

Cognitive Music Listening Space: A Multivariate Approach

Brendon Mizener¹, Mathilde Vandenberghe², Herve Abdi¹, & Sylvie Chollet²

¹ University of Texas at Dallas

² YNCREA

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. Brendon Mizener: Stimuli creation, Survey design & creation, Data collection & processing, Statistical analyses, Writing - Original draft preparation; Mathilde Vandenberghe: Original concept, Survey design & creation; Herve Abdi: Writing - Review & Editing, Statistical guidance; Sylvie Chollet: Original concept.

Correspondence concerning this article should be addressed to Brendon Mizener, 800 W. Campbell Rd., Richardson Tex. E-mail: bmizener@utdallas.edu

16

Abstract

17 This is my abstract

18 *Keywords:* keywords

19 Word count: X

Cognitive Music Listening Space: A Multivariate Approach

```

21 ## [1] "It is estimated that your iterations will take 0.67 minutes."
22 ## [1] "R is not in interactive() mode. Resample-based tests will be conducted. Please t
23 ## =====

```

24 Music listening is a complex cognitive activity that involves many judgments per
 25 second. Listeners continuously evaluate incoming information and compare it with that
 26 which came before. These judgments involve many different dimensions of music related to
 27 both the technical and affective aspects of this acoustic medium. Generally, however, in the
 28 literature, many researchers describe sound in only two dimensions; pitch and loudness
 29 (**Wallmark2019?**). Even music cognition researchers resort to isolating only one or two
 30 dimensions, such as pitch or rhythm (e.g., (**Krumhansl2000?**)). Although the use of more
 31 realistic stimuli has become more common [citation], researchers still tend to simplify stimuli
 32 so that they are isochronic or simplified in other ways. Additionally, performing traditional
 33 hypothesis testing in experiments that evaluate multiple dimensions quickly becomes
 34 cumbersome and inelegant, and it can be difficult to interpret results. Imaging studies,
 35 which have merit in their own right, similarly lack inclusion and require prohibitive amounts
 36 of resources, and are severely location-dependent. The gradual increase in complexity of
 37 behavior and cognition studies, coupled with the rise of questions about the universality of
 38 experience and the democratization of science, compels us to find novel ways of investigating
 39 the experience of music.

40 Another consistent limitation is population. In general, studies of this sort are limited
 41 to WEIRD people (Western, Educated, Industrialized, Rich, and Democratic), and also tend
 42 to lack representation in ethnic minority communities. By developing investigative
 43 paradigms that are accessible on mobile platforms, and that reduce participant demand as
 44 much as possible while maintaining rigor, we are likely to be able to reach a much greater
 45 subset of the population. If we are able to pair this kind of data gathering with appropriate

exploratory analysis, we can target much more effectively where we might investigate with more traditional hypothesis testing. The initial motivation for this paper was as precursor work to a study aimed at evaluating the cross-modal similarities between gustatory perception and auditory perception. While there exist various established gustatory ‘spaces’ onto which music perception might be mapped, for example in wine tasting [wine citations here], there is no similar space for music. The goal of this study was to evaluate whether or not there is a similar music perception space against which other sensory domains could be compared, and to select stimuli that could anchor the corners of such a space.

There is some previous work that has approached timbre specifically from a similar analytical perspective.

The current study aimed to evaluate a set of musical stimuli in two ways, by association with a list of adjectives, and by rating on a set of musical qualities. These will then be analyzed independently and then combined to see what effects can be deduced from the combination of the two.

Valence/arousal model

- as questions become more complex, the burden on researchers to define parameters in which to test and evaluate their results becomes much greater. Controlling for extraneous variables becomes a problem in and of itself.
- lacks generalizability because of narrowness
- rising question of universalities & cross-cultural perception in music
- how do we access a larger population
- This all results in a need for an experimental paradigm that is robust to violations in experimental procedure, accesses

69 CATA paradigm

70 While not invented by (**Katz1933?**), the Check-All-That-Apply (CATA) investigative
71 paradigm was used in that study to evaluate racial stereotypes among college students. As
72 an method it's not terribly common in the psychological sciences any more, but it has been
73 and continues to be used widely to "obtain rapid product profiles" (**Meyner2014?**) from
74 participants. It is also commonly used in sensory evaluation. In this method, participants
75 are asked to select any and all adjectives from a list that describe a given stimulus. This
76 allows researchers to collect a lot of data about a given stimulus without placing demand on
77 the participants. The onus is then on the researchers to aggregate and evaluate this data to
78 see what insights may be gained on the stimuli. Because of the

