TISMIR

The IsoVAT Corpus: Parameterization of Musical Features for Affective Composition

RESEARCH

ubiquity press

CALE PLUT
PHILIPPE PASQUIER
JEFF ENS
RENAUD TCHEMEUBE

*Author affiliations can be found in the back matter of this article

ABSTRACT

While there is a breadth of research in mapping Western musical features to perceived emotion within research in music and emotion, a critique of the field is that this breadth of methodologies lacks in inter-communication, which may reduce the generalizability of findings across the field. We consolidate previous research in this area to construct a parameterized composition guide that maps musical features to their associated emotional expression. We then use this guide to compose the "IsoVAT" dataset, a collection of symbolic MIDI clips in a variety of popular Western styles. This dataset contains a total of 90 clips of music, with 30 clips per affective dimension, organized into 10 sets of 3 clips. Each clip within a set is composed to express a low, medium, or high level of an affective dimension when compared to the other clips within the same set. We empirically evaluate the validity of our affective composition guide, to establish a ground-truth emotional expression in the dataset. Our validation reveals 19 sets where listener labels match the composed labels, 10 sets with listener labels that disagree with composed labels, and 1 clip that does not have clear agreement across the three study designs.

CORRESPONDING AUTHOR:

Cale Plut

Simon Fraser University, 8888 University Dr., Burnaby BC, Canada

cplut@sfu.ca

KEYWORDS:

Music; Affect; Emotion; Dataset; VAT

TO CITE THIS ARTICLE:

Plut, C., Pasquier, P., Ens, J., and Tchemeube, R. (2022). The IsoVAT Corpus: Parameterization of Musical Features for Affective Composition. *Transactions of the International Society for Music Information Retrieval,* 5(1), 173–189. DOI: https://doi.org/10.5334/tismir.120

1. INTRODUCTION

Music-emotion research (MER) is a broad interdisciplinary field that uses numerous approaches, models, and methodologies. In surveys of MER studies and stimulus selection for Western MER, criticisms concern a lack of internal coherence in terms of stimulus selection, emotion model, and definitions of musical features (Eerola and Vuoskoski, 2012; Warrenburg, 2020).

As we discuss in Section 2, these surveys identify musical stimuli as a potential confound in MER, noting that most studies use commercial recordings of existing music, selected by researchers or based on inclusion in previous studies. While using "real-world" music keeps external validity high, control over musical features is lost. To ensure that only desired features are altered, and to verify the emotional expression of musical stimuli, surveys suggest composing parameterically controlled music, as well as empirically ground-truthing the emotional expression of a musical dataset (Eerola and Vuoskoski, 2012; Warrenburg, 2020).

Another possible confound in MER datasets is the semantic gap between human perception of music and low-level features extracted from audio (Yang and Chen, 2011), and it has been suggested that audio features alone are insufficient for determining emotional expression (Panda et al., 2018). We address this confound by using symbolic representation of our corpus in MIDI, allowing for direct control over composition features.

In Section 2.5, we collate survey results from across MER (Juslin and Sloboda, 2011; Livingstone et al., 2010) to create a central set of musical features and their relationship to emotional expression. We delineate these features by whether they are primarily in the domain of musical composition, or expressive performance.

Section 3 describes and details our collated feature-affect guide. This guide presents musical parameters to control for emotional expression in music composition. To evaluate the guide, we interpret it to compose a musical corpus, and empirically evaluate the perceived emotions of the music. Because the guide is intended for use across a range of popular Western musical styles, we compose our music in a variety of popular Western styles.

We compose the "IsoVAT" corpus, manipulating the intended emotional expression of the music by manipulating the composition-related features of our guide. This corpus contains 90 4-bar musical clips, and is described in Section 4. The IsoVAT corpus is divided into three sets, expressing the isolated emotional dimensions of Valence, Arousal, and Tension. These clips are further divided into 10 sets of three, where each set contains clips expressing comparatively low, medium, and high levels of the associated dimension respectively. Each set shares instrumentation, genre, and tempo, to control for the possible effects of these features. A mix of popular, classical, and jazz genres is represented in the corpus.

We validate our musical set across three study designs in Section 5. The first study design, "2-rank", is discussed in Section 5.1.1. This design evaluates the clips as composed, with participants selecting the clips that they perceive the lowest and highest level of the associated affective dimension. The "1-rank" design is discussed in Section 5.1.2, and asks participants to rank 2-clip subsets of each set, selecting the clip that expresses a higher level of the associated affective dimension. Finally, the "Likert" design is discussed in Section 5.1.3, and asks participants to rate the degree of each clip's perceived affect from 1–7 along the associated affective dimension. Results across study designs show a surprising amount of variance, particularly the 1-rank design. The most stable evaluation occurs in the 2-rank design. We discuss the results of our empirical evaluations in Section 6.

While our results exhibit substantial variance, the corpus itself is also composed with several constraints that limit its emotional expression. Trends are generally shared between the 2-rank and Likert results that support the IsoVAT guide, though the 1-rank results are more varied. We combine all results to produce 29 ground-truthed ranked sets of 3 clips, with 1 set exhibiting too much variance to accurately ground truth. In Section 7, we musically analyze sets from the corpus whose ground-truth order is different than the composed order. We discuss common themes and elements that occur in these sets.

Overall, we investigate whether collected findings from previous MER literature can be used to express a desired affect while controlling a compositional process. In other words, we explore whether the study of affect and analysis of musical features can be applied to the creation of new music. In doing so, we find support for both findings and critiques of previous MER.

2 BACKGROUND AND MOTIVATION

2.1 AFFECT MODEL AND REPRESENTATION

Though the mechanisms are not fully understood, listeners perceive emotions in music, and music is commonly believed to be capable of evoking and inducing emotions in the listener. Affect models generally follow one of two approaches. Categorical models describe a set of basic universal emotions, from which all other emotions derive. Dimensional models describe emotions with two or three bipolar dimensions. The number of dimensions in a model is often derived from the application of the model, and the 3-dimensional model often contains some correlation between dimensions (Juslin and Sloboda, 2011; Schimmack and Grob, 2000).

Eerola and Vuoskoski (2012) describe a potential drawback to discrete emotion models, that they may produce Type 1 "false positive" errors and overconfidence. In a study with both categorical and dimensional models, Vieillard et al. (2008) find support for this, with

dimensional responses showing higher variability than categorical responses.

We use a 3-dimensional Valence-Arousal-Tension (VAT) model, similar to other 3-dimensional models (Wundt and Judd, 1902; Schimmack and Grob, 2000; Reisenzein, 2000). Table 1 provides an overview of common 2- and 3-dimensional emotion models used in previous MER. Tension is often discussed in music (Juslin and Sloboda, 2011), and therefore we include tension in our model. The most common other emotional models use valence/pleasure and arousal/activity (Yang and Chen, 2011; Eerola and Vuoskoski, 2011; Warrenburg, 2020).

2.1.1 Valence

Valence, sometimes called the "hedonic tone", is associated with the pleasantness or attractiveness of stimuli (Eerola and Vuoskoski, 2011; Schimmack and Grob, 2000). A stimulus that is pleasant or attractive has a high, positive valence. A stimulus that is unattractive or unpleasant has low, negative valence. Examples of high-valence emotions include joy, excitement, and triumph. Examples of low-valence emotions include sadness, fear, and disgust.

In music, positive valence is generally associated with major modes and consonant harmonies (Juslin and Sloboda, 2011). Harmonic consonance is not always well-defined, as it often depends on contextual elements like genre or historical context. An example of a high-valence piece used in MER is Vivaldi's *La Primavera*. A low-valence example is Barber's *Adagio for Strings*.

2.1.2 Arousal

Arousal is an emotional dimension associated with the energy or activity of stimulus (Eerola and Vuoskoski, 2011), and is occasionally called "energy arousal" (Schimmack and Grob, 2000). Arousal is a state of heightened activity, which may be positive or negative. Examples of high-arousal emotions include excitement, anger, and triumph. Examples of low-arousal emotions include satisfaction, depression, exhaustion, and relaxation.

In music, arousal is often associated with tempo and note density, increased volume, pitch level, and melodic direction (Juslin and Sloboda, 2011). For example, many loud pitches moving with an upwards contour will likely express positive arousal. Barber's *Adagio for Strings* expresses a low arousal and low valence. Mussorgsky's *Night on Bald Mountain* is a high-arousal piece used in previous MER studies.

2.1.3 Tension

Tension is a prospect-based emotional dimension associated with future or prospective events (Ortony et al., 1990). Tension can occur with both positive and negative valence. For example, excitement is a positively valenced tension – a subject believes that a future event is coming that will have a desirable effect. Fear is a negatively valenced tension – a subject believes a future event is coming that will have an undesirable effect. Examples of high-tension emotions include fear, excitement, and unease. Examples of low-tension emotions include satisfaction, sadness, and joy.

In music, tension is most often associated with harmonic instability, often described as dissonance (Juslin and Sloboda, 2011). Dissonances are a "clash" between notes that imply future resolution into consonance. Tension will generally increase as the implied resolution does not occur. Tension is also associated with melodic range and interval size (described in some literature as "interval pitch level", Juslin and Sloboda, 2011). *Night on Bald Mountain* expresses a high level of tension and arousal. Mozart's *Eine Kleine Nachtmusik* expresses a low level of tension.

