

# A Theory-Based Explainable Deep Learning Architecture for Music Emotion

(Authors' names blinded for peer review)

August 18, 2023

Music is used to evoke emotion throughout the customer journey. This paper develops a theory-based, explainable deep learning convolutional neural network (CNN) classifier to predict the time-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. Next, we design and construct novel CNN filters based on the physics of sound waves associated with the perceptual feature of consonance, which is 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) is related to human-interpretable features of music. The structure of the filters enables explainability while remaining flexible enough for the model to capture other music features that influence emotion. The classification performance of our proposed theory-based model is as good as state-of-the-art black-box CNN models. Finally, we illustrate an application involving digital advertising. Motivated by YouTube’s mid-roll ads, we first conduct a lab experiment in which we exogenously place ads at different times within content videos and find that ads placed in emotionally similar contexts increase ad engagement in terms of lower skip rates and higher brand recall rates. For a given ad, we use the model’s predictions to identify emotionally similar contexts in content videos.

*Key words:* audio data, deep learning, explainable AI, emotion, music theory, digital advertising

---

## 1. Introduction

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 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), marketers routinely use music to evoke emotion along the customer journey, from need recognition to purchase, in advertising, content marketing, and physical stores (Gorn 1982, Krishna 2012). In particular, music is almost universally used in advertising and advertisers spend considerable effort crafting it to elicit a desired emotion and marketing outcomes.<sup>1</sup> As such, a tool that can map music (or an ad with music) to its evoked emotion in listeners can be valuable to marketers. For example, marketers can use it to automate emotion-based contextual matching of ads and content at scale to increase ad engagement (Shukla et al. 2017). On music and video platforms, it can be used to automate creation of mood-based playlists and “next song” recommendations for users from their large catalogs.<sup>2</sup>

In this paper, we develop an explainable, theory-based deep learning convolutional neural network (CNN) model that takes music as input and predicts the sequence of emotions it evokes in a listener.<sup>3</sup> For many problems involving unstructured data (e.g., text, images, videos, music), deep learning models have been able to improve prediction accuracy, relative to more traditional feature-engineered models but as black-box predictors, their predictions are harder to explain. In this paper, we propose that by imposing structure from the theory of how music maps to emotion when designing the deep learning model, we can improve explainability relative to atheoretic deep learning models without sacrificing accuracy.

We now elaborate on the two key distinguishing features of the deep learning model. First and most important, we illustrate how music theory can be incorporated when designing

<sup>1</sup> A content analysis of over 3,000 ads showed that 94% of ads use music (Allan 2008). Further, over 75% of advertising hours in broadcast media uses music in some form (Huron 1989). As Huron (1989) states: “on a second-for-second basis, advertising music is the most meticulously crafted and most fretted about music. . .”

<sup>2</sup> 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.

<sup>3</sup> In the model, we characterize emotion using the well-established and widely used valence-arousal framework developed by Russell (1980) (see Figure 1 S6 for a visualization). 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. Many researchers have adopted the valence-arousal framework for music emotion classification (e.g., Panda et al. 2018, Yang and Chen 2011a, MacDorman 2007).

CNN filters for deep learning. Current atheoretic CNN implementations for music typically adapt the models developed for computer vision images where spatial contiguity in the image is meaningful. But spatial contiguity has less meaning for musical representation (which is based on representing the amplitude of sound waves of varying frequencies), so filters inspired by vision are inadequate and/or inefficient for music. We therefore design filters for musical constructs, like consonance, that have theoretical meaning based on acoustic physics and human auditory response and well-established empirical links to human emotion.<sup>4</sup>

Second, in contrast to deep learning models that are typically black-box predictors, our theory-driven filters make the model more explainable because there is a link between conceptually well-understood music characteristics and predicted emotion. To aid explainability, we visualize the link between the features learned by the theory-motivated filters and emotion in a post-estimation stage. For this, we adapt recent advances in the visualization of deep learning models for computer vision. The explainability allows us to go beyond model prediction, giving managers confidence that the model learns features of the music that can generalize outside of the training data. Explainability helps managers develop trust in the model and hence adopt it at scale.<sup>5</sup>

Finally, we illustrate the practical value of our model in 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, an ad can be placed in a number of different ad breaks. Given the scale, identifying the optimal locations in the videos for ad insertion requires automation. We examine whether matching the ad’s emotion with the emotion of the content at the ad insertion point can improve ad effectiveness (e.g., reduced ad skip, higher brand recall).<sup>6</sup> To test this, we conduct a lab experiment in which we insert ads at points in videos of varying emotional similarity. We find that greater emotional

<sup>4</sup> Consonance refers to a combination of notes that sound pleasant when played simultaneously. Dissonance, or the lack of consonance, refers to a combination of notes that sound harsh or jarring when played simultaneously (Müller 2015).

<sup>5</sup> In 2020, McKinsey conducted a survey that found that after cybersecurity and regulatory compliance, explainability is the risk firms are most worried about regarding AI. Source: <https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2020>

<sup>6</sup> We use the emotion elicited by only the music as a proxy (or the dominant modality) for the emotion elicited by the video. This is reasonable given that music is typically designed to elicit the intended emotion in videos (Bullerjahn and Gildenring 1994, Herget 2021). We leave the question of how other modalities in video (e.g., text of dialogues, images) jointly impact emotion along with music to future work.

similarity decreases ad skip and increases brand recall. We then input the audio of the ads and content videos into our model to predict emotion and use the predicted emotion to determine the most emotionally similar ad insertion point for each ad within each content video.

We provide an outline of the details of our model 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 explainable. Our methodological approach uses the raw sound wave of a music clip rather than pre-defined engineered features as the starting input. We transform the sound wave into a mel spectrogram, which reflects how humans hear and is the input into the CNN. Details of why and how we transform the audio are discussed in §3. The mel spectrogram is a two-dimensional image with frequencies along the  $y$ -axis on a log scale, time along the  $x$ -axis, and the square of the amplitude (i.e., volume) corresponding to the frequencies represented in color. This 2D representation is useful because it allows us to “read” some of the clip’s musical features, such as frequency range and note density. Recent research developing and applying CNN for music uses mel spectrograms as input (e.g., Pons et al. 2016, Chowdhury et al. 2019, Rajaram and Manchanda 2023).

CNNs are deep neural networks specially built for image processing, in which 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, and 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. Typical small square CNN filters (which account for contiguous areas of an image) cannot capture such constructs based on non-contiguous frequencies, 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. By

leveraging ideas from music theory to build the model, we set the stage for explainability post-estimation.

For explainability, we adapt Grad-CAM (Selvaraju et al. 2017), 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. Consider classifying images of wolves and huskies. Suppose a CNN trained to distinguish wolves from huskies highlights the background of the image as the main predictor because wolves are often photographed in snow while huskies in grass (Ribeiro et al. 2016). If the background, a spurious correlation, drives prediction, a new example of a husky playing in snow is more likely to be classified incorrectly as a wolf. With Grad-CAM highlighting the background, we learned that the model did not learn to distinguish the animals based on features of the animals themselves, putting the generalizability of the model into question. Knowing why a model makes various predictions helps assess model generalizability, which facilitates greater trust and hence wider adoption.

In the context of music, for a model to be explainable, it should relate the top-level outcome/label of interest (e.g., emotion) to a mid-level set of features (e.g., harmony, rhythm, pitch) related to music.<sup>7</sup> While low-level features (e.g., frequency, time) provide some degree of transparency, Fu et al. (2010) argue they do not have a clear explainable link to top-level labels. To overcome this challenge, we identify a mid-level music feature closely connected to emotion to motivate the design of our convolution filters. Combining the mid-level feature with a model visualization technique like Grad-CAM enables explainability of the CNN by connecting a feature with clear musical meaning to the emotion classification. We design the convolution filters to show which parts of a song relate to the mid-level feature of consonance and the final emotion classification. Our paper highlights a strategy for explainability based on constructs whose links with the classification have clear theoretical motivation.

It is important to note that the features learned using theory-based filters are different from handcrafted features (e.g., mel-frequency cepstral coefficients) that are used with traditional machine learning models (e.g., support vector machine). The construction of handcrafted features is mathematically predefined but the filters learn weights from the data to transform the music into features. Thus, the structure imposed by the filters enables

<sup>7</sup> Appendix Table B1 organizes features used in music classification by level of interpretability.

Grad-CAM visualization of consonance but the filters do not restrict the model to learn only consonance. There is enough flexibility in the model to learn other musical features that impact emotion, like whether high harmonics or low harmonics are present, loudness, and pitch (see Appendix Table A1 for music definitions). It is this flexibility that enables the model to classify music emotion accurately.

Summarizing, our key contributions are as follows. First, we develop a theoretically-motivated, explainable deep learning framework that allows us to model and predict time-varying emotional response throughout the duration of a music clip. Second, our approach integrates a theoretically-motivated feature of music, consonance, that is known to impact emotion. We design novel CNN filters that enable the visualization of consonance in a music clip and its connection with emotion. Third, 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 explainability of deep learning models.<sup>8</sup>

The rest of the paper is structured as follows. §2 overviews the relevant literature. §3 describes the deep learning pipeline that takes music audio waves as input and outputs a measure of emotion. §4 describes model training, §5 describes the application—ad insertion in online videos, and finally §6 concludes.

## 2. Related Literature

Our paper builds upon several streams of literature across different academic fields. We organize our discussion in four sections: (1) Listener Response to Music; (2) Machine Learning with Unstructured Audio Data; (3) Explainable AI; and (4) Ad Insertion in Videos.

### 2.1. Listener Response to Music

Music induces emotion, as shown by a wide literature using methods ranging from surveys to brain scans (Johnson-Laird and Oatley 2016). Researchers often measure emotion using Russell’s circumplex model (e.g., Yang and Chen 2011b), which maps emotion onto the two underlying dimensions of valence and arousal. Valence and arousal levels are in turn

<sup>8</sup> Researchers have advocated for theory-based deep learning models in domains outside of music (e.g., physics) to aid better representation learning of systems and increase robustness (Raissi et al. 2017, Vasudevan et al. 2021).

associated with certain musical feature settings. For example, in Western tonal<sup>9</sup> music “consonance usually signals stability and positive emotional valence, whereas dissonance signifies tension, instability, and negative emotional valence” (Gabrielsson and Lindström 2010). Similarly, other features like loudness and timbre are also associated with emotion (Gabrielsson and Lindström 2010, Gabrielsson 2016).<sup>10</sup> Loud music is associated with high arousal while soft music is associated with low arousal. In terms of timbre, music with many high harmonics is associated with high arousal while music with low harmonics is associated with low arousal. Since our focus is on the background music of content and ads, which typically falls under Western tonal music (Stoppe 2014, Nelson et al. 2013),<sup>11</sup> we focus on the musical associations between Western tonal music and emotion.<sup>12</sup>

These emotional responses have marketing implications, and a substream of literature focuses on the relationship between music characteristics and marketing outcomes. Bruner (1990) overviews how music elicits different moods, which in turn impact ad outcomes. Yang et al. (2022) 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 albums.

While this paper focuses on emotion, music can impact listeners through other mechanisms, such as through music associations and perceptions of time (North and Hargreaves 2010). Our research is broadly related to the literature on sensory marketing as music impacts consumers through the auditory sense (Krishna 2012).

## 2.2. Machine Learning with Unstructured Audio Data

While audio broadly includes both speech and music, our focus in this paper is on music. Traditional machine learning methods, like SVM, previously produced good classification

<sup>9</sup> Christensen (2006) writes, “tonality most often refers to the orientation of melodies and harmonies toward a referential (or tonic) pitch class.”

