Quantifying the temporal dynamics of music listening: A critical investigation of analysis techniques for collections of continuous responses to music

Finn Upham

Master of Arts

Music Technology

McGill University

Montreal,Quebec

2011-10-05

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Masters of Arts in the Area of Music Technology.

ACKNOWLEDGEMENTS

There are dozens people whose work contributed to this thesis via the experimental data. My thanks to the team who supported the Angle of Death project; to the those who performed the Boston Symphony Orchestra experiments; to Mark Korhonen for sharing his experimental data; and to the staff at CIRMMT and my past and present lab mates at MPCL who put on the Audience Response concert sessions. Many research partners helped grow my understanding of these data with their experience and questions about analysis techniques. My thanks to Eric Thule, who helped me explored many numerical tangents; to Dan Levitin, who first asked if we could count the events to which participants' might responsed in music; to the music theorists at the Schulich School of Music, for their insightful criticism of music cognition; and to my lab mates, who regularly identified holes in my ideas over the years. I am indebted to those who helped edit this document. My thanks to Marie-Hélène Cartier for translating the abstract; to David Sears, for understanding what I was trying to convey; to James McKinney, for turning the task of editing into a teaching opportunity; to Susan Upham, my patient mother, for catching innumerable errors; and lastly, to my supervisor, Stephen McAdams, for supporting this research and this document from beginning to end. His patience, generosity, and confidence gave these ideas a chance to grow, while his insightful pruning has vastly improved their stability and presentability.

ABSTRACT

Continuous response measurement offers a data-rich trace of a listener's experiences of music in time. Listeners' responses are most often studied in collections—each a set of time series of the same response measure to the same stimulus from multiple listenings. Inter-response variability and the challenges of time series analysis complicate the interpretation of these collections. This thesis describes traditional and novel methods of analyzing collections of continuous responses to music with the goal of identifying what information can be found in these collections before trying to establish possible relationships to the features of the stimulating music. Besides mathematical investigations of these analysis methods, their potential outcomes are assessed by applying each to forty experimental collections of continuous rating responses and four artificial collections of unrelated continuous rating responses. The traditional analyses studied include the average response time series and Pearson correlations between continuous responses as a measure of response reliability. The chapter on novel techniques introduces activity analysis and coordination tests, evaluates measures of the relative significance of time points in these collection, and applies cluster analysis in search of distinct patterns of response to the same stimuli. The results of these analyses suggest that though music does not provoke the same continuous response from all listeners, musical works can induce distinct and repeatable listening experiences which are measurable in collections of continuous responses.

ABRÉGÉ

L'évaluation des réactions continues permet d'obtenir un tracé riche en données de l'expérience des auditeurs par rapport à la musique au fil du temps. En règle générale, les réactions des auditeurs sont analysées par ensembles, c'est-à-dire par groupes de séries chronologiques portant sur de mêmes relevés de réactions au même stimulus provenant d'écoutes multiples. La variabilité entre les réactions et les défis inhérents à l'analyse des séries chronologiques rendent l'interprétation de ces ensembles encore plus complexe. La présente thèse décrit des méthodes traditionnelles et nouvelles d'analyse d'ensembles de réactions continues à la musique afin d'identifier quelles informations peuvent être recueillies dans ces ensembles avant de tenter d'établir des liens possibles avec les caractéristiques de la musique stimulante. En plus de létude mathématique de ces méthodes d'analyse, leurs résultats potentiels ont été évalués en appliquant chacune d'entre elles à quarante de ces ensembles d'évaluation de réactions continues ainsi qu'à quatre ensembles artificiels d'évaluations de réactions continues non apparentés. Les analyses traditionnelles étudiées comprennent les séries chronologiques moyennes et des corrélations de Pearson entre les réactions continues comme évaluation de la fiabilité de la réaction. Le chapitre portant sur les nouvelles techniques commence par une présentation de lanalyse de l'activité et des tests de coordination. Par la suite, il évalue les mesures de pertinence des repres temporels de ces ensembles, puis il rend compte de l'analyse par regroupements visant à identifier des modèles précis de réactions aux mêmes stimuli. Les résultats de ces

analyses sous-tendent que bien que la musique ne provoque pas la même réaction chez tous les auditeurs, l'œuvre musicale peut créer des expériences d'écoute distinctes et reproductibles pouvant être évaluées dans des ensembles de réactions continues.

TABLE OF CONTENTS

CHAPTER 1
Introduction

1.1 Continuous responses to music

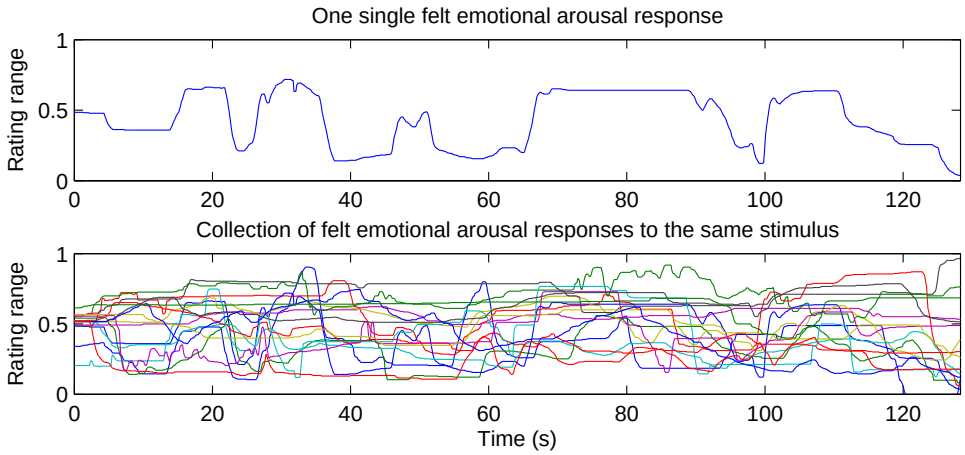

Figure 1–1: Above is a single response, one participant's rating of their felt emotional arousal while attending a live performance of a madrigal. Below is a collection of responses, 35-Arc2A, from 17 participants performing the same task to the same stimulus.

Continuous response data are generated by the rapid repeated sampling of a measure of response to a stimulus. These time series quantify reactions within the time span of seconds and minutes rather than days or years. Human responses can be measured through neurological or psychophysiological activity, through motion, or through self-evaluation. This thesis will address mainly the analysis of collections of the continuous responses to music as measured by continuous ratings—a type of behavioural measure—however many of the approaches described in the following

chapters can be adapted to explore data from other forms of behavioural response and continuous physiological measures.

Experiments measuring continuous responses to music often collect the same response measures from a number of people as they are presented the same musical performance. Others adapt the selection of stimuli to each participant while trying to capture the same kinds of responses (e.g., chills). In the latter case, it is not possible to compare the complete temporal profiles of participants' responses. Figure 1–1 shows a single response and the collection of responses to a live performances of the Renaissance madrigal *Il bianco e dolce cigno* composed by Jacques Arcadelt. Response collections such as the lower graph shown in figure 1–1 make it possible to consider how these responses resemble each other and how they might be related to the stimulus. Handling these collections of stimulus and measure related responses, rather than individual responses or sets of responses collected to different stimuli, is the focus of the work that follows.

1.1.1 Behavioural responses

In the last century, the most commonly collected continuous responses to music were through behavioural measures, and these measures are still quite popular. These time series have also been called self-report responses or self-evaluations because they require participants to assess and express their own experience moment-by-moment. Such tasks can be more challenging for participants to perform than post-stimulus judgments, and many studies report having participants practice the task before recording responses to the experimental stimuli.

Behavioural measures have used a number of different collection methods. Some reports are made on a continuous scale via positioning a physical object like rotating the pointer on a dial or shifting the knob on a slider. Continuous measures are also collected using virtual versions of these devices on computers or handheld devices with touch interfaces, with GUI sliders or two-dimensional rating fields from which cursor positions are recorded. These responses fall on continuous scales: participants report their position between extremes such as pleasant to unpleasant or excited to calm. Other continuous response measures are more categorical, using button interfaces to report isolated events (e.g., perceived phrase boundaries or experience of chills) or to choose between multiple categories (e.g., a list of words describing the music, or a list of elements of music to which they are attending.)

GUI interfaces are often more economical to use, requiring only specialised software to run on standard lab equipment. These, however, can be quite limited in the feedback they give to participants, requiring users to look where they are clicking, tracing, or dragging within a range set on a screen. Some physical devices incorporate haptic feedback via springs to help participants feel their position on the rating scale. One important example of this was the Simenon used by Frede Nielsen in his influential experiments in the early 80's. This device was a spring-mounted potentiometer that participants squeezed to indicate tension [Nie87].

1.1.2 Measures of experience vs measures of perception

Some behavioural responses ask participants to report their perception or assessment of the stimulus, called "perceived" responses. These measures are often of the emotion expressed by the music or performers. The other focus of continuous responses are of the subjective experiences of participants. "Experienced" responses are recordings of participants' subjective responses to the stimulus. Behavioural measures of experienced responses include reporting their emotional state while the music plays or what aspects of the music are attracting their attention. Other uses of continuous responses include judging the quality of performances, which may be treated as perceived or experienced, depending on the context. While the debate continues on how and whether music induces emotions in listeners, both types of response measures seem to be of theoretic interest.

1.1.3 Contrasting continuous measures

Studies have found significant differences in continuous ratings depending on what aspect of their response participants are measuring. A striking difference between ratings of "tension" and "aesthetic response" published by W. E. Fredrickson show strong contrasts in the average response time series to the same stimuli [Fre95]. In particular, the average time series had much greater variability for the ratings of tension than for that of the ratings aesthetic response. Whether this is due to different rating strategies or less contraditory responses is not clear.

1.1.4 Psychophysiological responses

Psychophysiological measures used to assess responses to music are usually of biosignals measured using non-invasive sensors such as surface electromyography

electrodes for facial muscle activation and finger cuff photoplethysmyography sensors for blood volume pulse (BVP). These measures are generally intended to measure aspects of participants' experiences as they are expressed physically, and often involuntarily, within the time during which the stimulus is presented. Common measures are skin conductance (SC), heart rate, skin temperature, facial eletromyographic activity (EMG) of the zygomaticus and the corrugator, and respiration. Some are reportedly related to arousal and attention, such as skin conductance and heart rate, while others are presumed to give signals of emotional responses, principally positive valence emotions through the smiling muscle zygomaticus and negative valence emotions through the brow-furrowing corrugator.

1.2 A history of analysing continuous responses to music

Continuous responses to music have been collected with excitement for more than half a century. One early experiment was a complex study of psychophysiological responses and verbal accounts of subjective experience recorded to polygraph paper [Fra56]. Several responses to the same stimuli were collected, a total of 38 across three groups with distinct levels of musical expertise. Most of the analysis was performed through study of individual responses, counting different behaviours as seen across groups. From the beginning, musical experience has been the first choice in trying to explain the variability of responses to music, over factors such as pre-listening mood, arousal, or stimulus familiarity.

Once signals began to be collected digitally, the analysis of the response average became common practice. From Nielsen's analysis of tension ratings

[Nie87] to the present day, this method of summarising a collection of responses is standard, and seen in two-thirds of the publications reviewed using continuous responses to music. Some publications interpret the plotted average time series visually, while others apply to it more sophisticated statistical and numerical techniques.

These responses are collected in time, but analyses have often been performed out of time. For many of the early studies trying to compare populations' responses, the simplest method of evaluation was to average each response longitudinally to a single datum [Fra56]. This turns the collections of time series into much more manageable sets of discrete points to be compared between different stimuli [VVS93], responses measures [Lyc98], and participant groups [Fre99b]. Work by Ruth Brittin and others defended using continuous ratings in this fashion by showing that the longitudinal averages were quantifiably different from post-stimulus rating judgments [BS95] [DC01]. This type of response- or participant-wise reduction is typical for most analyses of psychophysiological responses[Kru97] [Ric04], more popular than average time series. This is because many meaningful features in these signals are assessed over time periods of minutes [IM99] or seconds for particularly controlled stimuli, [SKS06] and these features could not be evaluated from sensor data averaged across responses.

According to Diane Gregory's review of the reliability of continuous responses [Gre95], the earliest published example of correlating continuous response time series was in a 1989 article by Deborah Capperella. Many experiments performed with the Centre for Music Research of Florida State University requested that

some percentage of participants provide a second continuous rating of the stimuli to test the reliability of participants results. The consistency of raters' responses would be assessed by averaging the Pearson product-moment correlation coefficients between the first and second ratings from each retest participant.[Mad97]. This assessment of reliability was also used to compare responses between participants [Kru96], average response time series of different collections to the same stimuli [Fre99a], and average response time series to continuous representations of aspects of the music [MGF97].

Correlations of continuous responses have continued to be used as statistical test of reliability despite Schubert's 2002 article highlighting the problem of serial correlation in time series [Sch02]. Rather than abandon the practice, wider awareness of time series analysis has resulted in researchers applying different manipulations of response data [CJSK+10] prior to correlations as well as more complicated correlation techniques being proposed [VKWL06].

Continuous responses have been evaluated and discussed in relation to the stimuli since the very first responses were recorded to music, but models of responses have been slow to come forward. An influential first step was the multiple regression technique Schubert proposed first in his thesis [Sch99a]. With a set of time series descriptions of the stimulus, these models try to explain variations in response values over time using a mixture of numerical representations of dynamic aspects of music. In the 2000's, many "one-of-a-kind" multiple regression models/analyses were published, most being variants on linear regression [KCJ05][LTE+08].

In trying to fit loudness curves and one-dimensional reductions of harmony to average response time series, many have confronted the issue of response lag. Cross-correlation techniques have been used to quantify what delay between stimulus data and response data result in the tightest fit. For behavioural measures in particular, changes in ratings appear to follow stimulus events by 1 to 3 seconds [Sch04], but the duration of this lag varies between participants and across stimulus effects. Attempts to model continuous responses from stimulus information invariably parametrize lag, usually per stimulus feature, to handle this issue in self-report measures of response.

Analysis techniques have diversified since the turn of the century. The introduction of functional data analysis (FDA) to this research in 2004 has been followed by applications in a number of papers, but other "novel" approaches have not spread as far. Models of averaged responses have been attempted using non-linear regression [FS09], multilinear modelling [DMR06], and neural networks [CC09]. Other models have tried to fit the variance in these collections with functional principal component analysis [LNVR07] and decision trees [TMCV06].

Instead of trying to explain the data from the stimuli, other new techniques have focused on exploring the robustness of continuous response collection results. Some have tried to identify moments of extreme responses and degrees of coordination in physiological and behavioural responses [GNKA07a] [Sch07], while others have tried to characterize the time course of continuous responses [BBL$^+$09] [TWV07]. Finding the reliable information in these complicated data sets can only improve discussions of collective and individual responses to music in time.

Continuous responses have been discussed in many neighbouring disciplines. This type of data have been recorded to audiovisual stimuli such as movie excerpts [LC10] and TV commercials [RT04] for marketing [WM08], emotion [MR09], and political communication [Iye11] research. Experiments using music as a stimulus has been published in music therapy, music education, music theory, music cognition, information theory, and psychology journals of international repute. From a database of 66 publications spanning 55 years, continuous response to music experiments have been reported in at least 24 journals, with the highest concentration of articles in Music Perception and the Journal of Research in Music Education. These two have been publishing articles on the topic steadily for more than 15 years. Similarly, researchers have been exploring continuous responses around the world. Authors in the article database worked from more that 50 research and educational institutions found in 15 countries across four continents. The diversity and distribution of the interested community likely accounts for both the breadth of analysis methods published and the rarity of debate on methods relative merits.

This thesis reviews traditional techniques and presents some novel methods for exploring collections of continuous responses in and of themselves. Before explaining these data in terms of the stimuli, we need to know what information they contain. By comparing the performance of these methods on several collections of responses, the advantages and limitations of these techniques can be more fully explored than is possible in the single data set discussions of the published articles mentioned above.

1.3 Introducing the data sets

The analyses that follow have been applied to data sets collected for a number of earlier experiments. The data sets are of continuous ratings, though the measures are not all the same. Each set includes participants' responses to more than one stimulus, and some contain different audiences responses to different performances or productions of the same musical work.

For the purpose of these analyses, the ranges of all behavioural response collections were linearly transformed to the interval [0,1], and responses showing no changes in ratings were excluded as "outliers" under the assumption that this meant the participant chose not to perform the task (see section 2.3). Below are explanations of the behavioural data sets used in this thesis. The appendix contains tables with more detailed descriptions of each collection. The data collections used in examples throughout this thesis are referred to by the names listed in these tables, of the form "16 - MorningV" and "35 - Arc2A" which specify the collection index and collection tag name.

Set AR1: collections 1 to 6

The first set of response collections was from the first of a three-session experiment testing the Continuous Audience Response System (CARS) at McGill University in 2009. In this session, 45 university community members rated continuously their felt emotions on one of three scales via iPod Touch devices. Fifteen participants rated the valence of their emotions (negative to positive), on a one-dimensional slider GUI. Fifteen participants rated the arousal of their emotions (weak to strong) on a similar one-dimensional slider. The last 15 participants rated

both dimensions simultaneously by reporting their felt emotions via finger position on a square surface 2D GUI with the horizontal dimension representing emotional valence and the vertical emotional arousal. For the purpose of the following analyses, the 30 arousal ratings and 30 valence ratings are collected together regardless of whether this was the only dimension of emotion the participant reported or one of two. Participants were given a chance using the reporting devices during a training piece, and then their responses were recorded to three stimuli of contrasting genres. The first piece was a Renaissance madrigal, the second one movement from a Romantic string quartet, and the last an Electroacoustic work for a digital instrument, the soprano T-stick. This data set is described in Table 5-1.

Set Kor: collections 7 to 18

The second set of data was collected by Mark Korhonen at the University of Waterloo [Kor04]. The responses recorded were two-dimensional reports of appraised emotion (valence $\times$ arousal), collected using Schubert's EmotionSpace computer GUI [Sch99b], from 35 participants of diverse musical background. The stimuli were six excepts of Classical music from the Naxos recording "Discover Classical Music, Vol. 1". The Kor data set contained 12 collections of 35 responses, with one valence and one arousal collection per stimulus. More details can be found in Table 5-2 of the appendix.

Set Moz: collection 19 to 26

The third set of data was collected from participants attending concert performances of the Boston Symphony Orchestra (one, a live concert, the other, a digital reproduction using stereo sound and HD video presented in a concert hall).

The stimuli were four orchestral excerpts by W. A. Mozart. Participants rated their felt emotional intensity on handheld slider potentiometers with the range marked from "Strong" to "Weak". These eight collections are described in Table 5-3.

Set AoD: collections 27 to 34

These behavioural data were collected as part of the Angel of Death project [MVV+04]. Responses were recorded during two live concerts, one in Paris, France, and the other in La Jolla, California; at each, two groups of participants continuously rated their experience through two versions of Roger Reynold's Angel of Death. One group rated the force of emotion (*force emotionnelle* in Paris) they experienced in response to the music, while the other rated resemblance (*familiarité*). This second measure was a subjective assessment of how much the current music resembled the musical materials presented in the piece up to that point in time. Table 5-4 of the appendix describes these collections in more detail.

Set AR3: collections 35 to 40

This fifth set of experimental response collections were collected at the last of the three-session CARS experiment. In a participating audience of 70, 30 rated continuously their felt emotions in response to the three live performances, using the two-dimensional GUI interface used in AR1. The stimuli were live performances of the three stimuli of AR1, each different interpretations than that of the recordings. There were technical problems with the data collection during this experiment, resulting in number of responses being lost. 17 complete responses

were collected for the first stimulus, 8 for the second and all 30 for the last. These six collections are presented in Table 5-5.

Set Rdm: collections 41 to 44

The last set of response collections are were constructions of unrelated responses. When responses of the same measure to the same stimulus show as such diversity, it is necessary to consider whether these responses may in fact be random, unrelated to the stimulus or task. To challenge this concern, these four unrelated response collections were assembled to explore whether and how analysis techniques differentiate these from the experimentally related responses gather in the collections 1 to 40.