2.2 THE NEED FOR PARAMETERIZED MUSIC-EMOTION DATASETS

Warrenburg (2020) highlights the importance of empirically ground-truthing datasets before use. 37% of musical stimuli in the "Previously Used Musical Stimulus" (PUMS) database was selected for a study due to inclusion in a previous study, without additional ground-truthing. Eerola and Vuoskoski (2012) advocate for creating parameterized musical stimuli, while

Model	# of Dimensions	Dimensions	Source
Wundt	3	Pleasure/Displeasure Arousal/Calmness Tension/Relaxation	Wundt and Judd (1902)
Circumplex (Russell)	2	Valence Arousal	Russell (1980)
2DES (2-Dimensional Emotion Space)	2	Valence Arousal	Schubert (1999)
PAD	3	Pleasure Arousal Dominance	Mehrabian (1996)
Schimmack and Grob	3	Valence Energy Arousal Tension Arousal	Schimmack and Grob (2000)

Table 1: Summary of common dimensional emotion models.

maintaining ecological validity. 33% of musical stimuli in Eerola and Vuoskoski's survey of MER was hand-selected by the researchers.

Most datasets in MER utilize commercial audio recordings as musical stimuli. Examples include *PMEmo*, drawn from Billboard Top 100 lists (Zhang et al., 2018), and *Emotify*, drawn randomly from a selection provided by the Magnatune company (Aljanaki et al., 2014). As mentioned in Section 1, using audio recordings reduces the extractable compositional features compared to symbolic music.

Multimodal approaches are one potential solution to the limits of audio feature extraction. Panda et al. (2013) present a dataset that includes audio, MIDI, and lyrics. This dataset is tagged with discrete emotion clusters using the MIREX classifications, and a classifier is trained on the dataset. The *EMOPIA* dataset is a multimodal audio and MIDI dataset in the "pop piano music" genre, annotated with emotional tags by the researchers (Hung et al., 2021).

Composers are occasionally asked to write custom music that expresses a particular emotion for a study, and listeners are generally able to identify the intended expression (Thompson and Robitaille, 1992; Vieillard et al., 2008). This shows support for the approach of composing custom music. However, these approaches generally leave the manipulation of the music mostly or entirely up to the composer's interpretation.

2.3 PARAMETRIC CO-CREATIVE COMPOSITION

Generative music is music that is partially or completely created with some automated process (Pasquier et al., 2017). One possible application of generative music systems is the co-creation of music with a human composer, though such systems often require the composer to use tools and techniques that may be technical and unfamiliar to them (Gerhard and Hepting, 2004), which may lead to frustration. Another difficulty in co-creation, as with MER in general, is the lack of agreed-upon musical parameters and terminology. One proposed solution is to automatically derive features from an input musical corpus (Paz et al., 2018).

Some generative music systems take a small, potentially single-piece corpus as input, and generate additional, similar music (Hernandez-Olivan and Beltran, 2021). *MidiMe* fine-tunes a VAE that is initially trained on Google's MusicVAE, and can be tuned based on a small corpus (Dinculescu et al., 2019). Two "inpainting" models take an input of two partial music clips, and output music that maintains the stylistic elements of the clips while musically transitioning between them (Pati et al., 2019; Hadjeres and Crestel, 2021).

Ens and Pasquier's *Multi-track Music Machine (MMM)* (Ens and Pasquier, 2020) has several inpainting capabilities. MMM optionally takes as input a single MIDI clip, which may be single-track or multi-track. Depending on the user's interaction, MMM can create music without an input, add additional instrumental lines to an input clip, and replace user-selected musical content with similar musical content, that musically fits into the input piece's musical context.

We believe that one possible future application of providing human-interpretable musical parameters that can integrate into a composition process is to allow for some degree of control over the output of a co-creative generative system. This approach could allow for affective control over a generative model without requiring a large, affectively tagged corpus of input music or formal definitions of musical parameters.

2.4 MUSICAL FEATURES AND ASSOCIATED EMOTIONAL EXPRESSION

As mentioned in Section 1, to produce a set of parameterically controlled musical clips, we first create a set of musical parameters to manipulate, which we call the *IsoVAT* guide, discussed in Section 3. This guide is intended to be flexible enough to apply to a broad range of Western musical styles, while providing enough detail to be consistently applied when interpreted. This guide is intended to be scalable with any instrumentation or degree of harmonic complexity.

To create this guide, we collate results from surveys on the emotional expression of musical features, and from cross-model surveys and studies of emotion. We find two meta-reviews of research into Western musical features and associated affect (Livingstone et al., 2010; Gabrielsson and Lindström, 2012). To achieve consensus, we include results only that are strongly present in both surveys. Gabrielsson and Lindström (2012) collect over 100 studies, differentiated by whether they are early studies using open-ended responses, multivariate listening studies, or post-2000 experimental studies. This survey does not translate affective models or terminology between each other, which means that an increase in tempo may increase both arousal and "excitement", a high-arousal emotion.

Livingstone et al. (2010) translate results from 102 studies to Schubert's "2 Dimensional Emotion Space" (2DES) model (Schubert, 1996), which contains both categorical and dimensional emotion descriptors. This study translates results from previous studies with a range of emotion models and musical features into a collated set of features and their associated expression.

Importantly, these surveys directly use the musical terminology from their surveyed sources. The lack of internal consistency and coherence in terminology has been previously mentioned as a critique of the broader MER field, and we therefore combine semantically related musical concepts into a single, unified vocabulary.

2.5 COLLATING RESULTS FROM VARIOUS EMOTION MODELS

As with musical terminology, previous musical feature-emotion surveys generally report emotions using the terminology and models of the surveyed source, which contributes to the lack of internal consistency in the field. As with musical vocabulary, we translate these emotional models into a single Valence-Arousal-Tension (VAT) dimensional model.

To accomplish this, we examine studies that translate between affect models, both within and outside of musical contexts (Eerola and Vuoskoski, 2011; Vieillard et al., 2008; Hoffmann et al., 2012). We identify 5 common categorical emotions with related dimensional mappings. While various studies use various scales (e.g. 1–5, 1–7, −5–5, 1–10), We normalize the data from these studies to a scale from −5–5, and average the normalized data to create a single value for each emotion, as shown in Figure 1. In our scale, a value of 0 indicates a neutral level of an affective dimension. As an example, we see that "Happiness", represented by a light-green diamond has a high valence value (>4), moderate arousal value (≈2.5), and moderately low tension value (≈−2.5).

As described in Section 2.4, Gabrielsson and Lindström (2012) delineate the sources of their surveyed studies based on the experimental methodology. We include studies that use multivariate analysis or empirical experiments. Livingstone et al. (2010) indicate which associations are found in at least 3 independent studies, and we consider only the features that meet this quantity. We adjust for small differences in language, e.g. various sources may describe "Articulation connectedness"}, "Articulation staccato" and "Articulation legato" as separate features. We simplify to a single feature when possible, such as "articulation connectedness". Other than combining feature descriptions, we avoid changing the terminology used in the sources.

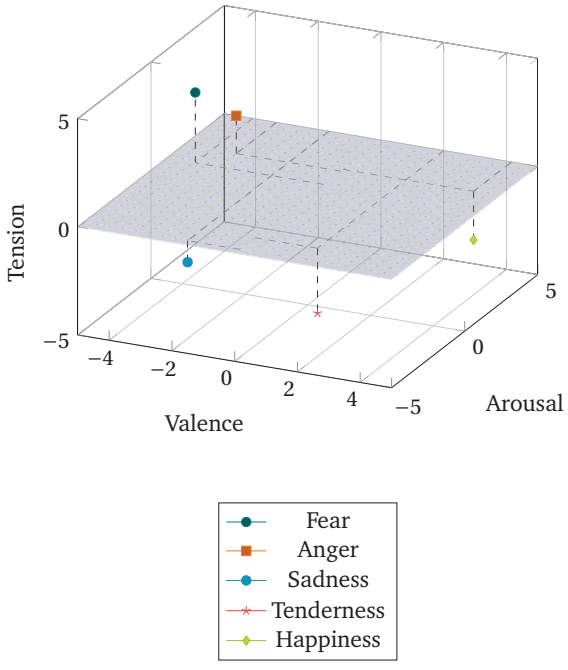


Figure 1: Discrete emotions placed in VAT space.

Figure 2a shows the dimensional mappings of composition-related features, and Figure 2b shows the mappings of performance-related features. We draw attention to the common trend in these results that all musical features have some correlative relationship with all three affective dimensions, indicating that musical features often produce multiple emotional correlations and expressions.

Figure 2 shows multiple directions to musically navigate the emotional space. To produce a composition guide for parameterically controlled emotional expression, we translate the data from Figure 1 one final time to an ordinal scale seen in Table 2.

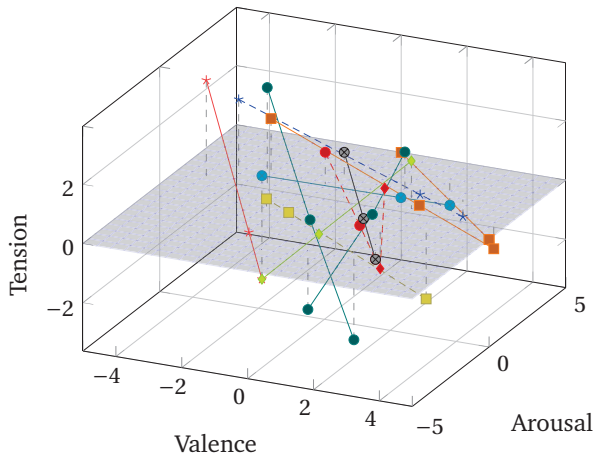
3. THE ISOVAT COMPOSITION GUIDE

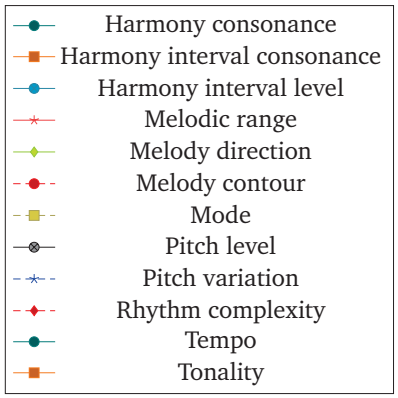
We collate the various MER sources into a central, unified model in both musical and emotional definitions, and present a set of feature-emotion mappings that is grounded in previous MER literature as much as possible, specific enough that various interpretations will result in relatively consistent emotional perceptions, and interpretable by a human composer during the composition task.