<sup>10</sup> See Appendix Table A1 for definitions of the musical features.

<sup>11</sup> Nelson et al. (2013) writes “Most films, however, are targeted at a broad, global audience with the implicit understanding that they share a common familiarity with Western tonal music, so the overtone series is the foundation of the cinematic musical language.”

<sup>12</sup> In Western tonal music, consonance and dissonance play a major role in creating emotion and we use the physical properties of consonance to construct the filters for our model. Many studies have found consonance to be inherent (Zentner and Kagan 1998) and universal across cultures (Bowling and Purves 2015) but other studies have found consonance to be culture-dependent and it remains an active area of debate and research (McDermott et al. 2016). It should also be noted that in some musical cultures, such as that of Indonesia, consonance and dissonance do not play a major role in inducing emotion. In such musical cultures, it would be more appropriate to develop a theory-based model that focuses on other musical features.

performance in many settings by using hand-crafted features, such as mel frequency cepstral coefficients (MFCCs). However, the performance of deep learning algorithms has overtaken almost all other methods in audio applications, similar to vision applications (Hinton et al. 2012). The crucial advantage that deep learning has is that features are automatically learned from data, rather than pre-specified (Choi et al. 2017a).

With deep learning, music is typically converted to a spectrogram, which is used as the input to the learning algorithm. A few researchers have attempted to build music-specific deep learning models rather than use CNN models designed for image recognition. To predict 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, Dubois et al. (2019) propose a deep learning model to predict sensory dissonance in piano chords. Using data with not only labeled emotion but also labeled mid-level features (e.g., melodiousness, articulation), Chowdhury et al. (2019) build a deep learning model that includes an interpretable mid-level feature layer that is then used to predict emotion. We contribute to the machine learning from audio literature by designing and developing novel CNN filters that are able to characterize consonance, and demonstrate how the filters are useful for prediction as well as explanation.

### 2.3. Explainable AI

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

Without explainability, it is challenging to trust AI systems since we do not understand when and how they may break (Meske and Bunde 2020, Tomaino et al. 2020). The example of the husky in grass and wolf in snow in the introduction highlights the potential pitfalls



of not having explainability. Only by knowing why the model makes the prediction it does, can we know whether the model is generalizable to data outside of the training data.

We approach explainability 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 filters place an explainable 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 to enable explainability, and not just ex post after training the model.

#### 2.4. Ad Insertion in Videos

We apply our theory-based music emotion model to an ad insertion application. There is an evolving literature examining video ads shown in streaming videos that tries to understand what ad characteristics, content video characteristics, and user characteristics impact ad performance. For a survey of this literature, see Frade et al. (2021).<sup>13</sup> The typical outcomes examined in video ads include ad skipping, brand recognition, intrusiveness, aided and unaided brand recall, ad acceptance, click through rate, and related metrics for ads accompanied by a call to action.

Much of the literature has studied only ad characteristics, without much focus on the context of the content video in which the ad appears. Studies that have focused on the impact of intrusiveness of ads, as measured by the number of ads within a break, find that brand recall is maximized with only one ad, rather than more (Bellman et al. 2012). Another experimental study examining ad length and content found that longer and more informative ads are considered less intrusive and received higher recall rates (Goodrich et al. 2015).

Papers exploring *both ad and content video characteristics* have studied the impact of ad relevance (e.g., showing a Ford ad in a video about Formula 1 racing) and ad congruence on performance. In terms of ad relevance, Zach et al. (2018) found that relevant advertising is important for users’ opinion of ads. Another study focusing on ad relevance found that for mid-roll ads, congruence between the ad product and the video content improves consumer receptivity, whereas the opposite is true for post-roll ads (Li and Lo 2015).

<sup>13</sup> For a survey focused on television advertising, see Wilbur (2008).

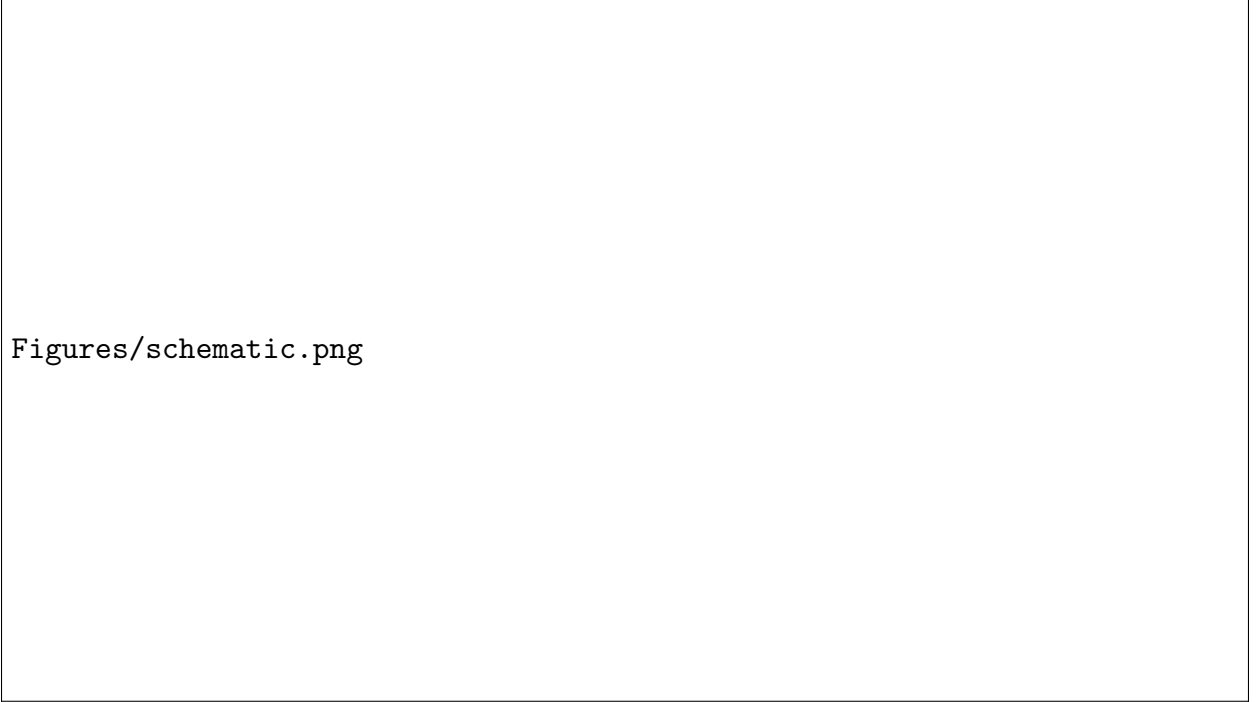
In terms of emotional congruence, results have been mixed. Belanche et al. (2017) found in an experiment that high-arousal ads are watched for longer in high-arousal content than in low-arousal content but found no such effect of congruence for low-arousal ads. On the other hand, although not in a video ad insertion context Puccinelli et al. (2015) found that consumers in a low-arousal state watch high-arousal ads for less time than moderate-arousal ads, supporting congruence. Kapoor et al. (2022) focused on the emotions of happiness and sadness and found in a field experiment that emotional contrast led to greater ad engagement. Overall, the results in the literature are mixed on whether emotional congruence improves ad performance.

One key difference between our setting and past research is our focus is on the time-varying emotion of the content video whereas past studies focused on the overall emotion of the content video for ad matching.

### 3. Model

We develop a CNN for emotion classification that includes several theoretically motivated components relating to the physics of sound waves and the perception of Western tonal music by listeners. We begin with an overview of the steps of our deep learning model that maps music to emotion in Figure 1. Step **S1** takes six seconds of raw audio sound wave data as input.<sup>14</sup> In Step **S2**, the music clip is converted to a short-time Fourier transform (STFT) spectrogram. In Step **S3**, we transform the STFT spectrogram to a mel spectrogram, which characterizes how the sound is perceived by the human ear. In Step **S4**, the mel spectrogram is used as a visual input to the 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. For performance comparison, we also consider atheoretical square and rectangular filters that are 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 into which the sound falls. Finally, stringing together the predictions of the six-second clips shows the emotion predictions of the music over time. Below we describe each of these steps in detail.

<sup>14</sup> We use six seconds of audio but the model can easily be adapted to incorporate other audio lengths.

**Figure 1 Music Emotion Classification Schematic**


Figures/schematic.png

**(S1) Physical Properties of Sound Waves:** 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 1 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 sound, which is based on the underlying frequencies.<sup>15</sup>

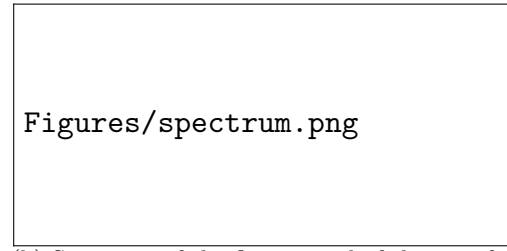
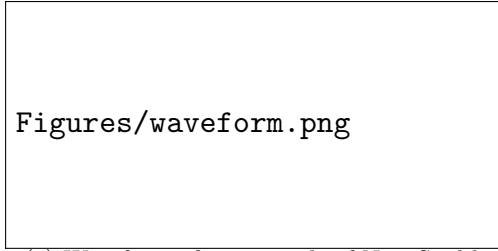
To get to musically relevant features, we need a representation of the different frequencies that the sound wave is composed of in terms of fundamental sine waves. This is because sine waves determine what humans hear and are at the foundation of musical concepts like pitch and harmony. The mathematical representation of this process is the Fourier transform, which decomposes a sound wave into its constituent sine waves. 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 2b shows the spectrum of the first second of music shown in the waveform Figure 2a. From the spectrum, we can identify the main frequencies that make up the sound and the magnitude of each frequency.

<sup>15</sup> Two different waveforms can map to the same sound.

**Figure 2 Example of Waveform and Spectrum**

(a) Waveform

(b) Spectrum



Notes: (a) Waveform of six seconds of New Soul by Yael Naim. (b) Spectrum of the first second of the waveform.

**(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.<sup>16</sup>

The parameters that go into generating an STFT spectrogram are the sampling rate, window type and size, and hop length.<sup>17</sup> 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:

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

where  $m$  is the time index,  $k \in [0 : K]$  is the frequency index, and  $i := \sqrt{-1}$ . 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:

$$S(m, k) := |X(m, k)|^2. \quad (2)$$

<sup>16</sup> 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 in terms of sample number rather than in seconds. Each second contains 44,100 samples.

<sup>17</sup> We set the sampling rate to 44,100 Hz, the window type to Hann, the window size to 4,096 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.

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  $2,049 \times 517$ .<sup>18</sup>

The STFT spectrogram of Figure 2a is represented in Figure 3a, with the x-dimension representing time, the y-dimension representing frequency, and color representing the power of each frequency bin at each time sample (red is high power and blue is low power). 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. In Figure 3a, the large patch of blue before second three indicates a lack of high frequencies early in the music clip.

**(S3) Mel Spectrogram based on Auditory Perception:** Humans are better at perceiving frequency differences at low pitches than at high pitches (Müller 2015).<sup>19</sup> With equal sensitivity across the frequency spectrum, the STFT spectrogram does not represent human hearing. 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.

