Appendix

A. Music Concepts

Table A1 Definitions of Music Concepts

Concept	Type	Definition
Frequency	Physical	The number of cycles a sine wave completes in a second, measured in Hertz (Hz)
Fundamental Frequency	Physical	Lowest natural frequency of a sine wave
Partial	Physical	Any of the sine waves that comprise sound
Harmonic	Physical	A frequency that is an integer multiple of the fundamental frequency
Spectrum	Physical	The range of frequencies contained in a signal
Musical Interval	Physical	Spacing between two sounds in frequency
Pitch	Perceptual	The attribute of sound that allows it to be ordered on a scale from low to high
Note/Tone	Perceptual	A pitched sound
Pitch Class	Perceptual	Set of all pitches that are an integer number of octaves apart
Harmony	Perceptual	Set of pitches played simultaneously
Tonalness	Perceptual	Music that has a specific note on which it is the most stable and at rest
Consonance	Perceptual	A combination of notes that sound pleasant when played simultaneously
Dissonance	Perceptual	A combination of notes that sound harsh or jarring when played simultaneously
Loudness	Perceptual	The intensive attribute of an auditory sensation, in terms of which sounds may be ordered on a scale extending from soft to loud
Timbre	Perceptual	The attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar

B. Music Feature Interpretability

Fu et al. (2010) use top-level labels to describe music constructs humans understand, like emotion and genre. To predict these labels, researchers use audio features, which can be divided by level of music understanding (i.e., interpretability). Low-level music features, obtained directly from the audio using simple mathematical transformations, are not closely connected to musical properties perceived by human listeners. More complex mid-level music features are usually based on transformations of the low-level features and are more closely connected to musical properties perceived by humans.

Table B1 Music Features by Interpretability Level (Fu et al. 2010)

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

C. Harmonics and Mid-level Musical Features

We explain how harmonics impact mid-level musical features that influence emotional response to music.

C.1. Harmonics and Consonance

Harmony captures the perception of simultaneous pitches and is characterized as being consonant or dissonant. In general, consonant sounds, such as the octave, are considered "pleasant or restful," while dissonant sounds are considered jarring (Sethares 2005).[1] Studies have revealed that consonance and dissonance are not binary categories, but rather opposite ends of a continuum.

Plomp and Levelt (1965) show that unison (two notes with identical frequencies) is the point of global maximum of consonance and specific other two-note frequency intervals form local maxima. 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. When a given sound has many overlapping harmonics, it is perceived as being consonant.

Consonance and dissonance are known to influence emotional response to music. Consonance is associated with positive valence emotion (e.g., tenderness) while dissonance is associated with negative valence emotion (e.g., fear). Consonance is also associated with low arousal emotion (e.g., contentment) while dissonance is associated with high arousal emotion (e.g., fear) (Gabrielsson 2016).

C.2. Harmonics and Timbre

Timbre is the quality of a sound that distinguishes it from another sound with the same loudness and pitch. Different musical instruments and voices create different patterns of harmonics varying in arrangement and strength when generating the same fundamental frequency, which create their timbre, or the "acoustic fingerprint" of an instrument or voice (Nelson et al. 2013). Tones with a strong emphasis on higher harmonics (i.e., high multiples of the fundamental frequency) are associated with high arousal while tones with suppressed higher harmonics are associated with low arousal (Gabrielsson 2016). Given that the filters are designed around harmonic frequencies, they are capable of learning different timbre patterns that help predict emotion.

C.3. Harmonics and Pitch

Pitch is the quality of a sound that allows it to be ordered on a scale from low to high. The fundamental frequency determines the perceived pitch, but the presence and alignment of harmonics affect how clearly the pitch is perceived. According to Gabrielsson (2016), "High pitch may be associated with expressions as happy, graceful, serene, dreamy, exciting, surprise, potency, anger, fear, and activity. Low pitch may suggest sadness, dignity/solemnity, vigor, excitement, boredom, and pleasantness." The filters are capable of learning different pitch patterns predictive of emotion.

[1] A classic example of a dissonant sound is the tritone. The tritone has been used in contemporary movies and music to provide a negative connotation or of something foreboding or fear-inducing (Lerner 2009).