Each of these unrelated response collections took one randomly selected response from each of the 40 behavioural response collections, resulting in a collection in which each response differs from all others in stimulus and/or measure of response. Two of these collections excerpt the first two minutes of the selected responses; two excerpt the last two minutes of each response. The responses were sampled at 1 Hz, using simple downsampling as described in the next chapter, so as to have all responses recorded at the same rate.

These collections took excerpts from either the beginning or the ends of responses to capture what seem to be an artifact of rating responses unrelated to stimuli. As will be demonstrated in the following chapters, rating response collections often show more agreement at the beginnings of responses than throughout the rest of these series. This may be the result of collection devices which are reset to a specific value between stimuli (say the middle of the rating range), or

a consequence of participants' need to listen for a few seconds to formulate their initial response. Whatever the cause, this higher degree of agreement may be a consequence of a stimulus beginning, rather than one stimulus in particular. Endings, in contrast, do not regularly show greater inter-response agreement than the middles of rating responses. By using unrelated response collections aligned by beginnings and by endings, the contribution of this "beginning" effect can be assessed in comparison the the endings collections and its impact can be considered on the experimental collections.

1.3.1 Notation and terminology

To discussion analysis techniques, it is necessary to establish a common language of description. This section presents the mathematical notation used through out this thesis to describe the numerical manipulations of data collections.

A collection of M continuous responses, represented by a capital letter such as X, is composed of synchronously sampled responses $\mathbf{x_r}$ such that $X := \{\mathbf{x_r}\}$, for $r \in \{1, 2, ..., M\}$. In sequence notation, the collection can also be expressed as the ordered set $X = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_r}, ..., \mathbf{x_M}\}$.

Each response $\mathbf{x_r}$, is recorded as a time series, $\mathbf{x_r} := \{x_{r,i}\}$ for $i \in \{1, 2, ..., N\}$, with N being the number of time points at which the response is sampled. Each response can also be written as: $\mathbf{x_r} = \{x_{r,1}, x_{r,2}, ..., x_{r,i}, ..., x_{r,N}\}$.

Lowercase bold font is used to notate both the single continuous response and the time series measured from it. Each datum, in lowercase italics, is a single response sampled at a time point, t_i, such that $x_{r,i} = \mathbf{x_r}(t_i)$. Unless otherwise specified, time points are sampled on a regular interval, $\Delta t = t_i - t_{i-1}$,

$\forall i \in \{2, ..., N\}$, and the sample frequency or sample rate, in Hz, is the inverse of this interval in units of a second.

The responses in a collection X are all sampled on the same one-dimensional measure. Though some experiments simultaneously collect multiple responses or dimensions of response from participants, analyses are generally applied to only one dimension of response at a time. While it is possible, and likely interesting, to study such responses as multi-dimensional time series, for the purpose of this thesis composite data sets are separated into distinct one-dimensional stimulus- and participant-related response collections. The response measure recorded in the series $\mathbf{x_r}$ takes values from $\mathbb{R}$ (the real numbers) though in many cases the range of values is restricted to a finite number of values, a finite continuous interval in $\mathbb{R}$, or transformed to have a particular distribution. Every response is a set of data in two dimensions, namely, time and response measure value. Thus a collection of responses is a set of data in three dimensions: participant or response number (ID), time of measurement, and response measure value. All of the statistics that follow depend on removing information from one or several of these dimensions. The notation used here, and the experiments from which the data are derived, treat the participants and time as independent variables, and the measured response values as the dependent variable to be interpreted.

Many figures in the following chapters will present the assessments or effects of analyses on every behavioural collection in figures similar to figure 1–2. These serve to show variation and trends in collection-wise characteristics and give a sense of the distribution of values or effects possible for response collections

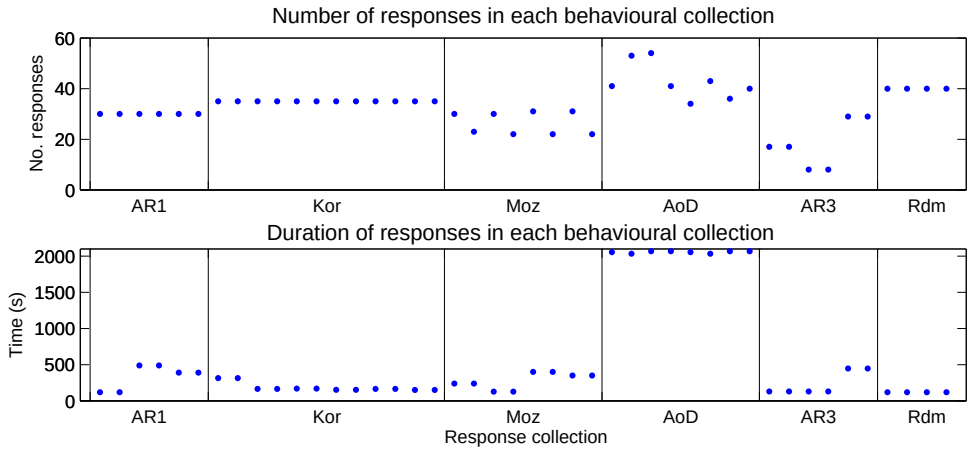

Figure 1–2: Example all-collections graph. Dots represent a datum describing each collection. One graph plots the number of responses included in each collection, while the other presents the duration of each collections' stimuli/responses time series in seconds.

similar to those used here. The x dimension presents the collections by index number, with data sets separated by vertical lines, and the y dimension is in units relevant to the statistics being reported. In figure 1–2, the response numbers graph shows that while most collections are made of responses collected from 30 to 35 participants, a few collections in one set (AoD) include upwards of 60 responses, while two in AR3 have fewer than 10. In the second graph, showing the duration of the responses recorded, the AoD set is again distinct, with an average duration above 2000 seconds per collection while the median response duration is 205 seconds. Other factors that distinguish these data sets are the participants whose responses were collected, the devices used to collect responses (either one-dimensional or two-dimensional, for example), and the focus of measurement (experience or perception). While the numerical analyses that follow treat all

the collections systematically, the distinctions between sets will be discussed throughout, as they are exposed by each analysis.

CHAPTER 2
Continuous response data preparation

Before proceeding to analyses, it is often necessary to massage a collection's responses to better expose pertinent information. These treatments remove information that is considered redundant or irrelevant to the research question.

2.1 Changing sample rate

Time series are, by and large, measurements sampled at a regular temporal interval. It is possible to collect data at thousands of samples per second, however human responses, particularly behavioural responses to music, are not expressed at such high frequencies. Unnecessarily high sample rates give a false sense of precision while complicating many types of analyses. Traditional signal analysis proposes reducing the sample rate of a signal to twice that of the highest relevant frequency, called the Nyquist frequency, to minimize redundancy. At least one study has tried to estimate such a threshold to downsample continuous response data from their quantification of the dynamic information in the stimulus [CJSK+10].

When downsampling, it is important to not go below the frequencies in which signal information is present. For example, the 0.1 Hz sample rate in figure 2–1 seems to skip relevant contour information while the 1 Hz rate is very close to the original 10 Hz signal, with a tenth of the points describing it. For behavioural responses, sample rates used are generally between 0.5 Hz and 10 Hz, while

physiological responses are collected and analyzed at higher frequencies, from 64 Hz to several kH.

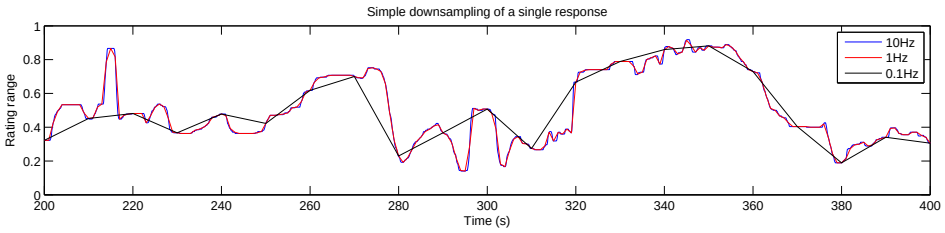

Figure 2–1: An excerpt of one response for the data collection 3-Sch1A from set AR1, sampled at 10 Hz and downsampled (simple) to 1 Hz and 0.1 Hz.

There are several ways to implement downsampling. Given a collection $X := \{\mathbf{x_r}\}$ with each response $\mathbf{x_r} = \{x_{r,i}\}$ for $i = \{1, 2, ..., N\}$ and a sample rate of $1/\Delta t$, downsampling is simplest when the new sample rate is an integer ratio of the original. When the rates are not so aligned, it is necessary to resample to the continuous responses by interpolating between the recorded points. Depending on the response, the best interpolation method may be step-wise (for some rating responses), nearest neighbour (for some categorical assessments), linear (for some other rating types) or higher-order, such as polynomial or spline (for smoother signals). The interpolation method may cause undesirable consequences for some analyses; it is necessary to determine whether connectedness, smoothness, or temporal accuracy is most important before choosing the resampling method.

If the sample rate is to be reduced from $1/\Delta t$ to $1/\Delta t'$ such that $\Delta t'/\Delta t = k$ and $k \in \mathbb{N}$ (k is a positive integer), the simplest method of downsampling a collection is to select every kth element in the response time series. Technically,

this downsampled collection X' would now consist of continuous responses $\{\mathbf{x'_r}\}$ such that $x'_{r,j} = x_{r,i}$ for $i = k * j$, $\forall\, j \in \{1, 2, ...\lfloor N/k \rfloor\}$.

In traditional signal processing, where the spectral content of time series is important, the simple downsampling approach is unsatisfactory because of the potential of aliasing; in highly oscillatory signals, it is necessary to filter the series to remove frequencies above the Nyquist limit to prevent them from inducing apparent lower frequency effects in the downsampled signal. In such cases, a downsampling of collection $X := \{\mathbf{x_r}\}$ by a factor of $k \in \mathbb{N}$ should first apply a low-pass filter $f(\mathbf{x})$ with cutoff of $2 * k/\Delta t$ Hz. This new collection, $X' := \{\mathbf{x'_r}\}$, would then be defined by $x'_{r,j} = f(\mathbf{x_r})_i$, when $i = j * k$ for $j \in \{1, 2, ..., \lfloor N/k \rfloor\}$.

When downsampling a very sparse series, it is sometimes relevant to extract a maximum value or some other statistic from the time interval to be represented by each sample of the new series. In such cases $x'_{r,j} = g(\{x_{r,k*j+l} \mid l = 1 - (k/2), ..., -1, 0, 1, 2, ..., k/2\})$, where g may be a function to determine the mean, the maximum, the minimum, the median, or some other characteristic of the time interval. The interval to be represented may be entirely forward or backward in time, or balanced around the associated time point as in the above formula. These more complicated methods of downsampling are a form of time series feature extraction, with the resulting time series reflecting some aspect of the initial series rather that trying to be the "closest" series at the new sample rate.

2.2 Filtering

Removing superfluous or irrelevant data from experimental results is often necessary. If information is being removed from a data set, the criteria must be

clearly articulated, justified, and systematically implemented. One type of data removal is filtering, a transformation which removes specific frequencies from a series.

Besides downsampling, it is fairly routine in signal processing to filter unwanted frequencies from signals when the interesting frequencies have already been identified. For physiological responses, filtering is necessary to remove sensor and measurement effects and to expose the relevant temporal variation in the signal. One aspect of filtering which is particularly important for the analysis of coordinated continuous responses is the filter's effect on the phase of frequency components. Standard linear filters attenuate unwanted frequencies, but in the process of reducing these frequencies' amplitudes, they also affect the amplitude and phase of the preserved frequencies. When filtering using cutoffs close to a rate of interest to temporal analysis, say below 3 Hz, the phase effects of filtering can move local peaks by tenths to tens of seconds. This scale of shifting can interfere with interpreting the alignment of responses with the stimulus and other signals. It is possible to avoid this by using filters designed to be phase-linear, because they preserve the phase of all frequency components. For those unfamiliar with filter manipulation, one way to perform a phase-linear filtering is to pass the signal through a standard linear filter twice, the second time in reverse linear time. This results in a sharper frequency cutoff having the desired effect on the amplitude of frequency components while reversing the phase effects of the first pass to leave all spectral components in their initial orientation.

2.3 Removing outliers

Similar to filtering, the purpose of removing outlier responses or data from a data set is to remove excessive and irrelevant variation from the experimental results. When a blood volume pulse sensor is attached too tightly or an EMG sensor is not grounded properly, the data collected deviates wildly from what is expected. Such outlier responses are easy to identify and discard, but other definitions of outliers are less obvious.

For behavioural responses, it is fair to exclude the data collected from participants who misunderstood the task or refused to perform it, but identifying such cases from the collected responses is not always straight-forward. Experiments can be designed to catch unreliable responses and evaluate response reliability: some experimenters ask participants to report whether their expressed responses matched their experience [MF93] [CS92], others have some of the participants respond multiple times to the same stimuli [Gre94], sometimes in different sessions [GNKA07b] [CJSK$^+$10].

After data collection, a few criteria have been used to exclude responses that appear to be too deviant. One type of outlier for rating data is the flat-liner: a response which appears not to change for minutes at a time [MVV$^+$04]. Post-experiment, it is not possible to discern whether the invariance in response is a true stasis in experience or a failure of the participant to perform the task. Rather than risk increasing variance with false data, such ambiguous rating responses are discarded because they are quantifiably different from the rest of the collection. Another method is to exclude participants with low average inter-response

correlations. Luck et al. excluded responses which, on average, correlated with other responses at $r < 0.2$ [LTE+08], again using demonstrable deviant criteria.

There is at least one instance of non-filter within-response outlier removal in the literature. In his work on modelling of continuous ratings of emotion in music, Mark Korhonen assessed the response collection for outliers by time point [Kor04]. For time points when less than 10% of response values were more than 2 standard deviations from the mean of that moment, those distant response values were excluded as outliers. Responses in which more than 10% of response values were excluded by the previous criterion were removed from the collection. Korhonen reported that this process reduced the variance in his data set of 6 collections by an average of 5% by removing less than 1% of the recorded data. While this process significantly reduces the measured variance in one experimental data set, it also carries the assumption that it is legitimate to discard these deviant responses in favour of a supposedly ideal central tendency to best describe these continuous responses. This position is challenged in Chapter 4.3, on clustering responses within data collections.

In the behavioural data sets used here, responses compromised by faulty equipment were discarded, and flat-liners showing no changes in ratings over the entire response were removed.

2.4 Normalization

For some continuous responses, the contour of the signal is more interesting than the values measured at each moment. For physiological responses such as heart rate, participants naturally have different distributions of values taken

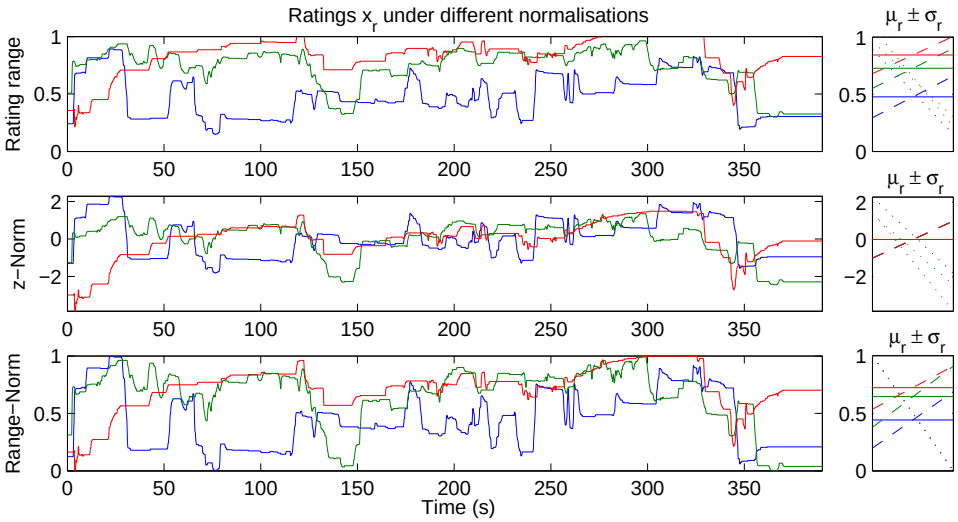

Figure 2–2: Effects of normalization. On three felt emotional arousal ratings from 5-Ste1A in AR1, the left plots show these on their original rating scale and after range and z-normalizations. Graphs to the right show the mean (solid horizontal lines), the mean +/− the standard deviation (ends of dashed lines), and the min and max values (ends of dotted lines) for each of the responses.

over time, and these inter-individual differences should be minimized before interpreting the aggregate. In rating responses, participants may be more or less expressive of the variation they feel or perceive, or they may hold a device differently, impairing the useful range of the instrument. These participant-wise differences are important for some analyses while in others, they are irrelevant to the information being sought. Normalization manipulates a response by scaling the range of values measured to simplify some kinds of comparisons between individual responses, while preserving the rank and contour of elements within each time series. There are two types of common linear transformations for normalization of responses within a collection: normalization of range—shifting and stretching

each response so that its maximum value is one and minimum value is zero—and z-normalization, which sets each response's distribution to have a mean of zero and variance of one.

From the initial response collection X, the range-normalized collection $Y := \{\mathbf{y_r}\}$ with $r \in \{1, 2, ..., M\}$ and $\mathbf{y_r} := \{y_{r,i}\}$ for $i \in \{1, 2, ..., N\}$ would be calculated as:

$$y_{r,i} = \frac{x_{r,i} - min(\mathbf{x_r})}{max(\mathbf{x_r}) - min(\mathbf{x_r})} \tag{2.1}$$

Thus each signal, unless it never changes value, will have an absolute minimum of 0 and an absolute maximum of 1. Figure 2–2 shows the effect of this kind of normalization on three continuous rating responses from the original rating scale of the top graph to the range-normalized result in the bottom graph. The small graphs to the right show how the means of each response (horizontal solid lines), and the standard deviations (dashed lines) shift with this transformation while the range of responses (dotted line) stretches the full range from 0 to 1. This type of normalization is convenient when contour is the most important aspect of the signal and different time series in the collection occupy arbitrary ranges. When responses each take values sufficiently uniformly over some interval on the measure range, this transformation is most appropriate. For rating responses where the midpoint of the scale is marked and the extremes are defined as opposite poles, it may be more appropriate to preserve the distinction of "above" or "below" this threshold. An alternative normalization for such situations, presuming this

midpoint has a value of 0 in the initial ratings, would be:

$$\mathbf{y_r} = \frac{\mathbf{x_r}}{max(\{|x_{r,i}|, \forall i\})} \tag{2.2}$$

Dividing by the maximum of the absolute value of each series stretches the response values into comparable ranges without shifting the centre.

For some physiological signals, the range and distribution of response values collected may vary greatly between participants, sometimes containing extreme values which dwarf the variation in most of the series. Rather than force all responses to the same range, these signals can be best standardized with z-normalization. If multiple responses are collected from the same participant over a single experiment session, it is useful to treat all of these data together to best capture the full range of the signal and thereby reduce the risk of amplifying insignificant jitter. From the initial response collection X, the data of the z-normalized collection Y would be calculated as

$$y_{r,i} = \frac{x_{r,i} - \mu_r(X)}{\sigma_r(X)} \tag{2.3}$$

This transformation does presume a normal distribution of values within each series. In figure 2–2, the middle graph shows the effects of this kind of normalization. The units of the z-normalized scale are standard deviations, as each response has been transformed to show the same spread of variance over the time series (overlapping dashed lines). It is worth considering the actual distribution of the measured values in the individual response series along with the intended analyses

when deciding whether and which form of normalization would be appropriate for preparing the response collection.