- 79 • what it is (cite Katz & Braly 1933, Meyners & Castura)
 - 80 – Often used in sensory discrimination, consumer product research
 - 81 – participant is instructed to select any and all qualities or descriptors that they
 - 82 associate with the presented stimulus.
- 83 • why is it good/how does it solve the problem
 - 84 – easy
 - 85 – low demand
 - 86 – robust to outlying responders
- 87 • why is it specifically useful for this paradigm (internet/online research)
 - 88 – Using a technique that minimizes participant effort while allowing for meaningful
 - 89 data gathering, in a context that necessitates data gathering from a distance
- 90 • what the data look like
 - 91 – After cleaning and preprocessing, the data for each participant will take the form
 - 92 of a pseudo contingency table. The difference here is that a contingency table is
 - 93 specifically when a participant selects only one option from a list for each
 - 94 stimulus, resulting in a table with one and only one one (1) per row. Because we

are using the CATA technique, a one (1) at the intersection of each row or column indicates that the participant selected that adjective or musical quality for that stimulus. A zero means that they did not. These individual tables are all compiled into a ‘brick,’ or three-dimensional array of data with Observations (stimuli) on the rows, variables (musical qualities or adjectives) on the columns, and participants on the third dimension, which we will refer to as ‘pages’ here. This brick is then summed across pages into a single table, so that any given cell contains the total number of times a participant selected a given adjective or quality to match with a stimulus. From this point there are two sets of data that can be analyzed. The first is the 3D array, which can be analyzed using various distance analyses, to evaluate differences between the participants using grouping variables extracted from the demographics surveys. The other is the pseudo contingency table, which can be analyzed using various multivariate techniques.

- what processing steps are needed

Present questions & methods of analysis

The initial motivation for this study came from a cross-modal study investigating the co-_____ between gustation perception, specifically beer, and music perception. Initial studies suggest that a wide variety in musical stimuli was necessary to determine any correlations or differences. As such, this study is designed to investigate whether a music cognitive listening space can be established using this paradigm to allow cross-modal comparison. Additional questions arise from the study itself: are there significant differences in how participants from different nationalities (and by extension musical cultures) perceive, or, more precisely, describe music? Are there parallels in how music non-specific descriptors and music-specific qualities are assigned to stimuli or evaluated? To analyze the data collected here, is there a cognitive space that can be deduced using this paradigm? - Is that cognitive space robust across nationalities (i.e., musical cultures) - Are there parallels

between the ways that music is rated using music-specific dimensions & adjectives - To analyze this data, we use hypothesis free multivariate analyses paired with inference methods like the bootstrap (resampling) and the permutation test (randomization). Although the analyses do not specifically test hypotheses in the same way an ANOVA or t test would, these inferences methods allow us to draw conclusions about statistical significance from them. - Correspondence analysis - brick excerpts (stimuli)/descriptors/subjects - pseudo contingency table - - PLSC - 2 contingency tables (2 data tables)

Arousal/valence model/other models?

- What is it?
- previous work on cognitive space in music perception

Where we got the words.

- revisit the IRB application

-

Multivariate methods?

Methods

Participants

Participants for this study were recruited in multiple ways, all of which represent convenience sampling. The participants in the United States were recruited using the traditional method of offering experimental participation credit, and also via social media. French participants were recruited by word of mouth, email, and social media. The only restrictions on participation were that the participant must have self-reported normal hearing. We recognize that although we suggest that data collected in this way have a much greater hypothetical reach, the data here represent a) a convenience sample, b) that is limited to participants that have access to the internet. Both of these specific limitations

could be remedied when designing and implementing future research.

In addition to the convenience sample described above, We sought out a specific population with ten years or more of music training to evaluate the stimuli on explicitly musical dimensions. We recruited this population for two reasons: firstly, as a validation step, to ascertain whether the stimuli truly reflected the composer’s intent. Secondly, we had the goal of evaluating how the musical qualities of the stimuli, as evaluated by the trained participants, correlated with the adjectives selected by those who participated in the adjectives survey. All recruitment measures were approved by the UT Dallas IRB.

