FROM ALGORITHM TO COGNITION: MUSIC ACOUSTIC FEATURES AND THEIR PERCEPTUAL CORRELATES

FROM ALGORITHM TO COGNITION: MUSIC ACOUSTIC FEATURES AND THEIR PERCEPTUAL CORRELATES

By MAYA FLANNERY, B.A.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

Descriptive note

McMaster University MASTER OF SCIENCE (2022)
Hamilton, Ontario (PNB)
TITLE: From algorithm to cognition: Music acoustic features and their perceptual correlates
AUTHOR: Maya Flannery, B.A. (McMaster University)
SUPERVISOR: Dr. Matthew Woolhouse
NUMBER OF PAGES: x, 42

Lay abstract

This thesis aims to increase our understanding of music preference by helping us more accurately describe differences in music. Previous research has defined differences in music by genre categories, such as Classical or Dance music. However, genres are often subjective, and even arbitrary, in their descriptions. Instead of genre, this thesis proposes that a new method of categorization is used: Music Acoustic Features. These features are not subjectively defined, they can actually be measured within a piece of music. Furthermore, these features can be modified and tested in experiments to see how listeners respond. Such future experiments will provide us with a better understanding of what kind of music people like and why.

Abstract

Music preference research seeks to explain the relationship between music listeners and their music. An important task of such research is to describe differences between types of music. Musical *genres* are often chosen to address this task. But they are inadequate as they require subjective interpretation by both participants and researchers, making results difficult to decipher. This thesis provides foundational work to establish *Music acoustic features* (MAFs). MAFs are intended to provide a reliable method of music classification and description for experimental research. First, a labelled set of 4800 musical stimuli representing six MAFs of varying levels were systematically produced. A program tool, Essentia, was then used to identify low-level audio features within the musical stimuli that correlate with MAF manipulations. The Essentia features (EFs) that best represented MAFs were identified and used to predict MAFs in 44 real-world music clips. An online study also collected ratings from participants ($N = 43$) for each of the 44 real-world clips. The results of MAFs predicted by EFs and MAFs rated by participants were compared for consistency. The MAF Tempo correlated strongest between predicted and rated MAFs in real-world music, followed by Dynamic, Texture, Articulation, Register, and Timbre. Based on the outlined process, MAFs were to shown to be manipulable for experimental analysis, measurable within real-world stimuli, and readily perceivable by music listeners. These three criteria firmly establish MAFs as a reliable method of music classification and description for use in experimental research. Furthermore, the process outlined here can easily be adapted to validate other potential MAFs that may exist in music. MAFs will improve future research by increasing the robustness and clarity of conclusions, and thus provide greater insight into how and why people listen to certain types of music.

Acknowledgements

I would like to thank my thesis supervisor, Dr. Matthew Woolhouse, for his guidance and inspiration throughout this work. I also thank my committee members, Dr. Daniel Goldreich and Dr. Michael Schutz, for their valuable feedback. Thank you to my fellow lab members, Joanna Spyra and Konrad Swierczek, for their feedback and interesting conversation over the last two years. Thank you to my friends and family for their support and encouragement. Most of all, thank you to my partner, Will, for his endless patience, support, encouragement, and praise over these (now many) years I have spent as a mature student pursuing my passion. It means the world to me.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

i. bpm: beats per minute

ii. EF: Essentia Feature

iii. MAF: Music Acoustic Feature

iv. MIR: Music Information Retrieval

v. MIDI: Music Instrument Digital Interface

vi. npm: notes per measure

vii. PR: primary response

Declaration of Academic Achievement

I declare that I (MF) am the author of this thesis. My supervisor, Dr. Matthew Woolhouse, provided feedback on conceptualization, experiment design, and analysis throughout this work.

I. Introduction

The scientific study of musical preference has long been hindered by a fundamental problem—we do not know how to describe music objectively. This claim may sound controversial: we can, for example, describe music as being of a certain genre, like Classical music; or as being appropriate for a specific time and place, like dinner music. But each of these descriptors are of abstract concepts that are subjectively interpreted by individuals. They often cannot be reliably defined and measured. Furthermore, these descriptors are difficult to manipulate, which is a basic requirement for experiments designed to determine causal relationships (Chapter V: Jhangiani et al., 2019). This thesis aims to address these problems by establishing a reliable method of music categorization and description for use in experimental research. This method has produced what we call *Music Acoustic Features* (MAFs). These features each satisfy three criteria: they are objectively measurable within music, they can be intentionally manipulated in laboratory experiments, and they are readily perceivable by individual listeners. MAFs will better suit the requirements of experimental research and allow for more robust, consistent, and insightful conclusions about music preference.

Background

Psychologists have explored *music preference* relationships for several decades. Broadly, explanatory models of music preference are broken into two factors: the listener, who might be categorized and described in a number of ways, such as age, gender, personality, or mood; and the music itself, which requires a categorization and description of some particular 'type' of music (e.g., Rock, Classical, Dance, etc.). Theories suggest that variation within in-

dividuals is related to listeners' preference for certain types of music (Schäfer & Sedlmeier, 2010). For example, extroverts might have a greater preference for Dance music than introverted people. In this instance, *personality* categorizes and describes the listener, and *genre* categorizes and describes the music.

Early research investigated music–listener relationships by designing experiments related to the example described above. Litle & Zuckerman (1986) analyzed participants' sensation-seeking personality trait and their preference for "established categories of music based on divisions in the recording industry" (p. 1), or simply, musical genre. They found that individuals with a high sensation-seeking trait preferred Rock music.

Studies continued to investigate personality–genre relationships in greater detail, but problems with using musical genres to classify and describe music quickly became apparent. Aucouturier & Pachet (2003) provided a detailed critique of genre categorization methods, noting that genres are far too broad and inconsistently defined, leading to "ungrounded projections of fantasies" (p. 83). Such inconsistencies are easily found within typical genres: Baroque and Classical music are both distinct musical genres defined by the year in which they were composed (Baroque between approximately 1600–1750, and Classical between 1750–1800). In contrast, Love songs are defined by their lyrical content; and Electronic music is defined by its instrumentation and method of production. With just these few examples, we can see that some genres are primarily defined by time-period, some by lyrical content, and some by sound characteristics. Musical genre, while well-known and recognizable by both musicians and nonmusicians, requires outside knowledge and subjective interpretation to be fully understood. Such interpretations provide acceptable utility for use in recording industries but are highly problematic in scientific research. Music must be operationalized (i.e., defined in a way that can be practically measured) in research. More importantly, without being properly operationalized, musical stimuli cannot be systemat-

ically manipulated in experiments; and therefore, causal relationships cannot be reliably inferred.

New methods of analysis have attempted to refine how musical genre could be used in research. *Music dimensions* were proposed by Rentfrow & Gosling (2003). They recorded self-reported music preference ratings of individuals, then performed a principal component analysis to determine broad musical categories, which they originally called "Music dimensions". The number of such categories has varied though, and depends on individuals' age, culture, geographic location, and the timing of any particular study (Rentfrow et al., 2011, 2012; Zweigenhaft, 2008). Most recently, these music dimensions have been labelled as: Mellow, Unpretentious, Sophisticated, Intense, and Contemporary (which form the acronym: MUSIC). Each of these dimensions consists of a few genres and have some very general musical qualities associated with them. For example, Mellow music consisted of Easy listening, Soft rock, and Electronica genres, and was described as relaxed, slow, and romantic. Unfortunately, these dimensions do little to provide a more objective interpretation of music than genre—they rely on subjective judgments that can vary widely between individuals. Furthermore, these dimensions do not lend themselves to experimental manipulation. For example, if 'Mellowness' is a subjective measure, then how can it be reliably varied as stimuli in an experiment?

Other alternative methods have also been proposed. *Psychological attributes* of music were introduced by Greenberg et al. (2016). They investigated 38 psychological attributes of music that reduced to three principal components in their analysis. The perceived attributes were musical characteristics such as aggressive, reflective, sad, and intelligent, and the three components were depth, arousal, and valence. While these attributes provided more consistent descriptions of music, many are still highly subjective. For instance, human judges must determine the "intelligence" of a piece of music, a process which is likely influenced by culture, age, and musical train-

ing. Subjective biases influence how psychological attributes describe music, which in turn, also affects how reliably these attributes could be manipulated in experiments.

One alternative stands out among these methods. A study by Eerola et al. (2013) used *musical cues* to investigate perceived emotion in music. They described musical cues as "properties inherent in the music itself" (p. 1), which consisted of Mode, Tempo, Dynamics, Articulation, Timbre, Register, and Musical structure. Moreover, they were able to manipulate these cues in a study to see how they affected the perception of Scary, Happy, Sad, and Peaceful music. Clear conclusions could be made from their investigation. For example, slow tempi were associated with the perception of Sad music, and fast tempi were associated with the perception of Happy music.