2.5 First-order difference, auto-regression residuals and other features

If the contour is more interesting than the original values of responses, one direct way to study it is through the first derivative of each continuous response. Using functional representations of the data, the derivative can be studied directly, as had been done for analyzing familiarity ratings in the Angel of Death project [MVV+04]. When working with a digital sampling of the response, the simplest numerical approximation of the derivative is the first-order difference series which reports the change of value from one sample to the next. From the initial response collection X, the first-order difference collection, $Y := \{\mathbf{y_r}\}$ with $r \in \{1, 2, ..., M\}$ and $\mathbf{y_r} := \{y_{r,i}\}$ for $i \in \{1, 2, ..., N-1\}$, would be calculated as:

$$y_{r,i} = x_{r,i+1} - x_{r,i} \tag{2.4}$$

First-order difference calculations using more points can also be used to approximate the derivative of continuous responses more precisely when the samples are sufficiently dense. For behavioural responses, however, the sample rate is too low to benefit from more sophisticated techniques. Similarly, higher order differences can be taken to approximate higher order derivatives, when the sample rate is sufficiently high for these estimates to be considered helpful.

Besides describing how a response changes over time, this transformation is a useful representation of continuous response because it can greatly reduce the effect of serial correlation in continuous responses. Time series data usually cannot

be analyzed by traditional forms of discrete statistics because they fail to satisfy the necessary requirement of independent sampling. In continuous responses, the response measured at time t_{i+1} is most often very close in value to the response measured at time t_i, and this proximity demonstrates that the values recorded in these series are not sequentially independent. When applying statistics to assess the distributions of these series, the serial relationship between data points violates the assumptions of common tests and this results in misleading estimates of significance.

Time series analysis has developed techniques to remove the problem of serial correlation from series while minimally compromising relevant information. For example, given a time series $\mathbf{x} := \{x_i\}$, auto-regression analysis constructs a linear model of a series' elements x_i from elements x_{i-j} for various integer values of j. The difference between such a model and the actual series, the residual, is often a (weakly) stationary series which supposedly only contains noise or information from outside the model.

The first-order difference transformation is equivalent to the residual of a basic first-order auto-regression model. More sophisticated auto-regression analysis may pull out higher-order temporal relationships, but for behavioural responses, the sequential subtraction at around 1 Hz does a great deal to reduce serial correlation. Figure 2–3 shows a continuous rating response and its 1 Hz first-order difference series. Their respective distributions, to the right, show how the difference series has values very close to zero more often than not and this series is visibly more stationary that the original. In the original series, most adjacent

Figure 2–3: One continuous response from collection 5-Ste1A and the first-order difference of the response when sampled at 1 Hz. The small plots to the right show the distribution of values of each series over their respective ranges.

time points are close to each other in value and rarely cross the mean of the series, while in the transformed series it is not easy to predict from one point whether the next will be of similar value, zero or of opposite sign. This representation of continuous response time series has been used by Schubert and others for regression and correlation analyses [Sch02], though others have not found it to be useful for relating responses to time series representations of stimuli [LTE+08].

Many other transformations are possible and necessary for physiological response analysis. There are several ways of estimating heart rate from blood volume pulse, and respiration rate from chest expansion, most involving the detection of minima or maxima and calculating their period from one cycle to the next.

2.6 Preparing collections for comparison

Numerically, when comparing time series in time, it is necessary to have them sampled at the same rate for the same number of samples. It is also relevant to expose the information that these series contain in similar ways. A series or collection of series which has been z-normalized may not be easily modelled by range-normalized time series because of the way the values are distributed. Similarly, a stationary series will not correlate well with a non-stationary series, even if the relationship would be clear after the removal of serial correlation. There are many factors that go into deciding the methods of data preparation, both for responses and stimuli representations, but among them should be consideration of the rate and character of information to be compared.

CHAPTER 3
Traditional analyses

There are some analysis techniques which have been applied to continuous responses for decades with some degree of success but little criticism. This chapter presents commonly applied methods for summarizing and comparing continuous response collections, with examples and comments derived from the application of these techniques to the six data sets.

3.1 Cross-sectional distribution time series: mean and dispersion

3.1.1 Normal statistics

Of all the possible methods for generating a single representative time series for a collection, the most popular for continuous responses is to take the average at each sample point. Figure 3–1 shows the distribution of response values recorded each time point which are expected to be represented by their average (red dashed line). For a set of continuous responses $X = \{\mathbf{x_r}\}$, the average time series of the collection $\mu_{\mathbf{I}}(X) := \{\mu_i(X)\}$ is defined, for all $i \in \{1, 2, ..., N\}$:

$$\mu_i(X) = \frac{\sum_{r=1}^{M} x_{r,i}}{M} \tag{3.1}$$

To distinguish this from means calculated over other dimensions, the result of equation 3.1 can be called the cross-sectional average as it is the average over successive slices of the collection, cross-sections of the responses per time point [Say89].

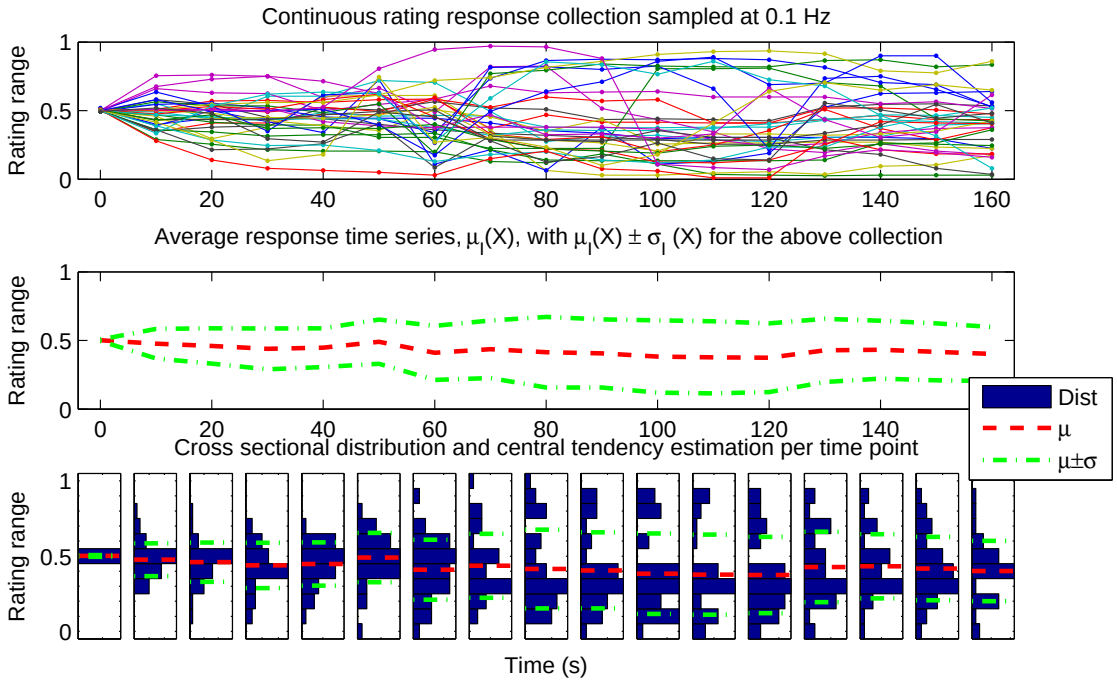

Figure 3–1: Example of cross-sectional distributions in the collection 10-AranjuezV of the Kor data set. Responses in the collection were downsampled to 0.1 Hz and plotted in the top graph with dots marking each datum. The middle graph presents the cross-sectional average with response spread measured by the cross-sectional standard deviation. The distributions of response values per time point (or cross-section) are plotted below with the related mean and standard deviation.

As a descriptor of the distribution, the standard deviation of the cross-section can also be calculated per time point, $\sigma_{\mathbf{I}}(X) := \{\sigma_i(X)\}$:

$$\sigma_i(X) = \sqrt{\frac{\sum_{r=1}^{M}(x_{r,i} - \mu_i(X))^2}{M}} \tag{3.2}$$

The standard deviation measures the average distance of elements in a set to their average. This statistic is very well understood and useful when the set it is

describing has a Gaussian distribution, a.k.a. is normally distributed. However, as shown in the cross-sectional distributions of the response collection in figure 3–1, the set of response values in each cross-section would not always pass tests of statistical normality. In such cases, the average and standard deviation time series are not the most useful descriptors of cross-sectional distribution, being only very rough measures of the central tendency and dispersion of responses at each time point.

3.1.2 Non-parametric statistics

To get around the problem of assuming normality, the non-parametric statistic of the median has been proposed as an alternative for these cross-sectional assessments. The cross-sectional median has been used instead of the mean in a few analyses [GNKA07a][Kor04]. The median time series of the collection $X := \{\mathbf{x_r}\}$ is here notated as $\mu_{\mathbf{1/2\,I}}(X) := \{\mu_{1/2\,i}(X)\}$. The cross-sectional median time series is defined, for all $i \in \{1, 2, ..., N\}$, as:

$$\mu_{1/2\,i}(X) = min_{r\in\{1,2,...,M\}} \left(\left\{ x_{r,i} \text{ such that } \frac{\|\{q \in \{1, 2, ..., M\} \mid x_{q,i} \leqslant x_{r,i}\}\|}{M} \geqslant 1/2 \right\} \right) \tag{3.3}$$

The median is the middle-most value of a set, or in this case, of a cross-section of responses.

A non-parametric alternative to the standard deviation time series is the interquartile range. Quartile statistics, notated as $\mu_{\mathbf{1/4\,I}}(X)$ and $\mu_{\mathbf{3/4\,I}}(X)$, are defined similarly to equation 3.3, only replacing the 1/2 threshold with their respective 1/4 and 3/4. To get a single statistic describing the dispersion in each cross-section, the interquartile range is simply the difference between these

statistics. For the purpose of comparisons, however, the author finds it more convenient to use the half interquartile range, $\mu_{3/4-1/4\,I}(X)/2$. For each time point i, this would be :

$$\mu_{3/4-1/4\,i}(X)/2 = \frac{\mu_{3/4\,i}(X) - \mu_{1/4\,i}(X)}{2} \tag{3.4}$$

Using again the downsampled response collection 10-AranjuezV, figure 3–2 shows how the normal statistics compare to the non-parametric statistics at each cross-section/time point.

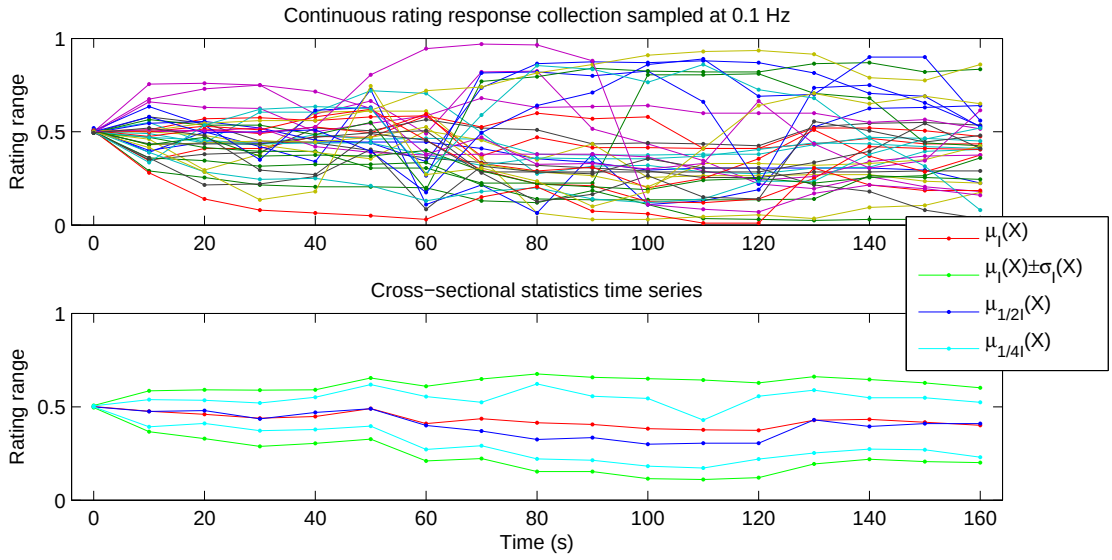

Figure 3–2: Example of cross-sectional distributions in a collection 10-AranjuezV. Bottom graph compares the parametric $\mu_I(X)$ and $\sigma_I(X)$ with the non-parametric $\mu_{1/2\,I}(X)$, $\mu_{1/4\,I}(X)$ and $\mu_{3/4\,I}(X)$.

These statistics, parametric and non-parametric, are meant to capture what is called the "central tendency" of the cross-sections of these responses. They select a representative value for each set, the average or median, and measure the quality

of that estimate according to their corresponding rules for describing dispersion, standard deviation and interquartile distance. These methods for summarising a response collection presume that the experimental results are noisy deviations from a single ideal response which they are designed to identify. Looking back at figure 3–1, it seems as though the distribution of responses per time point is bimodal from around 70 seconds to 120 seconds. In this stretch, the median performs better than the average because it represents common values by sticking with the dominant trend (below 0.5), but neither representation succeeds in warning of a possible split in response behaviour.

3.1.3 Effectiveness of cross-sectional distribution descriptors

The average time series is attractive as a first analytic step: it is expected to capture the dominant dynamics of response despite inter-subject differences. This series gives a rough idea of whether the ratings tend to be high or low and whether the responses shift strongly at some point in the stimulus. It is presumed to preserve variation in time that is representative of the participants' dynamic responses. This variation over time is an important clue for evaluating how effectively this statistic summarizes experimental data. The following analysis looks at trends across response collections in the distribution of values in the cross-sectional average and median time series, and in how they compare to those of the dispersion measures.

Figure 3–3 presents these summary time series for three behavioural response collections: one of perceived emotions, one of felt emotions, and another of unrelated continuous responses. The standard deviation series of these three contrast

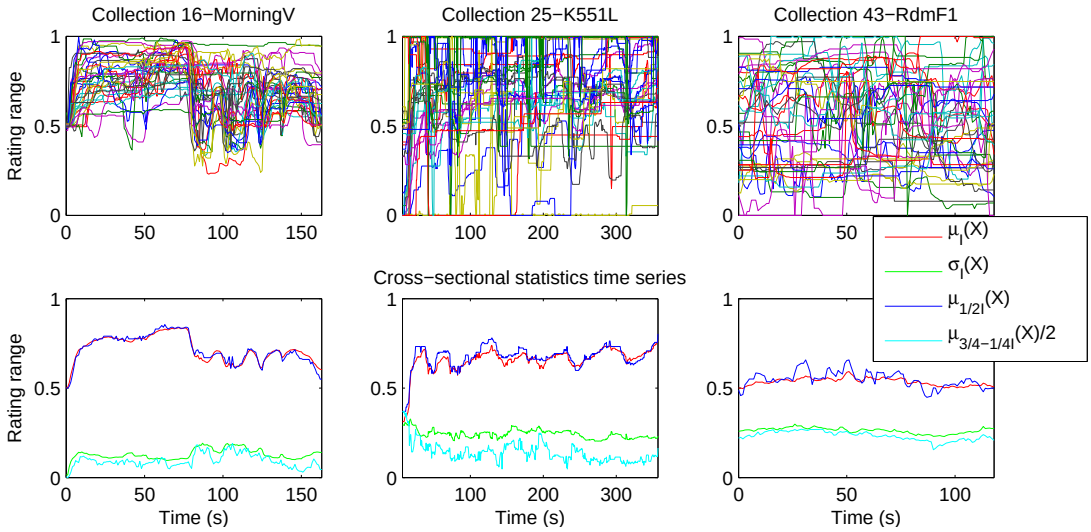

Figure 3–3: Central tendency summary time series for collections 16-MorningV in Kor, 25-K551L in Moz, and 43-RmdF1 in Rdm.

predictably. The variation is generally lower for ratings of perceived emotion than in the collection of felt emotion ratings, and the unrelated responses also have relatively high standard deviation values. The differences in the cross-sectional distributions of the felt and the random collections is more clearly marked by the half interquartile distance, in light blue. For the unrelated responses, the two measures of dispersion are very close, while the spread of responses in the felt response collection looks to be in a range more similar to the perceived responses when using the non-parametric measure.

In all three examples of figure 3–3, the range of the values taken by the average is a fraction of the range employed by the individual responses in the collection. In the felt emotional response series, the variation in time is mostly less than the standard deviation measured for most cross-sections. The unrelated

responses central tendency series are flatter still, with high standard deviation for all time points. Given that the mean time series is expected to be meaningless in the last case, are the mean time series of the other collections demonstrably more informative?

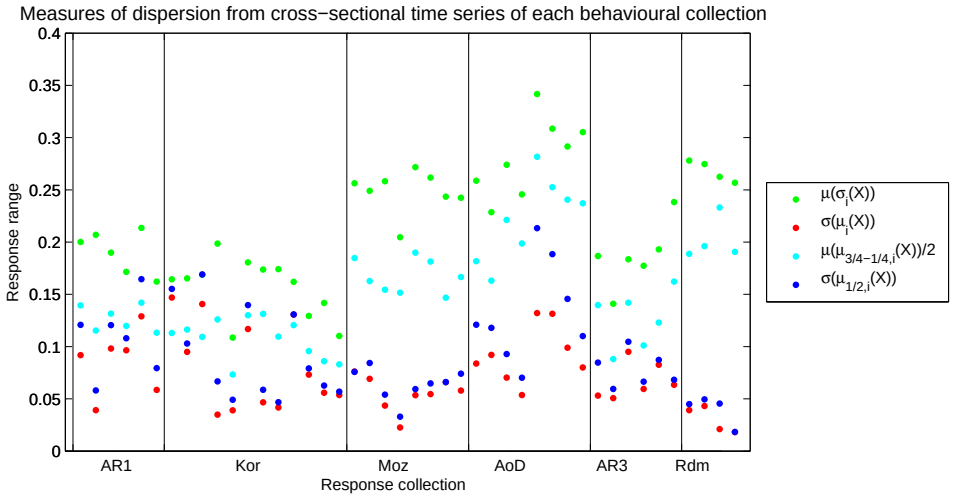

Figure 3–4: Dispersion statistics of cross-sectional distribution time series for all behavioural response collections. Measures of either standard deviation of central tendency time series or means of dispersion measure time series are compared on units of the rating ranges $[0, 1]$.

To get a sense of the behaviour of these statistics across all of the data sets, figure 3–4 presents four summary measures of dispersion as applied to each collection. The spread of values suggests many things, but three trends in particular relate to the discussion of the information contained in these central tendency statistics:

1. $\sigma(\mu_{\mathbf{I}}(X)) \leqslant \sigma(\mu_{\mathbf{1/2\,I}}(X))$: The variation of the cross-sectional average over time is generally smaller than that of the median. The median and other

percentile measures of distribution are increasingly popular statistics for continuous response analysis because they are presumed to be more robust to variation in outlier responses. For discrete data sets with non-parametric distributions, the median does indeed prove to be more robust to outlier data, but in a time series, the median does not show greater sequential stability than the mean. While the series are close to each other in the examples of figure 3–3, the median is noticeably steeper climbs and falls. The median time series changes value in two cases: when more than half of the responses change simultaneously in the same direction, and when one or more responses cross the median line (see equation 3.3). In the collections considered here, the second case is much more common than the first, resulting in a jagged median time series as it jumps between neighbouring response values from one cross-section to the next every time an individual response crosses current median. The mean, on the other hand, is deceptively smooth while it shifts slightly with every change in individual responses. Despite the consistent ordinal relationship of the variance of the mean being smaller than that of the median, these values do not correlate significantly across collections. Since the cross-sectional response values do not usually have a normal distribution, the median values may be more appropriate, and appear (when averaged) to be sensitive to distinct information in these response collections, although the details of the contour of this series should be interpreted with its own particular sensitivity to individual response behaviour in mind.