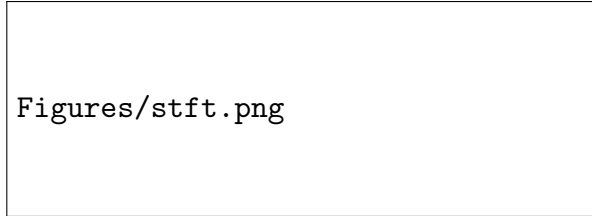
The additional parameter that goes into generating a mel spectrogram from an STFT spectrogram is the number of mel bands. The number of mel bands specifies the mel filter banks, which are the weights that map the STFT frequencies to the mel frequency scale. We use 256 mel bands. Figure 3b shows the mel spectrogram of Figure 2a, which more clearly displays the differences among the lower frequencies relative to Figure 3a.

<sup>18</sup> The dimensions of the spectrogram are obtained in a standard way as per Müller (2015) and come from the length of music, sampling rate, window size, and hop length. The time dimension is equal to length of music clip  $\times$  sampling rate / hop length = (6 seconds)(44,100 Hz)/512 samples = 517. The frequency dimension is equal to the window sample size / 2 + 1 = 4,096 / 2 + 1 = 2,049. A discrete Fourier transform (DFT), which underlies the STFT, generates redundant information when it transforms audio information from the time domain to the frequency domain and only the first half of the frequencies are meaningful. The DFT transforms the 4,096 samples to 4,096 frequency bins but we retain only half of them because the other half is redundant. The key idea is the larger the window sample size the greater the frequency resolution but the lower the time resolution.

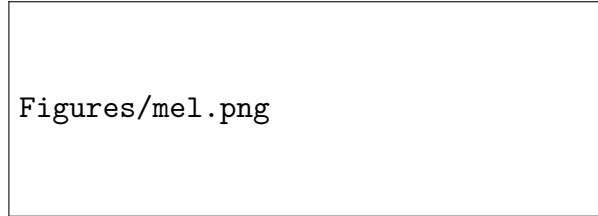
<sup>19</sup> Pitch is a subjective measure of frequency and is defined as the attribute of sound that allows it to be ordered on a scale from low to high. For a pure tone sine wave, the pitch and frequency are the same, and are determined by its fundamental or lowest frequency. However, they can differ for more complex and realistic sounds. In all cases, the higher the frequency, the higher the perceived pitch.

**Figure 3 Spectrograms**

(a) STFT Spectrogram



(b) Mel Spectrogram



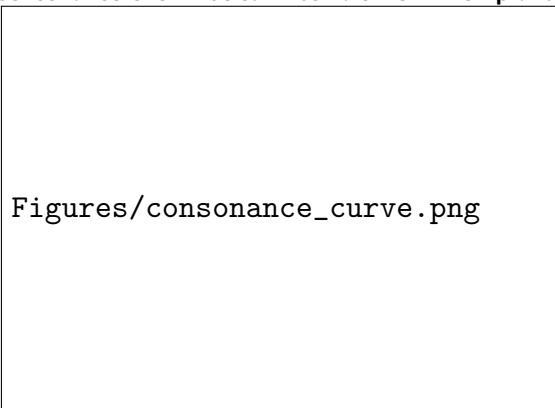
Notes: (a) STFT spectrogram of six seconds of New Soul shown in Figure 2a. The STFT spectrogram visualizes the time and frequency features of audio data. The x-axis represents discretized time, the y-axis represents discretized frequency, and color represents the squared magnitude of each frequency bin 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.

**(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 partial). For example, the harmonics of note A4 are  $f_0 = 440$  Hz,  $f_1 = 2f_0 = 880$ ,  $\dots$ ,  $f_n = (n + 1) \cdot f_0$ . 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.<sup>20</sup>

Consonance and dissonance have a strong connection with the physics of sound waves. We use this connection to make precise use of the theory in our deep learning model. Experimental studies have revealed that consonance and dissonance are not binary categories,

<sup>20</sup> 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 (Peretz and Zatorre 2003, Lerner 2009).

**Figure 4** 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. Points of greatest consonance occur at small integer ratios of the frequencies.

but rather the opposite ends of a continuum. When two notes with identical frequencies (i.e., unison) are played simultaneously, they are judged as highly consonant (Plomp and Levelt 1965, Rasch and Plomp 1999). As shown in Figure 4, borrowed from Plomp and Levelt (1965), unison is the point of global maximum of consonance and specific other two-note frequency intervals form local maxima. In general, consonance is associated with small integer 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.

Finally, we introduce the construct of pitch class. Western music is organized around 12 pitch classes: C, C $\sharp$ , D, D $\sharp$ , E, F, F $\sharp$ , G, G $\sharp$ , A, A $\sharp$ , B. Different combinations of these pitches lead to consonant and dissonant sounds. We use the 12 pitch classes to organize the filters.

We use the mathematically represented properties of consonance (i.e., overlapping harmonics) and the twelve pitch classes 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 the overall CNN architecture. The atheoretical square and rectangular filters and low-level frequency and time filters are discussed in §3.1.

**(S4) Filter Design and Construction:** CNNs have typically been used for problems involving visual image inputs. 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 a novel convolution filter design guided by music features associated with consonance, which impacts valence and arousal in terms of listener emotion. Consonance is associated with positive valence emotions (e.g., tenderness) while dissonance is associated with negative valence emotions (e.g., fear). Consonance is also associated with low arousal emotions (e.g., contentment) while dissonance is associated with high arousal emotions (e.g., fear) (Gabrielsson 2016).<sup>21</sup>

Since sounds with overlapping harmonics produce more consonant music in Western music, we propose consonance filters based on harmonics. The consonance filters use “blind-ers” 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 filters” 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  $\mathcal{H} := \{\omega_n\}$  of a tone with fundamental frequency  $f_0$  where  $n \in \mathbb{Z}^+$  contains frequencies:

$$\omega_n = n f_0. \quad (3)$$

*Pitch Class Blinders.* We use the set of frequencies defined in eq. (3) 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 (Table 1). In total, we construct 12 blinders (one for each pitch class).

**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.* Next we overview the steps to create the convolution filter for a given pitch class  $p$ .

- i *Calculate pitch class filter frequencies:* Beginning with the fundamental frequency  $f_0$  of pitch class  $p$ , calculate the set of harmonic frequencies. For example, the lowest C has

<sup>21</sup> Gabrielsson (2016) writes “In Western tonal music, consonance usually signals stability and positive emotional valence, whereas dissonance signifies tension, instability, and negative emotional valence.”



a fundamental frequency of 16.35 Hz, so the C harmonics frequencies are  $1 \times 16.35$  Hz,  $2 \times 16.35$  Hz,  $3 \times 16.35$  Hz, etc. The number  $n$  of harmonics is set to cover the entire range of the spectrogram ( $n = 803$  for harmonics).

- ii *Calculate frequency bands and STFT indicator column:* We create frequency bands to match the  $\omega_n$  associated with harmonics with the STFT bins. The underlying rationale is that the human ear cannot distinguish frequencies within a small band. 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. Since  $n$  is high for harmonics, we calculate three hertz bands centered around each frequency  $\omega_n$ . For each pitch class, create an STFT indicator column such that each element is equal to one when the center frequency of the STFT bin falls within one of the frequency bands. Create STFT indicator columns for each pitch class.
- iii *Construct mel blinders:* To convert the STFT blinders to mel blinders, we multiply the STFT indicator column by the mel filter bank to generate a mel weight column that has  $N_{mel} = 256$  dimensions. Repeating the mel weight column over the time dimension generates the mel blinders. We multiply the mel spectrogram by the mel blinders to generate the input to the CNN’s convolution layer. Appendix §C details the steps.
- iv *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 several architectural 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 designed for music. We design the convolutional and pooling layers to incorporate the physics of sound waves and music and emotion theory.

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 (the Online Appendix describes these standard modeling choices in detail). Dropout and early stopping prevent overfitting. Batch normalization and rectified linear activation unit (ReLU) facilitate model learning.

Since we seek to predict which of four valence-arousal quadrants a music clip falls into, our problem is a multiclass classification problem. The objective of the model is to learn parameters to minimize the cross-entropy loss between the predicted and actual outputs. Let  $k$  represent the quadrant,  $y_{ik}$  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:

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


The operation of the model for the consonance filters 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.
  - (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 one fully connected layer and apply softmax to output a probability distribution over the four emotion quadrants.
5. Output the quadrant with the max probability.

Figure 5 summarizes the overall architecture. Since our innovation is in the design of the consonance filters, we discuss the related model ingredients (i.e., convolution and pooling layers) below and summarize the standard CNN architectural decisions in the Online 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. The stride specifies how much the filter slides over the image before performing another operation. A small stride captures more fine-grained information but also requires more computational and resource costs. The output of convolution is a feature map. If height and width represent the  $y$ - and  $x$ -axes, channel can be thought about as the  $z$ -axis. It is common practice to learn convolution filters over multiple channels to learn different features.<sup>22</sup>

<sup>22</sup> The loss function incentivizes the model to learn different features over different channels. So for example, for images, one channel might learn to detect vertical edges, a second channel might learn to detect horizontal edges, and a third channel might learn to detect texture. The optimal number of channels is determined empirically.

**Figure 5 Model Architecture**


Figures/musicemocnn\_arch\_2.png

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 the fully connected layer. The final output is the valence-arousal quadrant prediction.

*Convolution Filter Design:* 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 explainability. A tool that provides 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 class by plotting the class’s gradients 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 explainable 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. If the frequencies were contiguous, they could easily be captured by a square or rectangular filter. The second

objective of explainability 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<sup>23</sup> since it produces a one-dimensional Grad-CAM heatmap over time.<sup>24</sup> We would expect that the resulting heatmap captures the concept of consonance over time since the CNN is shown only specific frequencies.

*Pooling.* The machine learning literature typically uses average and max pooling to summarize hidden layers. Pooling applies a function over all units within a specified shape (e.g., square) and is used for computational efficiency and to make the model invariant to small translations of the input (Goodfellow et al. 2016).

To summarize the features of a six-second clip, we need a summarization over time and over the pitch classes. We empirically test the four possible pooling combinations by training models according to the four architectures. As shown in Table K2 in the appendix, we find that average pooling across time and max pooling over pitch classes results in the best classification performance. We believe that this is because first, consonance over a time period is based on the average consonance during that period (and not the max). A clip that is perceived to be consonant throughout is more likely to be perceived to be consonant overall than a clip that is consonant at one point but otherwise dissonant. And second, we believe overall consonance is dominated by the most consonant pitch class. A clip with one highly consonant pitch class (i.e., many overlapping harmonics) is perceived to be more consonant than many pitch classes that are slightly consonant. We therefore use average pooling over time and max pooling over pitch classes in our main model specification.

While we use theory to inform the network architecture and impose structure on the filters, we do not restrict the model to only learn consonance. One way to think about the consonance filters is that they focus on specific non-contiguous parts of the image. The model is flexible enough to learn a wide variety of other features, such as pitch range and timbre. Returning to the wolf versus husky example (see Section 1), it would represent focusing on specific objects in the image (e.g., the animal’s face). The model is free to learn which features (e.g., eye color, nose shape) differentiate the animals.

<sup>23</sup> This convolution transforms the  $256 \times 517$  mel spectrogram to a  $1 \times 517$  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.

<sup>24</sup> Setting the filter width to one time sample produces a granular Grad-CAM heatmap but the tradeoff to doing so is it becomes more challenging to capture features related to rhythm. To better capture features related to the time dimension, we could use wider filters but this would reduce the explainability of the model.

**(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.

### 3.1. Benchmark Models for Comparison

In addition to CNNs that use the mid-level consonance filters, we train a number of benchmark deep learning models for comparison. The benchmark models can broadly be characterized as either atheoretical or musically-motivated but focused on low-level features.