Material

Surveys. There were two separate surveys presented to participants. One evaluated the musical stimuli on concrete musical qualities like meter and genre, the other evaluated the stimuli using adjectives, and used the CATA paradigm. Both surveys also captured participants’ demographic data, including age, gender, nationality, occupation, and musical experience. The musically trained participants completed the musical qualities survey (QS). The adjectives survey (AS) was administered without respect to participants’ musical training. As a result, there is a wide range of musical experience present in the sample. The words for the AS were selected using (Wallmark2019?) as a guide and in consult with a French professional musician. Some words were initially selected in French and some in English. In all cases, words were selected for which there was a clear French (vis-a-vis English) translation. The words and their translations are listed in [supplementary materials?]. The qualities assessed in the QS were selected from standard music-theoretical descriptors of music. For example, when rating the excerpts on tempo, participants were asked to rate the excerpt using the scale *Very Slow*, *Slow*, *Moderately Slow*, *Moderate*, *Moderately Fast*, *Fast*, and *Very Fast*. The full list of musical qualities and associated levels is in [supplementary materials?].

Stimuli. All stimuli were original stimuli composed for this study. They were designed to evaluate a number of musical dimensions and control for others (e.g., timbre). The stimuli were all string quartets, in order to control for the confounding factor that different instruments are fundamentally described in different ways. All stimuli were between 27s and 40s long, with an average length of 32.4s. The intent was to have all stimuli be around 30s long while preserving musical integrity. All stimuli were composed between April 13 and June 18, 2020.

Procedure

Participants were provided with a link to either the AS or the QS. Both surveys were administered using Qualtrics. After standard informed consent, participants were instructed to listen to the excerpts presented either using headphones or in a quiet listening environment. Participants in the AS condition answered a single question per excerpt, in which they selected any and all adjectives that they felt described the excerpt. Participants in the QS condition answered 10 questions per excerpt, rating the excerpts using the qualities and scales provided.

Data analysis

Raw data were cleaned and processed in Excel and R. This included translating all French responses to English for ease of analysis. Data were cleaned and transformed into a pseudo contingency table for each participant. These individual tables were all compiled into two ‘bricks,’ or three-dimensional arrays of data with Observations (stimuli) on the rows, variables (musical qualities or adjectives) on the columns, and participants on the third dimension, which we will refer to as ‘pages’ here. Each array was then summed across pages into a single table, so that any given cell contained the total number of times a participant selected a given adjective or quality for a given stimulus. The arrays were analyzed using distance analyses to evaluate differences between the participants using grouping variables extracted from the demographics surveys. The summed tables were analyzed using

Correspondence Analysis. Since we did not use *a priori* grouping variables for the excerpts, the summed tables were evaluated using hierarchical cluster analyses to see what groupings arose during evaluation. A final analysis (Partial Least Squares Correlation, PLSC) evaluated correlations between the two summed data tables to see what information was shared between the two tables. For each of these analyses, variance is extracted in the form of eigenvalues, which form the principal components of the maps shown below. The dimensions extracted in this process are by definition orthogonal and share no information. For each of the analyses below, we focus on the first two dimensions. Third dimensions are presented solely for the sake of illustration.

Correspondence Analysis. Correspondence Analysis (CA) is similar to Principal Components Analysis (PCA), except that it allows for the analysis of qualitative data. Because of the organization of the data, this technique allows for a biplot of both rows and columns in a single factor space. We can also calculate and identify which observations (stimuli) or variables (adjectives or qualities) contribute significantly to a given dimension in the factor space. In addition to factor plots, we used permutation tests and bootstrapping to project confidence intervals for the means of certain variables in the factor spaces, which allows for statistical inferences.

Partial Least Squares Correlation. Partial Least Squares Correlation (PLSC) is analyzes two data tables either on the same observations (rows) or variables (columns). In this case, the stimuli. This technique is commonly used in imaging studies to evaluate correlations between matrices of imaging data and of behavioral or task data. In this technique, we will be looking at covariance and correlation between the variables in the two tables.

Results

Experiment 1: Musical Qualities Survey

Participants. The scree plot below shows the eigenvalues for the distance analysis between musical experts. The usual guideline of analyzing only dimensions with eigenvalues greater than one seems prohibitive here, as all dimensions except the last have $\lambda > 1$. For the purposes of this experiment, we've opted to focus on the first two dimensions, with $\lambda = 9.36$ and $\lambda = 7.98$, respectively. This analysis revealed no significant difference between the experts when grouped either by nationality or by gender identity. Training level was not considered as a factor for this analysis because this survey was only distributed to participants with greater than 10 years of musical training.