Essentially the IsoVAT guide presents a method for applying MER to human composition in a relatively controlled way. While there is interpretation required to realize the IsoVAT guide into music, it provides a higher degree of musical specificity and control than previous similar approaches (Vieillard et al., 2008; Thompson and Robitaille, 1992). Because most MER uses Western tonal music, the IsoVAT guide is primarily useful when composing Western tonal music, both functional and non-functional.

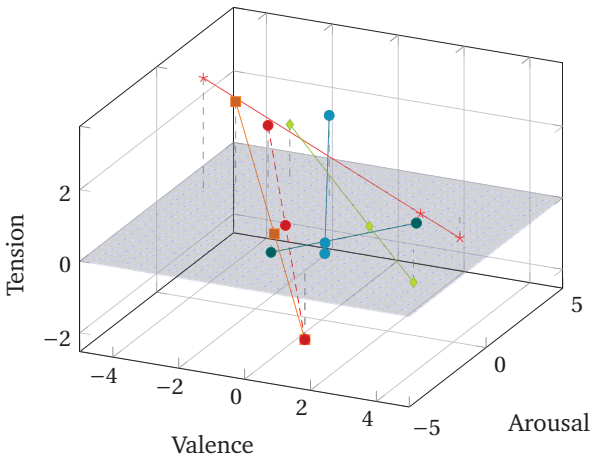
The IsoVAT guide can be understood as a set of *constraints*, to be used by the composer to express particular emotions. Composers often integrate them into their composition process. The IsoVAT guide is best classified as what Hasegawa (2020) describes as "relative material constraints". Composers are often familiar with including relative material constraints in their composition process, and many of the "rules" of tonal music can be classified as such (Hasegawa 2020).

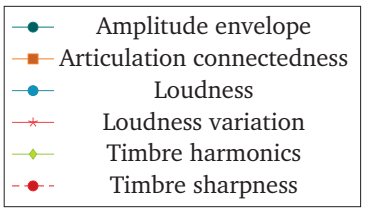
Table 2 shows our feature-affect composition guide. Table 2 reduces the data from Figures 2 into 3 ordinal relationships: A feature may be associated with a decrease (–), no change (0), or an increase (+) in perceived affect. We annotate the dimension that has





(a) Composition features





(b) Performance features

Figure 2: Musical features mapped to expressed affect.

Domain	Feature	Valence	Arousal	Tension	Description
Performance	Amplitude envelope roundness	0	0*		Increases as the amplitude envelope is more round/smooth
Performance	Articulation connectedness	+	−*	−	Increases as articulations are more connected/*legato*
Performance	Loudness	0	+*	0	Increases as overall volume increases
Performance	Loudness variation	−*	0	+	Increases as volume level peaks have greater difference
Performance	Timbre harmonics	+*	0	0	Increases with presence and strength of harmonics
Performance	Timbre sharpness	+	−*	−	Increases as higher harmonics are increasingly represented in timbre
Both	Tempo	0	+*	+	Increases as tempo increases
Composition	Harmonic consonance	+	−	−*	Increases as harmonies simplify — exact definition determined by genre
Composition	Interval consonance	+*	−*	0	Increases as melody uses simpler intervals — genre-dependent
Composition	Interval size	0	+*	−	Increases as intervalic distance is increased
Composition	Melodic range	−	+*	+	Increases as difference between high and low pitches in melody increases
Composition	Melodic direction	0	+*	0	Increases as melody motion is towards higher pitches
Composition	Melodic contour	0	+*	0	Increases as melody motion contains more internal distances between pitches
Composition	Mode	+*	0	0	Increases as mode is increasingly Major
Composition	Pitch level	0	+*	0	Increases as overall pitches are higher
Composition	Pitch variation	+*	0	0	Increases as distance of individual intervals increases
Composition	Rhythm complexity	0	0*	0	Increases as rhythms are moved away from standard “strong” beats such as 1 and 3
Composition	Tonality	+*	0	0	Increases as hierarchical tonal relationships are increasingly used

Table 2: Composition Guide for affective Western music.

the strongest affective association with an asterisk (*). While a composer may include all musical features when modifying affect, they may also select a subset of features to manipulate, as the surrounding musical context or genre conventions may reduce the composer’s freedom to modify all features.

As an example of how this guide might be used, if the composer wishes to express an increasing amount of musical tension, decreasing the harmonic consonance, decreasing the interval size and broadening the melodic range will express increasing tension. We use this guide to create sets of three clips that express differing levels of emotion, by manipulating these features in comparison to the other clips within the set. For example, a low-arousal clip may have a narrower melodic range, with a narrower melodic contour (moving in smaller intervals horizontally), that moves in a downward direction with lower pitch levels (tessitura), and narrower intervals in the accompanying harmony, compared to the moderate and high arousal clips within the set.

4 THE ISOVAT CORPUS

4.1 COMPOSITION

Our composer and first author of this paper composes a total of 90 4-bar musical clips, which we call the *IsoVAT* corpus. Our composer has a background in music composition and live performance in an array of popular Western styles. This background includes three years as a pianist and occasional band leader onboard luxury cruise ships, and 7+ years performing as a pianist and composer across the United States and Canada.

The duration of the clips was chosen to provide a single musical idea with a consistent emotional expression. Emotional perception of clips can be measured and modeled in two main ways: as a continuous time series, or as a single time-point (Kim et al., 2010). When classifying individual clips of music with emotional perception, listeners are able to identify the expressed emotion with as little as 1 second of music (Kim et al., 2010; Eerola and Vuoskoski, 2012).

The *IsoVAT* corpus can be divided by emotional dimension, and further grouped into 10 sets of 3 per dimension. Each clip, within each set, is composed and labeled to express a low, middle, or high level of the expressed affective dimension, when compared to the other two clips within the set.[1]

We notate sets using the shorthand {Dimension}-{Number}, and clips using the shorthand {Dimension}-{Number}-{Clip}, where V-6 indicates Valence set 6, and T-3-H describes the high-composed clip in tension set 3.

Each set of pieces shares an instrumentation and genre, drawing from a variety of Western popular styles. For example, V-2 uses a single Disco ensemble for all 3 clips. We isolate a single affective dimension at a time in the IsoVAT corpus, and therefore do not require consistency of genre and instrumentation across dimensions. We include an example of well-known artists within each genre, and note that we do not attempt to mimic these artists, but they serve as examples of the target genre.

Each set is composed to primarily express affect by manipulating the composition-domain features identified in Table 2. We avoid manipulating features that are not strongly associated with the chosen dimension when possible. While we identify a set of music performance features, we only manipulate the composition features to produce our dataset. The genre, instrumentation, and examples of the target genre of each clip is provided in Table 3.

Figure 3 shows a score reduction for A-7, written for jazz ensemble in the swing/bebop genre, to provide an example of how our composition guide is used. All scores have been written in the MuseScore 3 notation software.[2] The composition features where arousal is the strongest emotional association are: interval consonance, interval size, melodic range, melodic direction, melodic contour, and pitch level. We manipulate all of these features in A-7, as described in Table 4.

In Table 4, A-7-H uses a mix of dissonant and consonant intervals, ranging from moving by half steps to moving by minor sevenths. A-7-H's melody has a total range of 16 semitones, from E♭3-G4. To measure the direction, we describe the number of melodic peaks, or the number of times the melody changes direction. In A-7-H, the melody has 6 melodic peaks, the contour is jagged, and many direction changes involve large leaps.