Later work, by Grimaud & Eerola (2022), described musical cues as being structural and expressive cues used by composers and performers to encode emotion in music. Their study used an 'analysis-by-synthesis' method where participants varied the level of musical cues in real-time to express different emotions. Their findings also showed clear relationships between musical cues and emotion. For example, participants used fast tempo and bright timbre to express Joy and Surprise in musical excerpts. The approach in both of these studies showed that certain musical descriptors can be defined and reliably manipulated. Viewing the problem from a music production standpoint, that is, how composers and performers vary music to express emotion, provides a promising solution for music preference research.

There is growing evidence that musical cues could be used in music preference research. A study on listening behaviour supports this idea. Barone et al. (2017) investigated *acoustic features* of listeners' preferred and non-preferred genres. Their analysis consisted of data from a large music-download database (similar to Spotify). Users' preferred genres were inferred from their playlists. Acoustic features were computationally extracted from tracks in the database and consisted of features like Loudness, Tempo, and

Valence. Their results showed that tracks outside of a user's preferred genre were also preferred if they had similar acoustic features. This suggests that listeners might prefer specific features within music rather than broad music genre categories.

Another study, by Flannery & Woolhouse (2021), attempted to directly explore this possibility. Building on both the musical cues used in emotion research, and the acoustic features found in behavioural listening data, the idea that MAFs related to music preference was tested. It was found that, indeed, certain features in music were preferred over others (e.g., fast tempo over slow tempo), and personality factors interacted with these preferences (e.g., participants high in Extraversion preferred loud dynamic over soft dynamic, those low in Extraversion did not). While the results from this study showed that MAFs are a viable method in music preference research, further work is needed to formally establish how MAFs are defined and how they should be used.

The above approaches have started to satisfy some of the limitations of genre approaches: musical cues can be reliably manipulated and acoustic features can be objectively measured. However, some limitations remain. First, musical cues and acoustic features are only loosely connected. While Tempo appears in both methods, it is measurable and manipulable, other musical cues like Timbre do not have a clear objective measure. Likewise, many acoustic features, like Danceability, are not easily manipulable. Ideal features must be selected that have the advantage of being both objectively measurable and manipulable. Second, ideal features must be confirmed to be readily perceivable by participants to ensure the reliability of their effects. Intended manipulations, objective measures, and perceived effects should all correspond. The work that follows in this thesis addresses these limitations.

Present Study

The present study was designed to formally define and determine a series of features within music that can be used for experimental purposes. Based on the limitations of musical genre classification and the potential benefits of musical cue and acoustic features outlined above, we had determined that such ideal features must possess the following qualities:

1. The feature must be *manipulable*. It must be possible to systematically produce music that varies along a given measure.

2. The feature must be *measurable*. It must be possible to determine the level of a feature by analysis of a digital audio input.

3. The feature must be *readily perceivable*. Manipulated levels and objectively measured values must correspond with perceived differences in musical stimuli.

These criteria will ensure that features are properly operationalized. Each will be defined in how it is objectively measured, how it is to be manipulated in experimental procedures, and how it is interpreted by listeners in musical and psychological contexts. Since these features are rooted in both the acoustic properties of an audio signal and their musical interpretation, and given the terminology established by Flannery & Woolhouse (2021), these features are referred to as 'Music Acoustic Features' (MAFs).

The first requirement, that a MAF is manipulable, was used as the starting point of the investigation. We chose candidate features from the existing studies that successfully used musical cues as manipulable features (i.e., Eerola et al., 2013; Flannery & Woolhouse, 2021). The potential features were: Articulation, Dynamics, Mode, Register, Tempo, Timbre, and Musical structure. To facilitate the experiment's design, we chose to use features that could be manipulated in a continuous manner. For example, a stimulus's

Tempo could range from 60 to 120 beats per minute (bpm). Musical Mode is usually either *major* or *minor*, but is often difficult to define (sometimes the intended mode is not what is perceived by a listener). For this reason, Mode was removed as a potential feature. Musical structure was removed as well since it broadly captures many potential sub-features, such as phrasing and repetition, and would drastically increase the scope of our analysis. We identified one additional feature that had not been used previously, musical Texture. The resulting features of interest, and a basic description of how they sound, are a follows:

- Articulation: refers to how quickly a note is played. *Staccato* notes are played quickly, their duration is very small, so the notes sound short and disconnected from each other. In contrast, *Legato* notes are held until the next note sounds, so notes sound like they are smoothly connected[1].

- Dynamic: refers to how soft or loud the music is played. The intensity, or volume, of the music is affected, *Soft* dynamics contain notes that are low volume, and *Loud* dynamics contain notes that are high volume.

- Register: refers to the pitch, or frequency, of the notes played (either an average pitch, or pitch range). *Low* register contains low frequency notes, *High* register contains high frequency notes.

- Tempo: refers to the speed of the music and is measured in beats per minute (bpm). *Slow* tempi have few bpm, *Fast* tempi have many bpm.

- Texture: refers to the number of notes/instruments that are played. *Sparse* texture consists of few notes/instruments, *Dense* texture consists of many notes/instruments at once.

[1] A detailed description and musical examples of articulation can be found here: https://en.wikipedia.org/wiki/Articulation_(music).

- Timbre: refers to the type of sound an instrument makes. It can be described in many ways, but for the purpose of this experiment, timbre is described as the music's brightness. Some instruments sound *Dark* while others sound *Bright*.

The next requirement was MAF measurement. The fields of audio engineering and computer science have derived many features from the fundamental properties of sound waves. Furthermore, researchers have developed programming packages, libraries, and toolboxes to analyze digital audio files and extract these features (e.g., The Essentia library, Bogdanov, Wack, Gómez, et al., 2013; and MIR Toolbox, Lartillot & Toiviainen, 2007). In general, features are divided into *low-level* features, like the frequency of a sound; *mid-level* features, like the tempo of a musical excerpt; and *high-level* features, like the sadness of an excerpt. It is important to note though, that high-level features require advanced machine learning methods that potentially reintroduce human biases into their decisions, and may not be interpreted as completely objective (Alonso-Jiménez et al., 2020). Low- and mid-level features are reliably measured with traditional algorithms, however, and thus remain objective. We chose to explore the capabilities of the open source library Essentia. Essentia provides a collection of algorithms that extract low to high level features from digital audio files.

A drawback of using these low-level features is they are often difficult to interpret in a psychological context. To further complicate this problem, the low-level extraction algorithm from Essentia produces close to 450 features from a single audio stimulus. It is difficult to know which low- and mid-level features might correspond to the MAFs of interest. Thus, we developed a procedure to find Essentia features (EFs) that correlate with MAF manipulation. This procedure involved: generating a training dataset of stimuli with varying levels of each MAF, extracting low-level features using the Essentia library, and analyzing correlations between low-level features and MAFs (the procedure is detailed in the Methods section). The result of this procedure

produced a predictive model for each MAF that could be used to analyze other stimuli.

The final requirement was to show that each MAF is readily perceivable. A listening experiment was designed where participants listened to short musical clips and provided their subjective ratings of each MAF. To increase the generalizability of the analysis, the stimuli used in this experiment were musical clips taken from real-world recorded music.

Following on from the MAF criteria outlined above, this thesis considers the following hypothesis. MAFs are structurally similar to each other by virtue of the fact that: 1) they are manipulable; 2) they are objectively measurable; and 3) they are perceived in music by listeners (including non-musicians). In order to test this tripartite hypothesis, a four stage process was adopted:

1. Low-level EFs were identified that reliably covaried with manipulated levels of each MAF;

2. For each identified EF, the level of the corresponding MAF was predicted from a given digital audio stimulus;

3. In a listening task, human participants were required to rate the level of each MAF;

4. Using an identical stimulus set, MAFs predicted by EF models (from Stage 2) and MAFs rated by participants (from Stage 3) were correlated with one another.

II. Methods

The methods for this study are divided into four components: (1) Generate training stimuli, (2) Essentia feature selection, (3) Real-world music analysis, and (4) Listening task experiment.

Generate training stimuli

A collection of stimuli were generated containing each possible combination of MAF–level manipulation. First a base stimulus was written in the open source program, MuseScore (MuseScore Team, 2022). The base stimulus (see Figure 1) is written in 4/4 time lasting seven measures. It contains four separate voices, each with a monophonic melody (only one note per voice was played at a time, the voices were played simultaneously). The notes are written in the E mixolydian mode (closely related to the Major mode) and centred around C4 (262Hz, also known as 'middle C') with a range from E2 (41Hz) in the Bass voice to F#5 (740Hz) in the Soprano voice. A program script performed manipulations to the instrumentation, tempo, and dynamics.