D. Benchmark Models for Comparison

CNN with Square Filters. Image recognition CNN models typically use square filters that capture associations across two orthogonal spatial dimensions. Although mel spectrograms visualize music, the vertical and horizontal dimensions represent frequency and time rather than spatial dimensions. Thus, in music, the dimensions have very different meanings and resulting properties.

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

CNN with Rectangular Filters. It is possible that square filters might not be ideal to capture the features of audio data. Thus, we generalize this by using rectangular filters that are relatively more: (a) tall and narrow (specifically, replace each $k \times k$ convolution filter by a $2k \times k$ filter), and (b) short and wide (specifically, replace each $k \times k$ convolution filter by a $k \times 2k$ filter). We make these replacements for both the 5×5 and 3×3 filters that are used in our baseline square filter model.

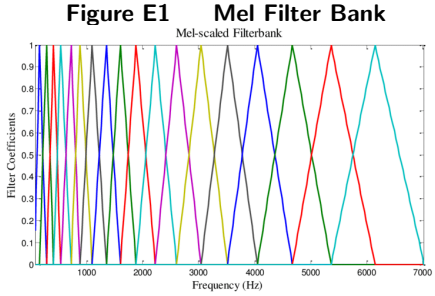
CNN with Time and Frequency Filters. We compare our proposed mid-level harmonics filters against low-level time and frequency filters proposed by Pons et al. (2016). We design a model that uses frequency filters, a model that uses time filters, and a model that combines the two types of filters. Frequency filters, which are tall and skinny, are designed to capture timbral features across the frequency spectrum, e.g., a specific combination of notes, while time filters, which are short and wide, are designed to capture temporal features, e.g., tempo. We allow the models additional flexibility by including an additional fully connected layer after pooling and before the final classification.

E. Consonance Blinders Transformation

We describe the process to transform the STFT blinders to mel blinders. To simplify the problem, let us assume that the frequency dimension of the STFT is continuous for now. For fundamental frequency f_0 we retain the frequencies f_0, $f_1 = 2f_0$, $f_2 = 3f_0$,..., $f_n = (n+1)f_0$. To account for human auditory perception, we allow for a band of frequencies centered around each frequency. The bandwidths are based on a constant bandwidth so our retained frequencies with bandwidth δ are of the form: $[f_n - \delta, f_n + \delta]$.

We allocate the power associated with the frequencies to the mel bands. The mel filter bank maps frequencies to (a maximum of 2) mel bands. Figure E1 shows the mapping of frequencies to 20 mel bands (in the paper we use 256 but it is more challenging to visualize). The top of each triangle represents the center of each band. Each triangle represents the weight each frequency contributes to a particular band. For example, the right most triangle maps frequencies ranging from roughly 5,400 - 7,000 Hz to the 20th mel band. Frequencies below 5,400 Hz receive zero weight. The triangles grow wider with higher frequencies because human hearing resolution is worse at higher frequencies. Let β_j represent the function that maps

Figure E1 Mel Filter Bank



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

$$\beta_j(x) = \begin{cases} 0 & \text{if } x < b^- \\ \frac{x - b^-}{b - b^-} & \text{if } b^- \leq x \leq b \\ \frac{b^+ - x}{b^+ - b} & \text{if } b \leq x \leq b^+ \\ 0 & \text{if } x > b^+ \end{cases} \tag{1}$$

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

$$M_{jn} = \int_{f_n - \delta}^{f_n + \delta} \beta_j(x) P(x) \phi(x) dx \tag{2}$$

where x represents frequency, $P(x)$ the power of x, and $\phi(x)$ the distribution over frequencies. We assume $\phi(x) \sim U[f_n - \delta, f_n + \delta]$. Since multiple frequency bands could contribute to a single mel band, we sum the contributions so the final power of mel band j is $M_j = \sum_n M_{jn}$. The set of power over all j comprises the mel blinders. They highlight which mel bands are input to the CNN and the weight of each band.

F. Mel - Harmonics CNN Architecture Choices

We describe a few relatively standard CNN modeling choices and their operationalization in the model.