Valence

#	Genre	Instrumentation	Genre example
1	Classical	Solo piano	J.S. Bach
2	Disco	Disco ensemble	Earth, Wind, and Fire
3	Swing	Jazz combo	Glen Miller
4	Rock/Pop	Rock band	Rolling Stones
5	Piano rock/Funk	Rock band (w. piano)	Stevie Wonder
6	Soft rock	Rock band	Grover Washington Jr.
7	60s rock	Rock band	Creedence Clearwater Revival
8	Latin	Jazz combo	Guido Guidoboni
9	Ragtime	Solo piano	Scott Joplin
10	Film	Orchestra	John Williams

Arousal

#	Genre	Instrumentation	Genre example
1	Classical/Romantic	Woodwind quintet	Debussy
2	Rock/Pop	Rock band	AC/DC
3	Rock/Pop	Rock band	The Doors
4	Hard rock/Metal	Rock band	Metallica
5	Piano rock	Rock band(w. piano)	Billy Joel
6	Rock/Hard rock	Rock band	ZZ Top
7	Swing/Bebop	Jazz combo	Dizzy Gillespie
8	Latin	Jazz combo	Bob Mintzer
9	Piano rock	Rock band(w. piano)	Elton John
10	Classical	Brass quintet	Malcom Arnold

Tension

#	Genre	Instrumentation	Genre example
1	Classical	Solo piano	Mozart
2	Classical	Brass quintet	Ligeti, Sousa
3	Surf rock	Rock band	The Surfaris
4	60s rock	Rock band	Pete Townshend
5	Bluegrass	Bluegrass ens.	Foggy Mountain Boys
6	Europop	Electro/Synth	Haddaway
7	Rock/Pop	Rock band	Grateful Dead
8	Stadium rock	Rock band	Bon Jovi
9	Folk rock	Rock band	Dolly Parton
10	Choral	SATB+Piano	Eric Whitacre

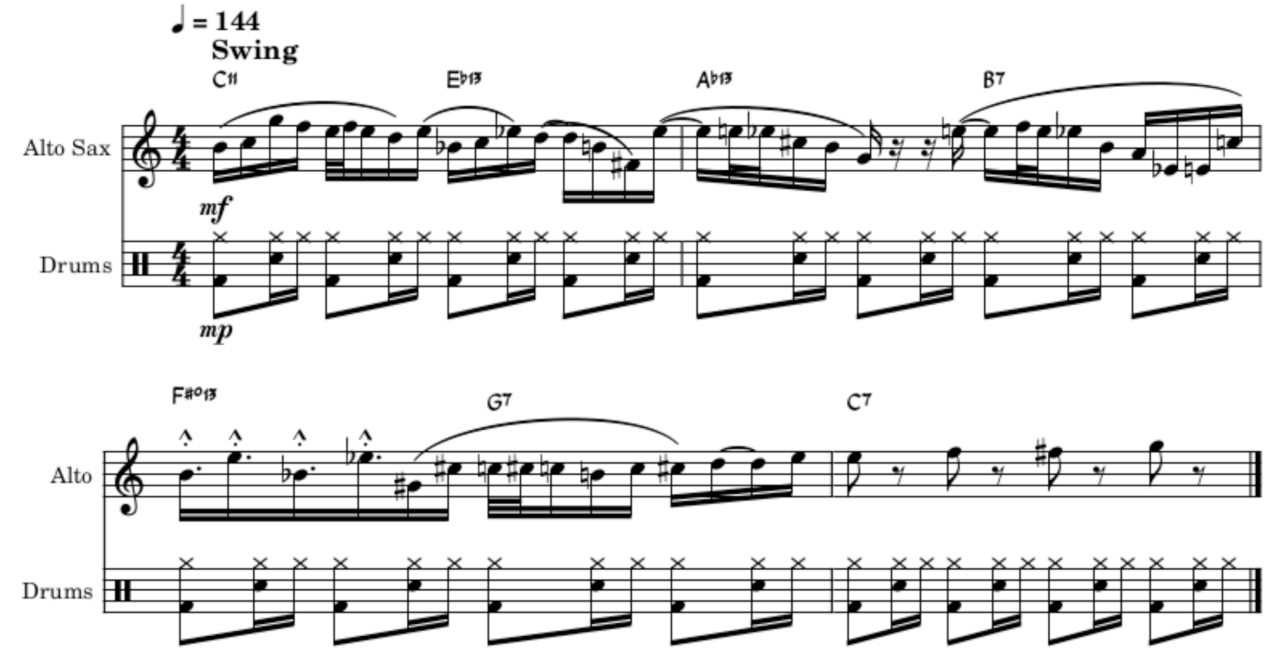
Table 3: Genre and Instrumentation of IsoVAT corpus.

4.2 AUDIO RENDERING AND INTERPRETATION

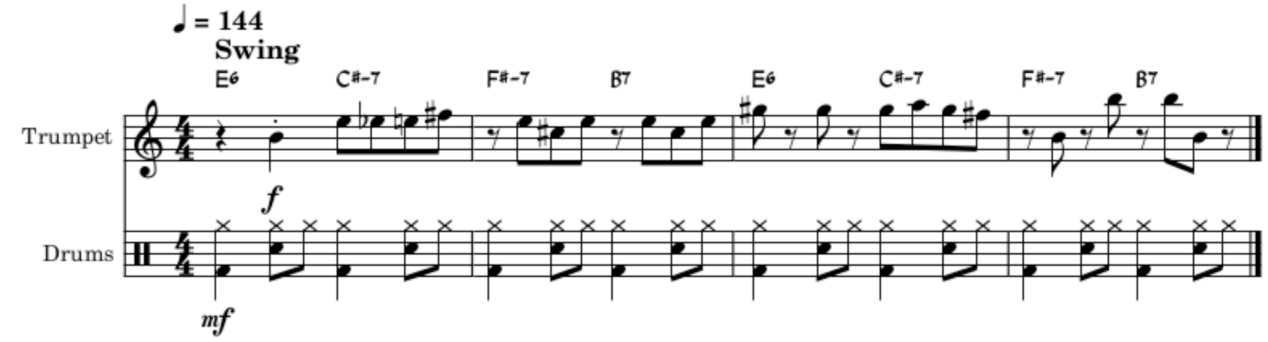
MIDI represents music as data that must be synthesized to produce audio. When a MIDI file is played, the sounds are determined by a sample-based soundfont. Soundfonts can be used to replicate the synthesis of a MIDI file consistently between computers. We use the "Arachno" soundfont,[3] to synthesize the IsoVAT corpus into audio that will be consistent across listeners.

5. GROUND TRUTHING EXPERIMENT

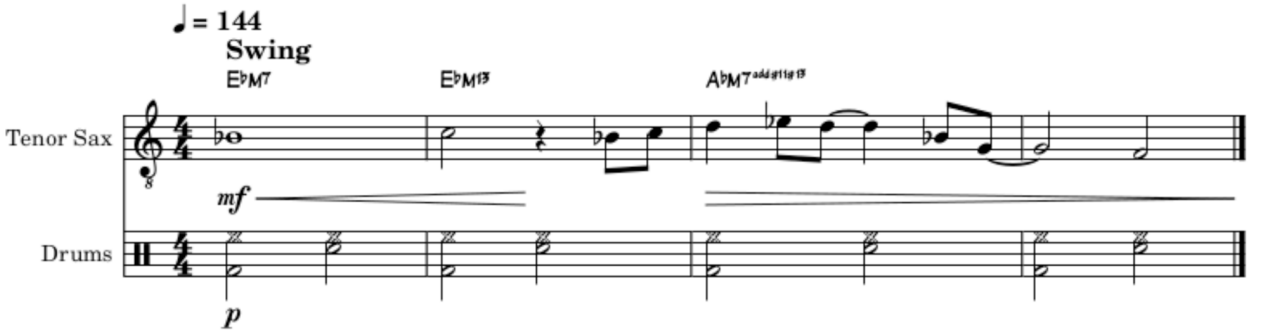
Each clip the IsoVAT corpus is labeled with the intended emotional expression in its set, e.g. "high", "medium", or "low". To evaluate the composition guide, we ground-truth order the dataset by empirically labeling the emotional perception that listeners report for each clip, with varying degrees of musical context.



(a) High



(b) Mid



(c) Low

Figure 3: Reduced score for Arousal set 7.

Label	Int. consonance	Int. pitch level	Mel. range	Mel. dir.	Mel. contour	Pitch level
Low	Mostly consonant	Steps, thirds	10 semitones	1 peak	Smooth, small	F3-E♭4
Mid	Generally consonant	Steps, thirds, fourths, fifths, octaves	12 semitones	2 peaks	Smooth, wide	B♭4-B♭5
High	Mix	Steps, thirds, fourths, tritones fifths, sixths, sevenths	16 semitones	6 peaks	Jagged, wide	E♭3-G4

Table 4: Arousal-manipulating features as manipulated in Arousal set 7.

In the 2-rank design discussed in Section 5.1.1, all three clips are heard for comparison, presenting the full composed context. In the 1-rank design in Section 5.1.2, clips are heard with one contextual clip, but not the other. In the Likert design in Section 5.1.3, clips are completely removed from their musical context, and rated on an absolute scale.

Across all empirical study designs, we collect 30 rankings per clip. Participants are recruited, and the study is performed, using Amazon's Mechanical

Turk platform. MTurk does not provide, nor do we collect, additional demographic information. Consent is obtained prior to participation. Participants may participate in each study design only once per dimension, but may participate in multiple study designs. Participants take an average of 20 minutes to complete all ranking tasks.

5.1 EMPIRICAL METHODOLOGY

5.1.1 As composed: "2-rank"

Each participant answers a total of 10 questions relating to a single affective dimension per task. 90 participants are paid US$0.10/ranking ($1.00/task). Consent is obtained before each task. Participants are provided a description of the randomly selected affective dimension that they will evaluate. One set of 3 clips is randomly selected to provide an example of low, medium, and high levels of the assigned dimension.

Participants complete 9 ranking tasks, with the remaining sets. Each task involves listening to 3 musical clips, and selecting the clips that express the highest and lowest level of their dimension. To ensure participant accuracy, an additional audio file that consists of a voice instruction to select a particular response is included. Participants who fail to correctly follow the speech instructions are removed from the study.

5.1.2 Pairwise "1-rank" of 2 clips

In the 2-clip/"1-rank" study, 180 participants are asked to perform a single ranking, rather than selecting a low and high. Participants listen to 15 pairs of clips, including an example. For each question, participants listen to two clips, and select the clip that they believe expresses their assigned dimension more strongly. The clips are drawn from a random selection within 5 sets, with participants performing 3 rankings per set, for a total of 15 pairwise combinations of subsets.