Texture. The first manipulation was Texture because it had to be produced manually. Since Texture is related to the number of notes and instruments sounding in a section of music, these elements were manipulated from the base stimulus. We created levels of sparse texture by removing notes and/or entire voices from the stimulus. Dense texture was created by adding notes. Figure 2 shows an example of resulting (A) sparse and (B) dense stimuli. Four additional levels of Texture were created from the base stimulus. To quantify the Texture of a piece, the total number of notes in the stimulus were divided by the number of measures, giving an average number

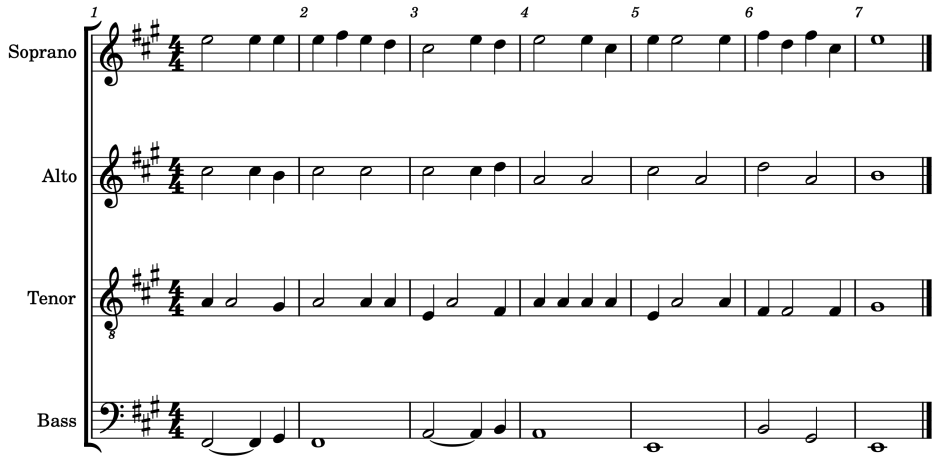

Figure 1: The base stimulus was written in the E mixolydian mode [indicated by 3 sharps (#)] in 4/4 time and lasts 7 measures. There are four voices: Soprano, Alto, Tenor, and Bass, which each indicate the general range of notes (pitches) for the voice (Soprano contain high notes, Bass contain low notes). While each voice only plays a single note at a time, all four play together at the same time. Tempo, dynamic, articulation, or instrumentation are not marked on the score as they are directly specified in subsequent steps.

of notes per measure (npm). The five levels contained: 3.86npm (the sparsest texture), 5.43npm, 7.14npm, 9.43npm, and 17.71npm (the densest texture).

Following the Texture manipulation, the five levels of texture were exported from MuseScore as Music Instrument Digital Interface (MIDI) files. MIDI files hold the basic information of the stimuli, such as note onsets, pitches, and tempo, and can be used as input to synthesized instruments to create digital audio (either as live sound played from a speaker, or saved in a file format). A python package, called Mido (MIDI Objects for Python, Bjørndalen, 2022), can modify MIDI content and was used to perform the remaining manipulations.

Articulation. Levels of articulation were created by modifying each note's start and stop commands in the MIDI files (as exported from the Texture manipulation). By default, the MuseScore MIDI output is Legato, so each note sounds until the next note begins. The note duration is 100% of the distance between notes. Differing levels of Articulation were created by multiplying the stop commands by a factor between 0.25 and 1 (which created shorter notes). Four levels were produced: 0.25 (Staccato), 0.5, 0.75, and 1.0 (Legato).

Dynamic. Levels of Dynamic were created by modifying the velocity commands in the MIDI files. The velocity command informs a synthesized instrument of how 'hard' an instrument is played (e.g., a low velocity piano note will be soft and quiet, a high velocity piano note will be loud and harsh). Velocity ranges between 0 (no sound) and 127 (as loud as possible). It was explicitly set for each level of dynamic: 50 (soft), 80 (medium), and 120 (loud).

Register. Levels of register were created by modifying the pitch commands in the MIDI files. In MIDI notation, each pitch is given a number (e.g., E2 = 40, C4 = 60, F#5 = 78). Five levels of Register were created by adding or subtracting from the values of each pitch: -12 (one octave below the base stimulus), -7, +0, +7, and +12 (one octave above the base stimulus).

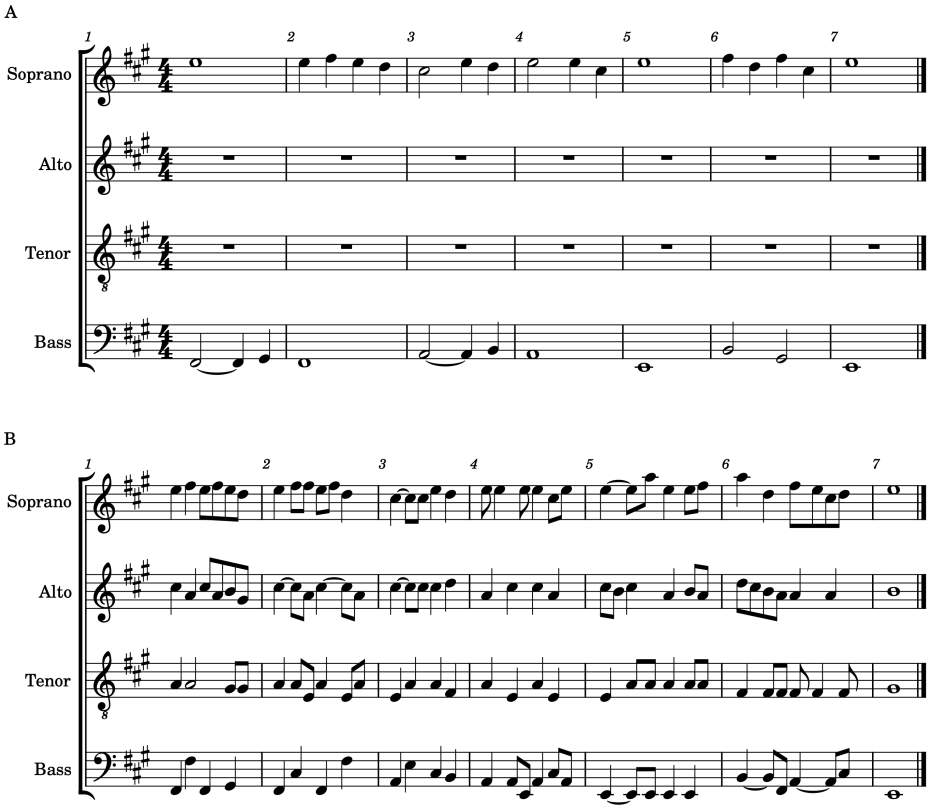

Figure 2: Two Texture manipulations from the base stimulus are shown. Part **A** is a sparse manipulation. The two inner voices and repeated notes were removed, leaving this level with 3.86 notes per measure. Part **B** is a dense manipulation. Several notes were added to all four voices, and produced a level with 17.71 notes per measure.

Tempo. Levels of Tempo were created by modifying the tempo command in the MIDI files. The tempo command indicates how quickly notes are played (expressed as microseconds per quarter note) in a section of music. Four levels of Tempo were explicitly set (bpm values were automatically converted to the MIDI time clock format): 70bpm (slow), 90bpm, 110bpm, and 130bpm (fast).

Timbre. The Timbre manipulations were performed last. After the MIDI commands were modified by the previous four MAF manipulations, they were rendered to audio files (in .flac format). The rendering process used the FluidSynth open source synthesizer (see https://github.com/FluidSynth/fluidsynth/wiki) with the General MIDI SoundFont[2]. The soundfont includes approximately 193 sampled instruments. Four instruments were selected based on perceived instrument brightness reported by McAdams (2019). The four levels of Timbre were: Trombone (darkest), Piano, Guitar, and Harpsichord (brightest).

The described procedure produced 4800 labelled stimuli (six factors with 3–5 levels each: 5 x 4 x 3 x 5 x 4 x 4) in audio (and MIDI) format that were used to identify EFs.

Essentia feature selection

Essentia is an open source music information retrieval (MIR) tool (Bogdanov, Wack, Gómez Gutiérrez, et al., 2013). It contains a collection of algorithms for analyzing digital audio content. These algorithms perform basic input and output of audio files, signal processing, filtering, and computation of low-, mid-, and high-level audio descriptors. Although some algorithms are capable of high-level descriptions of audio, such as genre and instrument iden-

[2]This synthesizer does not produce instrument sounds with the highest accuracy. It was a convenient choice, though, as it could be run inside of a Python script, which was necessary due to the quantity of stimuli.

tification, the focus of this study was on low-level features, such as descriptors that provide numeric values summarizing rhythmic, tonal, and spectral qualities of audio. Essentia provides a general function, called 'MusicExtractor', that can retrieve all low- and mid-level features from an audio file.

Using the provided MusicExtractor, each of the 4800 stimuli produced by the generation process were analyzed. This resulted in a labelled data set containing the unique MAF-levels of each stimulus, and the complete set of extracted EFs.

The correlation between each MAF and each EF was then analyzed with the linear model function in R (R Core Team, 2018). The R^2 value was then used to select correlated pairs for further analysis; if the R^2 value was above 0.25, it was plotted and visually inspected to confirm there was a linear trend between the values of the EF and the MAF.