Table F1 Standard CNN Model Elements

Element	Purpose
Pooling	Pooling applies a function over all units within a specified filter shape. We evaluate average pooling over time and both average and max pooling over pitch class. Max pooling over pitch class results in a higher F1 than average pooling so we use average pooling over time and max pooling over pitch class in our main model specification.
Batch Normalization	Batch normalization standardizes the inputs (mean zero, standard deviation one) in a mini-batch (data seen each time model parameters are updated). This procedure standardizes the inputs and helps achieve faster training, reduces overfitting, and stabilizes learning.
ReLU	Rectified Linear Activation Unit (ReLU) transforms the output of convolution to allow the model to learn nonlinear relationships.
Dropout	A form of regularization, dropout randomly removes neurons in specified layers of the neural network each mini-batch based on the specified dropout rate to prevent overfitting and obtain a more robust model.
Fully Connected Layer	Layer that connects every neuron in the hidden layer previous to every neuron in the next layer.

G. Dataset Merging Details

Our dataset is made up of the Soundtracks data combined with the DEAM classical and pop data. Soundtracks and DEAM were annotated on different numeric scales and in this section, we describe how we combine the two. For the Soundtracks data, perceived valence and arousal were annotated on a set of discrete emotions, as well as on bipolar scales (using adjectives), which the researchers transformed to a scale ranging from 1 to 7. The valence extremes were captured by the adjectives pleasant-unpleasant, good-bad, and positive-negative. The arousal extremes were captured by the adjectives awake-sleepy, wakeful-tired, and alert-drowsy. The midpoint of the numeric scale for Soundtracks is (4,4) for valence and arousal.

We observe that for the excerpts chosen to capture discrete emotions, the labels have high inter-rater consistency for all emotions except for surprise. We therefore use the discrete emotion labels to map these excerpts to the emotion quadrants. For the excerpts chosen to capture dimensional emotions and surprise, we use the valence and arousal labels to map these excerpts to the four quadrants.

For the DEAM data, perceived valence and arousal were annotated continuously on a scale of -10 to +10. Like Soundtracks, the DEAM creators also provided adjectives to describe valence and arousal. The valence extremes were extremely negative/unpleasant to extremely positive/pleasant with neutral in the middle. The arousal extremes were low arousal/calm to high arousal/activated/excited. Aljanaki et al. (2017) transform the data to range from -1 to +1. The midpoint of the numeric scale is therefore (0,0).

It is important that the Q1 to Q4 emotion labels are based on the scale used for each dataset. For example, Q1 captures positive valence-high arousal emotion. For Soundtracks, this maps to valence ≥ 4 and arousal ≥ 4 or discrete emotion = happy while for DEAM this maps to valence ≥ 0 and arousal ≥ 0. It would be incorrect to label Soundtracks using the DEAM scale and vice versa. Essentially, we have standardized the data so that the emotion quadrants mean the same thing for the DEAM and Soundtracks datasets.

H. Impact of Dataset Size

Table H1 Performance by Dataset Size for Mel - Harmonics Model

	Precision	Accuracy/Recall	F_1
25% of data	0.5069	0.4713	0.4571
	(0.0893)	(0.0711)	(0.0598)
50% of data	0.5129	0.5023	0.4976
	(0.0606)	(0.0570)	(0.0601)
100% of data	0.5224	0.5049	0.5057
	(0.0505)	(0.0478)	(0.0506)

I. Handcrafted Features

We use 11 handcrafted features highlighted by Panda et al. (2018) for predicting the emotion quadrants.

J. Confusion Matrices

The confusion matrices for the other models are available upon request from the authors.

K. Grad-CAM Visualizations

L. Application Study Details

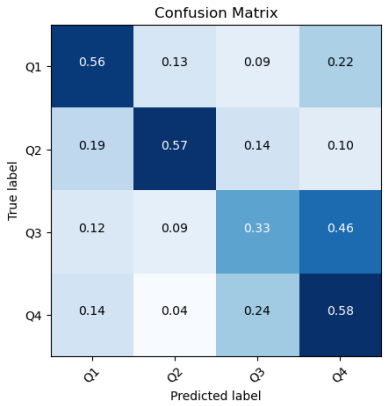
Below we provide more details for the ad insertion application.