The atheoretical models include CNN with (n-by-n) square filters, or with rectangular filters, both tall and skinny (2n-by-n) and short and wide (n-by-2n). These are borrowed from image recognition models. CNNs with square filters 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 some audio features but it is unclear what these are and how they relate to music emotion. Compared to square filters, rectangular filters of different shapes might allow us to capture features that span a larger portion of the frequency space or a larger portion of the time space.

The musically-motivated models 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). Tall and skinny (a-by-1) filters are designed to capture timbral features across the frequency spectrum, e.g., a specific combination of notes, while short and wide (1-by-b) 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 with an additional fully connected layer results in comparable performance. We provide additional implementation details for the benchmark models in Appendix §E.

## 4. Empirical Analysis

In this section, we begin by describing the dataset used to train the various models. We then report the performance of our proposed architecture, which uses consonance filters, and compare it against benchmark models proposed in the literature. Finally, we show how our model is explainable using gradient-based model visualizations and compare it to

visualizations generated by other CNN models which use atheoretical and low-level feature filters.

#### 4.1. Datasets

We combine two public datasets compiled by music emotion researchers, which serve complementary purposes in our analysis: Soundtracks (Eerola and Vuoskoski 2011) and the MediaEval Database for Emotional Analysis in Music (DEAM) (Aljanaki et al. 2017).

The Soundtracks dataset is comprised of 360 excerpts from movie soundtracks that range in duration from 10 to 30 seconds. One benefit of movie soundtracks is that they are composed to elicit emotion. The music clips are instrumental and do not contain any lyrics, dialogue, or sound effects. The clips were chosen to be unfamiliar to prevent song familiarity from impacting emotion tagging and to evoke only a single emotion over the length of the clip. University students and staff with musical expertise annotated the song emotions, and six annotators tagged each music excerpt.<sup>25</sup> Perceived valence and arousal were separately annotated on a scale of 1 to 9. Inter-rater agreement (Cronbach’s alpha) was 0.92 for valence and 0.90 for arousal. We split the excerpts into non-overlapping six-second segments. All six-second clips from the same excerpt have the same valence and arousal labels since these excerpts were chosen to only evoke a single emotion over the length of the clip. 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 high valence-high arousal, Q2 captures negative valence-high arousal, Q3 captures negative valence-low arousal, and Q4 captures positive valence-low 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—sadness, Q4—contentment.

The DEAM dataset is comprised of 1,802 mostly 45-second excerpts of royalty-free music. The music annotations were crowdsourced through MTurk and each excerpt was annotated by at least 10 workers. Perceived valence and arousal were annotated from a

<sup>25</sup> 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 DEAM. 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.

scale of  $-10$  to  $10$  every half second using a graphical interface. The clips were largely unfamiliar to workers. Inter-rater agreement was 0.28-0.66 for arousal and 0.20-0.51 for valence. In contrast to Soundtracks, the DEAM music was not chosen to elicit a particular emotion, which helps to explain the relatively low inter-rater agreement.<sup>26</sup> In addition, the DEAM music can vary in emotion over time. DEAM therefore complements Soundtracks by covering more of the valence-arousal space. We exclude music that pushes the boundaries of Western music theory like experimental and blues music. We divide each clip into non-overlapping six-second segments and average the 12 annotations taken every half second to obtain valence and arousal labels. To convert the continuous valence-arousal labels to the four discrete quadrants, we discretize the valence-arousal space around the midpoint  $(0,0)$ . We thus obtain the same Q1 to Q4 quadrants across both datasets.

Finally, we combine the Soundtracks and DEAM data. To improve data balance across the four emotion quadrants, we subsample the data so that no quadrant has more than 50% more clips than any other quadrant. In total, we have 2,176 six-second music clips distributed 28%, 18%, 28%, 27% over Q1, Q2, Q3, and Q4, respectively.

#### 4.2. Model Performance

We use precision, recall, and F1-score, standard measures in the machine learning literature, to evaluate our model. We calculate these metrics 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. The performance metrics are averaged over each fold of the held-out test data from stratified 10-fold cross-validation.<sup>27</sup> The standard deviations of the performance measures calculated over the ten folds are in parentheses.

In interpreting and assessing the results, it is important to consider the specific prediction task involved. Prediction tasks involving *subjective human response* (e.g., emotion recognition, humor detection) typically have lower prediction accuracy than more *objective* recognition tasks (e.g., object identification, instrument identification). This is because for subjective tasks, in the absence of objective ground truth, humans have heterogeneous

<sup>26</sup> The relatively low inter-rater agreement also highlights the inherent subjectivity of music emotion labeling.

<sup>27</sup> We create the folds at the song-level rather than the six-second clip-level to prevent data leakage. If clips from the same song are part of both the training and testing data, the training process may pick up some other elements of the song that can be used to predict emotion in the test data, leading to high accuracy, but lower generalizability.

**Table 2** Classification Performance - Deep Learning Models

	Precision	Accuracy/Recall	$F_1$
<b>Benchmark Models</b>			
<i>Atheoretic Filters</i>			
Mel - Square	0.5471 (0.0828)	0.5246 (0.0600)	0.5020 (0.0668)
Mel - Tall Rectangle	0.5500 (0.0647)	0.5376 (0.0618)	0.5216 (0.0670)
Mel - Wide Rectangle	0.5488 (0.0737)	0.5376 (0.0734)	0.5299 (0.0747)
<i>Theory-based Low-level Filters</i>			
Mel - Time	0.4002 (0.0572)	0.4110 (0.0439)	0.3698 (0.0491)
Mel - Frequency	0.5455 (0.0890)	0.5237 (0.0766)	0.5128 (0.0739)
Mel - Time-Frequency	0.5425 (0.0796)	0.5214 (0.0746)	0.5116 (0.0724)
<b>Proposed Theory-based Mid-level Filters</b>			
Mel - Harmonics	0.5570 (0.0873)	0.5434 (0.0780)	0.5380 (0.0791)

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. The measures are weighted by the proportion in each quadrant. There are 2176 clips distributed 28%, 18%, 28%, 27% over Q1, Q2, Q3, and Q4, respectively.

responses implying less agreement on the “ground-truth.” In such cases, even our best approximation is bounded above by this subjectivity, as with music emotion tagging.

First, we detail the classification performance of the benchmark models, which include a combination of atheoretic as well as low-level theory-based models. The atheoretic CNNs with square and rectangular filters obtain  $F_1$ s of 0.50-0.53. We note that a conceptual low-level CNN filter based on time does not perform as well as the atheoretic square filter (with  $F_1$  of 0.37 compared to 0.51). However, we do find the frequency filters perform almost as well as the atheoretic filters. The combination of the time and frequency filters performs comparably to the model using only frequency filters, suggesting that frequency-related features might play a larger role in eliciting emotion in short music clips relative to temporal features. Freq seems to perform well, but not time. So I have changed this text.

Next, we examine the performance of our proposed theory-based mid-level CNN filters. We find that the harmonics filters obtain an  $F_1$  no worse than the best benchmark models of 0.54 as well as comparable precision and recall measures. Another way to assess classification performance is to examine the confusion matrix. Online Appendix F shows the



confusion matrix of each model and Online Appendix G shows the performance broken down by quadrant.<sup>28</sup>

Our proposed harmonics filters perform well despite being part of a smaller neural network. The CNN with consonance filters has 100,000 trainable parameters while the CNNs with square and rectangular filters have nearly 5 million and 10 million, respectively. The CNN with frequency filters has 1.5 million, the CNN with time filters has 2.2 million, and the CNN with frequency-time filters has 3.8 million parameters. The consonance filters impose structure that has empirically been observed to relate to emotion on the input to the CNN. In doing so, the model is able to learn features relevant to emotion recognition with fewer model parameters. Smaller models with fewer parameters require less computation to train, reducing monetary and time costs associated with training.

Overall, the CNN that uses the mid-level consonance filters performs as well as those using atheoretic filters despite having far fewer parameters. The key distinction is that in contrast to the atheoretic filters, the proposed consonance filters are more explainable, and directly connected to theoretical concepts that help us better understand the explanation.

### 4.3. Model Explainability

A common concern about large and complex models including deep learning models is that they lack explainability, reducing trust in such models and making them difficult to fix when broken. A key benefit of our proposed theory-based consonance filters is that they enable post-hoc explainability when combined with gradient-based visualizations.

*Grad-CAM based Visualization:* 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 emotion classification decision for any input music clip. Grad-CAM uses the gradients of a target class (e.g., Q1—exuberance) that flow into a given convolutional layer to produce a heatmap that highlights the input regions that positively predict the target (Selvaraju et al. 2017).<sup>29</sup> Brighter portions of the heatmap capture the parts of the spectrogram with greater contribution to the classification.

<sup>28</sup> An alternative input to the mel spectrogram is the STFT spectrogram. Although the literature has most frequently used mel spectrograms since they are more reflective of how humans hear, we also assess the performance of using STFT spectrograms since they are more granular. We find that the average mel model performance measures are slightly higher than those from the STFT spectrograms.

<sup>29</sup> For example, when classifying an image as a shark a good image classifier might produce a heatmap that highlights sharp teeth.

Such visualizations have provided visibility into CNNs designed for image classification. It turns out that earlier convolution layers learn simpler features, like edges, while later layers learn more complex features, like shapes. Therefore, many researchers no longer consider image classification CNNs to be “black-box models.” However, the same cannot be said about CNNs applied to spectrograms. While edges and shapes define visual objects, which rely on spatial contiguity of objects, other types of features (e.g., harmony, tempo) define music. Therefore, visibility into image CNNs does not translate to visibility into music CNNs.

**Explainability of Consonance Filters.** Our theory-based architecture imposes useful structure that enables explainability. The CNN with consonance filters generates Grad-CAM heatmaps that are  $1 \times 517$  (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. When we stack the heatmaps associated with each pitch class, we get a heatmap in which the y-axis represents the 12 pitch classes and the x-axis represents time. Color represents the importance of the frequencies within each pitch class towards the classification into a given quadrant, with brighter colors representing greater importance.

*What should we expect to see based on theory?* 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 more consonant (Gabrielsson 2016). From a bottom-up perspective, music theory tells us that music with overlapping harmonics produces consonance. Combining these two perspectives generates hypotheses around what we should expect to see in the gradient-based visualizations.