2. $\mu(\sigma_{\mathbf{I}}(X))$ & $\mu(\mu_{\mathbf{3/4-1/4\,I}}(X)/2)$: Although the non-parametric measure of dispersion used here is always somewhat smaller than the standard deviation, the two values are highly correlated, Pearson $r = 0.94$ with $p \ll 0.001$. Of interest is the way these descriptors of distribution change across the different data collections. The unrelated response collections (Rdm, the right-most set in figure 3–4) show high cross-sectional dispersion, but not higher than all experimental response collections. If the dispersion measured in unrelated collections determined a threshold for "acceptable" inter-response disagreement, a fifth to a quarter of the actual response collections would exceed it. That some collections show greater variance than these random collections may be an indication that the single central tendency model is not appropriate for capturing trends in responses within each cross-section. Though not included in the graph above, the variance of the half-interquartile distance time series is also generally greater than its parametric cousin's. Its contour is similarly sensitive as the median time series, and should be interpreted for gross rather than fine contour information unless the behaviour of individual responses have been considered for context.
3. $\mu(\mu_{\mathbf{3/4-1/4\,I}}(X)/2)$ versus $\sigma(\mu_{\mathbf{1/2\,I}}(X))$: While the difference between the parametric measures of dispersion is more dramatic than the non-parametric in figure 3–4, the latter may prove to be more interesting. The largest differences between the standard deviation of the median time series and the average of the half-interquartile distance series are seen in the unrelated collections, where the spread of independently varying responses flatten the

central tendency series into stillness. But these two statistics are very close in a number of other response collections, some in which standard deviation of the central tendency time series exceeds the average of the dispersion measure time series. Such cross-overs seem to be determined by the greater variability of the median rather than exceptionally low interquartile ranges. Future analyses should explore the possibility that the responses in such collections are more consistent than those with greater distance between the standard deviation of the median time series and the average of half-interquartile distance time series.

The median and quartile time series are more explicit measures of the distribution of response values in each cross-section, and the longitudinal distributions of these time series do not seem to be any less reliable than their parametric counterparts for estimating the overall variability in these collections. When discussing the central tendency time series, the following pages of this thesis will most often refer to the parametric measures rather than the non-parametric, but only because they are more commonly used in the literature. The analysis of this section strongly recommends using the nonparametric time series, except, perhaps, for the purpose of standard correlations (which will be discussed later) and tight regression fitting because of the sensitivity of these summary series to crossings by individual responses.

3.1.4 A note on graphical assessment

Some papers stop the numerical analysis once the cross-sectional average response time series has been calculated and continue the investigation using visual

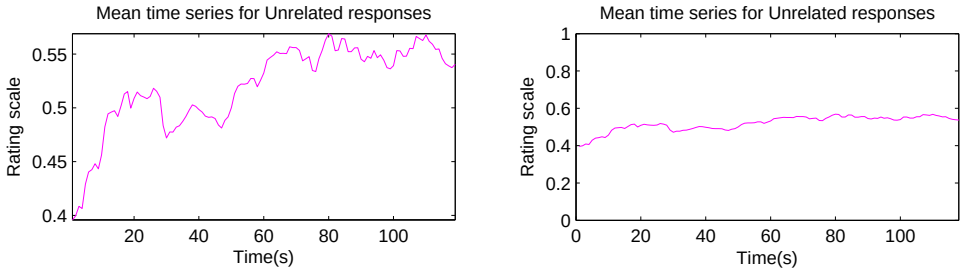

Figure 3–5: Average time series of 41-RdmI1, a collection of unrelated responses, shown on a subsection of the rating range and on the full range $[0, 1]$.

analysis of the plotted mean. One proponent of this tact was Frede Nielsen, who wanted to avoid the "atomistic approach" which compared aspects of the stimuli to the responses directly and tempted researchers to generalize relationships too far [Nie87]. Vision can be misleading when interpreting non-spatial data, but looking at the summary time series may stimulate analytic intuition. No matter the mode of analysis which follows the calculation of this summary of a collection, the random response collection mean shown in figure 3–5 is a reminder of how easy it can be to see significant variation in this time series when there is none. Plotting such series on a graph scaled to show the full rating range puts the details of its contour in perspective and can discourage investing too much before employing more impartial methods. At this time, there is no rule for what size of change in average can be trusted as significant, and until such limitations can be articulated, skepticism of these time series is reasonable.

3.2 Discrete statistics on continuous responses

Continuous response data are recorded as time series but they are not always analyzed in time. By summarizing each response series into one or a few discrete values, the issues of serial correlation and temporal alignment are conveniently put aside. Though these longitudinal summaries can be used to describe a collection of responses, most often they are calculated in order to compare collections of responses related by some experimental factors, such as stimuli, and differentiated by others, for example participants' musical expertise.

3.2.1 Response-wise statistics

For a collection X, statistics describing each response time series $\mathbf{x_r}$ are calculated to describe the collection in a distribution of points $f_R(X) := \{f_r(X)\}$ for $r \in \{1, 2, ..., M\}$. For example, the longitudinal averages of a collection form a set of the average value of each response. Formally, $\mu_R(X) := \{\mu_r(X)\}$ for $r \in \{1, 2, ..., M\}$ such that:

$$\mu_r(X) = \mu(\mathbf{x_r}) = \frac{\sum_{i=1}^{N} x_{r,i}}{N} \tag{3.5}$$

Such averages can be referred to as longitudinal or response-wise averages to make explicit the dimension of data being summarized. Though very similar to the calculation of the cross-sectional average time series described by equation 3.1, averaging per response across time results in a set of very different character. For one, μ_R is a set of independent samples: the value of μ_r does not determine the

value of μ_{r+1}, in contrast to μ_i which is a strong determinant of μ_{i+1}. Independence of samples makes the full suite of discrete statistical techniques applicable to this type of collection reduction.

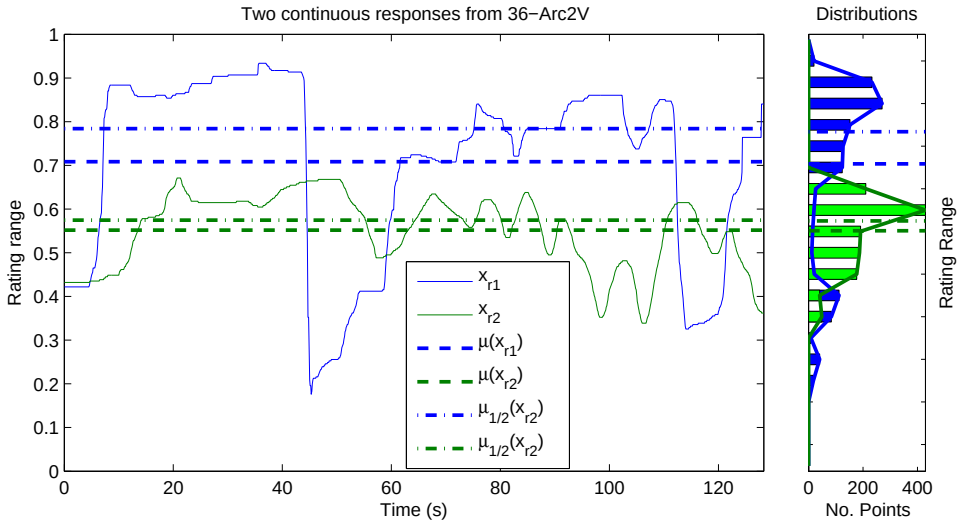

Figure 3–6: Calculation of the mean or median values per response on two example rating responses from collection 36-Arc2V from AR3.

Like the cross-sectional statistics, parametric measures of distribution on individual responses are often inappropriate because the sampling of values by a response on the rating range is not necessarily normal. Figure 3–6 shows how these responses can be bimodal or skewed away from symmetric distributions properly represented by the arithmetic mean. The median may fare better with its simpler relationship to the distribution, but either way, these central tendency distribution descriptors do not tell the whole story.

Once the calculation per response has been performed, the distribution of these longitudinal statistics are then subject to statistical analysis. Figure

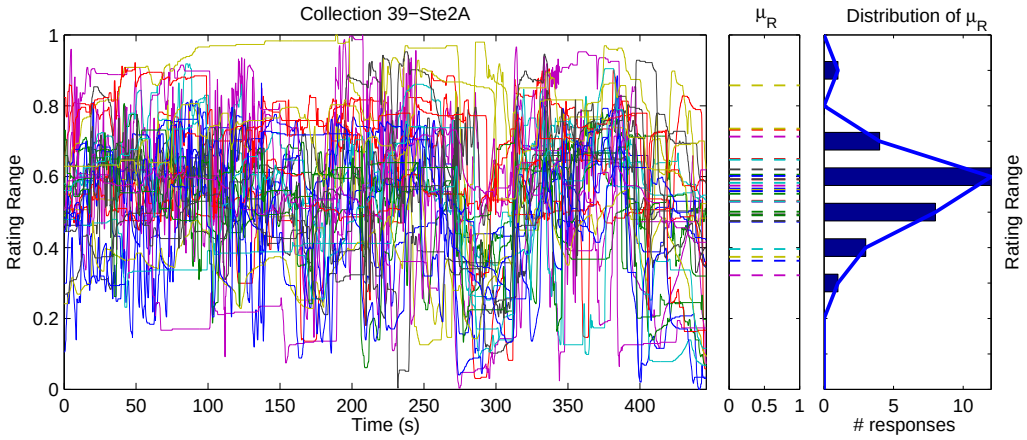

Figure 3–7: Collection 39-Ste2A, continuous ratings of felt emotional arousal from AR3, with response-wise averages (dashed lines) and the distribution of these longitudinal statistics for this collection.

3–7 illustrates the distribution of response-wise averages for a collection of felt emotional arousal ratings. The spread of these response-wise statistics suggests that distinct experiences may be presente in the collection.

3.2.2 Longitudinal distributions and interrupted time series

Other kinds of longitudinal descriptions of response evaluate how many samples, i.e. how much time, participants' responses spend in subsections of the range. One example of this in the literature collected categorical behavioural responses. The categorical measure of focus of attention asked participants to indicate which elements of the music they were attending to at any given moment [MG90] using a slider-potentiometer interface. The resulting collections were compared by evaluating how much time participants spent on Rhythm or Timbre from one stimulus to the next. For responses collected on a finite continuous scale,

a similar question would be what proportion of sample points per response are greater than or equal to 75% of the scale range.

Whether of average values or more complicated descriptors of response behaviour, these response-wise calculations have been very important for comparing responses of the same participants to different stimuli. This approach is used on physiological responses as well as behavioural responses [IM99][Kru97][Ric04]. Related to stimulus-wise comparisons, some studies cut long stimuli into chunks on the order of a minute to make comparisons between successive excerpts [IM99]. When responses are summarized over stretches of time, this approach is in some sense a kind of downsampling, with intervals between time points chosen to greatly reduce any expected effects of serial correlation.

Another short excerpt approach is to evaluate responses around or following some event. Events in the series can be stimulus defined, such as the onset of silences in Lisa Margulis's studies of tension through silence in classical piano repertoire [Mar07], or they can be participant defined, such as button presses indicating pleasure level in chill studies [SBL$^+$09]. While the responses are sampled in time, their relative positions are not considered in the analysis. One way to analyse the response around these events is called Interrupted Time Series Analysis [SD04]. This technique compares time series values before and after events to evaluate the consistency of effect and has been used to compare, for example the strength of changes in familiarity ratings at different types of section boundaries [MVV$^+$04].

3.2.3 Effectiveness of response-wise statistics

Longitudinal summaries of response collections can seem like a waste of resources. For behavioural responses in particular, the task of rating continuously is quite demanding on participants while the analysis makes use of only one datum per continuous response. Given the continuing use of post-stimulus Likert scale ratings, it has been necessary to consider whether or not these summarized time series are better descriptions of participants' responses to music than the less-technologically demanding alternative. While there are similarities between the responses from static post-stimulus measures and longitudinal averages of continuous measures, studies have found that they are not equivalent [BS95], particularly for stimuli which provoke more varied responses over the course of continuous responses [DC01].

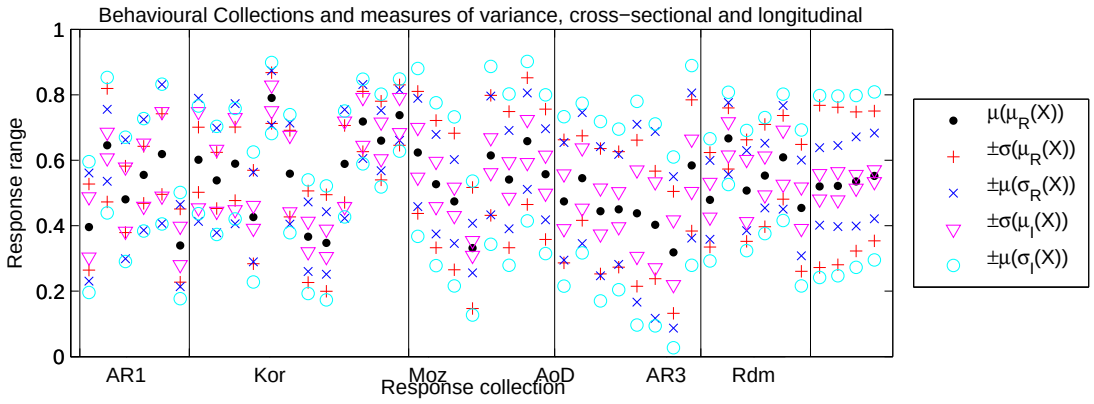

Figure 3–8: Average rating value, $\mu(\mu_R(X))$ for each behavioural collection with four measures of collection dispersion, both longitudinal and cross-sectional variation measures marking the "spread" of response values range around each $\mu(\mu(X))$.

To get a sense of how the longitudinal summaries describe behavioural collections of responses and how these compare to some of the cross-sectional statistics, figure 3–8 displays results of these descriptions of response collections from all six data sets. Each column of symbols describe the distribution response values in a single collection. Centred around the black dot of the average of the average are plus and minus the standard deviation of the response-wise averages in red crosses, the average of response-wise standard deviations with blue exes, the standard deviation of the average time series with pink triangles, and the average cross-sectional standard deviation with light blue circles. Like figure 3–4, there are a few trends in these statistics that deserve some attention:

1. Average rating value per collection, (n.b. $\mu(\mu_R(X)) = \mu(\mu_{\mathbf{I}}(X))$): In the collections considered here, this average central value ranges between 0.3 and 0.8 of the normalizes rating range. This spread of average values is consistent with the significant differences found between collections using ANOVA on response-wise averages reported in many papers to date. By the nature of these data sets, it would be very rare to have this statistic move closer to the extremes of the rating range. While the average of the average is a very poor representative of a continuous response collection, there are some tantalizing trends. For example, AR1 and AR3 are collections of the same measures of response from different participants on stimuli which were different interpretations of the same three musical pieces. Despite their differences, these means of means show the same pattern of values. Quantifying the

significance of such results is possible with relatively accessible discrete statistics when using the set of response-wise averages.

2. $\mu(\mu_R(X)) \pm \sigma(\mu_R(X))$: The measure of dispersion used to compare the longitudinal distributions is marked in figure 3–8 by red +'s. Adding this to the consideration of the differences in means' mean within datasets, there are some differences of distribution of responses which look significant, at least under these parametric assumptions.
3. Ranges of $\mu(\sigma_R(X))$ (x) and $\sigma(\mu_{\mathbf{I}}(X))$ ($\bigtriangledown$): For the most part, the average longitudinal standard deviation of responses is greater than the standard deviation of the cross-sectional averages. This demonstrates how the average time series is flatter than most responses it is meant to represent, its range reduced by the distance and asynchrony between individual responses. In the collections considered in figure 3–8, those of the Rdm set have the most dramatic differences in these two statistics, as expected. Collections with proportions of $\mu(\sigma_R(X))/\sigma(\mu_{\mathbf{I}}(X))$ similar to these of the unrelated sets should be treated with some skepticism, such as the 4th collection in the Moz data set, 22-K16R.
4. $\mu(\sigma_{\mathbf{I}}(X))$ (O) and $\sigma(\mu_R(x))$ (+): When the standard deviation of longitudinal averages is much smaller than the breadth of variation measured at the average time point, this suggests that an important difference between responses (and contributor to the second statistics) in this collection is the breadth of the rating range used in each, rather than disagreements as to where responses are centred on the rating scale. When these two statistics

are very close, this suggests that responses are occupying different parts of the rating scale while making use of similar proportions of the rating range. The distributions of response-wise averages are not so deviant as to be found significantly not-normal by the single sample Kolmogrov-Smirnov goodness-of-fit hypothesis test, however, as shown in figure 3–6, parametric statistics are not necessarily the most useful descriptors of the distribution of values in individual response time series. The medians of the medians for these behavioural response collections are almost always more divergent (further from 0.5) that the corresponding means' means. Selecting the right longitudinal summarizing function can effect differences measured between collections, and the parametric default may not be appropriate.

While statistical techniques can quantify the significance of the differences in this conveniently discrete representation of continuous response collections, these techniques do little to explain the results. Investigation into the time course of responses is necessary to explore how these differences in the distribution of response-wise means are manifest.

3.3 Correlations on continuous responses

Pearson correlations have been used in many publications to compare individual responses [Fre99b], average response time series, and continuous response to time series representations of stimuli. Earlier uses of the Pearson Product Moment Correlation Coefficient (PPMCC) were intended to assess the reliability of individual participants responses: some participants were asked to perform the rating task a second time and the correlations between these first and second

ratings were averaged across this subgroup of participants to gauge the validity of the task[Gre95]). In the mid-nineties, publications began to share average inter-response correlation values as a measure of how consistent responses were between participants so as to demonstrate whether "most listeners showed the same general patterns" of response [Kru96]. This practice of average inter-subject correlations was possibly borrowed from analysis techniques for discrete rating response sets [PK87]. Collections' average time series have been also widely used to assess the temporal relationship between continuous responses, exploring both the strength of covariance and cross-correlations to evaluate lags in between responses and stimuli. This next section will discuss the correlation of time series with a particular focus on inter-response comparisons.

3.3.1 Correlating time series

Mathematically, the correlation coefficient can be expressed and interpreted many ways [RN88], but the following definition is most pertinent to the discussion of time series. For two paired sets of values $\mathbf{x} := x_i$ and $\mathbf{y} := y_i$ for $i \in I = \{1, 2, ..., N\}$, the Pearson Product Moment Correlation Coefficient, r is defined as:

$$r_{\mathbf{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.6}$$

In words, $r_{\mathbf{x,y}}$ is a measure of the amount variation from their respective means that is shared between the paired values of the two sets. Though the index in the calculation orders the pairing of values between sets, the additive series is not sensitive to the sequence of the pairs and thus is blind to the order of time series. The comparison is also unitless: if one set has large variance and the other little,

this does not negatively effect the correlation because the measure is completely relative to the distribution with each set. $r_{\mathbf{x},\mathbf{y}}$ increases from pairings on the same side of their respective averages and decreases when they are opposite. When both sets have values close to their averages, these pairs have little impact on the correlation total.

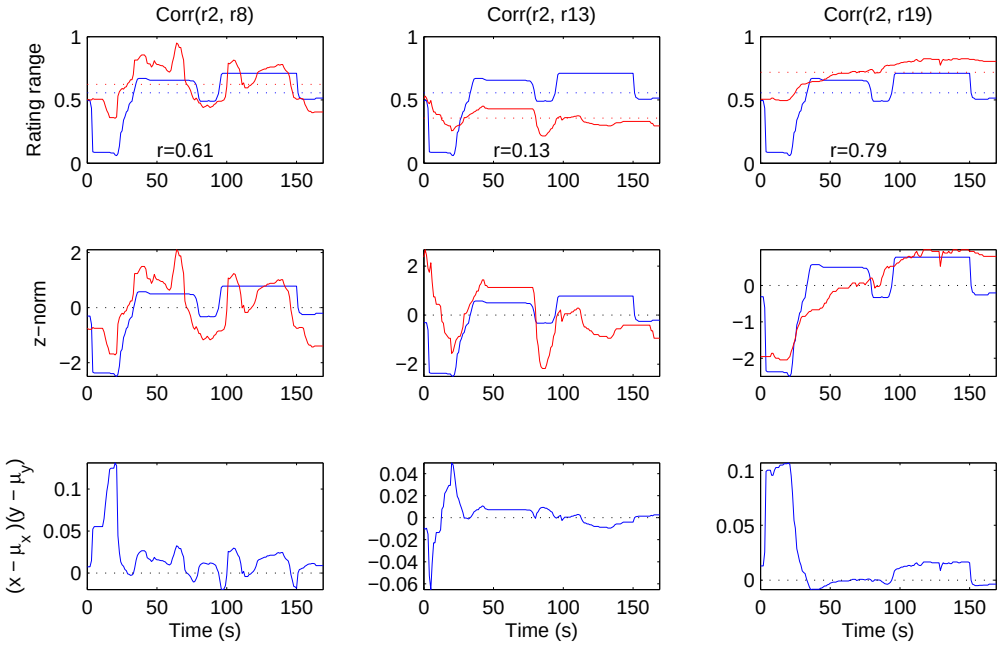

Figure 3–9: Example of three inter-response correlations between responses 2, 8, 13, and 19 from collection 12-fanfareV. The top row of graphs plot the two time series in the original rating scale, the middle show the z-normalized responses, and bottom presents the product-variance series for each pairing which is summed to determine the PPMC r values.