Table I1 Top Music Emotion Base Features from Panda et al. (2018)

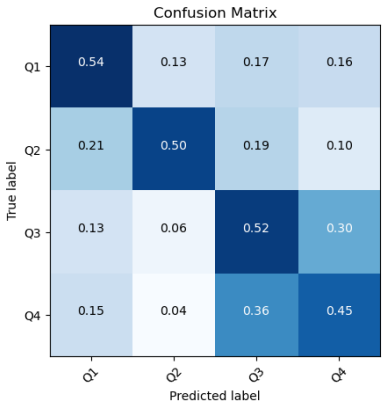
Feature in Panda et al. (2018)	Feature in MIR Toolbox	Musical Concept	Definition
FFT Spectrum - Spectral 2nd Moment (median)	Spectral Spread (median)	Tone Color	Standard deviation of the spectrum; a measure of the spread of the distribution
FFT Spectrum - Average Power Spectrum (median)	Spectral Centroid (median)	Tone Color	The geometric center of the spectrum distribution can be an indicator of the "brightness" or "sharpness" of the sound
FFT Spectrum - Skewness (median)	Spectral Skewness (median)	Tone Color	The third moment of the spectrum; a measure of the symmetry of the distribution
Spectral Skewness (std)	Spectral Skewness (std)	Tone Color	See above
Spectral Skewness (max)	Spectral Skewness (max)	Tone Color	See above
MFCC1 (mean)	MFCC1 (mean)	Tone Color	MFCC offers a description of the spectral shape of the sound
MFCC1 (std)	MFCC1 (std)	Tone Color	See above
Roughness (std)	Roughness (std)	Tone Color	An estimation of the sensory dissonance
Rolloff (mean)	Rolloff (mean)	Tone Color	Fraction of energy below specific frequency
Spectral Entropy (std)	Spectral Entropy (std)	Tone Color	Shannon entropy offers a general description of the spectral power distribution
Fluctuation (std)	Fluctuation (std)	Rhythm	Estimates the rhythm content based on spectrogram transformed by auditory modelling

Figure J1 Confusion Matrices

(a) Mel - Square Filters



(b) Mel - Harmonics Filters



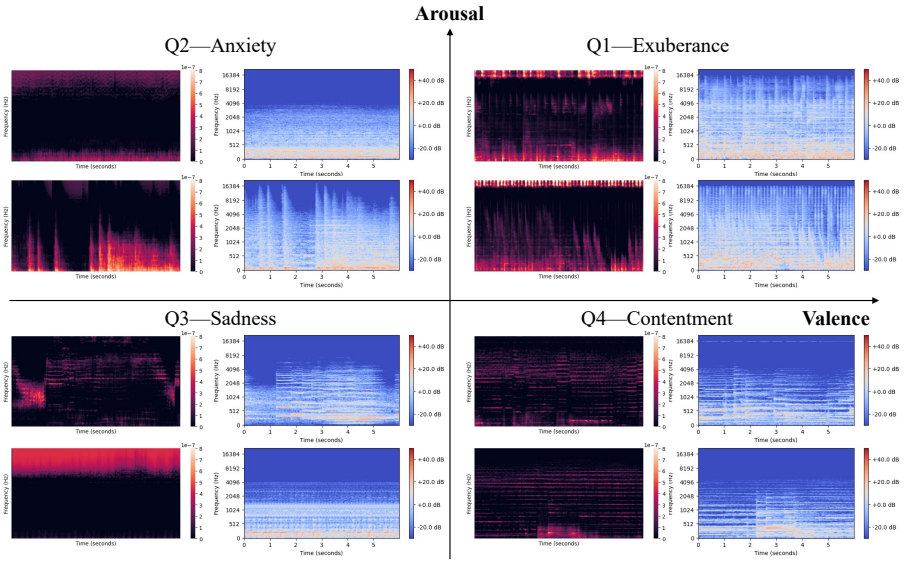
L.1. Incorporating Emotion Data from Images and Text

Although music is a key driver of emotion in video, other features like text (what is said), voice tonality (how it is said), and images can also influence emotion. We use Microsoft Azure to try to extract the emotion related to these other features to explore the impact of incorporating emotion data from other modalities beyond the background audio. Azure's Video Indexer emotion detection algorithm predicts emotion from speech text and voice tonality. The possible emotions include joy, fear, anger, and sadness. Azure's Face API detects emotion based on facial expressions from images.[2] The Face API treats the emotion prediction task