The parameters of the optimal linear model for each MAF were then recorded for use in the final two steps of the study.

Real-world music analysis

Clips of real-world music recordings were selected to be representative of the possible variation within MAFs. For example, there should be a range of Timbre from dark to bright, a range of Tempo from fast to slow, and so on. Since these clips were intended to be analyzed with both the EF models and in the listening task, they were selected to be relatively short and limited in quantity. Forty-four clips, ranging from 7 to 25 seconds in duration, were selected in total (listed in Appendix B: Tables 5 and 6). The stimuli were then analyzed by the MusicExtractor algorithm described in the previous section. Finally, the EF models were then used on the results of the MusicExtractor to predict the MAF-levels.

Listening task experiment

Participants

The listening portion of the study was conducted on the online platform, Pavlovia. Participants were recruited through McMaster University's Sona system and were reimbursed with course credit (some additional participants were recruited by email and word of mouth and were not reimbursed). Informed consent was obtained before participants started the experiment; and a debrief was provided, consisting of contact information and details about the study, when participants were finished. Ethics were approved by the McMaster Research Ethics Board (MREB: #2524).

Apparatus

The experiment was presented entirely online in a web browser. The JavaScript framework, JsPsych (De Leeuw, 2015), was used to code each portion of the experiment: 1) welcome/consent form, 2) demographic questionnaire, 3) training session, 4) listening trials, and 5) follow-up questionnaire and debrief. The questionnaires and stimulus response options are listed in Appendix A: Tables 2, 3, and 4. Participants were free to move through each step of the experiment at their own pace. Instructions advised that participants completed the listening tasks in a quiet area with either speakers or headphones. Participants could freely adjust the volume of stimuli and replay a stimulus unlimited times.

Procedure

Participants first answered a fifteen question demographic questionnaire (see Appendix A: Table 2). A brief training session then introduced participants

to each of the six MAFs. Each MAF was presented with a short description and two audio examples of contrasting levels of the feature. For example, the Register MAF included the text "This refers to your perception of the music's overall pitch height (i.e., frequency). Your responses can range from: very low to very high." Audio examples of very low register and very high register were presented below the text. Participants could refer back to the MAF terminology page and replay the examples, if needed, while completing the rest of the experiment. The listening task followed the training section and consisted of 44 trials. Each trial randomly presented one of the 44 real-world stimuli described in the previous section along with a response questionnaire (see Appendix A: Table 3). When the listening task was completed, a final questionnaire (see Appendix A: Table 4) was presented to gather information about the participant's understanding of MAFs and their confidence in responses.

III. Results

Essentia feature selection

The MusicExtractor algorithm yielded 451 features for each of the 4800 generated stimuli. Analysis of linear models for each MAF–EF combination returned 315 features with R^2 greater than 0.25. Twenty-five features correlated with Articulation, 7 with Dynamic, 33 with Register, 10 with Tempo, 27 with Texture, and 213 with Timbre.

Correlations for each MAF were plotted and visually inspected for linear trends. An example of Tempo is shown in Figure 3. Even though several EFs correlate with Tempo, only two demonstrated a clear relationship between the MAF and the EF. In this example, the `rhythm` EFs `onset rate` (EF_{OR}[3], $R^2 = 0.27$, $p < 0.001$) and `bpm histogram first peak bpm` ($R^2 = 0.53$, $p < 0.001$) were selected for further analysis. EF_{OR} best showed a distinct relationship with Tempo, low levels of EF_{OR} corresponded with slow Tempo and high levels of EF_{OR} with fast Tempo.

The five remaining MAFs were analyzed with the following results: Articulation negatively correlated with the `low-level` feature `pitch salience dvar` (EF_{PSD}, $R^2 = 0.43$, $p < 0.001$), high levels of EF_{PSD} corresponded to staccato (spiky) Articulation and low levels of EF_{PSD} to legato Articulation. Dynamic positively correlated with the `low-level` EF `loudness ebu128 integrated` (EF_{LEI}, $R^2 = 0.73$, $p < 0.001$), low levels of EF_{LEI} corresponded to soft Dynamic and high levels of EF_{LEI} to loud Dynamic. Texture negatively correlated with the `low-level` EF `spectral complexity dmean2` (EF_{SCD}, $R^2 = 0.33$, $p < 0.001$), high levels of EF_{SCD} corresponded to dense Texture and low level of EF_{SCD} to bright Texture. Timbre posi-

[3]The feature variable names relating Essentia output will be abbreviated throughout the Results and Discussion sections

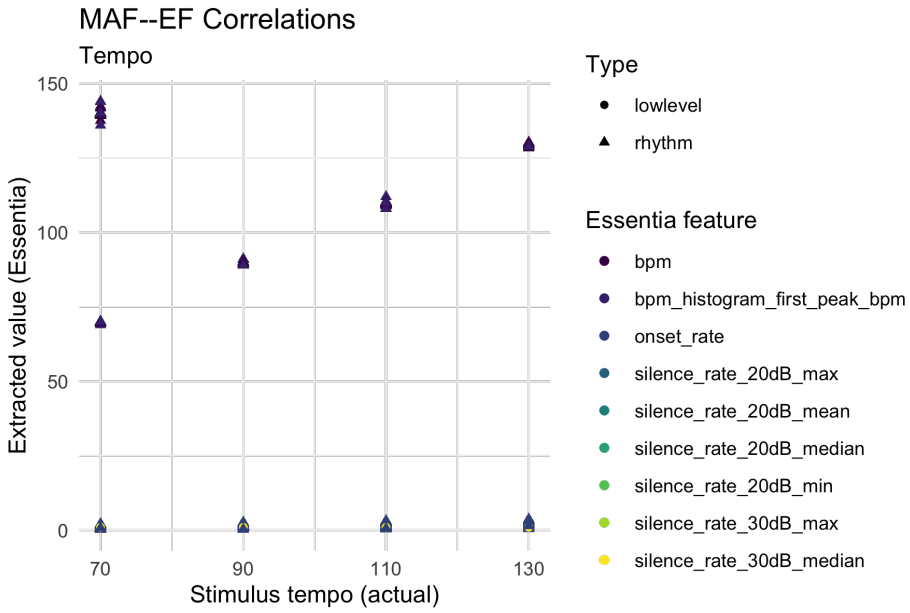

Figure 3: Relationships are shown between the manipulated MAF Tempo (X-axis) and the value of each EF (Y-axis). Colour indicates individual EFs.

tively correlated with the `low-level` EF `spectral spread median` (EF_{SSM}, $R^2 = 0.80$, $p < 0.001$), low levels of EF_{SSM} corresponded to dark Timbre (e.g., Trombone) and high levels of EF_{SSM} to bright Timbre (e.g., Harpsichord). Register required expanded analysis as the initial features performed poorly in subsequent analyses. It was found that the `low-level` EF `spectral centroid mean` (EF_{SCM}, $R^2 = 0.14$, $p < 0.001$), even though it had a relatively weak positive correlation, produced reliable predictions when analyzing the real-world stimuli. Low levels of EF_{SCM} corresponded to low Register, and high levels of EF_{SCM} to high Register. The resulting model coefficients are summarized in Table 1 and were used to predict features from the real-world stimuli in the next step.

Table 1: EFs that best correlated with MAFs are listed below. The parameter β_0 is the model intercept and β_1 is the EF coefficient. The listed models were used to predict MAF levels from the real-world stimuli set.

MAF	Essentia	B_0	B_1
Tempo	`rhythm_onset_rate`	72.776	17.783
Timbre	`log(lowlevel_spectral_spread_median)`	-27.080	2.014
Articulation	`lowlevel_pitch_salience_dvar`	1.044	-158.305
Register	`lowlevel_spectral_centroid_mean`	-6.955	0.006
Texture	`lowlevel_spectral_complexity_dmean2`	4.927	-0.829
Dynamic	`lowlevel_loudness_ebu128_integrated`	197.736	3.309

Real-world music analysis

Each of the 44 external audio stimuli were analyzed by the Essentia MusicExtractor used in the Section above (Essentia feature selection). The selected Essentia models (listed in Table 1) were then used to predict levels of each MAF according to Equation 1. In this notation, i refers to each of the six MAFs, and j refers to each real-world music clip. The resulting MAFs were

then normalized for each feature.

$$MAF_{ij} = \beta_{0i} + \beta_{1i} \times EF_{ij} \quad (1)$$

Listening task analysis

A total of 43 participants (ages ranged from 18 – 39 years, M = 21.95, SD = 5.13) completed the entire experiment (there are no missing data) with a median completion time of 29.06 minutes. There were no limitations on maximum experiment length and participants' web browsers were not restricted in any way, so some completion times were very high (1506 minutes). Other completion times were very low (12.39 minutes), near the minimum amount of time to listen to all stimuli (10.37 minutes). Four outliers were identified and considered based on completion time. Additionally, other outliers were considered based on reported confidence ratings. Participants were asked to rate their understanding of each MAF and their confidence in MAF ratings after completing the listening task (see Appendix A: Table 4). Two participants (non-musicians) reported zero's exclusively in both understanding and confidence, and that the training section was not helpful. The complete analyses were repeated on: the full data, completion time outliers (4 participants) excluded, poor confidence outliers (2 participants) excluded, and all outliers (5 participants) excluded. In all cases, the final results remained stable (R^2 did not change more than 0.07 in any condition). The following analyses are reported on the full data, no participants were excluded.