To illustrate what information is measured by the PPMCC in these continuous responses, figure 3–9 shows comparisons between one response and three others, all taken from behavioural response collection 12-fanfareV of perceived

emotion ratings (valence dimension). This set is among those with the highest average inter-response correlations analysed in this thesis. The actual r values for each pairing is printed in the top graph. The highest correlation is found between responses 2 and 19, and the z-normalisation shows why: both responses are at their lowest at the very beginning and are mostly above their respective means after second 100. In the first column, responses 2 and 8 correlate fairly well, with somewhat similar contours. Visual inspection of the top graph makes the proximity of the two series after around 75s to be a strong indication of relatedness. The bottom graph however, shows that the time points with the most positive impact on these PPMCC are between 0s and 25s, while the remanders of the series are less important because both responses are very close to their respective means and the mean crossings are not so well synchronized.

The pairing in figure 3–9 with the lowest correlation coefficient, between responses 2 and 13, is the most dramatic example of how the PPMCC fails to capture salient similarities between responses. Looking at the top graph, the contour of these responses share many traits. Near the beginning, both descend, though response 13 drifts gradually while response 2 jumps directly to its lowest values. At around 20s, both responses rise, with 13 again changing more gradually. These are both steady until around 80 seconds when they both dip for the same amount of time. And lastly, the two responses are fairly stable until 150s. Despite all of this common behaviour, the Pearson correlation is low because the measure depends on the distance from the longitudinal means of these responses—a

questionable statistic for the distribution of these values—rather than the contour information readily available to the eye.

At this level of detail, it is evident that the Pearson correlation coefficient does not capture much of what researchers find interesting in relating continuous responses. However, the average inter-response PPMCC is a popular measure of collection coherence, and a number of publications have shared suggestions on how such time-series correlations could be improved (rather than abandoned for something better). The following section presents some of the proposed imporvements and discusses what affect they have on this dubious statistic.

3.3.2 Qualifying inter-response correlations

The average inter-response correlation, also referred to as the average inter-subject correlation, is the average across the correlation values of the set of all pairwise comparisons between responses in a collection. For a set with M responses, there are $M * (M-2)/2$ unique pairings, so the average inter-response correlation, $\mu(\rho(X))$ can be calculated as:

$$\mu(\rho(X)) = \frac{\sum_{r=1}^{M} \sum_{s=r+1}^{M} \rho(\mathbf{x_r}, \mathbf{x_s})}{M * (M-1)/2} \tag{3.7}$$

When the two sets of data being correlated are independently sampled experimental measurements, the PPMCC can also be given a p-value to be judged against a threshold of significance such as $\alpha < 0.05$. Continuous response data do not satisfy the conditions for α to be meaningful, as will be explored later, but this same threshold will be used in the following figures to give a measure of the distribution of pairwise inter-response correlation values. Significance(*) here forward will be

starred to emphasis the assessments of "significance" are not in fact measures of statistical significance. The other description used to interpret the inter-response correlation distribution is the proportion of significant(*) and positive correlations, to give a sense whether there is strong disagreement between responses as would not be reflected in the average.

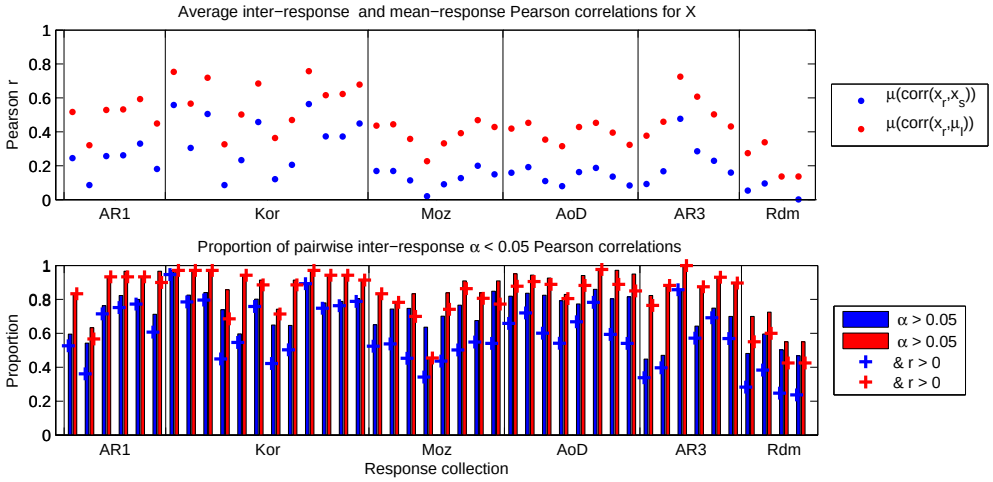

Figure 3–10: Comparing average Pearson Product Moment Correlation Coefficients (PPMCC) inter-response correlation values for the 44 behavioural response collections with the average correlation value between the responses and the collection's average time series. The top graph shows the average r values of the pairwise correlations. The bottom shows the proportion of these values which are (calculated as) significant with $\alpha < 0.05$ presented as bars, and plus signs (+) marking the proportion of significant and positive pairwise correlations for each collection and correlation condition.

The average inter-response correlation is sometimes replaced with the average mean-response correlation, literally the average r or ρ for all correlations between individual responses in a collection and the collections cross-sectional average. In figure 3–10, the results of this alternative (in red) are nearly parallel to the

average inter-response correlation (in blue). These two measures have a Pearson correlation of $r = 0.97$ with $p \ll 0.001$ but the average mean-response correlation is the greater value by 0.25, on average. The proportions of significant and positive correlations are also much higher for this second, smaller set, in part because the average time series contains variation from of all the responses it summarizes. The effect of this can be seen in the unrelated response of set Rdm: though the average inter-rating correlations are very close to zero, as would be expected from responses which were indeed unrelated, the average response to mean correlations are positive. The rare responses which correlate significantly(*) negatively with the average can be supposed to go strongly against the grain of the majority of their collections. Between the average inter-response correlation and the average response-average correlation, the former seems to be the more conservative measure of within-collection coherence.

PPMCC's measure the linear covariance of two sets, and sometimes the linearity assumption is too restrictive. A popular alternative to Pearson's r is Spearman's ρ, a statistic which does much the same calculation as the former, but after mapping each sets elements from their original values to their ordered rank, from smallest to largest. On these collections, the average pairwise inter-response Spearman correlation values (red in figure 3–11) are not very different from the average PPMCC's (in blue). These responses are all measured on a fixed finite interval, and so the differences between the Spearman and the Pearson correlations would be the result of how each interpret the already confined distribution of values within each responses. From section 3.2, we know that these responses are

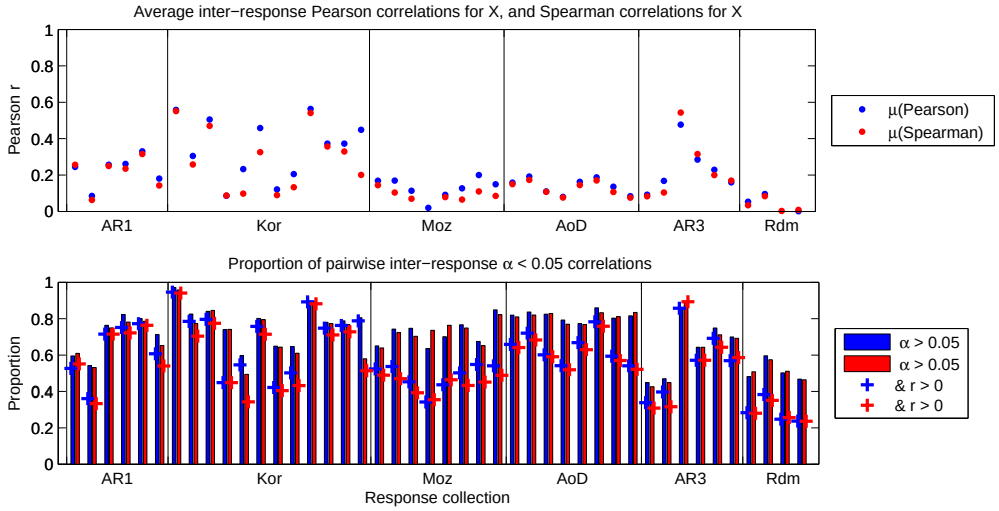

Figure 3–11: Comparing average inter-response PPMCC for the 44 behavioural response collections with the average Spearman's rank inter-response correlation value. Average correlation values per collection on top; proportion of pairwise correlations with $p < 0.05$ (bars) and proportion significant and positive (+) on bottom.

not usually normally distributed; when, for example, a response has a bimodal distribution of values sampled over time, a Pearson correlation will preserve the distance between these clusters while the Spearman correlation will erase it. For this reason, Spearman's ρ is a good default when comparing two sets of elements measured on different scales, such as continuous response ratings and the loudness of the stimulus. But given that the measures estimate very similar proportions of significant(*) pairwise correlations, it seems pertinent to go with the measure that preserve some characteristics of the responses' longitudinal distributions.

There has been some questioning of the validity of the first few seconds of behavioural responses. As many collection devices default to start at a certain point on the rating scale or field, the beginnings of ratings are typically characterized

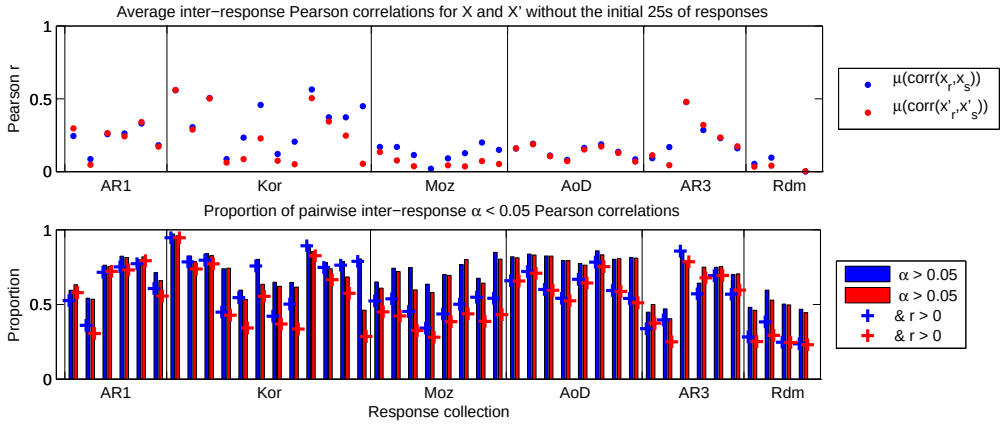

Figure 3–12: Comparing average inter-response PPMCC for the 44 behavioural response collections with and without the initial 25 seconds of responses. Average correlation values per collection on top; proportion of pairwise correlations with $p < 0.05$ (bars) and proportion significant and positive (+) on bottom.

by a large shift away from that default. Another issue is cognitive: it takes time for participants to develop and/or recognize their responses. In one study, participants were asked not to begin rating their felt emotional response to the stimulus before forming a clear idea of what emotion they felt, and on average, responses began 8 seconds after the beginning of the music [BBL+09]. This first disoriented section can bias correlations of responses which are otherwise very close to their averages (as seen in figure 3–9). To evaluate the importance of this beginning on the correlation values, figure 3–13 shows the average inter-response correlations for collections using complete responses and responses measured after the first 25 seconds (in red). For some collections, particularly those with very long stimuli as in AoD and the last collections in AR1 and AR3, removing the first 25 seconds does not change much in terms of significance(*) or average correlation. Some lose a proportion of significant correlations but those in the Moz set, for example, show

a drop in the number of *positive* significant pairwise correlations. Given that the effect of these first seconds varies from collection to collection, it may be an issue for some stimuli and data collection devices.

The principle reason Pearson correlations and other common discrete statistics cannot be applied to time series data is the fact that time series are not sets of independently sampled data. Serial correlation is evidence of this fact. To get around the problem of independent sampling, time series analysis has developed techniques for manipulating time series until the data look close enough to being relatively independent, with only new information at each point. A time series which reaches this degree of sequential unrelatedness is called "stationary". Continuous response ratings are not, on the outset, stationary, and they do not behave like most of the time series these techniques were designed to manipulate, but reducing serial correlation can improve (in part) the legitimacy of applying discrete statistical techniques, such as correlations, to time series data.

As mentioned in section 2.5, one method for reducing the serial correlation in time series that has been used in the continuous response to music literature is the first-order difference transformation. Figure 3–13 presents the PPMCC results for the normal behaviour collections and for their corresponding first-order difference collections (in red). While the transformed responses are much closer to behaving like sets of independent samples, the distributions of values are sharply concentrated around zero which makes most time points nearly irrelevant to the summation that generates the correlation. Thus, the correlation values are very low and never more than half of the pairwise comparisons are

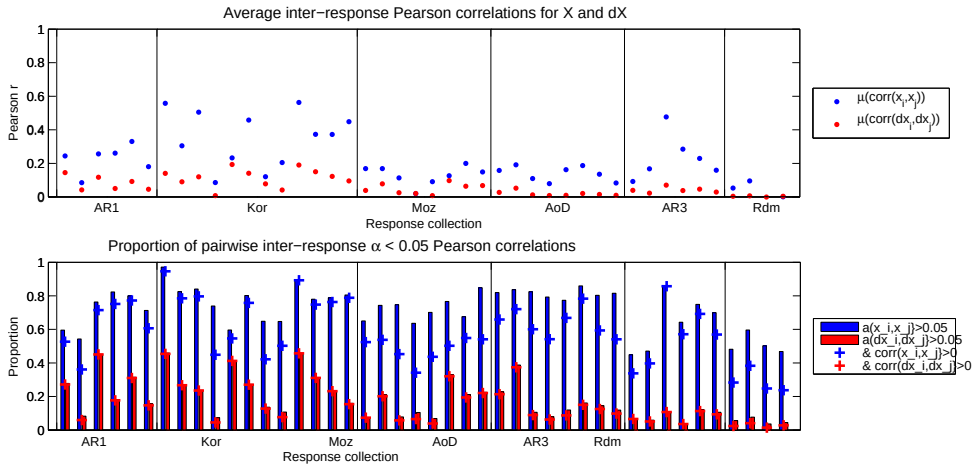

Figure 3–13: Comparing average inter-response PPMCC for the 44 behavioural response collections with that of the first-order difference series of these collections. Average correlation values per collection on top; proportion of pairwise correlations with $p < 0.05$ (bars) and proportion significant and positive (+) on bottom.

significantly correlated. It may be that other measures of similarity are more useful for interpreting these differenced data collections.

Another tactic to make a time series seem more like an independently sampled set is to downsample [CJSK+10], again using the assumption that if every sample point can be expected to carry new information, that is close enough. Figure 3–14 shows the performance of these average correlation assessments for response collections sampled at four different rates. For the most part, the correlations are very similar across sample rates, suggesting that indeed there was a great deal of redundancy in these serially correlated responses sampled at 1 Hz or 0.5 Hz. However, the proportion of the significant(*) pairwise correlations decreases consistently as the sample rate goes down, as expected.

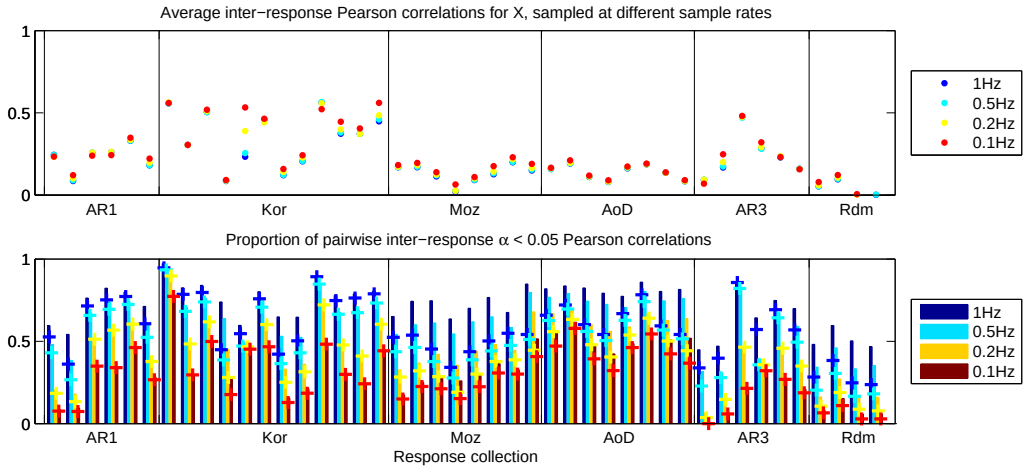

Figure 3–14: Comparing average inter-response PPMCC for the 44 behavioural response collections when sampled at different sample rates. Average correlation values per collection on top; proportion of pairwise correlations with $p < 0.05$ (bars) and proportion significant and positive (+) on bottom.

3.3.3 The insignificance of correlation significance

As previously mentioned, the commonly used measure of significance for PPMCC depends on assumptions that cannot be satisfied by time series data. When testing the significance of a Pearson correlation, three values are used: significance threshold, α, the correlation value, r, and size of the data set, N. For a set of a given size N and significance level α there is a value of r above which correlations are taken to be significant and below if which they are not. The test is however, most often expressed as the relationship between the correlation's p-value and α. The p-value is calculated from the cumulative distribution of Student's T with $N - 2$ degrees of freedom, and every added sample lowers p. When the samples are independent—say taken from different subjects—more samples means lower likelihood of getting a type I error. When sampling a continuous

response, the number of sample points is not a measure of how much independent information is contained in the series because any number of points can be used to represent the same continuous response. If we had an estimate of the amount of "independent" information in a response, regardless of how it was sampled, this quantity could be used to measure the significance of these correlations. The fact that the average inter-response correlations are mostly the same at 1 Hz and 0.1Hz suggests that the covariance of responses in these collections are described well enough by the lower sample rate and the correspondingly higher p-values are not unduly conservative.

The Pearson product-moment correlation coefficients measure variation from responses' respective longitudinal means—a volatile descriptor of these non-normal distributions of values—and it ignores the actual values measured on the rating scale. If researchers are interested in comparing the contour of these ratings, difference data or other derivative representations would be less effected by the mean estimate. If the actual rating values are of interest, a different measure of distance or difference between ratings, like the standard euclidean distance metric would be more practical.

The popularity of correlating continuous response data is likely due to the simplicity with which the supposed significance of the statistic can be calculated. However, like the average, convenience of calculation does not guarantee a meaningful result. Time series do not satisfy the assumptions of the standard p-value estimation used for independently sampled data sets. While the effect of serial correlation can be reduced within a given time series, the number of periodically

sampled points is still not analogous to the number of independent samples. The correlation between responses may be an interesting result for other interpretations of the calculation, but it is not an appropriate means of assessing the significance of coherence or agreement between responses in a continuous response collection or with other related time series.