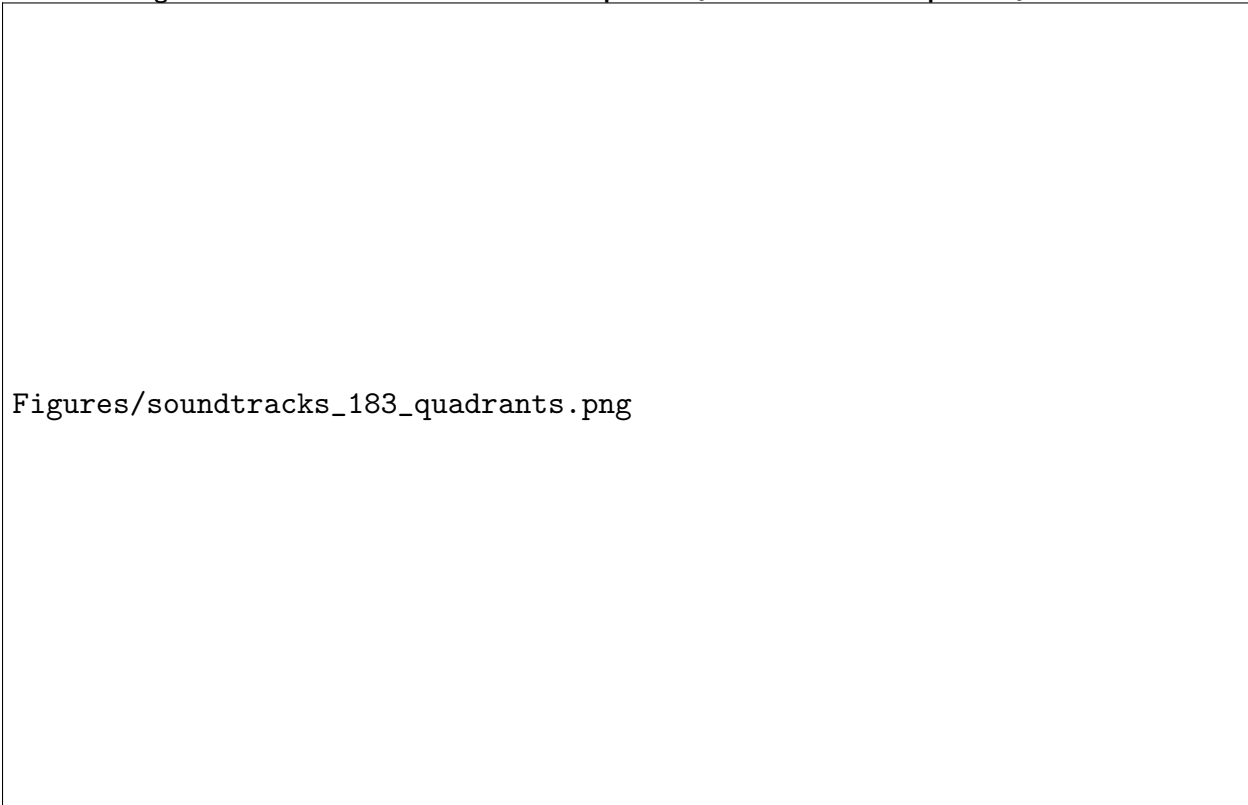
Since we limit the set of frequencies the CNN sees to consonant frequency combinations by placing structure (i.e., harmonics blinders) on what is input to the CNN, we expect the Grad-CAM heatmaps to pick up points of consonance to use in classification. Given the relationship between consonance and emotion, we expect consonant portions of music to greatly contribute to the classification of positive valence-low arousal (Q4—contentment) music and to minimally contribute to the classification of negative valence-high arousal (Q2—anxiety) music. Therefore, if the model indeed uses consonance for emotion prediction, we expect areas of the spectrogram with overlapping harmonics (i.e., consonant portions) to “light up” in the classification of Q4 but not the classification of Q2, and to

be somewhere in between for the positive valence-high arousal (Q1—exuberance) and negative valence-low arousal (Q3—sadness) quadrants. In other words, we expect brightness in the heatmaps to capture consonance.

To explore this, we take a music clip that contains a portion with only frequencies that follow a harmonic pattern, generating consonance, and show the Grad-CAM heatmaps for the four quadrants. The center image in Figure 6 shows the mel spectrogram of a Q4—contentment song that begins with a violin playing one note, changing notes, and then being joined by additional instruments shortly after three seconds. As can be seen in the mel spectrogram through the parallel horizontal bars at frequencies that are integer multiples of approximately 700 Hz between 0.25 and 2 seconds, and 800 Hz between 1.75 and 3.25 seconds, single violin strings produce a harmonic sound. Figure 6 also shows the Grad-CAM heatmaps for the classification of the clip into the four emotion quadrants using the CNN model with harmonics filters. The heatmap patterns align with what we would expect based on theory. First, the Q4 Grad-CAM heatmap shows that samples 20 to 280 (0.25 to 3.25 seconds) most greatly contribute to the Q4 classification, which correspond to periods of overlapping harmonics and therefore consonance, suggesting that brightness in the heatmaps captures consonance. Second, the lack of brightness in the Q2 heatmap further supports this idea. Finally, we see that Q1 and Q3 lie in between Q2 and Q4 in terms of heatmap brightness as predicted.

Next, we will analyze the heatmaps associated with the predictions made by the CNN model using harmonics filters across the Soundtracks and DEAM-classical music clips. Recall that Grad-CAM heatmaps help answer the question: Why was this clip classified into this emotion quadrant? From the theory, we expect that for fixed level of arousal, clips with positive valence have brighter heatmaps relative to clips with negative valence. We also expect that for fixed valence, low arousal clips to have brighter heatmaps than high arousal clips. We therefore hypothesize the following set of Grad-CAM brightness orderings based on the theory:  $Q1 > Q2$ ,  $Q4 > Q3$ ,  $Q4 > Q1$ ,  $Q3 > Q2$ .

Figure 7 shows a few prototypical Grad-CAM heatmaps generated by the harmonics filter and their associated mel spectrograms for each of the four quadrants. These heatmaps come from clips in the hold-out sets from ten-fold cross-validation. In general, we observe patterns in line with theory in that positive-valence heatmaps are brighter than negative-valence heatmaps and low-arousal heatmaps are brighter than high-arousal heatmaps. We quantify

**Figure 6 Harmonics Grad-CAM Heatmaps of a Q4—Contentment Clip over Quadrants**


Figures/soundtracks\_183\_quadrants.png

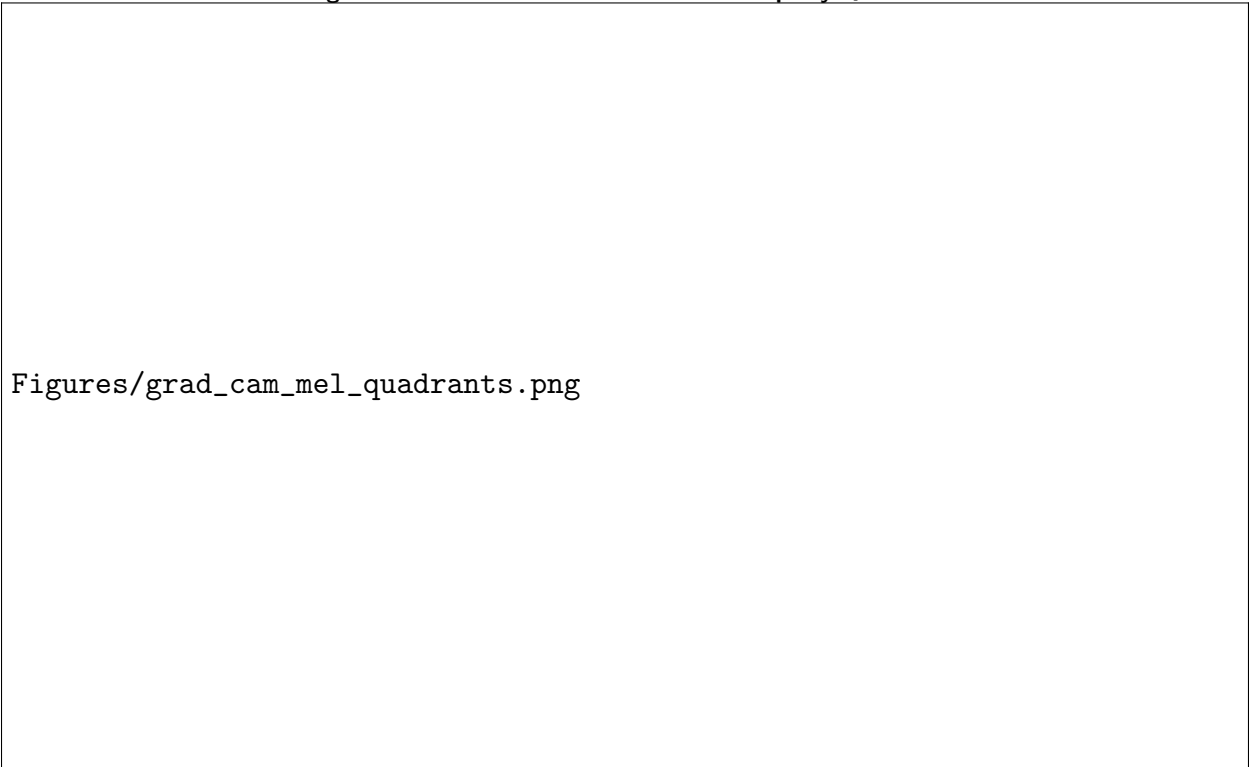
Notes: The center image is the mel spectrogram of Soundtracks track 183 seconds 0-6. The images in the four quadrants correspond to the Grad-CAM heatmaps for the classification of the clip into the four emotion quadrants using the CNN model with harmonics filters. [Add more notes!](#)

the heatmap brightness by summing up the heatmap values.<sup>30</sup> The average brightness levels are 322 for Q1, 114 for Q2, 356 for Q3, and 473 for Q4. The average brightness observed per quadrant is consistent with our hypothesized ordering, further confirming our consonance interpretation of the Grad-CAM heatmaps.

The explainability of the CNN with consonance filters builds trust in the model by providing transparency into what the model is learning. The heatmaps highlight areas of consonance, a mid-level feature not observable by eye,<sup>31</sup> and the patterns of brightness

<sup>30</sup> The Grad-CAM heatmap calculation steps can be found in Selvaraju et al. (2017). The heatmap is a linear combination of the post-convolution feature maps and their importance in predicting a target class. We quantify heatmap brightness by summing up the values from Eq. 2 in Selvaraju et al. (2017).

<sup>31</sup> 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). Music theorists have proposed a number of formulas to 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.

**Figure 7 Harmonics Grad-CAM Heatmaps by Quadrant**


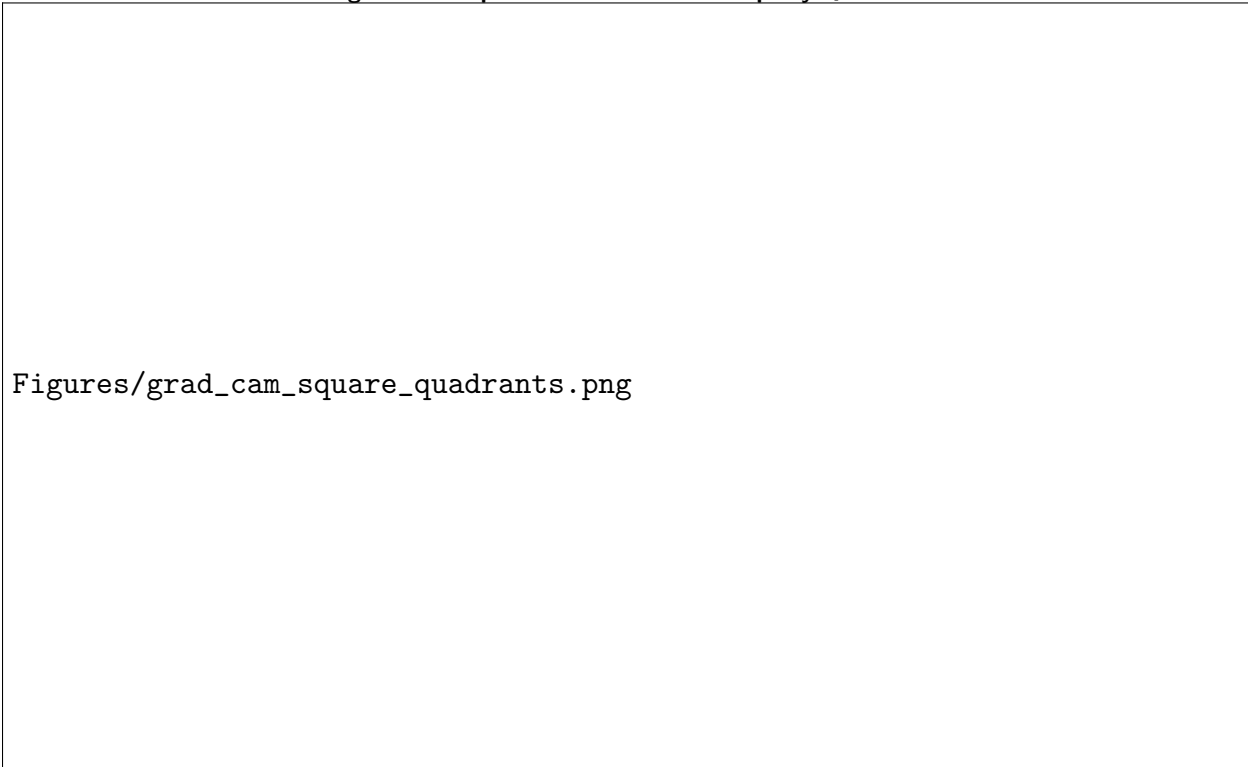
Figures/grad\_cam\_mel\_quadrants.png

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 following six-second clips are shown: Q1: DEAM song 1811 seconds 27-33 and DEAM song 1892 seconds 21-27; Q2: Soundtracks song 216 seconds 6-12 and Soundtracks song 142 seconds 0-6; Q3: DEAM song 2012 seconds 39-45 and DEAM song 159 seconds 27-33; Q4: Soundtracks song 33 seconds 12-18 and Soundtracks song 48 seconds 6-12.

align with the relationship between consonance and emotion. This provides us evidence that the model learns features predictive of emotion from the music and we show that it generalizes by showing the explainability on test data.