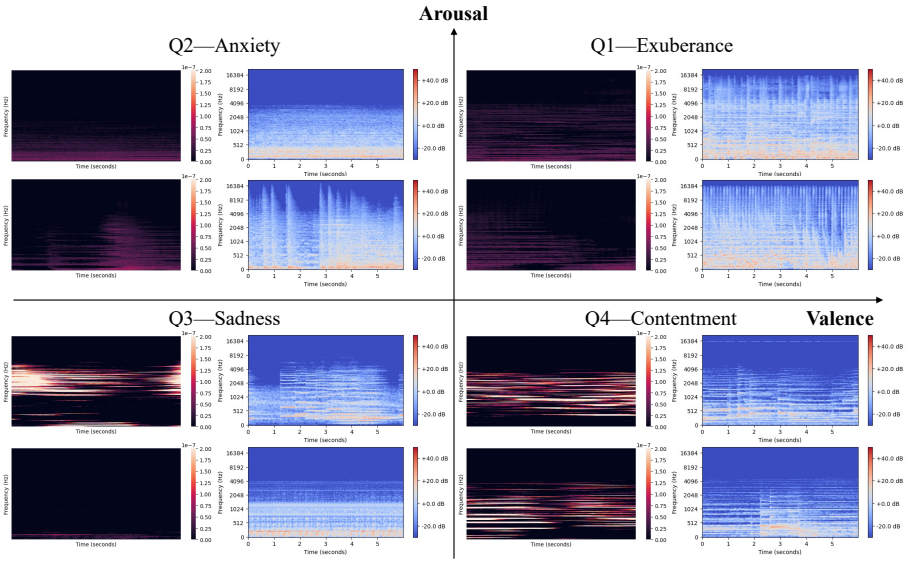
[2] Starting June 2022, users must apply to use the API due to accuracy concerns for specific demographic groups.

Figure K1 Frequency and Time Grad-CAM Heatmaps

(a) Frequency Heatmaps



(b) Time Heatmaps



Notes: Within each quadrant, the Grad-CAM heatmap (left) corresponds to the spectrogram (right). The heatmap covers the dimensions of the feature map after convolution. The clips are the same as those in Figure 6 of the paper.

as a multiclass problem so the probabilities over the eight possible classes (anger, contempt, disgust, fear, happiness, neutral, sadness, surprise) sum to one.

For the content videos, the Video Indexer predicted no speech emotion for two of the content videos since they do not contain speech. The first row in Figure L3 shows the distribution of speech emotion for the remaining two videos. Each distribution is defined as the average over the 30 seconds prior to the ad insertion time. The third row shows the human-tagged emotion distributions for comparison. For Run With

Table L1 Content Video Details

Title	Description	URL
Lost & Found	Two crocheted stuffed animals try to save each other. Animated and no speech. 6.6 min.	www.youtube.com/watch?v=35i4zTky9pI
Hope	A new hatched turtle learns about its surroundings and tries to get to the ocean. Animated and no speech. 6.2 min.	www.youtube.com/watch?v=1P3ZgLOy-w8
Unspoken	Two people get to know each other and develop a relationship through writing notes. Live-action and some speech. 5.7 min.	www.youtube.com/watch?v=8mpFYQbOCFo
Run With Me	A handicapped high school student participates in the 400m race to prove he doesn't need special treatment. Live-action and speech. 7.9 min.	www.youtube.com/watch?v=EisaD0ZsL3E

Note: The length captures the length of video shown to participants. Some videos are originally longer and we shorten them to start and end at natural times. Table K2 specifies the start and end times.

Table L2 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.

Table L3 Ad Details

Brand	Start Time	End Time	URL
Kit Kat	0:00	0:30	www.youtube.com/watch?v=4X_e3UWS9aA
Fragile Childhood	0:23	0:53	www.youtube.com/watch?v=XwdUXS94yNk
Eli Lilly–Cymbalta	0:13	0:43	www.youtube.com/watch?v=Nf6Mm__M5RU
Calm App	0:00	0:30	www.youtube.com/watch?v=LWisCdA5rB4

Me, Segment 3 aligns with the human-tagged emotion but Segments 1 and 2 less so. For Unspoken, Segments 4 and 5 are identified to be negative by the Video Indexer as well as by humans.