3.4 Conclusions on these traditional analyses

The results of this chapter's study of traditional techniques suggest a few recommendations of future analyses:

1. When not testing for normality of distribution, non-parametric statistics are recommended for capturing more reasonable summaries of the actual data, whether cross-sectionally or longitudinally.
2. When quantifying the dispersion of rating values across a collection of responses, it is worth considering both the longitudinal and the cross-sectional variability in response values with non-parametric measures to investigate the relative importance of different factors of variation.
3. A better measure of coherence or agreement between time series is needed to replace the often misused Pearson correlation.
4. These techniques are useful for exploring the data and generating basic contrasts between collections, but there is a lot that is lost in the reductions. Rather than ends, they should be means to directing more detailed investigations of continuous response collections so as to relate the measured results to theories and models of the experience of listening to music in time.

CHAPTER 4
Novel analyses

Many of the advantages of continuous responses are not employed in the traditional approaches described in the previous chapter. A collection of continuous responses to the same stimulus carries information about the individual responses and the timing of response variation. This level of temporal detail is of great interest to theories and models of music cognition. Rather than collapse responses by dimension, the analyses in this chapter ask if responses are coordinated, when responses appear to be coordinated, and which responses seem to be coordinated with each other. Audience activity analysis is a means of describing concurrent activity in responses of a collection and it enables measures of coordination of this activity. Event analyses are techniques which differentiate time points in the sampling series by the behaviour of responses in the collection. The last section of this chapter introduces cluster analysis to sort responses into groups of similar behaviour as elements in a set of time series.

4.1 Activity analysis and coordination testing

If participants all agreed in their responses to music, the average time series would be a perfect summary of a continuous response collection. One look at the top graph of figure 4–1, however, is often enough to convince most researchers that individual responses do not behave so consistently. For this collection, the standard deviation band around the mean time series engulfs more than half of the

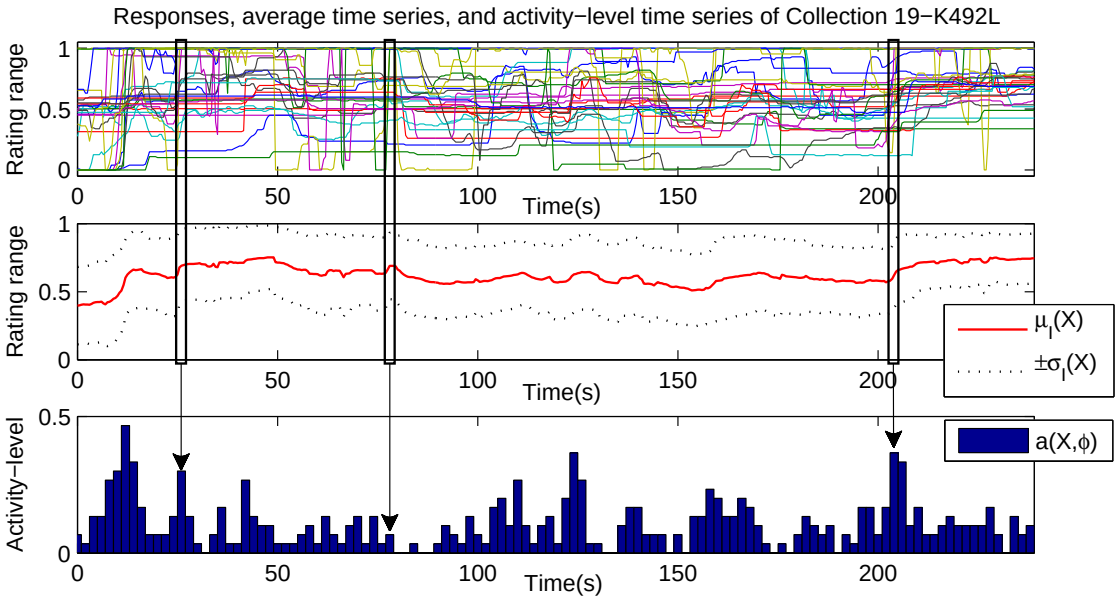

Figure 4–1: Summaries of collection 19-K492L of the Moz data set. On top, all responses; in the middle, the collection's average response time series +/- the standard deviation; below the activity-level time series, counting increases in ratings, measured in frames of two seconds. The boxes and arrows compare time frames with similar means and different levels of activity across the collection.

rating range at nearly every time point. Rather than apply statistics that presume a single ideal continuous response, activity analysis evaluates the concurrent activity across responses to the same stimulus. In figure 4–1, the response activity measured is increases in ratings in some time interval, and the last graph in the figure presents the proportion of responses showing this type of activity in successive two second time frames. Working with this representation of a collection of responses, this chapter section presents one method for answering the question of whether the responses of a collection are coordinated, i.e., are they related to each other. The Moz data set is used for most examples because it conveniently

contains two response collections for each of four stimuli, all orchestral works by W. A. Mozart. This technique for testing coordination demonstrates that some pieces provoke more coordinated responses than others, and some forms of response activity are more coordinated than others.

4.1.1 Activity basics

There are many reasons for individual responses to differ within a collection. First, the same music can evoke different experiences, depending on the listener's musical experience and state of mind. Second, participants also vary in reaction time, each relatively faster or slower to show the same kind of response. When responses are collected though a behavioural task, a third factor is that participants may perform the task differently by being more or less sensitive to changes, or by reporting different aspect of their experience. Also particular to ratings is a fourth confound: participants' attention may wander from the task they have been asked to perform. Participants can accidentally omit of changes in response, or they may inadvertently change their criteria for expressing their response over the course of an experiment. As a consequence of these performance concerns, participants may also attempt to compensate for deviations by making corrections not synchronized to the stimulus. Between legitimate differences in experience and all the other sources of noise, it is necessary to ask:"Is there any common experience of the musical stimulus to be found in these responses?"

A time-sensitive indication of a reaction to the stimulus is required to assess whether responses are coordinated, the timing of which could be compared between responses. Such an indication of a reaction to a stimulus is called response

activity. For continuous rating response collections, one kind of response activity would be increases in rating. An increase in rating is not guaranteed to be a sign that the participant's experience just changed because of the stimulus, but if the stimulus can motivate a change in experience, such a change would often be expressed through an increase in rating, and that is good enough for the tests that follow.

An activity indicator function, ϕ, determines whether a response $\mathbf{x_r}$ is active in some time interval $T = [t_a, t_b]$:

$$\phi_T(\mathbf{x_r}) = \begin{cases} 1 & \text{if } \mathbf{x_r} \text{ is active in the interval } [t_a, t_b] \\ 0 & \text{if } \mathbf{x_r} \text{ is not active in the interval } [t_a, t_b] \end{cases} \tag{4.1}$$

For increasing activity as measured here, ϕ detects whether a response has increased by at least 2% of the rating range over the course of interval T:

$$\phi_T(\mathbf{x_r}) = \begin{cases} 1 & \text{if } \mathbf{x_r}(t_b) - \mathbf{x_r}(t_a) \geqslant 0.02 \\ 0 & \text{if } \mathbf{x_r}(t_b) - \mathbf{x_r}(t_a) < 0.02 \end{cases} \tag{4.2}$$

Similarly, a decreasing activity indicator may be defined, here after represented by ψ, which would detect a decrease of at least 2% of the rating scale over the interval T.

To measure coordination across a collection X of M responses, the activity-level of the collection for some time interval T is the proportion of responses showing some activity, ϕ.

$$a_T(X, \phi) = \frac{\sum_{r=1}^{M} \phi_T(\mathbf{x_r})}{M} \tag{4.3}$$

To assess the variation of activity-levels over the course of time, we cut the duration of the recorded responses into a sequence of non-overlapping time frames of size ΔT on which the activity-level can be measured. Since responses are thought to lag behind the stimulus by 1 to 3 seconds [Sch04], two second time frames are used in the following examples to capture some temporally related activity.

A collection of responses, X, is measured at time points $\{t_1, t_2, ..., t_i, ..., t_N\}$ with a sample period of Δt. A sequence of time frames $\{T_j\}$ for $j \in J = \{1, 2, ..., N_J\}$, $N_J = \lfloor N * \Delta t/\Delta T \rfloor + 1$, are defined by a downsampling of points $\{\tau_j\}$ and $\{\tau'_j\}$ from the sequence $\{t_i\}$ such that $T_j = [\tau_{j-1}, \tau_j]$ and centred on time point τ'_j. The sequence of $\{\tau_j\}_{j=0}^{N_J}$ is defined as $\tau_j = t_{\lfloor (j-1/2)*\Delta T/\Delta t \rfloor}$ with $\tau_0 = t_1$ and $\tau_{N_J} = t_N$, while $\{\tau'_j\}_{j=1}^{N_J}$ is then defined as $\tau'_j = t_{\lfloor (j-1)*\Delta T/\Delta t \rfloor}$. The time frames can be defined more simply, but for the sake of graphical interpretation, this centring of time frames is less confusing to the eye. All time frames except the first and the last are of duration ΔT. The activity-level, for some activity ϕ, can now be measured on each of these time frames, making the activity-level time series for the collection X, $\mathbf{a}(X, \phi, \Delta T) := \{a_j(X, \phi, \Delta T)\}$ for $j \in J$ such that:

$$a_j(X, \phi, \Delta T) = \frac{\sum_{r=1}^{M} \phi_{T_j}(\mathbf{x_r})}{M} \tag{4.4}$$

The building of this series is hinted at in figure 4–1 and shown more explicitly in the first two graphs in figure 4–2. The boxes define the time frames in which the activity-level is measured, here increasing rating activity ϕ. This time series representation of the audiences' response to music exposes different information

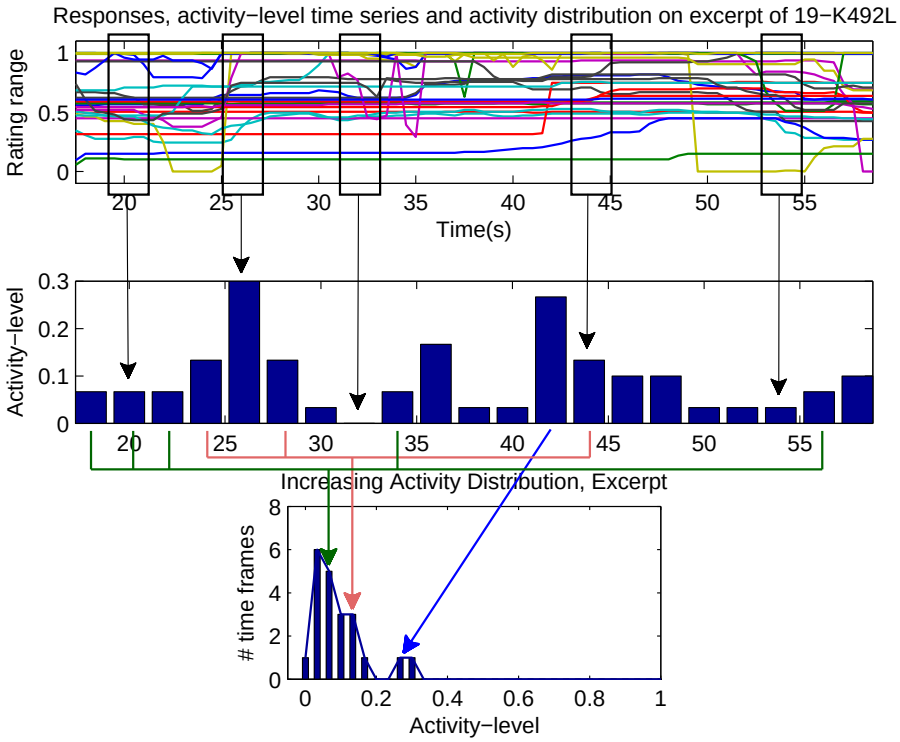

Figure 4–2: Steps to generating the activity distribution, $d_{\Phi,X,\Delta T}$, on an excerpt of 19-K492L. From the individual responses to the activity-level time series $\mathbf{a}(X, \phi, 2s)$, and down to the distribution of time frames at different activity-levels over this excerpt of the collection.

than that of the average time series. Looking at the activity shown in the bottom graph of figure 4–1, a series typical for the felt emotional intensity rating data collections of the Moz data set, it is possible to see how rarely even a third of responses show synchronized increases. In relation to the stimulus, a four minute overture, there are spikes of simultaneous activity in time to the first fortissimo and a couple other moments, but most peaks in synchronous rating changes are common to fewer than a quarter of participants' responses. Correspondingly, it

is also rare for there to be no ratings showing increases in these two second time intervals. The lowest quartile of activity includes up to 5% of participants sharing the same activity.

The audience activity time series shows a greater contrast from moment to moment than is measurable in the average time series. Time points with the similar cross-sectional averages, or even intervals with similar changes in cross-sectional average ratings, may reflect very different degrees of activity. In figure 4–1, activity around seconds 25, 80, and 208 have been boxed for comparison. The activity indicator ϕ democratizes the summary representation by allowing each participant's reported response to have equal weight rather than favour a few more dramatic raters over the more conservative.

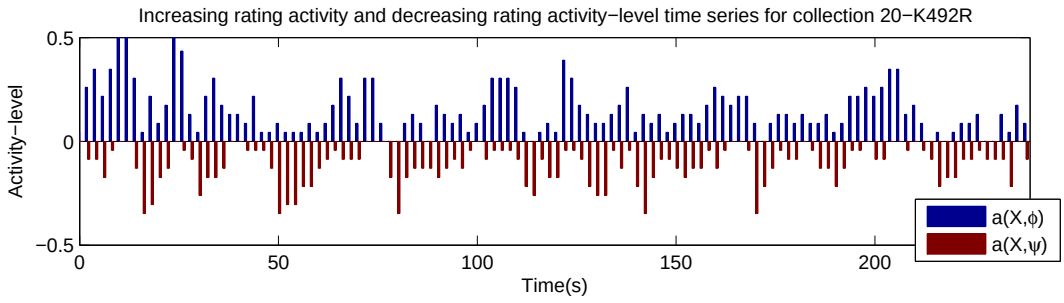

Figure 4–3: Comparing two activity-level time series of the same collection: increases in ratings above zero, $\mathbf{a}(X, \phi, 2s)$, and decreases in ratings below zero, $\mathbf{a}(X, \psi, 2s)$, for collection 20-K492R.

Visual inspection of both increases and decreases in ratings show that participants' rating changes can be contradictory. Figure 4–3 shows the proportion of participants decreasing ratings below the proportion of participants showing increases in the same time interval. For most of the samples, some participants

show increases in ratings while others show decreases. In some rating response collections, there appears to be alternation between increasing and decreasing activity but with some overlap. This fact does not imply that participants are having opposite responses to the stimulus. Rather, with variable reaction time and the width of the time frame, the disagreement could also be caused by different moments or by attention to different aspects of the work. Still, the contradictory responses, usually suppressed in the cross-sectional average time series, may be cognitively and musically significant.

Just as the activity-level time series shows information not in the average, the average time series presents information not obvious from activity. Particularly for responses using categorical distinctions or quadrant representations of emotion, changes in ratings do not show *which* emotions are being reported. Other definitions of activity could be used to consider the concentrations of agreement for categorical purposes. Regardless, the coordination of activity in a collection measures, in some sense, the robustness of the average time series.

4.1.2 Activity distributions

The activity-level time series show variation in concurrent activity across the responses of the collection from one time frame to the next, but the range of activity levels shown in these time frames is limited. In the examples presented, the highest concentrations of activity peak at half of the collection's responses showing simultaneous activity, while in the frames with the lowest activity levels, complete stillness is more rare than the single dissident response. Like the experimental collection seen in figure 4–3, the activity of a collection of

unrelated responses, shown in figure 4–4, has some variation in activity-level from frame to frame—moments of relatively high and relatively low activity. If the majority of responses to the same stimulus do not actively agree on when their respective experiences change, and there are nearly always a few responses showing activity in any given time frame, it is necessary to consider the possibility that responses to the same stimulus may not be measurably coordinated. Comparing

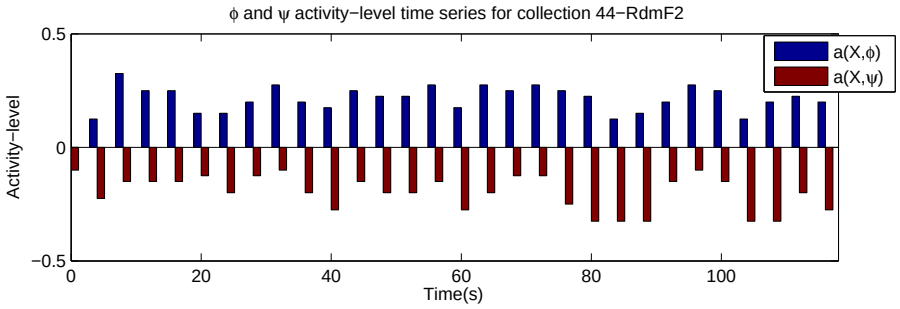

Figure 4–4: Two activity-level time series for collection 44-RdmF2, increases in ratings above zero, $\mathbf{a}(X, \phi, 2s)$, and decreases in ratings below zero, $\mathbf{a}(X, \psi, 2s)$.

experimental responses directly to random collections of time series cannot answer this question efficiently. A more practical method for evaluating the coordination in a collection is to consider the distribution of activity-levels over the time course of the stimulus. The longitudinal distribution of activity-levels is calculated by counting how often time frames show different proportions of responses being active. The activity distribution for activity ϕ on response collection X in time frames of size ΔT, notated as $d_{X,\phi,\Delta T}$, would be a sequence of $M+1$ values defined as:

$$d_{X,\phi,\Delta T}(m/M) = \|\{j \in \{1, 2, ..., N_J\} \text{ such that } a_j(X, \phi, \Delta T) = m/M\}\| \qquad (4.5)$$

for $m \in \{0, 1, ..., M\}$. Figure 4–2 traces the construction of an activity distribution between the middle and lower graph. The resulting histogram defines the increasing activity distribution of this excerpt from the 19-K492L collection of continuous responses.

If every response in these collections was the result of the same expression criteria and experience, the activity distribution would be bimodal: many time frames in which no or very few responses showed activity, some showing nearly all responses being active, and very few frames with activity-levels in between these extremes. In figure 4–5, the left graph is an example of how the activity distribution of such a harmonious response collection would look. If every response in a collection was motivated by completely unrelated experiences, each active independent of others and of the stimulus, the distribution of activity would be more like the right graph of figure 4–5: few instances of very low or high activity and most samples showing middle to low proportions of participants simultaneously active.

If the activity measured on a collection resulted in a distribution like that of the right of figure 4–5, this would be an indication that the activity is not more coordinated than the output of a random process. Whether or not the responses in the collection were confident reflections of individual listeners experience of the music, we cannot draw conclusions about the variation in activity-levels in relation to the stimulus if the activity distribution fails to show more coordination than a random process because we have no reason to expect that the same pattern of variation would be repeated.

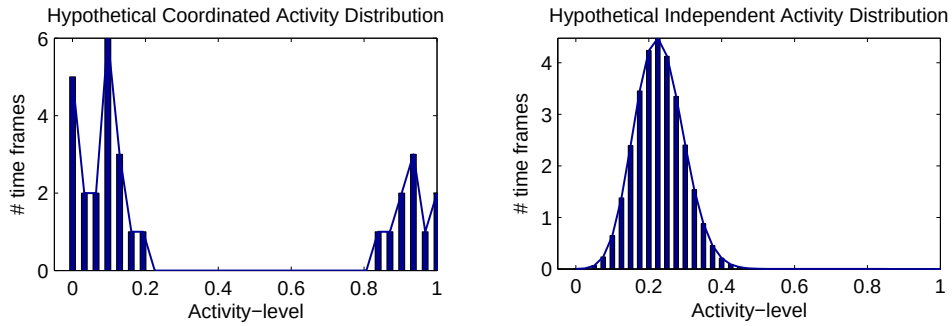

Figure 4–5: Hypothetical activity distributions. On the left, the activity distribution (histogram of time-frames with different activity-levels from the activity-level time series) for a highly coordinated collection of responses. On the right, the activity distribution expected for a collection of unrelated continuous responses.