Musicians were classified by participants' self-reported years of musical training[4] which ranged from 0 to 12 years (M = 2.02). If participants had more than two years of training they were classified as a *musician* (n = 15), training under two years were classified as *non-musician* (n = 28). Aver-

[4]Participants were asked: "Have you ever had any formal music training? ... if yes, enter how many years of training."

age response patterns were identical between musicians and non-musicians for 64% of ratings. The following analyses are reported on combined data. However, differences between musicians and non-musicians are shown in Figures 4, 5, and 6.

Participants' responses were analyzed by stimulus and MAF. For each stimulus, the frequency of each selected MAF-level was counted, this accommodated multiple responses by a single participant for a single stimulus–MAF combination (e.g., a participant could choose both 'very low' and 'low' Register options for a stimulus). For each stimulus–MAF combination, the sum of participant responses were calculated and the levels of each MAF with the highest values were selected as the primary response for each stimulus. For example, the Tempo response pattern of a single stimulus is shown in 4. No participants chose 'Very slow' or 'Slow' for this stimulus, and the most frequent response for both musicians and non-musicians was for the 'Fast' level of Tempo. Thus, the primary response (PR) rating was fast Tempo.

Comparison: EF predictions – MAF ratings

The final step of the analysis compared the primary response (PR) ratings by participants to the predicted MAF levels by the EF models. A linear regression of MAF_{EF} predicted by MAF_{PR} was used to examine each MAF relationship. The resulting R^2 values were used to determine how closely the objective EF measure and the subjective PR measure aligned in predicting levels of MAFs.

Articulation. The relationship between MAF_{PR} and MAF_{EF} (as predicted by the `low_level` EF `pitch salience dvar`) is shown in Figure 5. This plot shows that when participants rated stimuli on the 'Spiky' end of the response scale (X-axis), the stimuli were predicted to have low Articulation values (Y-axis). The reverse was true when participants rated stimuli as 'Smooth', higher Articulation values were predicted. Articulation predicted

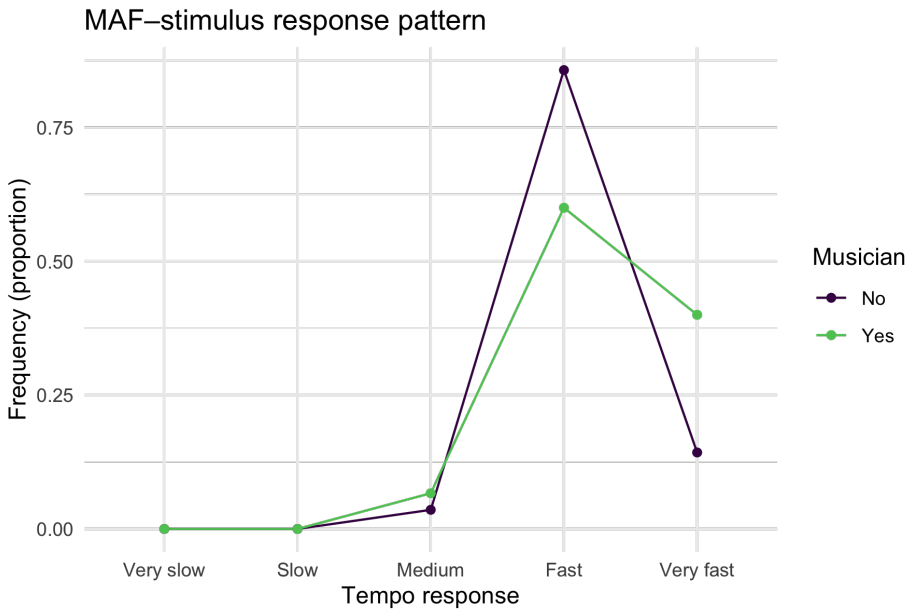

Figure 4: Tempo responses for a single stimulus. The five possible responses are shown on the X-axis, and the proportion of responses are shown along the Y-axis. Proportions are shown by musician and nonmusician status in colour. For this particular stimulus, no 'Slow' or 'Very slow' responses were recorded. The maximum response was 'Fast' for both musicians and nonmusicians. The maximum response for each MAF–stimulus were used in subsequent analyses.

by participant rating was significant, $F(4,\ 88) = 4.719$, $p = 0.002$, with adjusted $R^2 = 0.139$.

Dynamic. The MAF–EF (the `low_level` EF predicted by `loudness ebu128 integrated`) relationship for Dynamic is shown in Figure 5. Low dynamic ratings corresponded to low dynamic predictions, and high ratings with high predictions. Dynamic predicted by participant rating was significant, $F(4,\ 92) = 20.903$, $p < 0.001$, with adjusted $R^2 = 0.453$.

Register. The MAF–EF (the `low_level` EF predicted by `spectral centroid mean`) relationship for Register is shown in Figure 5. Low register ratings corresponded to low register predictions, and high ratings with high predictions. Register predicted by participant rating was significant, $F(4,\ 88) = 4.595$, $p = 0.002$, with adjusted $R^2 = 0.135$.

Tempo. The MAF–EF (the `rhythm` EF predicted by `onset rate`) relationship for Tempo is shown in Figure 6. Slow tempo ratings corresponded with low tempo predictions, and fast tempo with high predictions. Tempo predicted by participant rating was significant, $F(4,\ 89) = 35.846$, $p < 0.001$, with adjusted $R^2 = 0.60$.

Texture. The MAF–EF (the `low_level` EF predicted by `spectral complexity dmean2`) relationship for Texture is shown in Figure 6. Sparse texture ratings corresponded to low texture predictions, and dense ratings with high predictions. Texture predicted by participant rating was significant, $F(4,\ 86) = 13.712$, $p < 0.001$, with adjusted $R^2 = 0.361$.

Timbre. The MAF–EF (the `low_level` EF predicted by `spectral spread median`) relationship for Timbre is shown in Figure 6. Dark timbre ratings corresponded with low timbre predictions, and bright timbre with high predictions. Timbre predicted by participant rating was significant, $F(3,\ 93) = 5.016$, $p = 0.003$, with adjusted $R^2 = 0.112$.

