as M_j is:

$$M_j = \sum_n M_{jn}. \tag{12}$$

The set of power over all j comprises the mel blinders. They highlight which mel bands are input to the CNN and the weight of each band that is seen. An alternative way to implement the transformation is with discrete STFT bins. In this case, the frequencies above would be replaced by the center frequencies of each STFT bin and the integral would instead be a summation.

C. Music Feature Interpretability

Table C1 Music Features by Interpretability Level (Fu et al. 2010)

Feature Type	Musical Construct	Examples
Top-level labels	Emotion	Valence/arousal
	Genre	Pop, rock, country
	Instrument	Piano, violin, flute
Mid-level features	Harmony	Chord sequences
	Rhythm	Beat histogram
	Pitch	Pitch histogram, chroma
Low-level features	Frequency (timbral)	MFCC, zero crossing rate
	Time (temporal)	Amplitude modulation, statistical moments

D. CNN Architecture

The following table describes standard CNN modeling choices and their purpose.

Table D1 Standard CNN Model Ingredients

Ingredient	Purpose
Convolution Filter Channels	Convolution filters can be learned over multiple channels to learn different features. If height and width represent the y- and x-axes, channel can be thought about as the z-axis. The loss function incentivizes the model to learn different features over different channels. The optimal number of channels is determined empirically.
Batch Normalization	This procedure standardizes the inputs and serves several purposes. It helps achieve faster training, reduces overfitting, and ensures that none of the activations become outliers. A mini-batch represents the subset of data points the model sees each time it updates the model parameters; these are used rather than the full training set for memory and computational efficiency. Batch normalization standardizes the inputs (mean zero, standard deviation one) in a mini-batch so that the distribution of the inputs does not dramatically shift each mini-batch, stabilizing the learning process.
Rectified Linear Activation Unit	Activation functions are used to transform the output of convolution in CNNs to allow the model to learn nonlinear relationships in the data. ReLU activation takes the max of 0 and the input to the activation function. An important advantage of ReLU relative to a sigmoid or hyperbolic is that only a subset of neurons are simultaneously activated. ReLU is commonly used with CNNs because it does not suffer from the vanishing gradient problem like other activation functions and can therefore accelerate model learning.
Dropout	Dropout randomly removes neurons in specified layers of the neural network each mini-batch based on the specified dropout rate. To prevent overfitting and obtain a more robust model, we use dropout as a form of regularization. Dropout reduces the capacity of the network, reducing the likelihood of input memorization (learning features specific to the training set), and instead forces it to learn more robust features (Nielsen 2015). Specifically, we test dropout rates of 0%, 25%, 50%, and 75% and find that 50% results in the best model performance.
Fully Connected Layer	A fully connected layer connects every neuron in the hidden layer previous to every neuron in the next layer. We test including one, two, and three fully connected layers at the end and find no difference between using two and three layers and therefore opt for two.

Table D2 CNN with Square Filter Architecture

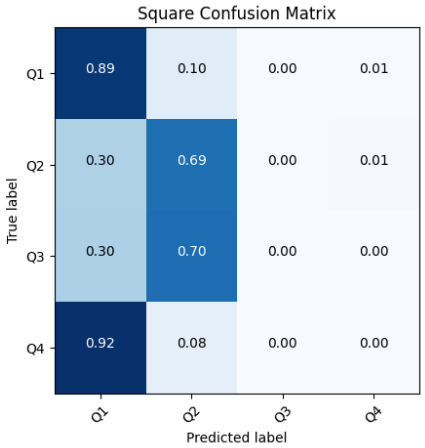
Layer	Size	Kernel Size	Stride	Pad
Input Image	128×517			
Convolution1		$5 \times 5 \times 64$	2	2
Batch Norm1				
ReLU1				
Convolution2		$3 \times 3 \times 64$	1	1
Batch Norm2				
ReLU2				
MaxPool2		2×2	1	0
Dropout2	dropout rate = 0.5			
Convolution3		$3 \times 3 \times 128$	1	1
Batch Norm3				
ReLU3				
Convolution4		$3 \times 3 \times 128$	1	1
Batch Norm4				
ReLU4				
MaxPool4		2×2	1	0
Dropout4	dropout rate = 0.5			
Convolution5		$3 \times 3 \times 256$	1	1
Batch Norm5				
ReLU5				
Convolution6		$3 \times 3 \times 256$	1	1
Batch Norm6				
ReLU6				
Convolution7		$3 \times 3 \times 384$	1	1
Batch Norm7				
ReLU7				
Convolution8		$3 \times 3 \times 512$	1	1
Batch Norm8				
ReLU8				
Convolution9		$3 \times 3 \times 256$	1	0
Batch Norm9				
ReLU9				
AveragePool9			1	0
FC10	256×4			

Note: The kernel size dimension is height × width × number of channels.