This 1-rank design provides some musical context for each ranking, as each clip is evaluated compared to a single other clip. However, it does remove part of the contextual musical information, as clips are composed in sets of 3.

5.1.3 Individual Likert scale rating

We evaluate the corpus via a single 7-point Likert scale. For this study, 180 participants listen to three example sets, arranged in sets of 3. Participants then listen to 14 individual clips, drawn randomly from the corpus, and provide a single rating from 1 (expresses a very low level of the affective dimension) to 7 (expresses a very high level of the affective dimension).

This design evaluates each clip in isolation, completely out of their composed context. While we are primarily investigating the parameterization as a contextual, ordinal guide, we expect that when removed entirely from context, participants will use a more absolute scale that will somewhat align with the contextual composition.

6 EMPIRICAL RESULTS

6.1 2-RANK

Results are analyzed to view the inter-rater agreement and ranked order of each set. Inter-rater reliability ranges from 56–76% in the ground-truth order. Agreement between composed labels and responses are between 39–76% compared to a random chance of 17%. 20 out of 30 sets are ground-truth ordered with the same labels as composed, with 6 valence sets, one arousal set, and 3 tension sets showing disagreement between composed labels and ground-truthed order. Shapiro-Wilk and Anderson-Darling tests are performed across participant data. In the 2-rank responses, participant responses demonstrate a normal distribution for all three dimensions.

Table 5 shows confusion matrices for responses as ground-truth ordered, and in their composed order, showing means and standard deviations for the low and high-selected clips. Because participants only select the low and high clip, the mid-level means are inferred. For example, we can see that for valence, clips labeled as the lowest by the ground-truth ranking are ranked as the lowest clip an average of 59% of the time, with a standard deviation of 15%. Figure 4 presents the same data without the inferred middle values.

Ground truth order

(a) Valence

Label \ Response	Low	Mid	High
Low	0.59 ± 0.15	0.28	0.13 ± 0.07
Mid	0.27 ± 0.08	0.48	0.25 ± 0.07
High	0.17 ± 0.07	0.26	0.57 ± 0.15

(b) Arousal

Label \ Response	Low	Mid	High
Low	0.64 ± 0.10	0.29	0.07 ± 0.05
Mid	0.28 ± 0.10	0.57	0.15 ± 0.07
High	0.07 ± 0.05	0.17	0.76 ± 0.12

(c) Tension

Label \ Response	Low	Mid	High
Low	0.66 ± 0.13	0.27	0.07 ± 0.10
Mid	0.25 ± 0.10	0.56	0.19 ± 0.11
High	0.08 ± 0.07	0.19	0.73 ± 0.20

Composed Labels

(d) Valence

Label \ Response	Low	Mid	High
Low	0.44 ± 0.21	0.24	0.32 ± 0.23
Mid	0.22 ± 0.17	0.41	0.37 ± 0.21
High	0.21 ± 0.12	0.40	0.39 ± 0.23

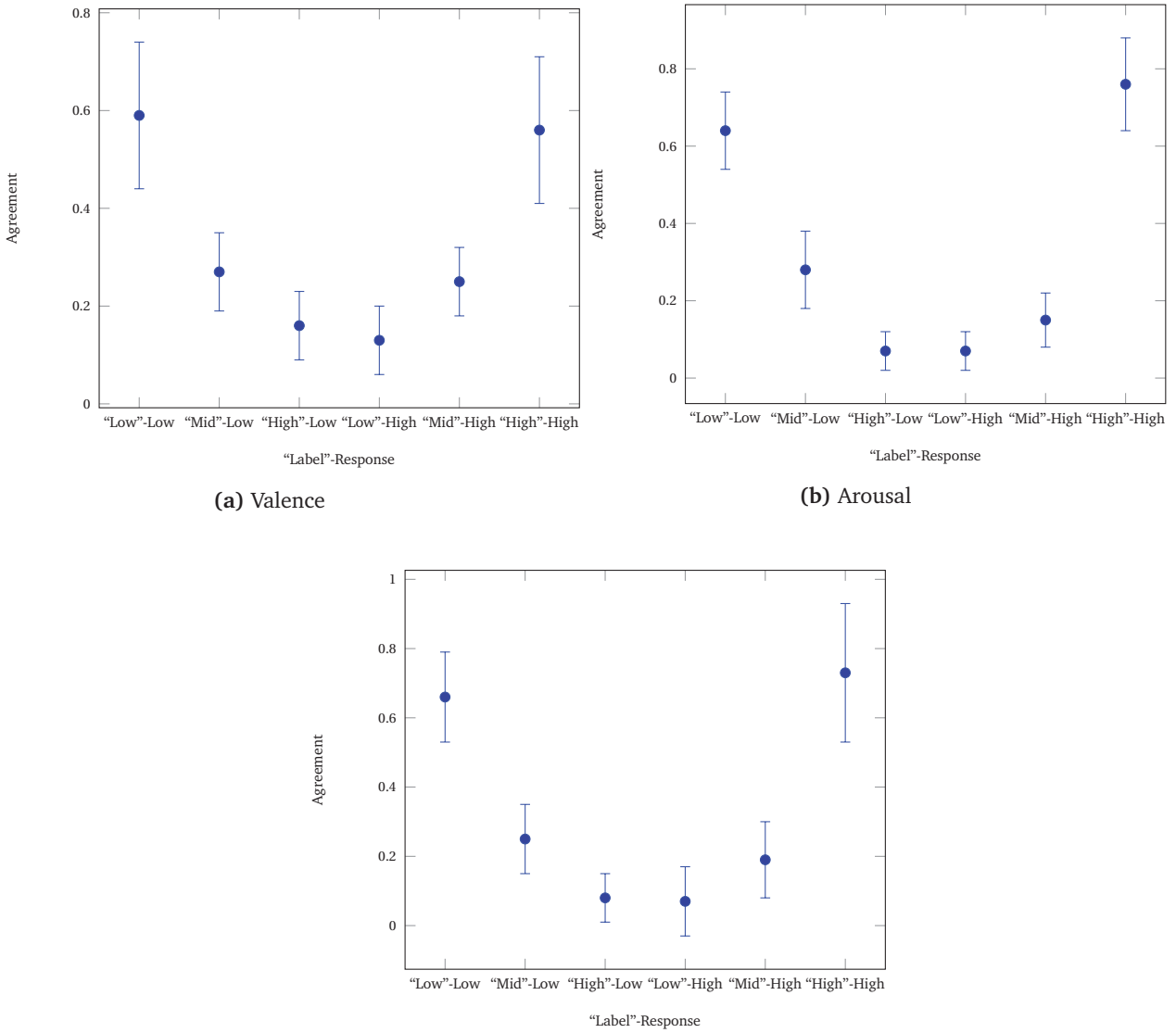
(e) Arousal

Label \ Response	Low	Mid	High
Low	0.63 ± 0.11	0.30	0.07 ± 0.06
Mid	0.28 ± 0.12	0.62	0.10 ± 0.07
High	0.12 ± 0.08	0.12	0.76 ± 0.12

(f) Tension

Label \ Response	Low	Mid	High
Low	0.60 ± 0.20	0.32	0.08 ± 0.11
Mid	0.23 ± 0.15	0.54	0.23 ± 0.21
High	0.15 ± 0.19	0.17	0.68 ± 0.25

Table 5: Means and standard deviations in confusion matrices for 2-rank study results.



(a) Valence

(b) Arousal

(c) Tension

Figure 4: Means and standard deviations for each clip's ground-truth label-response pairing.

6.2 1-RANK

Participants agree an average of 58.5% of the time across all 30 sets and pairwise comparisons. Divided by dimension, these agreements are Valence: 59.2%, Arousal: 59.4%, and Tension: 56.8%. Participants agree with the composed rankings 43.0% for Valence, 48.3% for Arousal, and 49.6% for Tension. This surprisingly underperforms random chance of 50%. We believe that this is due to the presence of incomplete musical context acting to confound listener perceptions. When musical clips are completely removed from context in Section 6.3, participant agreement with composed labels increases compared to the 1-rank study design. Table 6 shows the agreements by dimension and pairwise comparison. As in the 2-rank study, participant responses are normally distributed among answers.

We draw attention in Figure 6 to the lack of additional clarity in the comparisons of High-Low pairs compared to the intermediary comparisons. The high and low clips are expected to express the ends of an affective dimension, controlled for other musical factors, and we expect these end points to be more clearly differentiated than when one clip expresses a moderate level of the affective dimension.

Dimension	Labels	Ground truth		Composed labels	
		M	SD	M	SD
Valence	H-L	0.60	0.07	0.43	0.11
Valence	H-M	0.58	0.05	0.44	0.07
Valence	M-L	0.59	0.07	0.42	0.08
Arousal	H-L	0.58	0.52	0.49	0.10
Arousal	H-M	0.59	0.05	0.53	0.10
Arousal	M-L	0.61	0.07	0.43	0.11
Tension	H-L	0.55	0.05	0.51	0.07
Tension	H-M	0.56	0.04	0.49	0.07
Tension	M-L	0.59	0.06	0.49	0.11

Table 6: Agreement values from 1-rank study for ground-truthed order and composed labels.

6.3 INDIVIDUAL LIKERT RATING

While Likert scales are commonly analyzed as interval data, we follow suggestions to treat Likert scales as ordinal data (Wu and Leung, 2017). We compute the median absolute deviation for each clip. As with the other study designs, responses are normally distributed. We compute Cohen's kappa for each set of three clips to determine whether at least one clip within the set expresses a significantly different emotional level than the other members of the set. These statistics are shown in Table 7 for each set.