The Face API predicted no emotion for two of the content videos since they do not contain human faces but instead contain animated animal faces. The second row in Figure L3 shows the distribution of facial emotion for the remaining two videos. Each facial emotion distribution is defined as the average over the 30 seconds prior to the ad insertion time. For the most part, the faces are predicted to be either neutral or happy. For Run With Me, we see that there are some similarities with human-tagged emotion in that Segments 3, 4, and 6 are higher in facial happiness and higher in Q1.

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

Figure L1 Human-Tagged Emotion Distributions of Content Videos

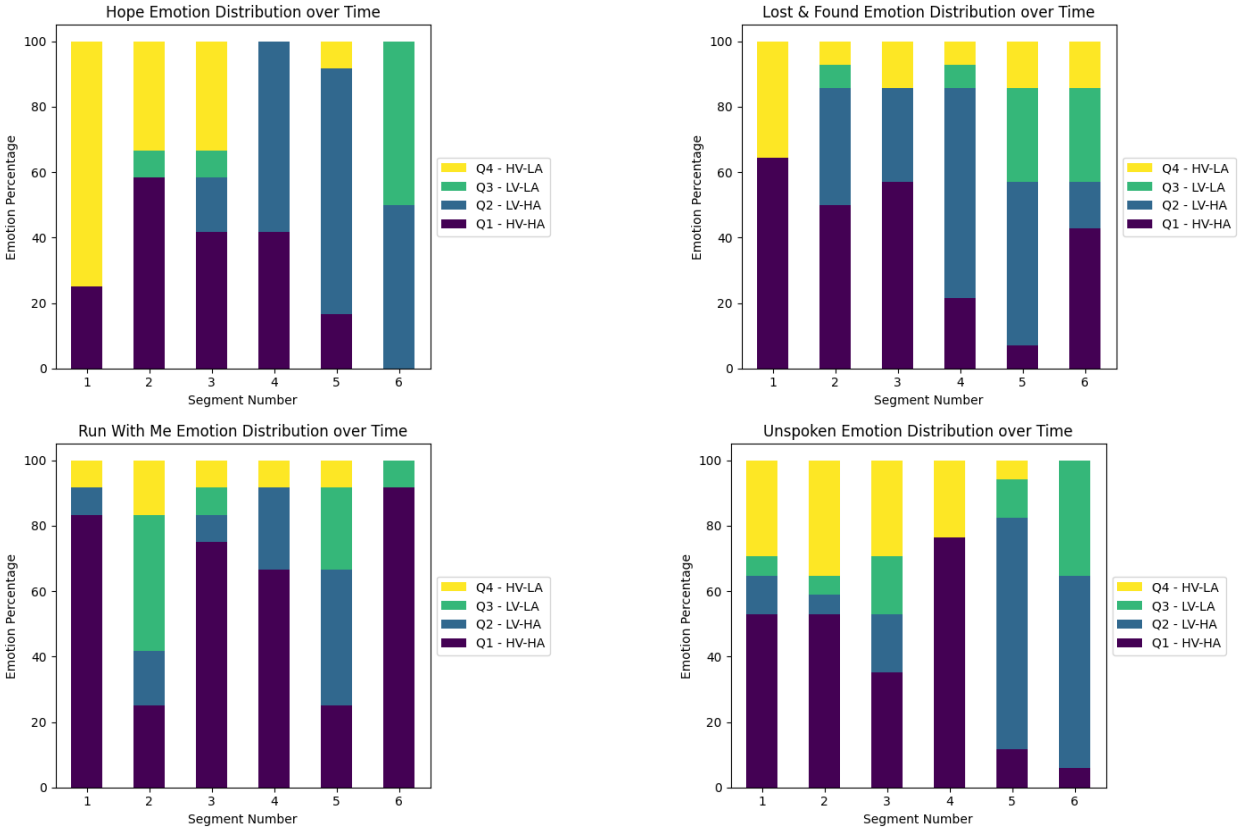
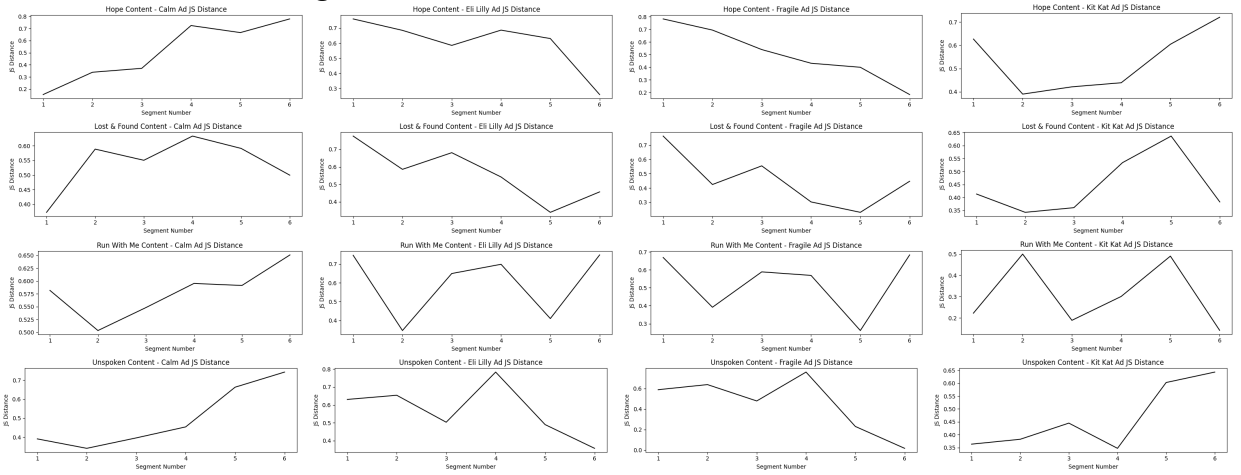