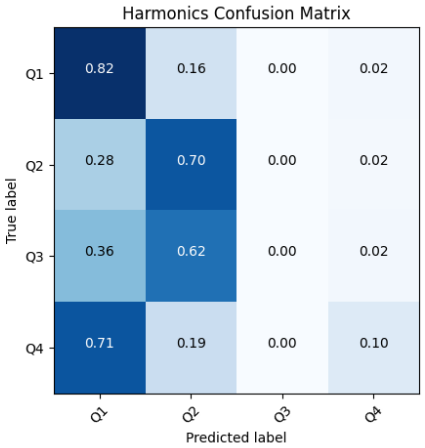
E. Model Confusion Matrices

Figure E1 Confusion Matrices

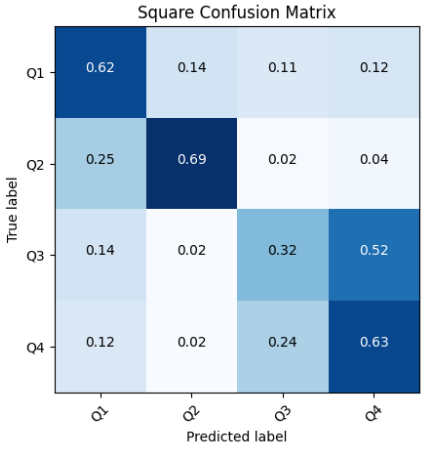
(a) Square Filters—Soundtracks Data



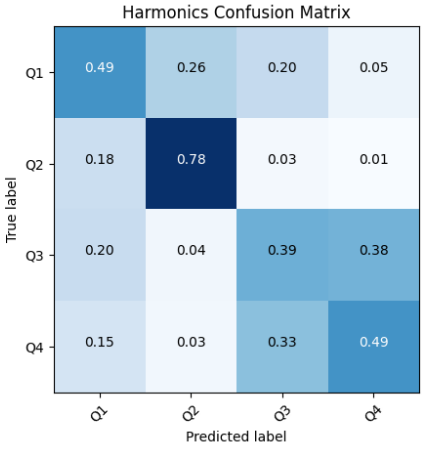
(b) Harmonics Filters—Soundtracks Data



(c) Square Filters—4Q Data



(d) Harmonics Filters—4Q Data



F. Classifier Performance for Combined Datasets

Table F1 Combined Soundtracks and 4Q Classification Performance

Features	Model	Precision	Recall	F_1
Current Atheoretical Filters				
MFCC	SVM	0.4460	0.4452	0.4321
		(0.0487)	(0.0519)	(0.0550)
Square	CNN	0.5431	0.5317	0.5123
		(0.0942)	(0.0759)	(0.0870)
Current Theoretic Low-Level Filters				
Frequency	CNN	0.4687	0.4666	0.4347
		(0.0805)	(0.0663)	(0.0809)
Time	CNN	0.2766	0.3420	0.2559
		(0.1680)	(0.0663)	(0.0923)
Time-Frequency	CNN	0.4747	0.4777	0.4500
		(0.0752)	(0.0682)	(0.0820)
Proposed Mid-Level Consonance Filters				
Harmonics	CNN	0.5089	0.5174	0.5022
		(0.0753)	(0.0677)	(0.0722)
Octaves	CNN	0.4348	0.4317	0.4062
		(0.0610)	(0.0632)	(0.0686)
Fifths	CNN	0.4228	0.4301	0.4028
		(0.0382)	(0.0414)	(0.0449)