In terms of agremeent with the composed labels, 1 valence set, all 10 arousal sets, and 5 of the tension sets agree with the composed labels when ties are broken towards the ground-truth order from the 2-rank study. When ties are broken towards composed order, 4 valence sets, 10 arousal sets, and 7 tension sets agree with the composed order. Overall, this means that when ties are broken towards the ground-truth order, Likert data agrees with composed order in 16 clips, and when ties are broken towards the composed order, Likert data agrees with the composed order in 21 clips. This data does not include the 3 excluded example Valence sets, as discussed below.

We additionally look at significant differences of Likert ratings within each set, measuring whether the clips within the set are significantly differentiated from each other in terms of median Likert scores. 2 sets expressing Valence levels, all 10 sets expressing Arousal levels, and three sets expressing Tension levels show significant differences. Of these sets, 8 arousal sets and one tension set show significant differences and agree with the composed set.

Our Likert data does not include 3 Valence sets — we use data from previous evaluations of the corpus to select example sets that provided the most clear data. Due to only three sets of valence clips matching composed labels in the 2-rank design, we do not collect data on these example clips. For arousal and tension, example clips could be included without reducing the number of evaluated clips, as there are enough sets that match the composed label that we randomly sample from possible example clips, and gather data on the others. We note that this creates a further reduction in the accuracy and agreement of the valence clips in this design. Our Likert data also does not have information on the lowest expression of each dimension within set 1, due to a coding error.

6.4 AGGREGATING GROUND TRUTH FROM MULTIPLE STUDIES

We collate trends between designs to ground-truthed order each set of clips. Only one set in each dimension have a ground-truth order that is consistent across all three studies, coincidentally set 10. For V-10, the ground-truth order is different than composed. Because only 3% of the corpus has agreement between all three study designs, we require agreement between at least two ground-truth study designs. Trends are mostly shared between the 2-rank and Likert designs. In the event of a tie within the Likert responses, the ground-truth 2-rank order breaks the tie. The valence sets that serve as an example for the Likert study are assumed to agree with the 2-rank ground-truth order for the purposes of deriving an aggregate ground-truth order.

Our results are presented in Table 8. Results are given with composed labels, the ground-truth orders derived from each study design, and the aggregate derived ground-truth order. Green ★ cells indicate agreement with the composed labels. Blue ◊ cells represent order agreement with at least one other study, in disagreement with composed labels. Yellow □ cells represent no agreement with composed labels or other study design orders. Finally, red cells with either △ or ▽ indicate an irreconcilable loop in the collected pairwise comparisons. Irreconcilable loops may occur as forwards loops with ground-truth ranking of H->M->L->H, indicated by △, or reverse loops as L->M->H->L, indicated by ▽. While neither loop type can be turned into a ground-truth order, the reverse loop is the more serious change from the composed order, as only a single pairwise ranking agrees with the composed label. In a forward loop, only a single pairwise ranking does not agree with the composed label.

As an example for reading Table 8, V-4 has an order of M->H->L in both the Likert and 2-rank responses, and an order of M->L->H in the 1-rank order. The Likert order shows a significant difference between at least one of the clips and the rest of the set, and contains a tie between the Medium and High clips. This tie is broken in favour of the 2-rank order, which aligns the Likert order with the 2-rank order, producing an overall ground-truth order of M->H->L.

In terms of study designs, the 2-rank study design that evaluates the expression of clips as composed has the highest inter-rater agreement at 76%, and the highest agreement with the composed order. The 1-rank study demonstrates the most variance, and in some cases the composed labels do not outperform random chance. In terms of dimensions, the arousal sets have the highest inter-rater agreement, and the valence sets generally have the lowest inter-rater agreement. This is consistent with previous research in MER. These trends are reversed in parts of the 1-rank design, though the trends in the 1-rank design are smaller and more varied than the other two designs.

V-2 is the only set that exhibits no agreement across at least two designs. In the remaining 29 sets, 27 of the ground-truth orders are derived from agreement between the 2-rank and Likert orders. T-6 and T-9 derive their ground-truth order from agreement between the 1-rank and Likert orders. Both of these sets' 2-rank ground-truth order agrees with the composed labels.

Valence						
#		Low	Mid	High	ChiSquare	p
1	Med		3	3	0.06	0.79
	Abs dev		1	1		
2	Med	5	4	5	1.11	0.57
	Abs dev	1	1	1		
4	Med	4	5	5	17.17	<0.01*
	Abs dev	1	1	1		
5	Med	4	4	3	0.26	0.87
	Abs dev	1	1	1		
6	Med	4	4	4	0.65	0.72
	Abs dev	1	1	1		
8	Med	5	5	4	13.19	<0.01*
	Abs dev	1	1	1		
10	Med	5	5	6	3.29	0.19
	Abs dev	1	1	1		
Arousal						
#		Low	Mid	High	ChiSquare	p
1	Med		2.5	4	4.80	0.03*
	Abs dev		1	1		
2	Med	4	4	6	24.67	<0.01*
	Abs dev	1	1	1		
3	Med	3	3.5	5	5.96	0.05*
	Abs dev	1	0.5	1		
4	Med	4	4	6	30.92	<0.01*
	Abs dev	1	1	1		
5	Med	3	3	4	9.94	<0.01*
	Abs dev	1	1	1		
6	Med	4	4	6	17.39	<0.01*
	Abs dev	1	1	1		
7	Med	4	4.5	7	13.21	<0.01*
	Abs dev	1	0.5	1		
8	Med	4	4	5	28.70	<0.01*
	Abs dev	1	1	1.5		
9	Med	3	3	4.5	12.33	<0.01*
	Abs dev	1	1	1.5		
10	Med	3	4	4	8.05	0.02*
	Abs dev	1	1	1		
Tension						
#		Low	Mid	High	ChiSquare	p
1	Med		3	5	3.55	0.06
	Abs dev		1	2		
2	Med	4	4	4	0.54	0.76
	Abs dev	1	1	1		
3	Med	3.5	4	5	6.29	0.04*
	Abs dev	0.5	1	1		
4	Med	4	4	5	2.72	0.26
	Abs dev	1	1	1		
5	Med	3	5	6	5.95	0.05*
	Abs dev	1	1	1		
6	Med	5	4.5	4	0.35	0.83
	Abs dev	1	1.5	1		
7	Med	4	5	4	2.17	0.33
	Abs dev	1	1	1		
8	Med	4	4	4	1.31	0.51
	Abs dev	1	1	1		
9	Med	4	6	5	8.89	0.01*
	Abs dev	0.5	1	1		
10	Med	4	4	6.5	5.63	0.06
	Abs dev	1	1	1		

Table 7: Medians, Absolute deviations, and Chi Square tests for Likert responses.

#	Composed labels	2-rank order	1-rank order	Likert Order	Aggregate G-T Order
			Valence		
1	H-M-L	(H-M)-L ⋆	L-M-H □	(H-M) ⋆	H-M-L⋆
2	H-M-L	M-L-H □	L-M-H □	(L-H)-M □	—
3	H-M-L	H-M-L ⋆	L-H-M □	Example —	H-M-L⋆
4	H-M-L	M-H-L ◇	M-L-H □	(M-H)-L* ◇	M-H-L◇
5	H-M-L	M-L-H ◇	Reverse Loop ▽	(M-L)-H ◇	M-L-H◇
6	H-M-L	M-L-H ◇	L-H-M □	(M-L-H) ◇	M-L-H ◇
7	H-M-L	H-M-L ⋆	M-L-H □	Example —	H-M-L⋆
8	H-M-L	L-M-H ◇	H-L-M □	(L-M)-H* ◇	L-M-H◇
9	H-M-L	H-M-L ⋆	M-L-H □	Example —	H-M-L⋆
10	H-M-L	H-L-M ◇	H-L-M ◇	H-(L-M) ◇	H-L-M◇
			Arousal		
1	H-M-L	H-M-L ⋆	L-H-M □	H>M* ⋆	H-M-L⋆
2	H-M-L	H-M-L ⋆	M-H-L □	H-(M-L)* ⋆	H-M-L⋆
3	H-M-L	H-M-L ⋆	Loop △	H-M-L ⋆	H-M-L⋆
4	H-M-L	H-M-L ⋆	H-L-M ◇	H-(M-L)* ⋆	H-M-L⋆
5	H-M-L	H-M-L ⋆	H-L-M ◇	H-(M-L)* ⋆	H-M-L⋆
6	H-M-L	H-M-L ⋆	Loop △	H-(M-L)* ⋆	H-M-L⋆
7	H-M-L	H-M-L ⋆	H-L-M □	H-M-L* ⋆	H-M-L⋆
8	H-M-L	H-M-L ⋆	L-H-M □	H-(M-L)* ⋆	H-M-L⋆
9	H-M-L	H-M-L ⋆	R. Loop ▽	H-(M-L)* ⋆	H-M-L⋆
10	H-M-L	H-M-L ⋆	H-M-L ⋆	(H-M)-L ⋆	H-M-L⋆
			Tension		
1	H-M-L	H-M-L ⋆	H-L-M □	H-M ⋆	H-M-L⋆
2	H-M-L	H-L-M ◇	L-M-H □	(H-L-M) ◇	H-L-M◇
3	H-M-L	H-M-L ⋆	M-H-L □	H-M-L ⋆	H-M-L⋆
4	H-M-L	H-M-L ⋆	H-M-L ⋆	H-(M-L) ⋆	H-M-L⋆
5	H-M-L	H-M-L ⋆	L-M-H □	H-M-L* ⋆	H-M-L⋆
6	H-M-L	H-M-L ⋆	L-M-H ◇	L-M-H ◇	L-M-H◇
7	H-M-L	M-H-L ◇	R. Loop ▽	M-(H-L) ◇	M-H-L◇
8	H-M-L	(M-L)-H ◇	L-M-H □	(M-L-H) ◇	M-L-H ◇
9	H-M-L	H-M-L ⋆	M-H-L ◇	M-H-L* ◇	M-H-L ◇
10	H-M-L	H-M-L ⋆	H-M-L ⋆	H-(M-L) ⋆	H-M-L⋆

Table 8: Central comparison of ground truth order by study design and final ground-truth order, see Section 6.4

V-8 and T-6 are ground-truthed in the reverse order compared to the composed labels. Three sets demonstrate "major" differences in order, where the low or high clip is re-ordered to be on the opposite end of the order. Five sets demonstrate "minor" differences in order, where the low or high clip is swapped with the medium clip. In no sets is the composed "low" clip ranked as the highest without also reversing the "high" and "mid" order.