Before measuring the difference between the actual activity distribution and that of the closest random process, it is necessary to generate this random model's distribution. The model distribution for a collection's activity describes the expected activity distribution if all responses were independently active—each showing activity at random rather than in response to the shared stimulus. The actual average rate of activity per response per time frame, $p_{\phi,X,\Delta T}$, can be calculated for each collection and activity indicator:

$$p_{X,\phi,\Delta T} = \frac{\sum_{j=1}^{N_J} a_j(X,\phi,\Delta T)}{N_J} \tag{4.6}$$

With this estimated parameter, the random model distribution is calculated for each possible level of activity $\{0, 1/M, 2/M, ..., 1\}$. In most the collections analyzed here, M is around 30, but for the sake of flexibility, the expected activity distribution has been calculated discretely. Under a binomial assumption, the probability that a time frame T_j would have an activity-level of m/M, $m \in$

$\{0, 1, 2, ..., M\}$, would be:

$$P\left(a_j(X, \phi, \Delta T) = m/M\right) = \frac{M!}{m!(M-m)!} p_{X,\phi,\Delta T}^{m} (1 - p_{X,\phi,\Delta T})^{M-m} \tag{4.7}$$

With N_J time frames, the expected frequency distribution of activity-levels in the time series $\mathbf{a}(X, \phi, \Delta T)$ for a collection of independently active responses would be:

$$e_{X,\phi,\Delta T}(m/M) = N_J * P\left(a_j(X, \phi, \Delta T) = m/M\right) \tag{4.8}$$

This equation defines the random model activity distribution: the distribution expected if all responses were independently active of each other. The binomial assumption presumes the average rate of activity per response is a sufficiently close approximation of the activity rate for each response, and that all response activity series are independent of each other.

The lower left graphs in figures 4–6 and 4–7 show activity distributions of two response collections plotted against their respective closest model activity distributions. Though the experimental activity distributions are not bimodal, the small differences between the actual distributions and the random model may be significant. More low activity-level and high activity-level time frames with correspondingly fewer middling activity-level time frames suggest that the activity of responses in a collection is not completely independent.

4.1.3 Goodness-of-fit test

A common hypothesis test for evaluating a distribution of observations in terms of a theoretical frequency distribution is Pearson's Chi-square goodness-of-fit test [DM88]. Given a significance level α, this test accepts or rejects the null

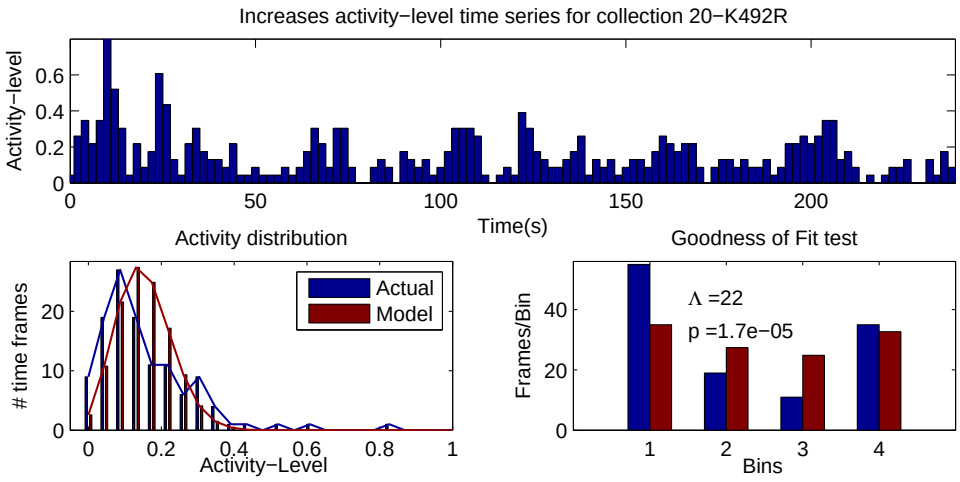

Figure 4–6: Testing the coordination of increases in responses in collection 20-K492R, X, of the data set Moz. Above, the ϕ activity-level time series on two second frames, $\mathbf{a}(X, \phi, 2s)$; bottom left is the activity distribution $d_{X,\phi,2s}$ with the random model distribution $e_{X,\phi,2s}$; bottom right are the bins for comparing these distributions and the results of the goodness-of-fit test with 2 d.f. The results reject the null hypothesis.

hypothesis that a set of observations, each falling into one of K bin, could have been sampled from a predetermined distribution function. In this case, we want to test whether $d_{X,\phi,\Delta T}$, the frequency of activity-levels observed in the time frames of the $\mathbf{a}(X, \phi, \Delta T)$ are distributed differently from that of the random model $e_{X,\phi,\Delta T}$.

Comparing the distributions at all activity levels m/M for $m \in \{0, 1, 2, ..., M\}$ could easily overfit the data, particularly since there are so few responses at the highest values. Instead, the distributions can be simplified to a small number of bins with comparable numbers of time frames expected in each, according to the random model. These bins must cover all categories of observations and be mutually exclusive. For K bins, B_k with $k \in \{1, 2, ..., K\}$, they are defined such

that $B_k \subset \{0, 1/M, 2/M, ..., 1\}$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\bigcup_{k=1}^{K} B_k = \{0, 1/M, 2/M, ..., 1\}$. A last condition on the bins, to satisfy the conditions of the χ^2 test, is that no bin be expected to contain less than 5 observations: $\sum_{b \in B_i} e_{\Phi,X}(b) \geqslant 5$ [DM88]. In the tests that follow, they satisfy a more strict standard of $\sum_{b \in B_i} e_{\Phi,X,\Delta T}(b) \geqslant N_J/2K$ for all $k \in \{1, 2, ..., K\}$, and the number of bins K is usually 3, 4, or 5. The lower right graph of figure 4–6 shows a bar graph with the number of time frames falling into each of the four bins used for the χ^2 test on this activity distribution.

To test the bin-wise fit of the model on the experimental data, their weighted squared-difference is calculated to give the test statistic, Λ, from which the p-value of the difference is estimated.

$$\Lambda_{e,d} = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} = \sum_{k=1}^{K} \frac{\left(\sum_{b \in B_k} d_{X,\phi,\Delta T}(b) - \sum_{b \in B_k} e_{X,\phi,\Delta T}(b)\right)^2}{\sum_{b \in B_k} e_{X,\phi,\Delta T}(b)} \tag{4.9}$$

Depending on the significance threshold, $\Lambda_{e,d}$ determines whether or not the model distribution/null hypothesis can be rejected for the degrees of freedom of the test. The degrees of freedom for this test is $K - 2$, i.e., the number of bins less one less the number of estimated parameters. The test statistic Λ is (nearly) asymptotically χ^2_{K-2} [CL54], the chi-square distribution with $K - 2$ degrees of freedom, and it is on the basis of this relationship that the p-value is determined for this test. In relating the test statistic Λ to distribution of χ^2_{K-2}, we can estimate the likelihood of finding a test statistic of equal or more extreme value under the null hypothesis, i.e., the p-value.

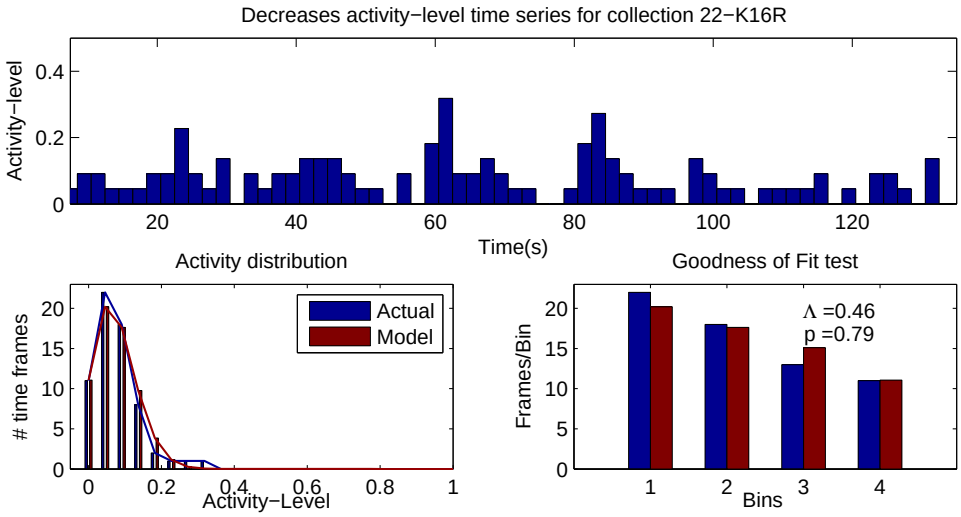

Figure 4–7: Testing the coordination of decreases in responses in collection 22-K16R, X, of the data set Moz. Above, the ϕ activity-level time series on two second frames, $\mathbf{a}(X, \phi, 2s)$; bottom left is the activity distribution $d_{X,\phi,2s}$ with the random model distribution $e_{X,\phi,2s}$; bottom right are the bins for comparing these distributions and the results of the goodness-of-fit test with 2d.f. The results fail to reject the null hypothesis.

In hypothesis testing, the p-value is not generally used for purposes beyond determining the outcome of the test. In the context of many tests, however, it is not uncommon to specify the performance of test statistics according to multiple significance levels at different orders of magnitude. In the following tests, $\alpha < 0.01$ is the significance level, with one chance in a hundred of falsely rejecting the null hypothesis. In cases when the p-value is less than 10^{-4}, or 10^{-8}, we may interpret these as very rough indications of how very different the actual activity distribution is from the random model.

Many of the behavioural response collections tested show sufficient coordination in the ϕ and ψ activity distributions to reject the random model (see figure

4–13), but there are many for which the test does not. As shown in Fig. 4–7, the distribution activity-levels for decreases in responses, ψ, for collection 22-K16R is very close to that of the random model. The goodness-of-fit test statistic fails to reach the significance threshold with a p-value above 0.5. If the null hypothesis fails to be rejected by the goodness-of-fit test, this does not mean that the responses really did respond randomly. Rather, it suggests that the actual activity distribution is not differentiable from the random model; we cannot presume that the variation of activity-levels across time frames is coordinated by the stimulus to such a degree that the pattern would be repeated in another set of responses to the same stimulus.

4.1.4 Joint activity coordination and contingency tables

Testing one type of activity per collection gives some information about the coordination of responses in time, but there are many possible measures of activity for any type of response. By considering different activity-level time series in relation to each other, more evidence of response coordination may be found. For example, the apparent inverse parallel between increasing and decreasing rating activity time series of a single collection, shown in figure 4–3, suggests to the eye that the responses mostly alternate in activity. If these time series were demonstrably alternating, this could support the study of the contour of average time series as representative of the collection's dynamic experience.

Like the simple goodness-of-fit test, the joint activity distribution test compares the actual joint distribution of two activity series to a model distribution function. This model, defining the null hypothesis of this test, is constructed

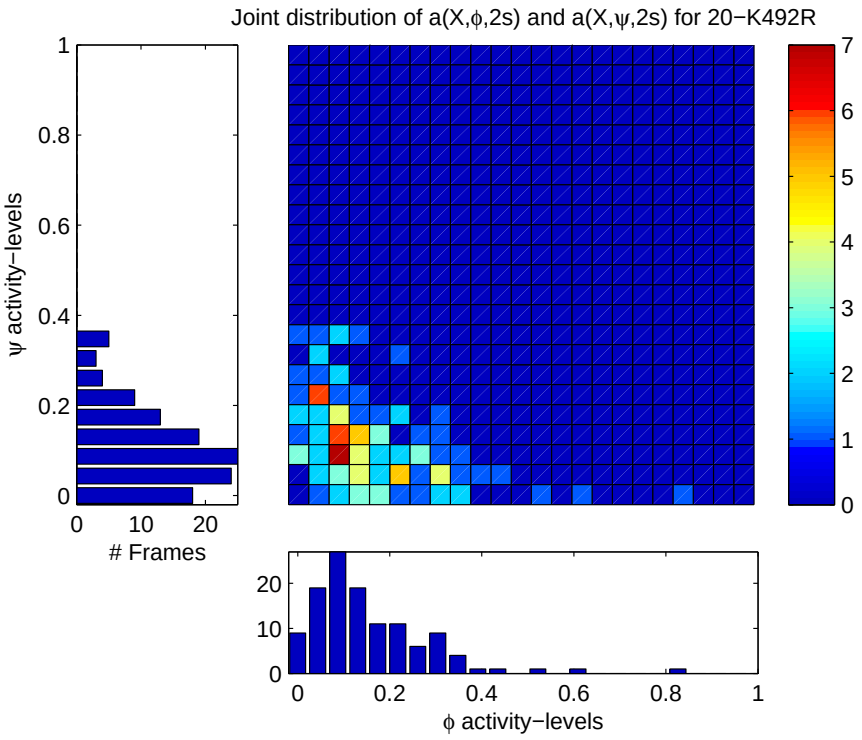

Figure 4–8: Composition of the joint activity distribution. The bar graphs are of the actual activity distributions for activity indicators ϕ and ψ measured on two second frames from the series seen in figure 4–3. The colours of the matrix in the middle show the actual number of time frames with any combination of the two activity-levels.

from the actual activity distributions of each activity-level time series under the assumption that the two series are independent.

Comparing two forms of activity on the same collection, X, the actual joint distribution, $d_{X,\phi,\psi,\Delta T}(m_1/M, m_2/M)$ counts the number of time frames with a ϕ activity-level of m_1/M and a ψ activity-level of m_2/M. The result is a two-dimensional matrix of the concentration of time frames across activity-levels, plotted using colour to show the number of time frames per cell in figure 4–8.

When the two activities are non-exclusive, i.e., when it is possible for a single response to show both kinds of activity in the same time frame, the model joint distribution is simply the product of the two simple distributions to reflect the null hypothesis of independence:

$$e_{X,\phi,\psi,\Delta T}(m_1/M, m_2/M) = d_{X,\phi,\Delta T}(m_1/M) * d_{X,\psi,\Delta T}(m_2/M)/N_J \tag{4.10}$$

When the two activities measured on the same collection are exclusive, such as increasing and decreasing ratings, the random model is modified by applying a binomial assumption to one of the distributions so as to offset the combinatorial limits of the finite population.

$$e_{X,\phi,\psi,\Delta T}(m_1/M, m_2/M) = d_{X,\phi,\Delta T}(m_1/M) * d'_{X,\psi,\Delta T}\left(m_2/(M - m_1)\right)/N_J \tag{4.11}$$

If the audience activity time series were indeed independent of each other, i.e. uncoordinated, we would expect the joint distribution to form a round blob of activity towards the lower range of both distributions, as in the left bottom corner in the left graph of figure 4–9. If the audience activity time series were

identical or varying in parallel, the activity would make a diagonal strip across the joint distribution matrix, as in the middle graph of figure 4–9. And if the two series alternated in degree of activity, one being quiet while the other is active and vice versa, the distribution would cluster to the edges of the matrix as seen in the right most graph of figure 4–9, with few time frames counted down the middle. Either of the last two would be examples of two activities jointly coordinated in the collection X.

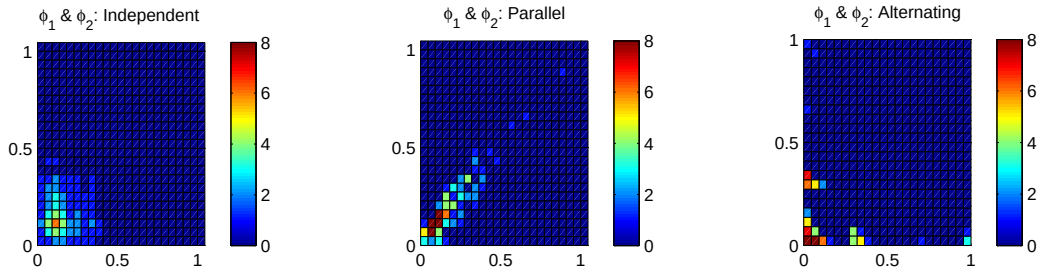

Figure 4–9: Three hypothetical joint activity distributions. To the left, the distribution expected if the two activities are completely independent and unrelated; in the middle, the joint distribution if the two activities are highly related and often happen together; to the right, the joint distribution if the two activities are related and tend to alternate.

The test for these joint distributions is called an r x c contingency table, a common method for assessing the independence of two categorical variables of a data set [DM88]. The hypothesis uses the same test statistic as the goodness-of-fit test, here called $\Lambda_{e,d}$, to measure the difference between the observed and expected number of samples falling into each cell of the contingency table. To compare the actual joint distribution to the model of independent activity using this hypothesis test, it is again necessary to aggregate these many levels of activity into a smaller

number of categories in both dimensions of the joint distribution. Both activity-level distributions are cut into three categories of relatively low, middle and high activity-levels. The cuts are made to aggregate around a third of the time frames of the activity-level time series into each category, when considering only one type of activity. The two dimensions combine to form a 3 x 3 contingency table, with close to a ninth of the total number of frames expected in each cell of table. Similar to the simple activity coordination test, if the expected number of cells in one time frame is less than 5, this might threaten the asymptotic assumptions used to estimate the significance of test statistic. When that happens, one of the activity distributions is reduced to two categories for a 2 x 3 contingency table. Now we have cells $B_{k_1k_2}$ of two matrices of size $K_1 \times K_2$ summarizing the expected and actual joint distributions, with elements $b \in B_{k_1k_2}$ of the form $b = (m_1/M, m_2/M)$. The test statistic, $\Lambda_{e,d}$, is then given by:

$$\begin{aligned} \Lambda_{e,d} = & \sum_{k_1=1}^{K_1} \sum_{k_1=2}^{K_2} \frac{(O_{k_1k_2} - E_{k_1k_2})^2}{E_{k_1k_2}} \\ = & \sum_{k_1=1}^{K_1} \sum_{k_1=2}^{K_2} \frac{\left(\sum_{b \in B_{k_1k_2}} d_{X,\phi,\psi,\Delta T}(b) - \sum_{b \in B_{k_1k_2}} e_{X,\phi,\psi,\Delta T}(b) \right)^2}{\sum_{b \in B_{k_1k_2}} e_{X,\phi,\psi,\Delta T}(b)} \end{aligned} \tag{4.12}$$

The degrees of freedom on this test is $(K_1 - 1)(K_2 - 1)$, and the p-value of the difference between the distributions, as measured by Λ is again evaluated using the χ^2 distribution with $(K_1 - 1)(K_2 - 1)$ degrees of freedom [DM88]. Again, p-values below $\alpha = 0.01$ are presumed to be significant.