G. Grad-CAM Visualizations

Figure G1 Harmonics Grad-CAM Heatmaps

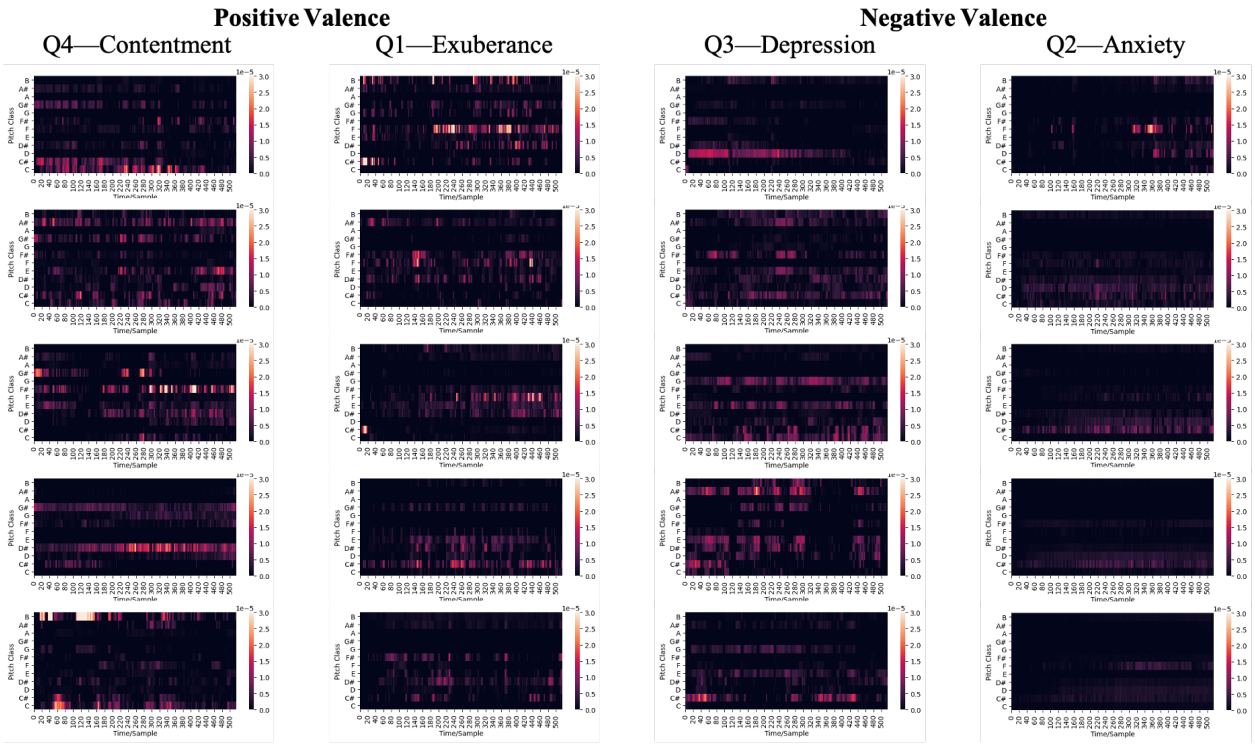


Figure G2 **Octaves Grad-CAM Heatmaps**

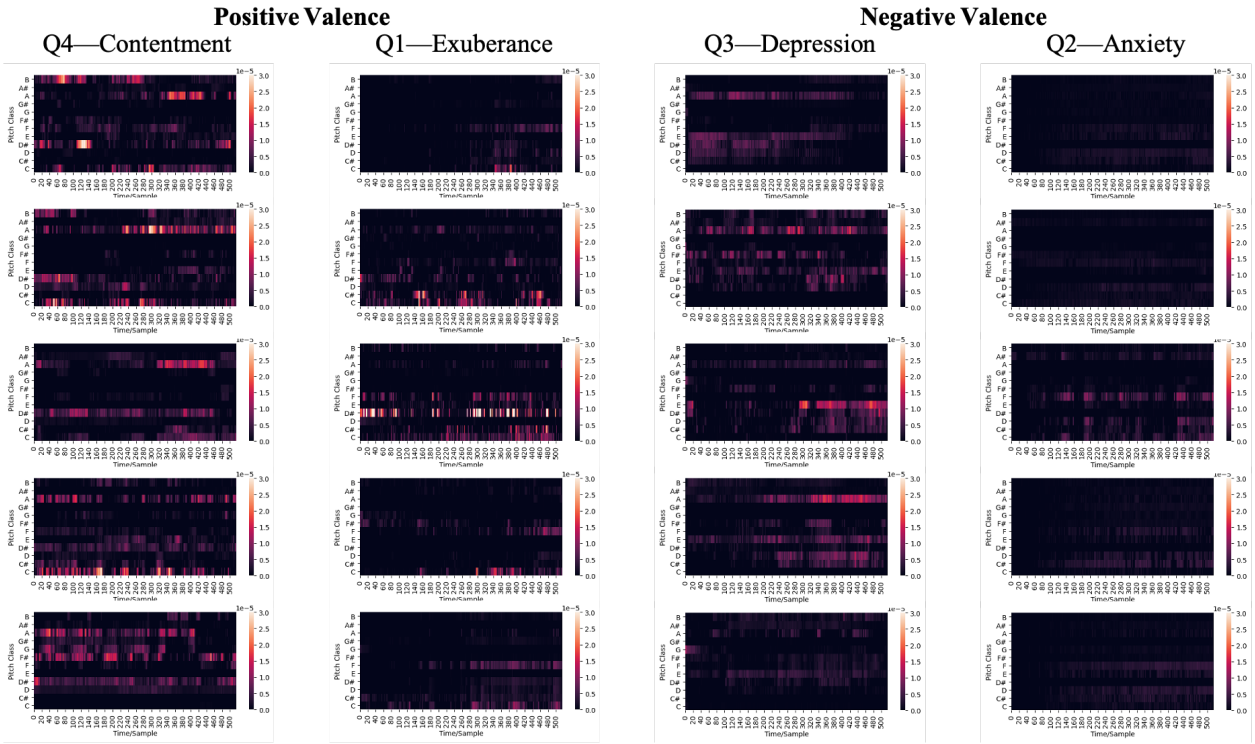