**4.3.1. Explainability of Atheoretic and Low-Level Filters.** Gradient-based visualizations can also be produced for the other filter types. We find that given their low-level focus without specific theory to guide our expectations, 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,

**Figure 8 Square Grad-CAM Heatmaps by Quadrant**


Figures/grad\_cam\_square\_quadrants.png

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 clips are the same as those shown in Figure 7. The bright portions of a heatmap capture the parts of the mel spectrogram that most greatly contribute to the classification of the specified emotion. The heatmap covers the dimensions of the final feature map after convolution. The x-axis captures time and the y-axis captures frequency.

shown in Figure 8, provide an idea of the range of frequencies and times contributing to the classification of the target class, it is otherwise not clear how to interpret the heatmaps, and what to conclude.

The heatmaps based on frequency filters highlight which frequencies contribute most to a particular classification (see Appendix Figure I1). The heatmaps based on time filters highlight which times contribute most to a particular classification (see Appendix Figure I2). Similar to square filters, it is not clear how to interpret the heatmaps, making it challenging to understand what musical features the models learn.

Thus, viewed broadly, our contribution here in incorporating the theory-based consonance filters is to provide explainability, while obtaining similar performance as atheoretical filters.

#### 4.4. Improving Accuracy with Handcrafted Features while Retaining Explainability

Despite the success of deep learning across many problem domains, approaches using feature engineering remain in common use. Feature engineering has long been used to define, transform, and evaluate features based on specialized domain knowledge (for example, for text analysis see Scott and Matwin (1999)). In general, deep nets are more flexible, since they learn the features that are relevant for prediction during the model training process, rather than hardcoding specific features. However, a tradeoff is that much of the performance gains with deep learning often come from large amounts of training data (Sun et al. 2017). In settings with smaller datasets, handcrafted features have been used in combination with deep learning to improve prediction (Tianyu et al. 2018, Min et al. 2018, Agajanian et al. 2019).

Time-varying music emotion datasets are relatively small. Even though a lot of raw music data is available, music tagging is labor-intensive. We therefore examine a variety of models combining our theory-based deep learning models with handcrafted features used widely in the literature.<sup>32</sup> Some handcrafted features, like mel frequency cepstral coefficients (MFCCs), are more atheoretical in that they were not designed for music but were instead designed for speech. Other features, like tempo, are more theory-based in that they were designed with music in mind.

First, we train two random forest models using only handcrafted features to show the baseline performance of the features. One model uses MFCCs, which describe the overall spectral envelope shape of audio, and have proven successful in music classification, including genre and emotion classification [ADD CITE](#). We use 13 MFCC coefficients as well as their first and second derivatives, resulting in 39 features. An alternative model uses a combination of 11 handcrafted features highlighted by Panda et al. (2018) for their ability to predict the four emotion quadrants<sup>33</sup>. Table ?? summarizes the Top Handcrafted Features features. Table 3 displays the performance of the two random forest (RF) models. The model using MFCCs as explanatory variables obtains an F1 of 0.50, outperforming the model using the Top Handcrafted Features, which obtains an F1 of 0.39.

<sup>32</sup> We thank the review team for the suggestion to incorporate handcrafted features into the model to maintain explainability but improve model accuracy.

<sup>33</sup> Panda et al. (2018) comprehensively considers a wide set of handcrafted features, and identifies the Top Handcrafted Features which most impact emotion classification into quadrants based on valence and arousal. Most of the identified features are low-level features that capture tone color or timbre. Please refer to Table ?? for categorization of low-, mid-, and high-level features.

**Table 3** Classification Performance - Incorporating Handcrafted Features

Features	Model	Precision	Accuracy/Recall	$F_1$
<b>Benchmark Models with Handcrafted Features</b>				
MFCCs	RF	0.5097 (0.0946)	0.5046 (0.0907)	0.4967 (0.0881)
Top Handcrafted Features	RF	0.4005 (0.0523)	0.4023 (0.0424)	0.3947 (0.0495)
<b>Combined Theory-based Mid-Level Filters + Hand-crafted Features</b>				
Mel - Harmonics + Top Handcrafted Features	CNN + RF	0.5557 (0.0782)	0.5532 (0.0756)	0.5469 (0.0746)
Mel - Harmonics + MFCCs	CNN + RF	0.5583 (0.0785)	0.5547 (0.0764)	0.5490 (0.0750)

*Note:* RF = Random Forest

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

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

## 5. Application: Emotion-based Ad Insertion in Content Videos

Our proposed theory-based deep learning model can be used in a number of real-time *emotion-based matching* applications by quantifying the valence and arousal of music clips. In an illustrative application, we demonstrate its value in determining the optimal emotion-based ad insertion point for an ad within a content video (e.g., YouTube video) with time-varying emotional content. The problem of optimal ad insertion is of significant managerial importance because of the rapid proliferation of user-generated content (UGC) on video platforms. The scale of the matching problem is very large, with a platform like YouTube needing to match billions of ads and content videos daily.



Such content-based matching is increasing in importance due to recent trends, with limitations of data available to platforms and advertisers due to privacy concerns.<sup>34</sup> 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 and it is therefore 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.

**Why Emotion-based Targeting?** Emotion can be contextually inferred entirely from video/ad content and does not require other behavioral and demographic variables, which are increasingly subject to privacy concerns and becoming less feasible to use. Marketers who had grown accustomed to achieving high performance using user targeting have found these strategies to decrease in effectiveness in recent times.<sup>35</sup> Emotion-based targeting is currently used in practice by many organizations. For instance, The New York Times experimented with targeting ads to news stories that elicit different emotions (called “Project Feels”). They found that targeting ads based on the emotion elicited by the article being read had a dramatic improvement on ad performance with an average lift on impressions of 40%. The firm now charges premium prices for such emotion-targeted ads (Edmonds 2019). This suggests that it is a topic that is worthy of study across multiple modalities (e.g., text, audio, video).

Throughout a content video, emotion often varies over time and so the various ad insertion slots will differ in emotion.<sup>36</sup> Since ads also often elicit emotion, we seek to understand how to match the ad to the insertion slot based on emotion. While marketing researchers have considered the overall emotion of content videos for ad matching (Coulter 1998, Kamins et al. 1991, Puccinelli et al. 2015, Kapoor et al. 2022), our focus is automatically identifying the optimal ad insertion position within content videos that vary in emotion.

<sup>34</sup> Many web browsers have eliminated third-party cookies (source: <https://www.mediapost.com/publications/article/346034/baking-up-new-strategies-for-a-post-cookie-world.html>). 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>).

<sup>35</sup> <https://www.cnn.com/2021/03/11/why-facebook-is-so-upset-about-apple-idfa-change-insiders-spill.html>

<sup>36</sup> According to YouTube documentation, ad insertion points are “placed at natural breaks in your videos to balance viewer experience and monetization potential” (Source: <https://support.google.com/youtube/answer/6175006>).

Should an ad that is primarily high-valence, high-arousal be placed at a primarily high-valence, high-arousal part of the content video or a primarily low-valence, low-arousal part of the content video? 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 emotional congruence is more effective (Lee et al. 2013, Kamins et al. 1991), including studies of matching in persuasion (Teeny et al. 2021) and fluency (Hertwig et al. 2008). On the other hand, studies have also 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 emotional contrast may be more effective (Kapoor et al. 2022).

Our goal here is twofold. We first seek to understand whether emotional congruence or contrast between the ad and content video ad insertion point increases ad engagement. To answer this empirical question, we conduct a lab experiment in which we exogenously insert ads into content videos at different ad insertion points. We characterize each ad and each content video ad insertion point by their respective emotions, which are based on human tagging. This allows us to create a measure of emotional distance between each ad and each content video ad insertion point. We measure ad skip and recall and see whether and how emotional distance influences these ad engagement measures.

Then, knowing the more effective ad matching strategy, we see whether our proposed theory-based deep learning model can select emotionally appropriate ads based on predicted emotion relative to other deep learning models. Given our focus on music and its effects, we treat the emotion evoked by the background music of a video as a proxy for the emotion evoked by the overall video.<sup>37</sup> In this setting of automated contextual targeting, the explainability of our model increases managerial trust in the tool to make reasonable decisions that generalize across a range of different content videos and ads outside of the initial training setting and thus increases confidence in adopting it.

### 5.1. Experiment: Is Emotional Congruence or Contrast More Effective?

Below we describe the design of the experiment, its operationalization, and the experimental results.

<sup>37</sup> Film professionals and researchers recognize the importance of music in driving emotion. Nelson et al. (2013) write “Music plays many roles in film, but it is possible to categorize all of them into two primary functions: creating consonance or dissonance to highlight the film’s emotion or narrative.”

**5.1.1. Experimental Setup.** The objective of the experiment is to understand how emotional distance between ad and content impacts ad effectiveness. As such, we vary the ad insertion points within a content video, where the insertion points vary in the evoked emotion. The outcome variables of interest are ad skip (as a proxy for attention and interest) and brand recall (as a proxy for memorability). We use a full factorial design across four ads, four content videos, and six ad insertion points per content video—yielding 96 experimental cells.

We develop a Qualtrics survey that shows an ad partway through a content video, mimicking the concept of YouTube’s mid-roll ads. Six seconds into the ad, similar to YouTube, a “Skip Ad” button appears, allowing participants to skip the remainder of the ad. Upon watching the ad to completion or skipping it, the content video picks up where it left off. Each participant sees only one content video and one ad. After participants finish watching the video, they are asked a number of questions about the content video and the ad. In particular, the survey asks participants to recall the company, brand, or cause that advertised within the video.

**5.1.2. Content Videos and Ads.** We select a diverse set of content videos and ads to show within the Qualtrics survey. We select content videos that a) contain background music at least some of the time and b) are long enough to vary in emotion over time and allow for multiple ad insertion points. The videos range from 5.7 to 7.9 minutes in length, include both animated and live-action videos, and include videos with and without speech. Appendix Table 4 provides details on the four content videos.