Figure L2 JS Distance between Ads and Content Videos

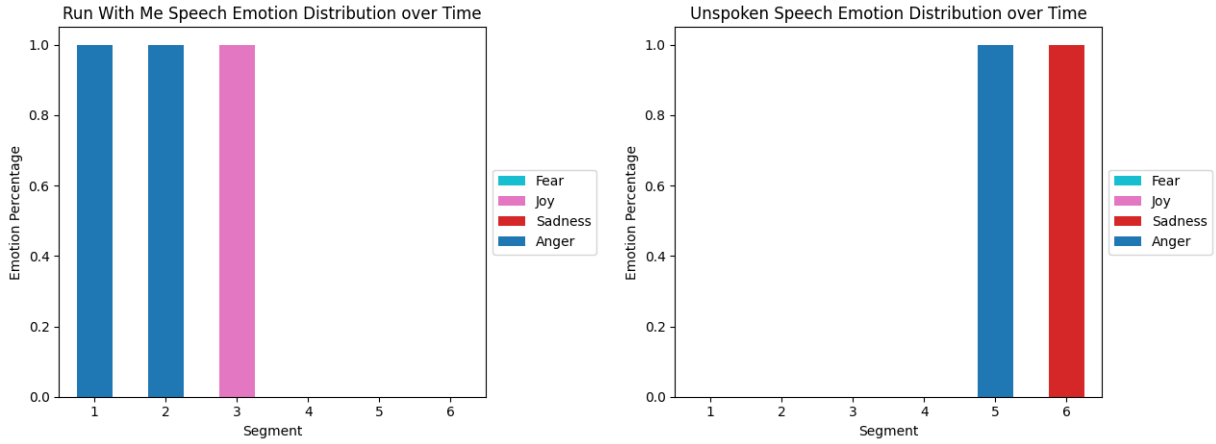


recall rate for each model. Table L4 compares these measures from using face and music emotion versus using only music emotion for the two content videos Unspoken and Run With Me.