Excerpts.

- first three dimensions explain ~40% of variance
 - first two explain ~30% of variance (find which elements contribute the most, display in table?)
 - First dimension is dominated by tempo and articulation, and the loud and soft levels of the dynamics quality
 - second dimension is dominated by meter and genre
 - third dimension is dominated by jazz genre and harmony variables (this excerpt is an outlier), other significant contributions come from dynamics and density, and to a lesser extent some contour.

Experiment 2: Musical Adjectives Survey

Participants. The distance analyses between participants in either survey revealed significant group differences between French and American participants in the AS.

Excerpts.

- first three dimension account for ~80% of the variance. Both of the first two

dimensions carry information from both the valence and arousal dimensions, suggesting that they aren't orthogonal.

- first dimension explains 50+%,

- * strong contributions in the positive direction are Sad, Dark, Melancholy, Slow, Mysterious, Solemn, Disturbing.

- * strong contributions in the negative direction are Fast, Dancing, Happy, Colorful, and Bright.

- Second dimension 20+%

- * strong + contributions: Aggressive, Fast, Disturbing, Mysterious, Solemn, Complex

- * strong - contributions: Warm, Soft, Slow, Happy, Round, Light.

- The third and fourth dimensions may represent an eigenplane, suggesting two dimensions that are very closely related in variance extracted and information represented.

Experiment 3: Combined Surveys

Participants. PLSC: - The

Discussion

- Because of the nature of the CA, the musical qualities survey is not robust to outliers.
- hierarchical clustering revealed that there are differences in groupings between the musical qualities and the adjectives judgments.
- valence/arousal model is very clear in the adjectives data
- which qualities dominate which dimensions in the musical qualities survey? Why? Is it because of clear subjective/objective agreement on some specific musical qualities?

Limitations & future directions

- continue to evaluate more participants from more varying backgrounds

- 269 • increase sample size between highly trained musicians and others
- 270 • add more dimensions (melodic complexity)

References

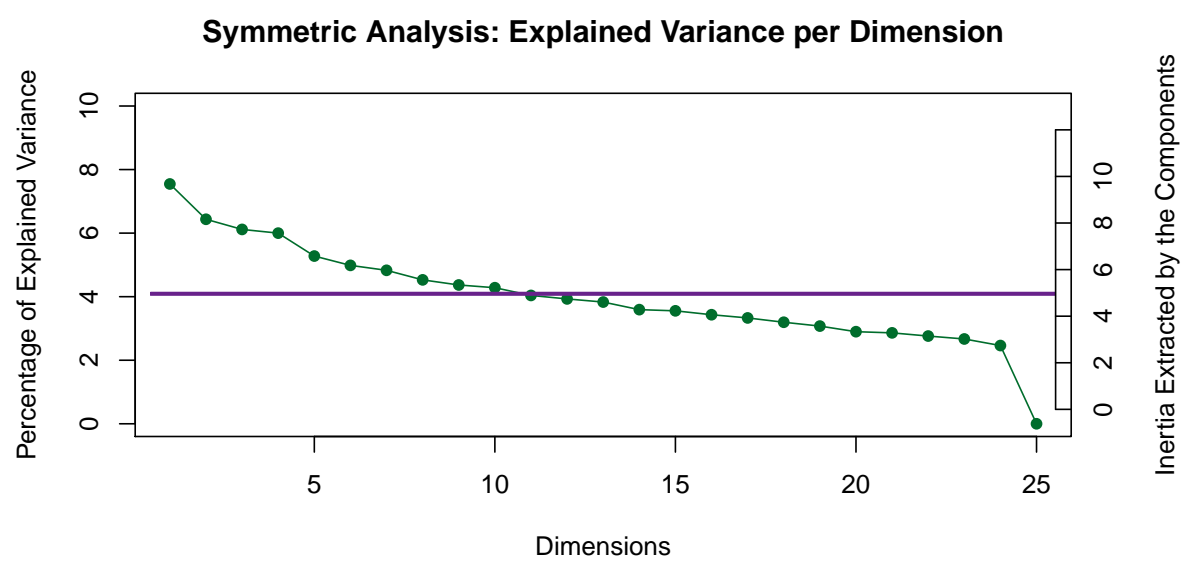


Figure 1