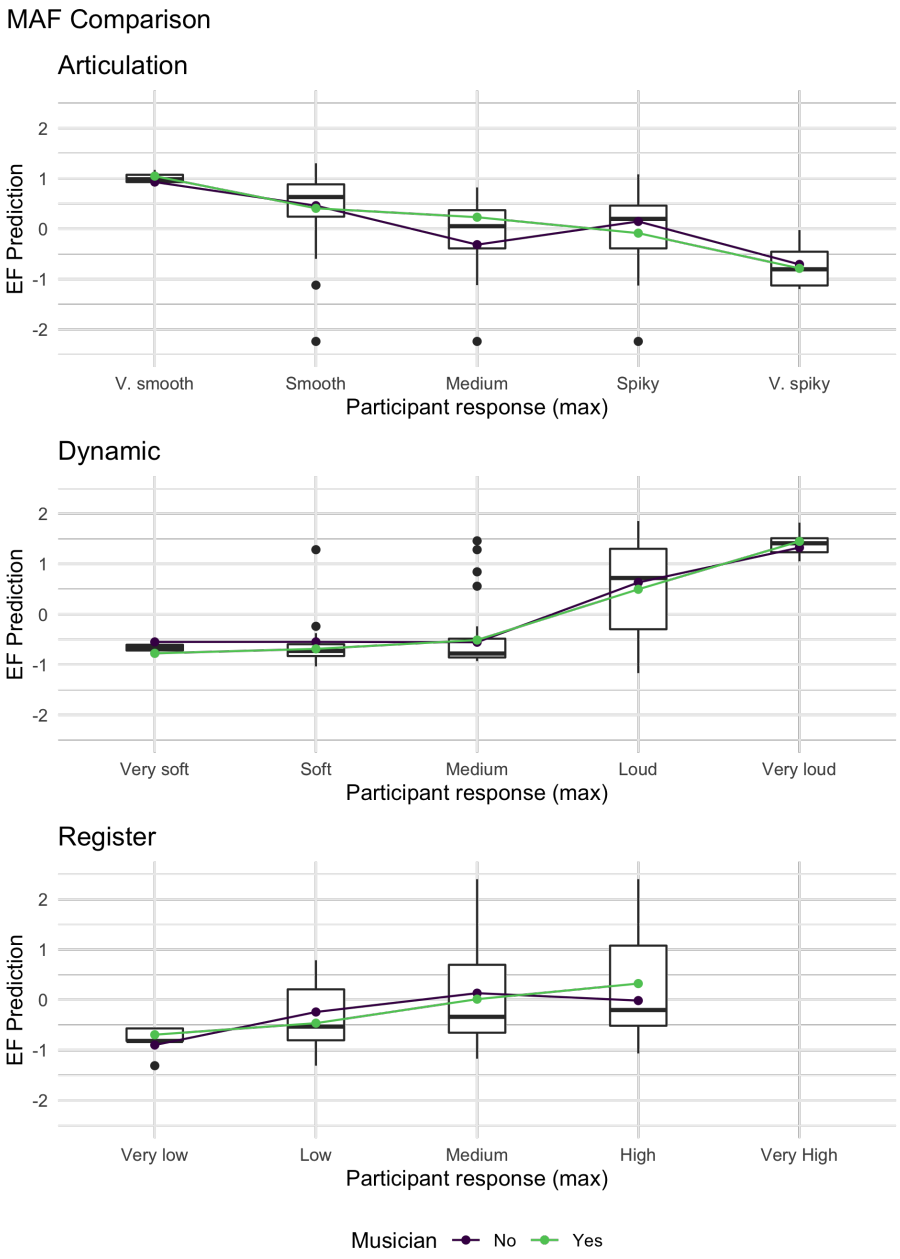

Figure 5: MAFs as predicted by EF models (Y-axis) compared to the most commonly selected participant responses (X-axis) for Articulation, Dynamic, and Register. The mean EF value is shown for musicians and nonmusicians in colour.

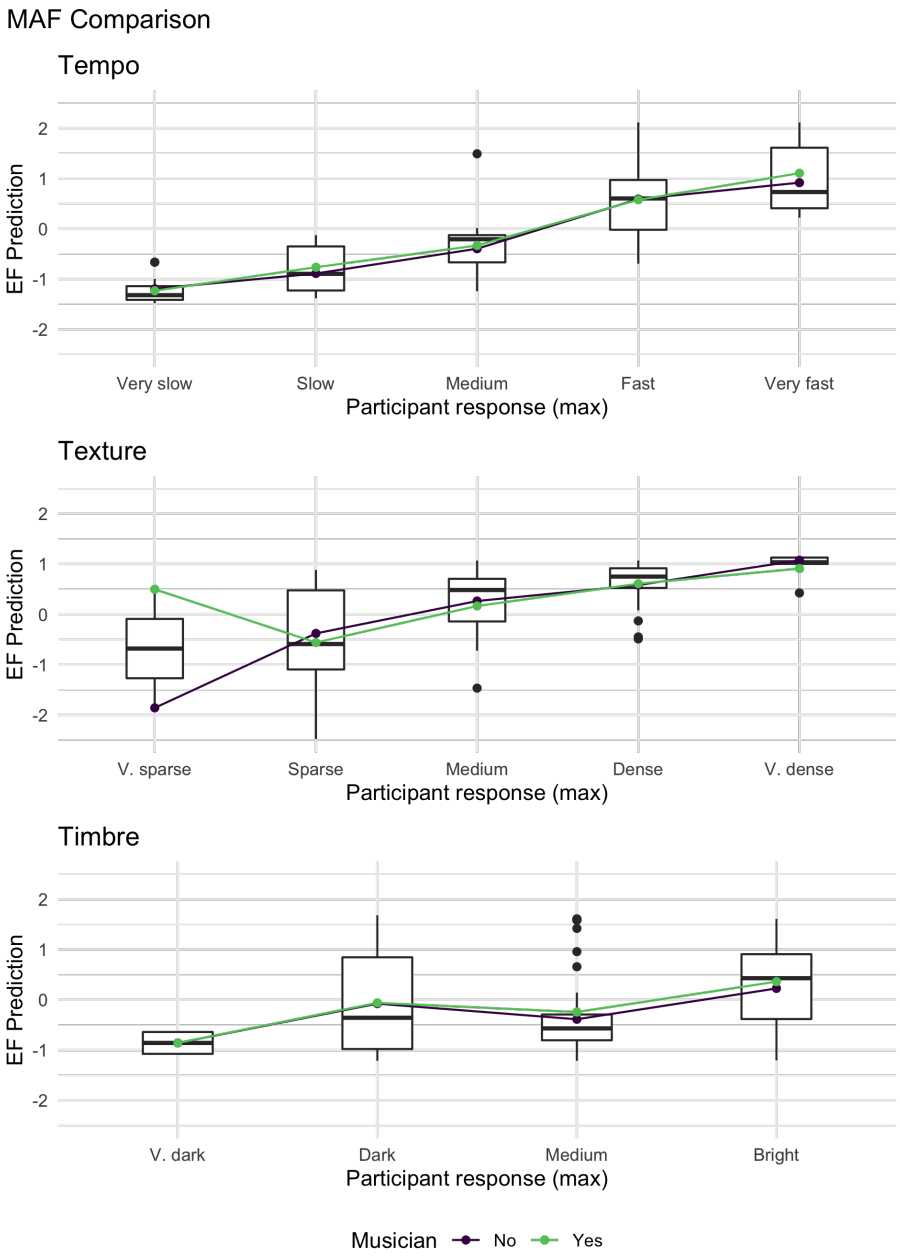

Figure 6: MAFs as predicted by EF models (Y-axis) compared to the most commonly selected participant responses (X-axis) for Tempo, Texture, and Timbre. The mean EF value is shown for musicians and nonmusicians in colour.

IV. Discussion

The purpose of this study was to establish MAFs for use in the experimental study of music preference. The criteria for a MAF was to be objectively measurable, manipulable, and readily perceivable by listeners. Six potential MAFs were tested through this multi-step study. The first step systematically generated a labelled stimulus set. Each factor–level combination was produced resulting in 4800 stimuli. The second step used the MIR tool set, Essentia, to extract low-level features from the labelled stimulus set. Correlations of these low-level features to each labelled stimulus were then analyzed to determine which best-predicted the level of each MAF. In the third step, EFs that best predicted MAFs in the example stimuli were used to predict MAFs in forty-four real-world music clips. In the last step, participants in an online study rated MAFs for the same forty-four stimuli. The results of EF predictions and participant ratings on the real-world stimuli were then compared for consistency.

The first step of the study showed that MAFs are manipulable. Articulation was varied from 'smooth' to 'spiky' by modifying MIDI start and stop messages. This manipulation is analogous to how composers indicate articulation in written music and how performers vary articulation while playing instruments. Staccato is the musical term from short spiky notes. They are played by quickly starting and stopping a note (e.g., picking a guitar string and dampening the string so it does not ring until the next note is played). Legato is the musical term for long smooth notes and are played by allowing notes to sound until the next note is played (e.g., a guitar string rings until the next note is sounded). Dynamic was varied from soft to loud by modifying MIDI velocity messages. Composers mark musical notes, phrases, and sections with dynamic markings, which are then read and performed appropriately. *Piano* means soft, notes are played gently at low volume (e.g., gently

pressing a piano key produces a soft, low volume note). *Forte* means loud, and are played by sharply producing notes at a loud volume (e.g., forcefully hitting a piano key produces a sharp, loud volume note). Register was varied by adding or subtracting from a base MIDI note value. MIDI note values translate directly to musical note values. Tempo was varied by modifying the MIDI tempo value. Composers usually describe the speed (e.g., *Lento*, meaning slowly; or *Allegro*, meaning fast, quickly, and bright) or indicate tempo in beats per minute (e.g., 50 beats per minute) at the beginning of a piece and indicate when it should be changed throughout. Texture was varied by changing the number of note onsets per measure, which approximates how composers change texture in music[5]. Sparse texture in music includes few instruments and few notes, while dense texture includes comparatively more notes and instruments. Timbre was varied by modifying the synthesized instrument used to render MIDI output. Composers often write with instrumentation in mind (e.g., solo flute with piano accompaniment), and instruments have associated timbrel qualities (e.g., piano timbre is dark compared to harpsichord, which is bright).

The second step of the study showed that MAFs are objectively measurable. Each MAF was correlated to a number of low-level EFs, and a single EF could be identified as a best predictor: Articulation by pitch salience, Dynamic by EBU128 loudness, Register by spectral centroid, Tempo by rhythm onset rate, Texture by spectral complexity, and Timbre by spectral spread.

The third step of the study showed that MAFs can be predicted in real-world music using specific EFs. It also provides evidence that supports the assumption made in the stimuli generating process: that MAFs naturally fluctuate in music—composers and performers vary these features between and within their music. While the variability of MAFs is captured, this step by itself does not verify the accuracy of the EF predictions. Since the

[5]Texture is not being referred to here in the music technical sense (e.g., monophonic, homophonic, polyphonic, etc.).

real-world music clips were not systematically generated with corresponding labels (as in step one), there is no 'correct' level for comparison. Instead, the EF predictions are compared to the subjective ratings by listeners in step four.