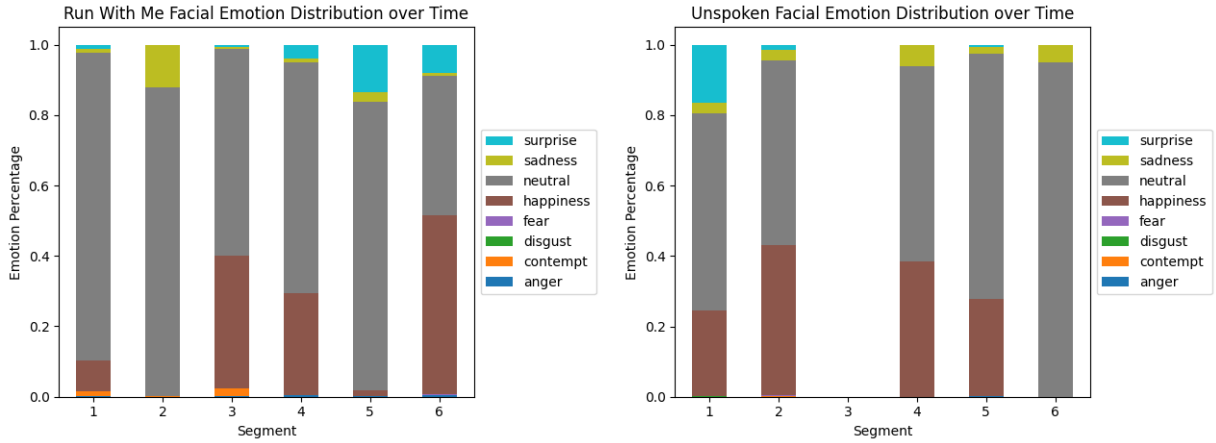
Including face emotion slightly improves the recall rate but hurts the skip rate. Overall, there is potential in incorporating emotion information from images and text but the existing tools are limited in their ability to extract emotion information from short clips (i.e., first six seconds of ads) and animated videos. The

Figure L3 Content Speech and Facial Emotion Distribution

(a) Speech Emotion - Video Indexer



(b) Facial Emotion - Face API



(c) Human-tagged Emotion

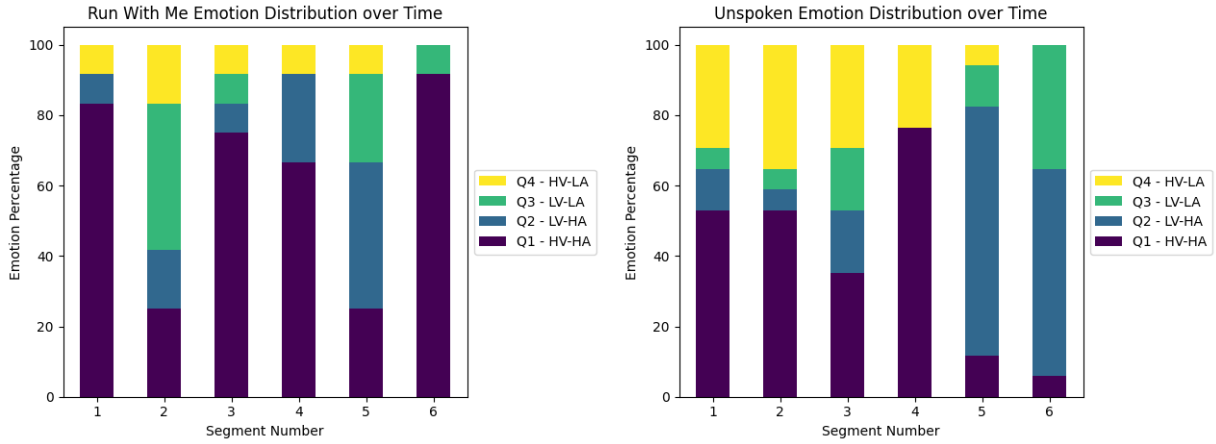


Table L4 Face and Music Emotion vs. Only Music Emotion - Unspoken and Run With Me

Feature	Model	JS Distance	Skip Rate	Recall Rate
Music Emotion				
Mel - Harmonics	CNN	0.429	42.0%	48.9%
Mel - Harmonics + Tempo Features	CNN + RF	0.395	40.7%	48.8%
Face and Music Emotion				
Mel - Harmonics	CNN	0.459	46.1%	49.9%
Mel - Harmonics + Tempo Features	CNN + RF	0.445	47.6%	49.0%

results suggest that audio models like the one included in Azure's Video Indexer service could benefit from incorporating music emotion classification.

References

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

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

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

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

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

Nelson DJ, Grazier R, Paglia J, Perkowitz S (2013) *Hollywood chemistry: When science met entertainment* (American Chemical Society).

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

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

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

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