7 MUSICAL ANALYSIS OF POTENTIAL CONFOUNDS

We compose the IsoVAT corpus based on our composition guide, and therefore evaluate the guide based on the results of the ground-truth experiments for the corpus. We musically analyze the sets that do not agree with the composed labels, and the set that does not have a ground truth consensus. We identify four features that are found in pieces whose ground truth order disagrees with the composed labels.

7.1 SEQUENCES

V-5, V-6, and V8 are ground-truth ordered with the composed mid clip being moved to the high position. V-5-M and V-6-M are shown in Figure 5. These clips use a falling-fifths progression, starting in minor, changing the sonority outlined by each chord to fit in the mode. V-6-M begins on an EbM7 chord to allow for the melodic pickup. V-8-M uses a slightly more intricate i-vii^{o7}/VI-VI-VI^{+9} sequence.

Tonality, pitch variation, mode, and interval consonance are the features primarily associated with valence. In these sets, we begin our sequence on a minor chord, and assume that alternating the sonority expressed in each chord between minor and major would provide modal ambiguity through the sequence, expressing a moderate level of valence. While these sequences could resolve to Major or minor end points, we believe that participants identified the tonal centre of each sequence as Major. This may indicate that the harmonic motion in 4th and 5th based sequences may be primarily perceived as an increase in tonal hierarchies, associated with positive valence.

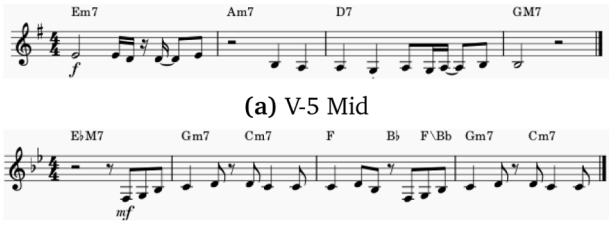


Figure 5: Reduced score for mid clips in valence sets 5 and 6.

7.2 HARMONIC COMPLEXITY AS DISSONANCE

Pieces using complex Major harmonies such as Major 7ths, 9ths, and 13ths tend to be ground-truth ordered in disagreement with the composed labels. V-5-H, V-6-H, and V-8-H use Major 7ths and 9ths, and are ground-truth ordered in the lowest position. V-8-H is shown in Figure 6. V-4-H contains Major 7ths, but fewer other complex harmonies than sets 5, 6, and 8, and is ground-truth ordered in the middle position.

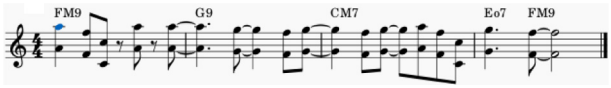


Figure 6: Reduced score for Valence 8-High.

While these features mainly affect valence, harmonic complexity may also affect perception of tension. In T-7, the medium and high-composed clips are swapped in the ground truthing. The mid-composed clip uses a half-diminished 7th chord in place of an expected V chord, to disrupt an otherwise stable vi-iv-V progression, while the high-composed clip uses unresolved suspended 4ths and dominant 7ths in an outlined V chord. The jarring chord substitution with a less stable chord may have overpowered the tension building from unresolved dominant-tonic motion.

Harmonic complexity may also be more vulnerable to cross-cultural effects, as the most widespread and popular Western music is comparatively harmonically simple. While "dissonance" is often described as a clash between notes, the specific intervals or chords that are considered dissonant depend on features such as genre and historical context — in an extreme example, early organum choral music only considers octaves, fourths, and fifths as "consonant" (Rich, 1998).

7.3 DENSITY

T-7-H and T-8-H use short, uneven chords that move towards an unresolved dominant chord. In both sets, clips with longer, more sustained notes are ranked higher in perceived tension. While silence and uneven rhythms are often used to create tension in film scores, lowered density in pop music may be directly associated with lower tension. This relationship is not entirely consistent across ground truth designs, and this association may be weak.

A similar effect occurs in V-10 between the low- and medium-composed pieces. The V-10-L is a fast, aggressive, minor piece, while V-10-M is slower and uses more ambiguous and shifting harmonies and orchestrations.

7.4 GENRE

T-2 and T-6 are ground-truth ordrered in disagreement with the composed labels. In T-6, the ground truth ranking is based on an agreement between the 1-rank and Likert order, with the 2-rank order agreeing with the composed labels. T-6 appears to be confounded by strict adherence to triad-based harmonies in "Europop". This genre mostly uses simple harmonies and consistent rhythms, which limits the dissonances that can be used without violating genre conventions. The lowered expressive range may have produced too little difference between the component pieces to produce a consistent ranking.

T-2 appears confounded for the opposite reason — the genre of "classical" is broad enough that sub-genre differences may create additional confounds. T-2-L is stylistically similar to a Sousa march, emphasizing I-V tonal relationships with triads, while T-2-M is much more harmonically complex, and uses more inversions to create rising lines with some dissonances (e.g. a $V\,sus\frac{6}{4}$ chord), with a more contemporary, impressionistic style. These subgenre differences may confound the perception of tension.

Features discussed as previous possible confounds also often occur as part of genre conventions. For example, Disco and Latin music commonly makes heavy use of Major 7th and 9th chords. Bridge sections with instrumental breaks are common in rock/pop to build excitement towards a final chorus. Genre conventions most commonly affect tension, possibly due to the shifting definition of "dissonance" in different genres.

7.5 SET WITHOUT GROUND TRUTH ORDER

V-2, the first set to be composed, is the only set that receives a different ranking in all three of the listener evaluations, and is shown in entirety in Figure 7. V-2 is Disco-genre, and contains genre-specific complex Major harmonies as well as sequences.

8. POTENTIAL APPLICATIONS

In addition to evaluating the IsoVAT composition guide, the IsoVAT corpus could be used to assist in affectively tagging larger, curated datasets. As discussed in Section 2, parameterically controlled composition and ground-truthing are recommended when selecting stimuli for MER. When training ML systems from user feedback, the use of clear examples and known "gold standards" is recommended to ensure accuracy in participant

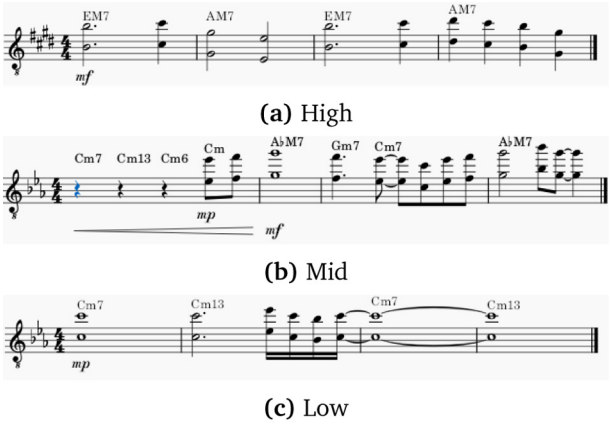


Figure 7: Reduced score for valence set 2.

responses — the ISOVat corpus provides a set of parameterically controlled and ground-truthed clips that express particular emotional relationships.

As we mention in Section 2.3, one potential application of the IsoVAT guide is to be used to provide some control over an input corpus for small-batch co-creative generative music. As the IsoVAT dataset is already composed based on the guide and has a ground-truth order, one immediate possible future project is to evaluate how much emotional expression is maintained when the IsoVAT dataset is used as an input for a generative system. Additionally, further examination of the IsoVAT guide could explore its use in parametric co-creative generative processes.

9. CONCLUSION AND FUTURE WORK

We present a composition guide in Section3 for composing affective music using valence, arousal, and tension, based on previous MER. We use this guide to compose a corpus of 90 musical clips that express emotion based on this guide, as described in Section4. We empirically produce a ground-truth ranking of the emotional perception of 29 out of 30 sets, with 19 sets ground-truthed in the order as labeled.

There are several areas for future work with this guide and corpus. While all identified musical features show correlative relationships in multiple affective dimensions, we only evaluate manipulation in one dimension at a time. Cross-dimensional validation would be useful to further evaluate the guide.

Our inclusion criteria for the IsoVAT guide was primarily determined by the source literature. This leads to ambiguity in the relationship between closely-related features such as "tonality" and "mode", or "melodic direction" and "melodic contour". We concur with Eerola and Vuoskoski's suggestion of future research into isolating and controlling these particular relationships (Eerola and Vuoskoski, 2012).