The fourth step of the study showed that MAFs are readily perceivable. Listeners provided subjective ratings for MAFs within the external stimuli. As noted in the third step, this provides evidence that MAFs vary between and within music, but does not provide any accuracy relating to a 'correct' level for a particular stimulus. The accuracy of subjective ratings are compared to the EF predictions in step three.

The final comparison of MAF ratings to EF predictions ties the three criteria of MAFs together. Each MAF has an established method of production relating to real-world music and performance techniques that can be approximated by the manipulation methods described in this study. The EF predictions, which are trained on a ground-truth of systematic MAF manipulation; and the MAF ratings, which are the result of a perceptual process by listeners; converge for the real-world music clips. This shows that both objective and subjective measures can be used to quantify MAFs in real-world music.

Revisiting the hypothesis outlined in the Introduction, we conclude the following:

1. *Low-level EFs were identified that reliably covaried with manipulated levels of each MAF;*
   This stage was supported, a subset of low-level features significantly correlated with each manipulated MAF.

2. *For each identified EF, the level of the corresponding MAF was predicted from a given digital audio stimulus;*
   This stage was supported, optimal EFs were chosen that produced a model which predicted MAFs from real-world music.

3. *In a listening task, human participants were required to rate the level of each MAF;*
   This stage was supported, participants were able to consistently rate MAF levels in a listening task with real-world music.

4. *Using an identical stimulus set, MAFs predicted by EF models (from Stage 2) and MAFs rated by participants (from Stage 3) were correlated with one another.*
   This final stage was also supported, MAFs predicted by EFs and MAFs rated by participants significantly correlated in their analyses of real-world music.

The criteria specified to establish MAFs are satisfied by the results of this study. MAFs were successfully produced with varying levels in the stimulus generation procedure. MAF levels were objectively measurable from these stimuli and in real-world music. MAF levels were also perceivable by participants who listened to real-world music. Lastly, each of these criteria were shown to be linked together, when MAFs are manipulated in music, both objective measures and perception consistently follow.

Limitations and future directions

There were a number of challenges to overcome in designing this study. First, the number of stimuli required to represent the full range of levels within multiple MAFs increased exponentially as factors were added. In this case, six factors, with just a few levels each, resulted in 4800 stimuli for a single example. The stimulus generation process was automated for five of the six features and helped make this process manageable. It did, however, take considerable computational resources. It would be beneficial to further improve the generation process to create a more varied ground-truth training dataset. Additional levels of Timbre should be added. The four used in

this study were chosen to represent a range of brightness, but Timbre varies in many dimensions (e.g., descriptors other than bright/dark have been explored: warm/cold, full/thin, soft/hard, etc., Wallmark & Kendall, 2018) and instruments regularly combine to produce unique Timbres in real-world music.

The number of base stimuli should also be increased. The example used in the study was a relatively simple two- to four-voice harmony. The use of complex stimuli that vary dynamics and combine different instruments would more accurately represent real-world music. Similar to improving the generation process, increasing the number and variety of training stimuli would also contribute to more robust predictive EF models.

A second challenge was in selecting models that predicted MAFs from EFs. The models used in this study were simple linear regression models that predicted a MAF from a single EF. More sophisticated algorithms exist that likely produce better predictions than single parameter linear regressions. For example, methods such as Convolutional Neural Networks, Decision Trees, and k-Nearest Neighbours have been trained on low-level features of music, or even only the spectrogram of a piece of music, to predict musical genre with classification accuracy above 90% in some cases (e.g., Bahuleyan, 2018; Ndou et al., 2021). These methods could be trained to, instead, predict MAFs in a similar manner. Increasing the accuracy in which MAFs are identified in music by EFs would strengthen Step three of the methods in the present study, and potentially improve the overall relationship with participant ratings.

The training portion of the experiment procedure may not have been adequate. While the majority of participants reported that the training was "Very helpful" ($n = 23$), their understanding and confidence ratings were relatively low (averages per MAF ranging from 1.42/5 to 3.02/5). The listening task explicitly asked participants to rate each MAF and was likely overwhelming for many participants. This may have caused participants to

guess in some cases. In future experiments, it might be better to use a dissimilarity rating task where participants only rate one MAF at a time. In such a task, participants simply rate how similar or different two stimuli are. Analyses of participants' responses then determine how differing levels of the stimuli were perceived. By simplifying the listening task, participants may provide more reliable data that would strengthen Step four of our methods and improve the overall relationship with EF predictions.

Despite the challenges listed above, this study has established a reliable method to assess MAFs. Moreover, the methods described here were designed to be modular, where each of the four steps can be improved upon individually and their overall improvements compared. For example, if the EF selection step was modified to include a new machine learning algorithm, the entire analysis could be rerun to compare how the MAF–EF relationship is affected. Furthermore, new potential MAFs, other than the original six outlined here, can be easily tested using the present procedure.

The MAFs developed here, and the potential to identify more MAFs with further investigation, provide a compelling tool for future research. Since MAFs have been linked to objectively measurable features extracted by Essentia, any existing audio can be analyzed. It might be possible to describe abstract music concepts, such as genre, in a more objective way. For instance, how are MAFs commonly used in Rock music compared to Classical music? Are there certain MAFs that are similar and others that are drastically different? Perhaps this may explain why an individual only likes *some* Rock music and *some* Classical music—they like the MAFs of specific songs.

MAFs can also be used to describe music in other ways. A large portion of digital music has considerable data associated with it (e.g., release year, sales, region, popularity, etc.) which can be included with MAF analyses. What MAFs are present in the most, or least, popular songs? How have such MAFs evolved over the years and decades? Answers to questions like these can provide insight into how people have listened to music over time and

place.

Lastly, since MAFs are manipulable, the insights gained through the type of analyses described above can be empirically tested. If we hypothesize that some factor(s) of individual people, for example personality traits, are linked to preference for specific MAFs, we can test our predictions in a well-controlled experimental setting.

Conclusion

This study focused on a long-standing issue in music preference research: the inability to classify and describe music. Genre has typically been used for this task, but interpretation of genre requires subjective and arbitrary decisions that are problematic for operationalization in music research. Furthermore, experimental manipulation of genre cannot be reliably performed, which limits the strength of conclusions researchers can make about its effects. *Music acoustic features* were proposed as a solution to these problems. They are objectively measurable from the information within digital audio stimuli (which can include real-world music). They are manipulable, as individual MAFs can be varied in level through compositional, performance, and digital audio techniques. And they are readily perceivable, listeners can explicitly identify and rate MAFs within real-world music. Furthermore, identification of perceived MAFs and objectively measured MAFs converge when compared for a single audio stimulus.

By establishing MAFs based on these criteria, we can design experiments to determine causal effects of music. We can also be confident that the musical stimuli we base our conclusions on are objectively described and have musical and psychological meaning to listeners. By focusing on MAFs as a foundational tool in research, we can gain new insight into how subtle differences in music affect our listening experiences.

V. References

Alonso-Jiménez, P., Bogdanov, D., Pons, J., & Serra, X. (2020). Tensorflow audio models in essentia. *Icassp 2020-2020 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp)*, 266–270. https://doi.org/https://doi.org/10.1109/ICASSP40776.2020.9054688

Aucouturier, J., & Pachet, F. (2003). Representing musical genre: A state of the art. *J. New Music Res.*, *32*(1), 83–93. https://doi.org/10.1076/jnmr.32.1.83.16801

Bahuleyan, H. (2018). Music genre classification using machine learning techniques. *Arxiv Preprint Arxiv:1804.01149*. https://doi.org/10.48550/arXiv.1804.01149

Barone, M. D., Bansal, J., & Woolhouse, M. H. (2017). Acoustic features influence musical choices across multiple genres. *Frontiers in Psychology*, *8*, 931. https://doi.org/10.3389/fpsyg.2017.00931

Bjørndalen, O. M. (2022). *Midi objects for python (mido) v1.2.10 [computer software]*. https://github.com/mido/mido/tree/stable

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J. R., & Serra, X. (2013). *Essentia: An audio analysis library for music information retrieval*. 493–498.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., & Serra, X. (2013). Essentia: an open-source library for sound and music analysis. *Proceedings of the 21st Acm International Conference on Multimedia*, 855–858. https://doi.org/https://doi.org/10.1145/2502081.2502229

De Leeuw, J. R. (2015). Jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music:

Contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology*, *4*, 487. https://doi.org/10.3389/fpsyg.2013.00487

Flannery, M. B., & Woolhouse, M. H. (2021). Musical preference: Role of personality and music-related acoustic features. *Music & Science*, *4*, 20592043211014014. https://doi.org/10.1177/20592043211014014

Greenberg, D. M., Kosinski, M., Stillwell, D. J., Monteiro, B. L., Levitin, D. J., & Rentfrow, P. J. (2016). The song is you: Preferences for musical attribute dimensions reflect personality. *Social Psychological and Personality Science*, *7*(6), 597–605. https://doi.org/10.1177/1948550616641473

Grimaud, A. M., & Eerola, T. (2022). An interactive approach to emotional expression through musical cues. *Music & Science*, *5*, 20592043211061745. https://doi.org/10.1177/20592043211061745

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., Leighton, D. C., & others. (2019). *Research methods in psychology*. Kwantlen Polytechnic University. https://doi.org/10.17605/OSF.IO/HF7DQ

Lartillot, O., & Toiviainen, P. (2007). A matlab toolbox for musical feature extraction from audio. *International Conference on Digital Audio Effects*, *237*, 244. https://doi.org/https://dafx.labri.fr/main/papers/p237.pdf

Litle, P., & Zuckerman, M. (1986). Sensation seeking and music preferences. *Personality and Individual Differences*, *7*(4), 575–578. https://doi.org/https://doi.org/10.1016/0191-8869(86)90136-4

McAdams, S. (2019). The perceptual representation of timbre. In *Timbre: Acoustics, perception, and cognition* (pp. 23–57). Springer.