**Table 4 Content Video Details**

Title	Length (min)	Description	URL
Lost & Found	6.6	Two crocheted stuffed animals try to save each other. Animated and no speech.	<a href="https://www.youtube.com/watch?v=35i4zTky9pI">https://www.youtube.com/watch?v=35i4zTky9pI</a>
Hope	6.2	A new hatched turtle learns about its surroundings and tries to get to the ocean. Animated and no speech.	<a href="https://www.youtube.com/watch?v=1P3ZgL0y-w8">https://www.youtube.com/watch?v=1P3ZgL0y-w8</a>
Unspoken	5.7	Two people get to know each other and develop a relationship through writing notes. Live-action and some speech.	<a href="https://www.youtube.com/watch?v=8mpFYQb0CFo">https://www.youtube.com/watch?v=8mpFYQb0CFo</a>
Run With Me	7.9	A handicapped high school student participates in the 400m race to prove he doesn’t need special treatment. Live-action and speech.	<a href="https://www.youtube.com/watch?v=EisaD0ZsL3E">https://www.youtube.com/watch?v=EisaD0ZsL3E</a>

For each content video, we fix six ad insertion points (Time 1, ..., Time 6) that are roughly one-minute apart and occur at natural changes in the audio and images, similar

to YouTube’s approach.<sup>38</sup> Appendix Table 5 specifies the six ad insertion times for each content video.

**Table 5 Ad Insertion Times**

Video	Start	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	End
Lost & Found	0:16	1:15	2:00	3:01	3:48	5:11	6:05	6:50
Hope	0:18	1:20	2:10	2:55	3:45	4:45	5:59	6:30
Unspoken	0:01	1:01	1:57	2:54	3:26	4:12	5:14	5:43
Run With Me	4:30	5:37	7:44	9:00	9:45	10:45	11:40	12:24

*Note:* Times are minute:second and based on time since 0:00 rather than time since Start.

Finally, we select four ads to be shown at the ad insertion points. We select four ads such that each of the four valence-arousal quadrants is represented by the primary emotion of the first six seconds of one ad. Importantly, we are interested in the emotion of the first six seconds, because this is the length of time viewers on YouTube see an ad for before having the option to skip. It is the effect of the interaction of the content video emotion with the initial ad emotion that we seek to understand. Appendix Table 6 provides details on the four ads. The ads include background music, are 30 seconds long,<sup>39</sup> and cover a range of industries.

**Table 6 Ad Details**

Brand	Start Time	End Time	URL
Kit Kat	0:00	0:30	<a href="https://www.youtube.com/watch?v=4X_e3UWS9aA">https://www.youtube.com/watch?v=4X_e3UWS9aA</a>
Fragile Childhood	0:23	0:53	<a href="https://www.youtube.com/watch?v=XwdUXS94yNk">https://www.youtube.com/watch?v=XwdUXS94yNk</a>
Eli Lilly–Cymbalta	0:13	0:43	<a href="https://www.youtube.com/watch?v=Nf6Mm_M5RU">https://www.youtube.com/watch?v=Nf6Mm_M5RU</a>
Calm App	0:00	0:30	<a href="https://www.youtube.com/watch?v=LWisCdA5rB4">https://www.youtube.com/watch?v=LWisCdA5rB4</a>

**Emotion Tagging:** Since we seek to measure emotional distance, we must first characterize the emotion of the content videos and ads. Within the experiment, we characterize emotion using human tagging because human-tagged emotion is the ground-truth. To obtain the emotion tags, we recruit survey participants on Prolific, an online survey platform. We show respondents either the content videos in segments (as defined by Appendix Table 5) or the first six seconds of the ads and ask them about their valence and arousal

<sup>38</sup> According to YouTube’s documentation, “YouTube’s advanced machine learning technology looks over a large volume of videos and learns to detect the best places for mid-rolls. This is done by evaluating factors like natural visual or audio breaks” (Source: <https://support.google.com/youtube/answer/6175006?hl=en#zippy=%2Cfrequently-asked-questions>).

<sup>39</sup> For ads originally longer than 30 seconds, we use a 30-second clip to keep the ad length consistent in the experiment across cells.

levels after watching each clip.<sup>40</sup> With multiple tags per clip, each clip can then be characterized by the distribution of emotion over the four valence-arousal quadrants.

Figure 9 shows the emotion distribution over video segments for the content video titled Hope. Looking at Segment 1, all respondents found the first video clip high valence (Q1 and Q4) and 25% found the clip high arousal (Q1) while the remaining 75% found the clip low arousal (Q4). Overall, viewers largely felt high valence during the first three segments, high arousal (Q1 and Q2) during Segment 4, anxious during Segment 5 (Q2), and negative (Q2 and Q3) during Segment 6, demonstrating large variation in emotion over time. Appendix Figure J1 shows the emotion distributions of all four content videos.

**Figure 9 Human-Tagged Content Emotion Distribution**

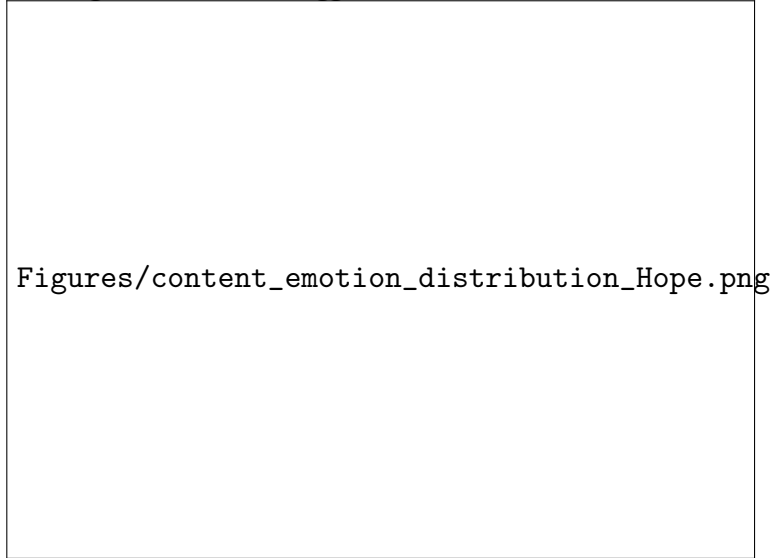
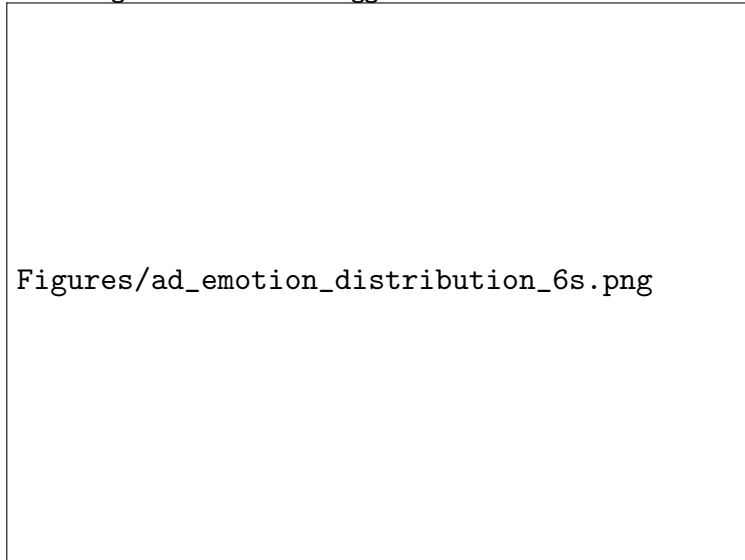


Figure 10 displays the emotion distributions of the first six seconds of each ad. As can be seen, the four ads were selected so that each emotion quadrant would be represented.

**5.1.3. Emotional Distance Measure.** To provide a measure of emotional distance between an ad and an ad insertion point in the content video, we calculate the Jensen-Shannon (JS) distance between their probability distributions over the four valence-arousal quadrants. Let  $P_t$  represent the content video emotion probability distribution at time  $t$  and  $Q$  the ad emotion probability distribution. JS distance is defined as:

$$JSD(P_t||Q) = \sqrt{\frac{1}{2}D(P_t||M) + \frac{1}{2}D(Q||M)} \quad (5)$$

<sup>40</sup> Each content video received 12-17 tags and each ad received 15-19 tags. Valence and arousal are measured on a scale of 0 to 100 so we convert the valence and arousal levels to valence-arousal quadrants.

**Figure 10 Human-Tagged Ad Emotion Distribution**

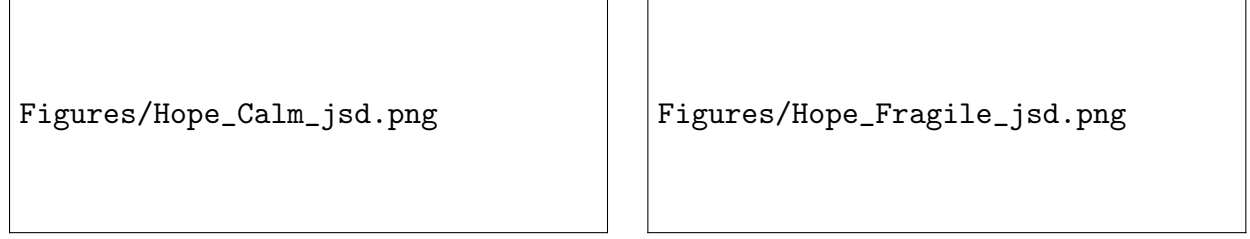
where  $M = \frac{1}{2}(P_t + Q)$  and  $D$  is the Kullback–Leibler (KL) divergence. KL divergence is in turn defined as:

$$D(P_t || Q) = \sum_{x \in X} P_t(x) \log \left( \frac{P_t(x)}{Q(x)} \right) \quad (6)$$

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 content video emotion at time  $t$ .

Figure 11 plots the JS distances between the ads Fragile Childhood and Calm and the content video Hope over Hope’s six content segments (i.e., time). For Calm, the minimum distance (i.e., greatest emotional similarity) occurs after Segment 1 and the maximum distance (i.e., greatest emotional contrast) occurs after Segment 6. While for Fragile Childhood, the minimum distance occurs after Segment 6 and the maximum distance occurs after Segment 1. Appendix Figure J2 plots the JS distances of the 16 potential combinations of content videos and ads. These plots show that there is large variation in emotional distances between the ads and the content. The JS distances in the data range from 0.017 to 0.783.

**5.1.4. Outcomes of Interest** Emotional distance is our independent variable and our dependent variables of interest are ad skip and brand recall. Both are “revealed preference”-type metrics and of significant interest to advertisers. Ad skip captures whether someone

**Figure 11 JS Distance between Content Video and Ads**

voluntarily continues to watch the ad. Brand recall captures whether someone paid attention to the ad. We also examine ad view time in Appendix XYZ.

We record whether a participant skips the ad using the Qualtrics survey. An ad skip occurs if the participant presses the “Skip Ad” button within five seconds of its appearance.<sup>41</sup> This definition allows us to capture the emotion interaction of the content video at the time of ad insertion and the first six seconds of the ad.

We capture recall in the set of questions asked to participants after watching the content video with the ad inserted at some point. In the Qualtrics setup, the brand shows as video information when the ad begins and disappears after three seconds. Therefore, even if a participant skips the ad they are still exposed to the brand.

**5.1.5. Experimental Results.** Each participant is randomly assigned to one of the 96 experimental cells.<sup>42</sup> Across all content videos, ads, and ad insertion points, 1,413 participants on average skipped 42.0% of ads (viewed 58.0%) and correctly recalled 45.2% of brands. As expected, the correlation between skip and recall is negative, with a value of  $-0.234$  (p-value  $< 0.01$ ).