While we compose discrete clips that express comparative emotions, the IsoVAT guide could also be used to compose music that expresses relative changes in emotion over time. We reccommend further investigation in this area as well.

Some of the variance in our results may be a result of participants being unfamiliar with Western music. We do not screen participants for familiarity with Western music, and we do not collect demographic or location information. Amazon's MTurk Platform does not provide any additional participant information. Participants may therefore be from regions or locations where Western music is not dominant.

Without further information or replicating the study with collection of demographics, we cannot determine how large of an effect cross-cultural issues may cause. We acknowledge this as a limitation of this research, but overall believe that the results of this study are valid. Non-Western listeners are often able to identify expressed emotion in Western music (Balkwill and Thompson, 1999; Fritz et al., 2009). However, further investigation into cross-cultural effects is also recommended.

As mentioned in Section 1, we attempt to address two identified possible confounds in previous MER studies. Because of the lack of internal consistency in terminology of musical features or affect, one criticism of MER is that the research may lack practical applicability. We use previous research to create a composition guide, and then use the guide in a composition process, to investigate the practical applicability of the research. While previous research has investigated composing for a desired affect, we are not aware of any previous studies that follow the suggested parameterized, literature-informed approach to composition, as we do (Eerola and Vuoskoski, 2012; Warrenburg, 2020). In doing so, we also produce a symbolic corpus with ground-truth emotion labels.

Overall, we also show support for both the criticisms and suggestions of MER. While we attempted to directly address the ambiguity of musical feature definitions, there is still a degree of ambiguity in our guide. While we collect various emotion models into a single VAT description, other emotion models may be used for different applications. While we present one possible application of previous research in this area, there are many other possible approaches as well.

NOTES

1 MIDI and rendered wav files are available on GitHub (https://github.com/CalePlut/IsoVAT_Dataset).

2 MuseScore 3 is available at https://musescore.org.

3 The Arachno soundfont is available at http://www.arachnosoft.com/main/soundfont.php.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Cale Plut orcid.org/0000-0002-2937-1900
Simon Fraser University, 8888 University Dr., Burnaby BC, Canada

Philippe Pasquier orcid.org/0000-0001-8675-3561
Simon Fraser University, 8888 University Dr., Burnaby BC, Canada

Jeff Ens orcid.org/0000-0003-0673-4286
Simon Fraser University, 8888 University Dr., Burnaby BC, Canada

Renaud Tchemeube orcid.org/0000-0001-8337-489X
Simon Fraser University, 8888 University Dr., Burnaby BC, Canada

REFERENCES

Aljanaki, A., Wiering, F., and **Veltkamp, R.** (2014). Collecting annotations for induced musical emotion via online game with a purpose Emotify.

Balkwill, L.-L. and **Thompson, W. F.** (1999). A crosscultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception*, 17(1):43–64. DOI: https://doi.org/10.2307/40285811

Dinculescu, M., Engel, J., and **Roberts, A.** (2019). MidiMe: Personalizing a MusicVAE model with user data. In *Workshop on Machine Learning for Creativity and Design, NeurIPS*.

Eerola, T. and **Vuoskoski, J. K.** (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49. DOI: https://doi.org/10.1177/0305735610362821

Eerola, T. and **Vuoskoski, J. K.** (2012). A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal*, 30(3):307–340. DOI: https://doi.org/10.1525/mp.2012.30.3.307

Ens, J. and **Pasquier, P.** (2020). MMM: Exploring conditional multi-track music generation with the transformer. *arXiv preprint arXiv:2008.06048*.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A. D., and **Koelsch, S.** (2009). Universal recognition of three basic emotions in music. *Current Biology*, 19(7):573–576. DOI: https://doi.org/10.1016/j.cub.2009.02.058

Gabrielsson, A. and **Lindstrom, E.** (2012). The role of structure in the musical expression of emotions. In Juslin, P. N. and Sloboda, J. A., editors, *Handbook of Music and Emotion: Theory, Research, Applications*, pages 367–400. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199230143.003.0014

Gerhard, D. and **Hepting, D. H.** (2004). Cross-modal parametric composition. In *Proceedings of the International Computer Music Conference*.

Hadjeres, G. and **Crestel, L.** (2021). The piano inpainting application. *CoRR*, abs/2107.05944.

Hasegawa, R. (2020). Creating with constraints. In Donin, N., editor, *The Oxford Handbook of the Creative Process in Music*. Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780190636197.013.17

Hernandez-Olivan, C. and **Beltran, J. R.** (2021). Music composition with deep learning: A review. *CoRR*, abs/2108.12290.

Hoffmann, H., Scheck, A., Schuster, T., Walter, S., Limbrecht, K., Traue, H. C., and **Kessler, H.** (2012). Mapping discrete emotions into the dimensional space: An empirical approach. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3316–3320. IEEE. DOI: https://doi.org/10.1109/ICSMC.2012.6378303

Hung, H.-T., Ching, J., Doh, S., Kim, N., Nam, J., and **Yang, Y.-H.** (2021). Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. *arXiv preprint arXiv:2108.01374*.

Juslin, P. N. and **Sloboda, J. A.,** editors (2011). *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., and **Turnbull, D.** (2010). Music emotion recognition: A state of the art review. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 937–952.

Livingstone, S. R., Muhlberger, R., Brown, A. R., and **Thompson, W. F.** (2010). Changing musical emotion: A computational rule system for modifying score and performance. *Computer Music Journal*, 34(1):41–64. DOI: https://doi.org/10.1162/comj.2010.34.1.41

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292. DOI: https://doi.org/10.1007/BF02686918

Ortony, A., Clore, G. L., and **Collins, A.** (1990). *The Cognitive Structure of Emotions*. Cambridge University Press.

Panda, R., Malheiro, R., and **Paiva, R. P.** (2018). Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, 11(4):614–626. DOI: https://doi.org/10.1109/TAFFC.2018.2820691

Panda, R. E. S., Malheiro, R., Rocha, B., Oliveira, A. P., and **Paiva, R. P.** (2013). Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In *10th International Symposium on Computer Music Multidisciplinary Research (CMMR 2013)*, pages 570–582.

Pasquier, P., Eigenfeldt, A., Bown, O., and **Dubnov, S.** (2017). An introduction to musical metacreation. *Computers in Entertainment*, 14(2):1–14. DOI: https://doi.org/10.1145/2930672

Pati, A., Lerch, A., and **Hadjeres, G.** (2019). Learning to traverse latent spaces for musical score inpainting. *CoRR*, abs/1907.01164.

Paz, I., Nebot, A., Mugica, F., and **Romero, E.** (2018). Modeling perceptual categories of parametric musical systems. *Pattern Recognition Letters*, 105:217–225. DOI: https://doi.org/10.1016/j.patrec.2017.07.005

Reisenzein, R. (2000). Wundt's three-dimensional theory of emotion. In Balzer, W., Sneed, J. D., and Moulines, C. U., editors, *Structuralist Knowledge Representation*, pages 219–250. Brill.

Rich, A. (1998). Harmony before the common practice period. In *Encyclopedia Britannica*. Encyclopedia Britannica, Inc. https://www.britannica.com/art/harmony-music/Harmony-before-the-common-practice-period.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178. DOI: https://doi.org/10.1037/h0077714

Schimmack, U. and **Grob, A.** (2000). Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14(4):325–345. DOI: https://doi.org/10.1002/1099-0984(200007/08)14:4<325::AID-PER380>3.0.CO;2-I

Schubert, E. (1996). Continuous response to music using a two dimensional emotion space. In *Proceedings of the 4th International Conference on Music Perception and Cognition*, pages 263–268.

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, 51(3):154–165. DOI: https://doi.org/10.1080/00049539908255353

Thompson, W. F. and **Robitaille, B.** (1992). Can composers express emotions through music? *Empirical Studies of the Arts*, 10(1):79–89. DOI: https://doi.org/10.2190/NBNY-AKDK-GW58-MTEL

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., and **Bouchard, B.** (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition and Emotion*, 22(4):720–752. DOI: https://doi.org/10.1080/02699930701503567

Warrenburg, L. A. (2020). Choosing the right tune: A review of music stimuli used in emotion research. *Music Perception*, 37(3):240–258. DOI: https://doi.org/10.1525/mp.2020.37.3.240

Wu, H. and **Leung, S.-O.** (2017). Can Likert scales be treated as interval scales? A simulation study. *Journal of Social Service Research*, 43(4):527–532. DOI: https://doi.org/10.1080/01488376.2017.1329775

Wundt, W. M. and **Judd, C. H.** (1902). *Outlines of Psychology*. W. Engelmann.

Yang, Y.-H. and **Chen, H. H.** (2011). *Music Emotion Recognition*. CRC Press. DOI: https://doi.org/10.1201/b10731

Zhang, K., Zhang, H., Li, S., Yang, C., and **Sun, L.** (2018). The PMEmo Dataset for music emotion recognition. In *Proceedings of the 2018 ACM International Conference on Multimedia Retrieval*, pages 135–142, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3206025.3206037

TO CITE THIS ARTICLE:
Plut, C., Pasquier, P., Ens, J., and Tchemeube, R. (2022). The IsoVAT Corpus: Parameterization of Musical Features for Affective Composition. *Transactions of the International Society for Music Information Retrieval*, 5(1), 173–189. DOI: https://doi.org/10.5334/tismir.120

Submitted: 14 October 2021 **Accepted:** 16 June 2022 **Published:** 14 November 2022