MuseScore Team. (2022). *Musescore v3.6.2 [computer software]*. https://musescore.org/en

Ndou, N., Ajoodha, R., & Jadhav, A. (2021). Music genre classification: A review of deep-learning and traditional machine-learning approaches. *2021 Ieee International Iot, Electronics and Mechatronics Conference*

(Iemtronics), 1–6. https://doi.org/10.1109/IEMTRONICS52119.2021.9422487

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, *84*(6), 1236–1256. https://doi.org/10.1037/0022-3514.84.6.1236

Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: A five-factor model. *Journal of Personality and Social Psychology*, *100*(6), 1139–1157. https://doi.org/10.1037/a0022406

Rentfrow, P. J., Goldberg, L. R., Stillwell, D. J., Kosinski, M., Gosling, S. D., & Levitin, D. J. (2012). The song remains the same: A replication and extension of the music model. *Music Perception*, *30*(2), 161–185. https://doi.org/10.1525/mp.2012.30.2.161

Schäfer, T., & Sedlmeier, P. (2010). What makes us like music? determinants of music preference. *Psychology of Aesthetics, Creativity, and the Arts*, *4*(4), 223–234. https://doi.org/10.1037/a0018374

Wallmark, Z., & Kendall, R. A. (2018). Describing sound: The cognitive linguistics of timbre. *The Oxford Handbook of Timbre. Advance Online Publication. New York, Ny: Oxford University Press. Https://Doi. Org/10.1093/Oxfordhb/9780190637224.013*, *14*.

Zweigenhaft, R. L. (2008). A do re mi encore: A closer look at the personality correlates of music preferences. *Journal of Individual Differences*, *29*(1), 45–55. https://doi.org/10.1027/1614-0001.29.1.45

VI. Appendix A

Table 2: Participants answered the following demographic questions before beginning the experiment.

Item	Question
00	What is your age?
01	What is your handedness?
02	Do you have colour blindness?
03	What is your gender?
04	Have you ever had any formal dance training?
05	Approximately how often do you dance?
06	What style(s) do you dance?
07	Have you ever had any formal music training?
08	Do you play a musical instrument and/or sing?
09	What is your principle instrument?
10	At what age did you start playing?
11	Including time spent rehearsing, approximately how many hours a week do you play/sing?
12	What type of music do you usually play?
13	How many hours per week do you listen to music?
14	What types of music do you listen to?

Table 3: Participants completed the following questionnaire while/after listening to each musical clip. The instruction, "Please rate the musical clip by selecting from the options below. Note: you may select more than one response", was displayed at the top of the page. A prompt and five options were listed for each MAF. Participants could select more than one option.

Item	MAF	Prompt	Options
00	Articulation	The articulation is:	Very smooth, Smooth, Medium, Spiky, Very spiky
01	Dynamic	The loudness is:	Very soft, Soft, Medium, Loud, Very loud
02	Register	The register is:	Very low, Low, Medium, High, Very high
03	Tempo	The speed is:	Very slow, Slow, Medium, Fast, Very fast
04	Texture	The texture is:	Very dense, Dense, Medium, Sparse, Very sparse
05	Timbre	The sound-colour is:	Very dark, Dark, Medium, Bright, Very bright

Table 4: Following the training and listening portion of the experiment, participants answered questions relating to their understanding and confidence of ther responses.

Item	Question	Options
00	How difficult was the listening task?	Very easy, Easy, Average, Difficult, Very difficult
01	Did you find the musical term definitions and examples helpful?	Not at all helpful, Somewhat helpful, Very helpful, They made no difference, I skipped the examples
02	How would you rate your understanding of each of the musical terms? Speed Loudness Register Articulation Texture Timbre	Poor, Below average, Average, Good, Excellent
03	How confident were you in the responses you provided for: Speed Loudness Register Articulation Texture Timbre	Not confident, Somewhat confident, Confident, Very confident

VII. Appendix B

Table 5: Real-world stimuli used in comparison analyses (continued in Table 6). Songs were clipped from the 'Start' to 'End' point (in seconds). For each clip, MAFs were predicted by Essentia feature models and were rated by participants in the listening task. Songs can be accessed online by the provided links (append the link as: https://youtu.be/<link>).

ID	Song	Start	End	Link
01	Frozen Crown - Battles In The Night	5	22	kVHZ6yVOkUE
02	Ariana Grande - Positions	0	7	xuOOAQoDKN0
03	Erik Nielsen - Sketches III Staccato	5	15	9R9xRrXJRMQ
04	Blur - For Tomorrow	198	218	J77C9ODFflw
05	Lady Gaga - Alejandro	23	33	06MV87zoZaY
06	Abhi Mujh Mein Kahin - Sonu Nigam	97	109	3Iq3j3L06rQ
07	Lord Huron - The Night We Met	0	15	KtlgYxa6BMU
08	Jaco Pastorius - Full album	6	22	pvjHT8Lepz8
09	Jaco Pastorius - Full album	1049	1057	pvjHT8Lepz8
10	Jaco Pastorius - Full album	1605	1615	pvjHT8Lepz8
11	Bach - Organ Sonata No4	0	13	h3-rNMhIyuQ
12	Grandson - Blood Water	57	67	LsHYFQgQxpw
13	In Flames - The End	72	80	yafxUluB6DA
14	Interstellar - Cornfield Chase	0	16	mykUt3yhZWA
15	Interstellar - Cornfield Chase	32	40	mykUt3yhZWA
16	Iced Earth - Phantom Opera Ghost	344	353	9ZyE2V-BQqc
17	Iced Earth - Phantom Opera Ghost	363	373	9ZyE2V-BQqc
18	Dunwich Beach - Autumn	10	25	cznwjb859PE
19	Cannons - Hurricane	0	10	LZ2kSbSrDLs
20	Pacifica Quartet - Bartok String Quartet No. 4	13	25	aBs53SlEkso
21	Sapna Jahan - Nigam, Mohan	342	362	iFq73v_cdTk
22	In Flames - Moonshield	0	16	AmcC9aJkBlw

Table 6: Continuation of stimuli described in Table 5.

ID	Song	Start	End	Link
23	Cannons - Bad Dream	120	137	Fz4axHOyccQ
24	Yes - Relayer	9	24	quPoq2699Xo
25	Snarky Puppy - What About Me	15	30	fuhHU_BZXSk
26	Epic Low Brass - The Rains of Castamere GoT	55	76	z9WAHOZaKTw
27	Billie Eilish - bad guy	0	14	4-TbQnONe_w
28	Brian Eno - Thursday Afternoon	15	32	TTHF2Dfw1Dg
29	Agar Tum Saath Ho - Yagnik & Singh	75	87	OGIOfNvr4fo
30	Apocalyptica - Path	14	28	m9xqO9kKqyk
31	Yes - To Be Over	0	15	bf52nD8ELcc
32	System Of A Down - Atwa	0	15	nVZ8tRlcZhA
33	Jhene Aiko - The Worst	0	22	npB9gNLC2_g
34	Opeth - Prologue April Ethereal	30	55	_gVqVYeztDk
35	Rone - Bye Bye Macadam	141	161	kfoJUeyMsOE
36	Rone - Bye Bye Macadam	0	20	kfoJUeyMsOE
37	Robert Miles - Children	0	13	CC5ca6Hsb2Q
38	Robert Miles - Children	210	220	CC5ca6Hsb2Q
39	Mcbaise - Water Slide	0	11	n11j73OYqOY
40	Mcbaise - Water Slide	178	198	n11j73OYqOY
41	Rone - Origami	1	16	hVv331iLMXM
42	Rone - Origami	58	71	hVv331iLMXM
43	Lorn - Acid Rain	0	17	nxg4C365LbQ
44	Lorn - Acid Rain	50	59	nxg4C365LbQ