To determine the impact of emotional distance on skip and recall, we regress a binary indicator for skipping and a binary indicator for correctly recalling the company, brand, or cause on JS distance, controlling for covariates. We estimate the following regression equation:

$$f(y_{ijt}) = \alpha + \beta JSD_{ijt} + \gamma X_{ijt} + \epsilon_{ijt} \quad (7)$$

<sup>41</sup> This is different from YouTube’s definition of skip rate (i.e., 1 - 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.

<sup>42</sup> Participants were asked to watch a 5-8-minute video and then 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 97%, and have completed at least 50 previous Prolific tasks. Prolific workers who tagged the content video and ad emotions were excluded from participating in the experiment. The survey took on average 11.6 minutes to complete and each participant was paid \$1.80 for their time. Participants that failed the attention checks were not considered in the analysis.

where the outcome  $y_{ijt}$  represents a skip indicator or a correct recall indicator for ad  $i$  at insertion point  $t$  within content video  $j$ ,  $JSD_{ijt}$  represents the JS distance between the emotion of the content video at point  $t$  and the ad,  $X_{ijt}$  represents covariates (i.e., ad, content, time of ad insertion point), and  $\epsilon_{ijt}$  the error term.

For ease of interpretation, we assume a linear probability model for skip and recall and report the least squares coefficients in Table 7. Appendix Table J1 reports the coefficients assuming a logit model for skip and recall and the results are robust to this alternative model. Table 7 Columns (1) and (4) do not include any covariates. We find that greater emotional distance increases the probability of skipping and decreases the probability of recalling the brand. These effects are statistically significant at the 0.05-level. Columns (2) and (5) include content video fixed effects, ad fixed effects, and a linear time trend for ad insertion time since the times differ across content videos. The results remain robust to including these covariates. Finally, Columns (3) and (6) include a second-order polynomial in time and the results do not qualitatively change.<sup>43</sup>

**Table 7 Effect of Emotional Distance on Ad Engagement**

Outcome:	<i>I(Skip)</i>			<i>I(Recall)</i>		
	(1)	(2)	(3)	(4)	(5)	(6)
JS Distance	0.188** (0.075)	0.196** (0.078)	0.206*** (0.079)	-0.179** (0.077)	-0.166** (0.077)	-0.168** (0.077)
Time		0.001*** (0.001)	0.001*** (0.001)		-0.001 (0.001)	-0.001 (0.001)
Time <sup>2</sup>			-0.001 (0.001)			0.001 (0.001)
Content FE	N	Y	Y	N	Y	Y
Ad FE	N	Y	Y	N	Y	Y
R <sup>2</sup>	0.004	0.035	0.036	0.003	0.092	0.092

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01; 1,413 observations for all regressions; robust standard errors used

Taken altogether, the JS distance coefficients suggest that emotional congruence between the ad and the content ad insertion point is more effective than emotional contrast for ad engagement. The significant positive coefficient for skip suggests that the more emotionally distant, the greater the likelihood of skipping the ad within the first 11 seconds. The significant negative coefficient for recall suggests that the more emotionally distant, the

<sup>43</sup> We also assess the impact of emotional distance on log view time in Appendix Table J2. We focus on skip because we are interested in understanding the interaction of the content video and the beginning of the ad, for which ad skip is a better proxy. We find that the impact of emotional distance on log view time is significant at the 0.05-level. Consistent with the skip results, greater emotional distance reduces log view time.



lower the likelihood of being able to correctly recall the brand. In terms of ad insertion time, the later an ad is shown the more likely it is skipped.

## 5.2. Ad Insertion Automation

The experiment establishes the relationship between emotional distance and ad engagement. Human-tagging of emotion for all video content and ads is not a viable strategy for determining emotionally similar ad insertion points so we demonstrate the use of our theory-based deep learning model and the benchmark deep learning models in determining the lowest emotional distance ad insertion point. While the emotional distances in the experiment were based on human-tagging, we now incorporate the models trained on the Soundtracks and DEAM data to *predict* the emotion distributions of the ads and content and calculate model-predicted emotional distances. Then using the experimental results, we compare ad skip and recall outcomes from showing each ad at the most emotionally similar ad insertion point as selected by our model and the benchmark models.

**5.2.1. Calculating Model-predicted Emotional Distances.** We begin by transforming the first six seconds of audio of each ad and 30 seconds of audio before each ad insertion point in the content videos into mel spectrograms.<sup>44</sup> We use the models trained on the Soundtracks and DEAM data to predict the emotion distribution of each six-second clip.<sup>45</sup> Using the 24 content emotion distributions and four ad emotion distributions, we calculate the JS distances between the ads and the content at the six ad insertion points.

For each combination of model, content video, and ad, we determine which ad insertion point is the most emotionally similar. For example, we find that for the content video Hope and ad Fragile, the Mel - Harmonics model suggests that the sixth ad insertion point is the most emotionally similar to the ad. The model-predicted emotional distance is 0.272 but the true emotional distance based on human-tagging is 0.183. Table 8 compares the average JS distances based on human-tagging of the ads selected by the various models.

<sup>44</sup> For the content videos, we break the 30-second clips into five six-second clips.

<sup>45</sup> For the deep learning models, the final softmax layer generates a probability distribution over the four valence-arousal quadrants. Instead of selecting the highest probability quadrant, we retain the probability distribution. For the 30-second content video clips, we average over the five predicted emotion distributions associated with each six-second clip.

**5.2.2. Skip and Recall Rates.** From the experiment, we have the skip and recall rates for the 96 experimental cells. For each model, we average the skip and recall rates of the most emotionally similar ad insertion point for each of the four ads in each of the four content videos. Table 8 shows the average skip and recall rates averaged over the 16 content video and ad combinations.

**Table 8 Ad Insertion Automation Results**

Feature	Model	JS Distance	Skip Rate	Recall Rate
<i>Atheoretic Filters</i>				
Mel - Square	CNN	0.568	41.2%	45.1%
Mel - Tall Rectangle	CNN	0.536	46.7%	46.2%
Mel - Wide Rectangle	CNN	0.533	45.7%	45.5%
<i>Theory-based Low-level Filters</i>				
Mel - Time	CNN	0.456	42.2%	46.5%
Mel - Frequency	CNN	0.506	42.9%	45.2%
Mel - Time-Frequency	CNN	0.456	42.2%	46.5%
<i>Proposed Theory-based Mid-level Filters</i>				
Mel - Harmonics	CNN	0.448	42.1%	47.1%
Mel - Harmonics + Top Handcrafted Features	CNN + RF	0.407	39.3%	50.3%

The proposed theory-based mid-level filter models combined with our experimental findings suggest using ad insertion points that are relatively higher in emotional similarity (i.e., low JS distance). These ad insertion points generate relatively favorable skip and recall rates. The results suggest that our model can be useful in emotion-based ad insertion.

**5.2.3. Incorporating Other Video Modalities.** Our primary analysis has focused on using emotion evoked from music, and characterizing that into four quadrants based on valence and arousal. However, with videos, emotional content may be present across multiple modalities (e.g., facial expressions, the text of speech). Multimodal emotional content can also be used to predict emotional distance. When the videos have human faces, we can use publicly available tools to estimate emotion from facial expressions. Similarly, emotional content can also be obtained from voice tonality and speech text.

In our application, we observe that not every content video has speech or human faces and the same to be true for the first six seconds of the ads, implying this approach is not always feasible. Including face emotion improves the recall rate when used in combination with the mel - harmonics model but hurts the skip rate when used in combination with the mel - harmonics + Top Handcrafted Features model. Appendix J.1 details the analysis that includes face emotion.

Overall, there is potential in incorporating emotion information from images and text but the existing tools are limited in their ability to extract emotion information from short clips (i.e., first six seconds of ads) and animated videos. However, this is a moving target and as these methods steadily improve these findings could well change.

### 5.3. 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 decreases ad skipping and increases brand recall. We develop a tool to facilitate the determination of emotion based on the background music of videos. We show that it performs as well as atheoretical CNN models in terms of approximating human-tagged emotion but is explainable.

We map music to emotion to determine optimal emotion-based ad matching for a given ad insertion point within a content video. In practice, we expect emotion to be used as a variable in addition to other ad targeting variables, such as past ad engagement behavior. A question that arises is why not directly try to predict ad skip and recall from the raw music instead of mapping through emotion. One of our key goals is to obtain explainability and without music theory to inform the model we cannot understand what the model uses for prediction, limiting confidence in the generalizability of such a model.

We demonstrate the value of our model 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, one quarter of songs are skipped in the first five seconds, 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.<sup>46</sup> Our model can quantify the emotional match between the end of one song and the beginning of the next to allow for continuity (or contrast) in the listener’s emotional experience. The classifier can also be used in contexts that match music with other forms of unstructured data. 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 our model can be used for the video ad. More broadly, any setting that involves emotion and requires music choice (e.g. call waiting music) could benefit from a music emotion classifier.

<sup>46</sup> <https://www.theguardian.com/music/2014/may/07/one-quarter-of-spotify-tracks-are-skipped-in-first-five-seconds-study-reveals>

## 6. Conclusion

Our research contributes to the literature that studies consumer response to unstructured data. Unlike other unstructured data, such as text (e.g., Liu et al. 2019, Toubia et al. 2021, Wang et al. 2021), images (e.g., Burnap et al. 2019, Liu et al. 2020, Dew et al. 2022, Zhang and Luo 2022, Troncoso and Luo 2020, Huang et al. 2022, Sisodia et al. 2023), and video (e.g., Zhang et al. 2020, Yang et al. 2021, Chakraborty et al. 2022, Kapoor et al. 2022), music has received relatively little attention in the marketing literature. Music is pervasive in customer interactions with firms. From music in ads to hold music for call centers, from workout 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 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 music. We develop novel filters to capture musical features associated with consonance using structures that have a foundation in the physics of sound waves (harmonics) and integrate them into CNN models. The filters are not only important in predicting emotional outcomes, but are also explainable 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 explainable.

In terms of explainability, we exploit specific elements of music theory to construct filters to capture musical features associated with consonance. We visualize the model using Grad-CAM (Selvaraju et al. 2017), which provides a visual representation of the areas in an image (spectrogram for sound) that most greatly contribute to the classification of a particular target class. 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).

We use our emotion classification method in an application where we match the time-varying emotion in a content video with the emotion of a short ad. We find that matching the ad emotion to the content emotion improves ad engagement by lowering skip rate

and increasing brand recall rate. As user-specific tracking due to privacy concerns decline, digital advertising will need to move away from demographic and behavioral targeting to strategies that are based on contextual targeting, such as emotion. We also note that besides advertising, our framework can be used in other settings, such as playlist formation and music therapy.

We conclude with a discussion of some limitations and suggestions for future research. First, we focus on relatively short music clips in our data and model. This choice is motivated by the application of ad insertion, where typically a video ad is played and the user is allowed the choice to skip the ad after watching it for six seconds. While the method in principle does apply to clips of any duration, in practice we might consider altering the architecture of the deep learning model to include temporal dependencies. Second, incorporating listener heterogeneity based on demographics could further improve the model’s predictive accuracy. More broadly, Mehta et al. (2008) posits that ads shape consumer preferences through informative, persuasive, and transformative effects. Emotion relates to the persuasive effect but developing explainable models that capture other features related to the informative and transformative effects will further push the frontier of ad analytics. 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.

### **Funding and Competing Interests**

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.