Figure G3 **Frequency Grad-CAM Heatmaps**

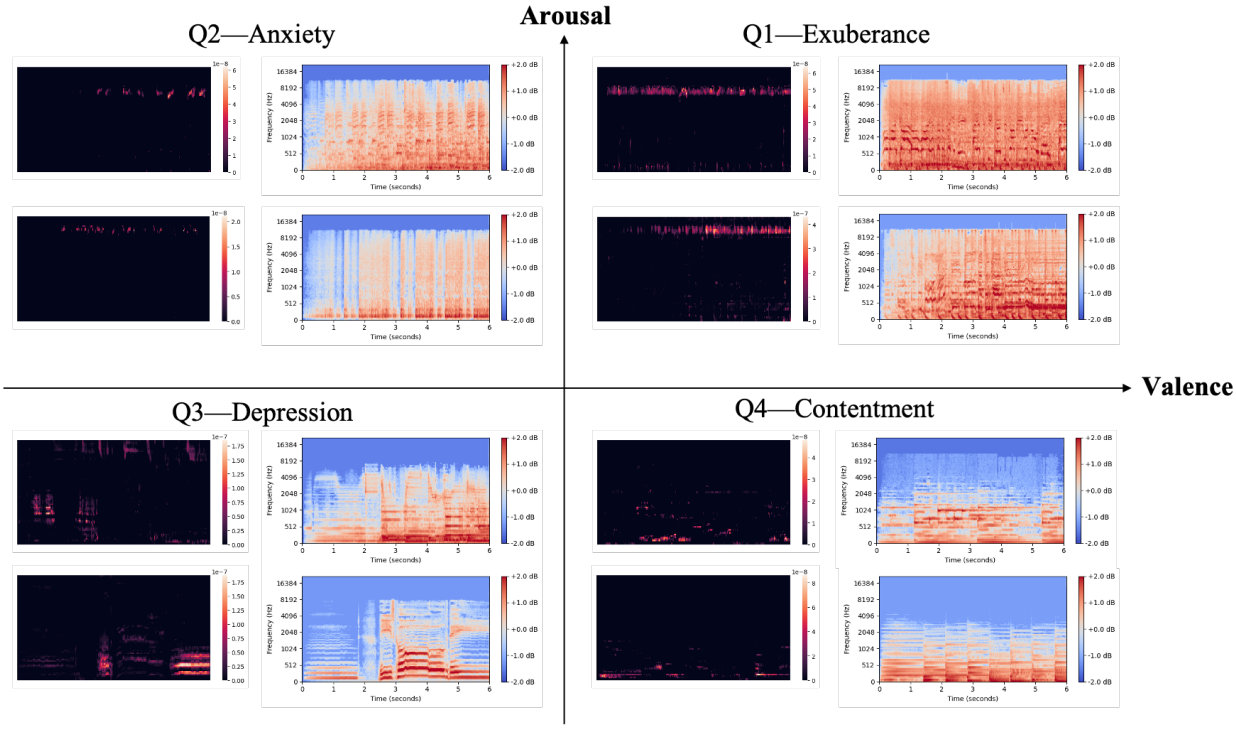
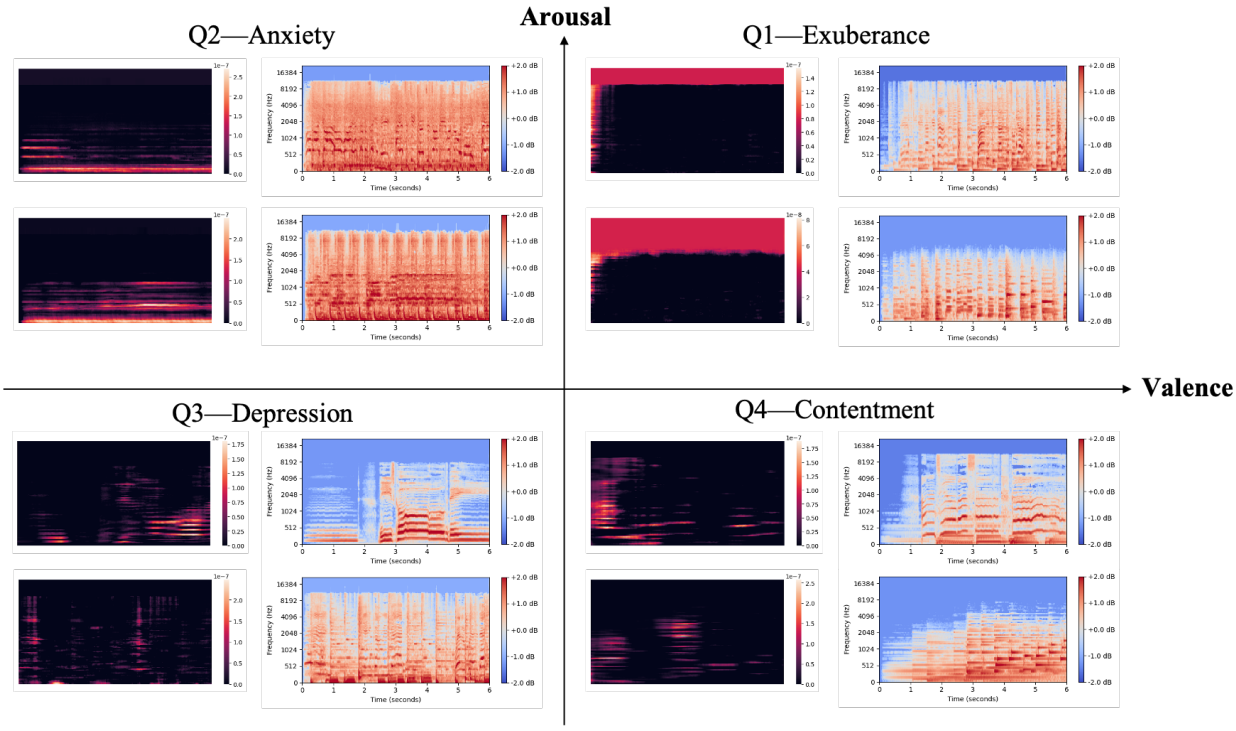


Figure G4 Time Grad-CAM Heatmaps



H. Application Study Details

Table H1 Number of Survey Participants Used in Study

Task	Number of Participants
Placement Emotion Tagging	49
Ad Emotion Tagging	44
Ad Insertion Experiment - Most and Least Distant	201
Ad Insertion Experiment - Additional Distances	346
Total	640