Figures 4–10 and 4–11 show the results of these contingency table tests of independence on joint distributions of activities of the same collection. Figure 4–10 shows the distribution of rating increases and decreases for one audience's ratings

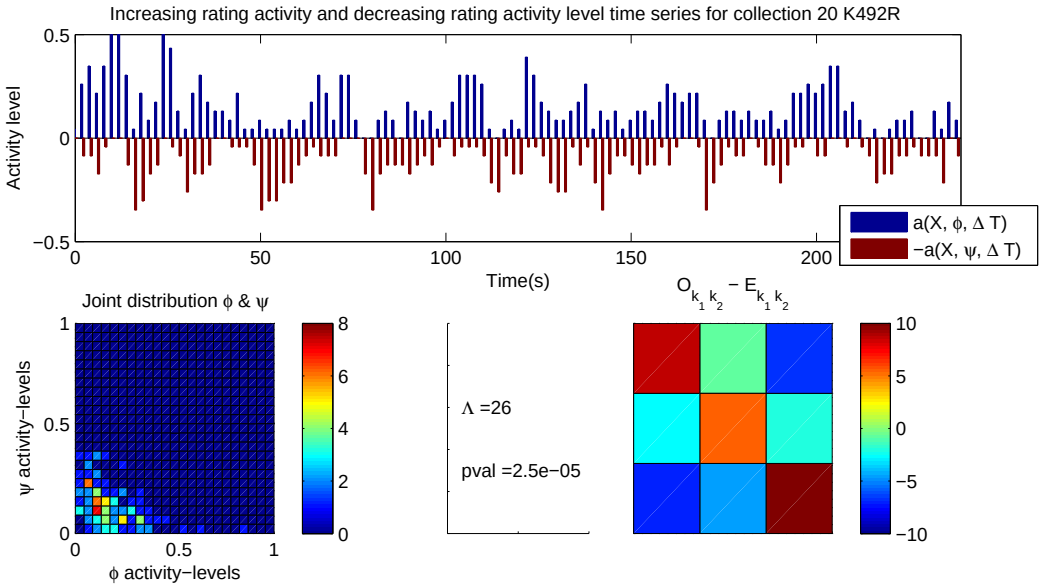

Figure 4–10: Test of independence of increasing and decreasing activity in 20-K492R. Above, the rating increases activity-level time series is plotted with the rating decreases (below zero) on two second time frames. Bottom left shows the actual joint activity distribution matrix, bottom right is the difference between the actual distribution and the independent model as evaluated on the contingency table, and in the middle is the test statistic and corresponding p-value for 4 d.f.

of experience emotional intensity during the presentation of a recordings of the Marriage of Figaro overture. The joint distribution shows a high concentration in the mid-low range for both distributions, as we would expect from a random distribution, but there many samples also along the lower and left edges of the joint distribution matrix. The right graph of figure 4–10 shows the difference between the number of time frames falling into each cell of the contingency table and the number expected by the assumption of independence. The test statistic $\Lambda = 27$ given the four degrees of freedom in this test, maps to a p-value which comfortably rejects the null hypothesis for a significance level of $\alpha = 0.01$. The

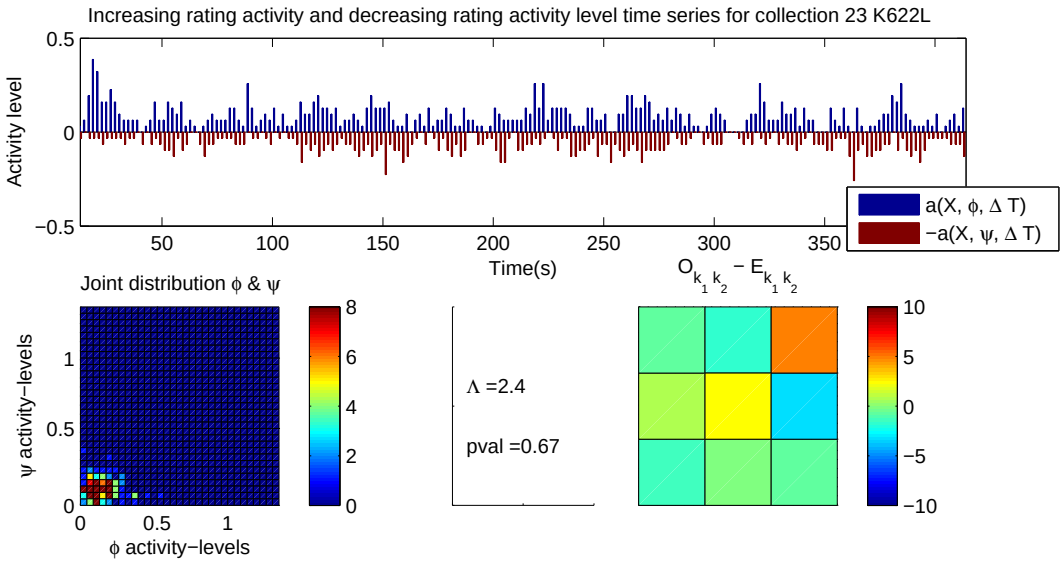

Figure 4–11: Test of independence of increasing and decreasing activity in 23-K622L. Above, the rating increases activity-level time series is plotted with the rating decreases (below zero) on two second time frames. Bottom left shows the actual joint activity distribution matrix, bottom right is the difference between the actual distribution and the independent model as evaluated on the contingency table, and in the middle is the test statistic and corresponding p-value for 4 d.f.

colours of differences per table cell show that fewer samples were both low-low and high-high in the actual joint activity distribution than expected by the independent model, while more time frames showed low-high and high-low joint activity. This pattern suggests that the coordination between the two forms of activity is a tendency to alternate.

Not all activity series are so well-behaved. The joint activity for 23-K622L, a collection of responses to the live performance of the Adagio movement of Mozart's famous Clarinet Concerto, is presented in figure 4–11. The joint distribution is

rather concentrated in the corner and the test of independence fails to reject the null hypothesis that these two forms of activity are unrelated.

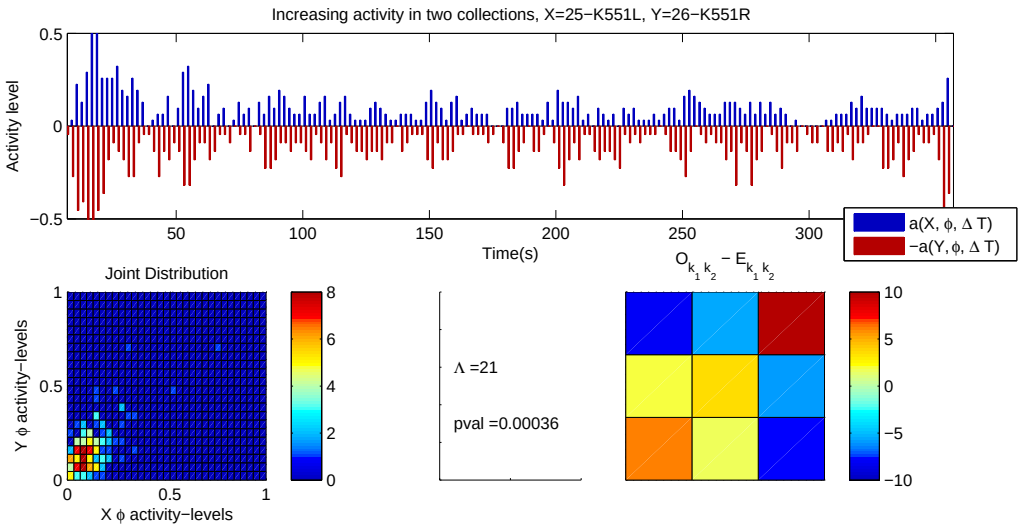

Figure 4–12: Test of independence of increasing activity in two collections of responses to the same stimulus: 25-K551L and 26-K551R. Above is plotted the rating increases activity-level time series of the first collection against the rating increases of the second (below zero) on two second time frames. Bottom left shows the actual joint activity distribution matrix, bottom right is the difference between the actual distribution and the independent model as evaluated on the contingency table, and in the middle is the test statistic and corresponding p-value for 4 d.f.

When two collections are related by stimulus, the activity-level time series for the same form of activity can be compared between the two series, and the calculation of the joint distribution is simplest because the measured activities are not exclusive. The top graph of figure 4–12 shows the activity-level time series for increases in responses over two second time frames are presented from two collections of responses to the same performance (one live, one recorded) of the Finale to Mozart's the Jupiter Symphony, K551. The joint distribution map shows

the time frames lifting away from the lower and left sides and collecting along the diagonal. This difference is caught by the contingency table test, with more time frames with similar activity-levels and many fewer of contrasting activity-levels than expected from the independent model. The test soundly rejects the null hypothesis, while the joint distribution pattern confirms that some of the audiences' temporal pattern of increasing activity is shared. This test result is a strong indication that stimuli can provoke repeatable patterns of response when considering the aggregate of many participants.

4.1.5 Coordination tests for all collections

The section to follow presents the results of coordination tests on increasing and decreasing activity in all 44 continuous response collection. The individual and joint activity tests applied and the resulting p-values (so as to manage the different degrees of freedom dependent on the number of time frames) are reported on a logarithmic scale.

Considering the results of the coordination tests on all of the data sets, in the top graph of figure 4–13, it is interesting to note that most data sets contain both collections with strong rating change coordination and collections which fail to reject the random and independent null hypotheses. Data sets AR1, Kor, and Moz use the same sets of participants for all or half of the collections in the set. This suggests that the differences in coordination of activity from one collection to the next is determined by the stimuli rather than variables associated with the participants.

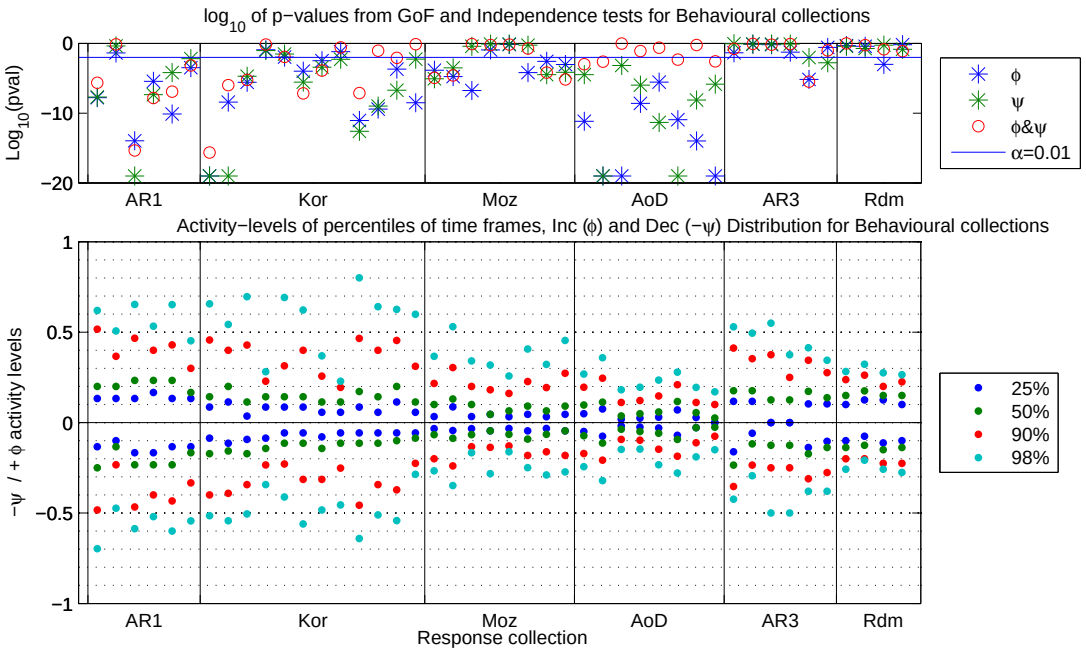

Figure 4–13: Activity coordination test summary for increases in ratings activity, ϕ, decreasing activity, ψ, and the joint activity of ϕ & ψ for the 44 behavioural response collections. In the top graph, stars represent the p-values of the goodness-of-fit tests, o's the p-values from the contingency table tests. When the dots are below the significance threshold (here 0.01) the null hypothesis of random or independent activity may be rejected. The lower graph describes the distributions of activity-levels across the two second time frames of each collection by presenting the quartile and extreme activity-level values for both increasing and decreasing activity by column.

Some collections show strong agreement in both increasing and decreasing activity and also show coordinated joint activity (note that the statistic does not specify whether these series are coordinated together or alternate in degrees of activity). Others fail one, two, or in many cases, all of the activity coordination tests attempted. Those collections which show coordination within each activity series but fail to reject the hypothesis of independence between increasing and decreasing activity require further investigation. This is mostly an issue with the low activity and long excerpt collections in the AoD set. It is reassuring to note that the random collections mostly fail to reject the null hypothesis in the goodness-of-fit tests. Their activity distributions, in the lower graph of figure 4–13, are similar to some of the experimental collections which also failed to show coordinations.

The lower graph of figure 4–13 presents some information on the distribution of these activity-levels across these behavioural response collections. On average, for either form of activity, around a quarter of time frames have less than 10% of responses being active, and the medians (green dots) show similar consistency. The distributions vary the most for the 90th percentile and 98th percentile activity-levels. In the AR1 and Kor data sets, both having been reported using two-dimensional graphical interfaces, a number of collections have between 5-10% of their two-second time frames with activity-levels above 0.5. In the other sets, the instances of a majority of participants showing the same activity in any two-second window is much lower. The fact that most moments of change in responses measured across a collection are driven by a minority of participants should inform

our interpretation of collection summaries such as the cross-sectional average time series.

4.1.6 Conclusions

Audience Activity explores the diversity in ratings and other continuous responses by focusing on events in individuals' continuous response time series and by evaluating the coincidence of such with those of other participants. Despite wide inter-subject variability in continuous ratings, collections of responses often show measurable coordination through the distribution of activity across time and responses. Responses of the same participants to different stimuli result in different degrees of coordination in all of the experimental data sets. Coordination tests show that responses are not acting completely independently, at least not in all collections. However, the distributions of activity show that the responses showing relatively high levels of active agreement in two-second time frames are rarely even half of those in the collection. Though the collections studied here were continuous ratings, other types of responses can be studied with similar or different definitions of response activity.

4.2 Event analyses: determining which moments to study

Continuous responses are interesting to music cognition because music happens in time and our experience of it is admittedly dynamic. Collections of responses make it possible to compare time points within the series, measuring whether a moment of music is aligned with normal or extreme response behaviour. Analyses of the time varying qualities of the musical experience have often been studied through the lens of correlation and regression to aspects of the stimulus, techniques which attempt to fit the full duration of these responses. Theoretic speculation on the temporal dynamics of music listening have taken inspiration from visual interpretation of the collections' cross-sectional averages, but there exists more reliable methods of distinguishing time points of a collection using systematic criteria.

One reason to distinguish between time points of a collection is to discard those which do not yield reliable (read replicable)information. The problem of inter-response variability calls for a measure to assess the reliability of the average response given the cross-sectional distribution of values at each moment. Such a test, designed by Emery Schubert for rating responses, is also intended to provide researchers with "a form of visual display from which meaningful conclusions can be reported" [Sch07].

Another type of time point differentiation is exceptional event finding. Tests of this type isolate moments when responses deviate from some expected behaviour: they identify time points or intervals that stand out from the usual noise. Studies like Sloboda's 1991 survey show that music listeners remember strong

responses at specific moments in music [Slo91]. If their memory of experience is accurate and if the experience is common, exceptional event analysis could expose these affective moments by their traces across collections of continuous responses. One method for identifying these events was first described in [GNKA07a], and discussed in more detail in the following pages.

4.2.1 Second-order standard deviation test

Inter-response variability has been fodder for skeptics of continuous responses to real musical stimuli. While the standard deviation time series, on its own, has not helped the interpretation of other cross-sectional statistics, researchers have been looking for a similar simple calculation to determine when they can have confidence in their results.

To assess the validity of the average time series from one moment to the next, Schubert proposed in 2007 a new method for assessing the significance of time points in collection of continuous responses. This method defines "significant" time points to be those which show relatively low cross-sectional dispersion when compared to other cross-sections of the collection [Sch10]. The threshold is set with respect to the distribution of the standard deviation time series, $\sigma_{\mathbf{I}}(X)$. Published uses of the test have defined the threshold according to equation 4.13, taking the average standard deviation, $\mu(\sigma_{\mathbf{I}}(X))$, plus or minus some factor k of the standard deviation of the standard deviation time series, $\sigma(\sigma_{\mathbf{I}}(X))$. For a collection X of M responses, each time series of length N, with a cross-sectional standard deviation time series $\sigma_{\mathbf{I}}(X) = \{\sigma_i(X)\}$ for samples of index

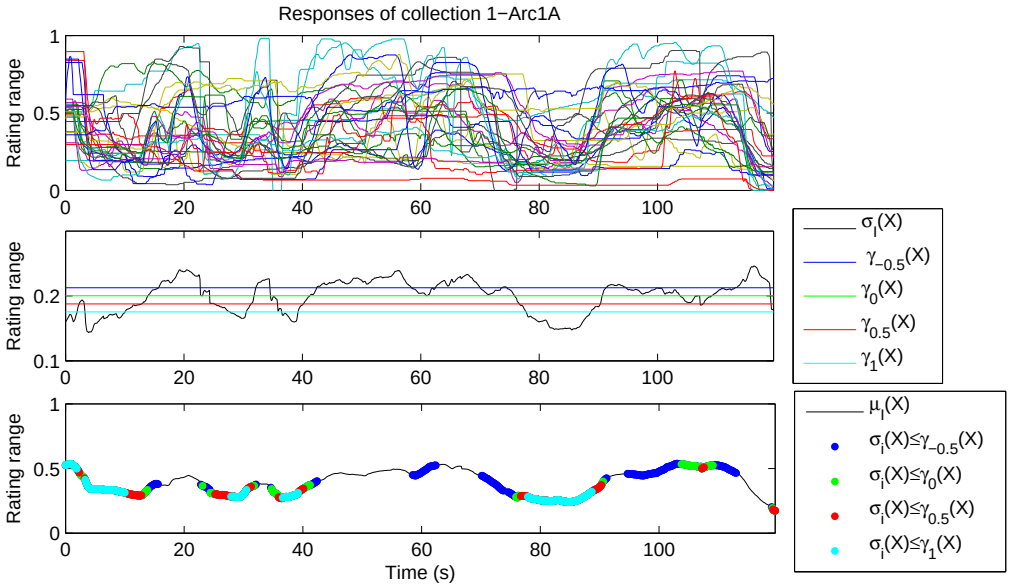

Figure 4–14: Example of second-order standard deviation test for significant events. Above, the collection of responses, middle, the cross-sectional standard deviation time series with different possible thresholds, bottom, cross-sectional average time series with "significant" time points highlighted according to the different possible thresholds. Note: the dots in the lower graph overlap; by the thresholds in the middle graph it can be assumed all points which are green are also dark blue, all red points also green and all light blue points also red.

$i \in \{1, 2, ..., N\}$, the variance threshold $\gamma_k(X)$ would be defined as:

$$\begin{aligned} \gamma_k(X) = &\quad \mu(\sigma_{\mathbf{I}}(X)) - k\sigma(\sigma_{\mathbf{I}}(X)) \\ = &\quad \frac{\sum_{i=1}^{N} \sigma_i(X)}{N} - k * \sqrt{\frac{\sum_{i=1}^{N} (\sigma_i(X) - \sum_{i=1}^{N} \sigma_i(X)/N)^2}{N}} \end{aligned} \tag{4.13}$$

for some $k \in \mathbb{R}$. Thus, a time point, t_i, is significant, according to this measure, if $\sigma_i(X) < \gamma_k(X)$, or more explicitly if:

$$\sqrt{\frac{\sum_{r=1}^{M} (x_{r,i} - \mu_i(X))^2}{M}} < \mu(\sigma_{\mathbf{I}}(X)) - k\sigma(\sigma_{\mathbf{I}}(X)) \tag{4.14}$$

The use of the longitudinal standard deviation of the cross-sectional standard deviation time series gives this technique the name of second-order standard deviation threshold [Sch07]. In figure 4–14, the application of this threshold is demonstrated on collection 1-Arc1A of the AR1 data set. The middle and bottom graphs are the summary time series of the collection plotted at the top of the figure. On the middle graph, the cross-sectional standard deviation time series is cut by thresholds determined by the second-order standard deviation equation for different factors of k. As according to equation 4.14, the time points for which this standard deviation time series fall below these lines are deemed "significant" by the test, and these points are highlighted by coloured dots on the average time series. By using the standard deviation as a measure of variability, this thresholding test depends on the legitimacy of the assumption that the cross-sectional average, $\mu_i(X)$, is an appropriate statistic at all moments t_i, i.e., that the cross-section has a normal distribution. Because of this, the test in some sense quantifies the validity of this average time series as representative of the collection moment by moment.

At this time, there is no recommended formula for the proportions, though published uses have set k to 1, 0.5 [Sch07], and -0.5 [SSM$^+$09]. In a first paper describing the technique, Schubert admits that this measure is entirely relative and would always define some time points as significant [Sch07]. Acknowledging the parametric assumptions of the mean and standard deviation, he also suggests other descriptors of the cross-sectional distribution might be appropriate, such as the median and interquartile values [Sch10].

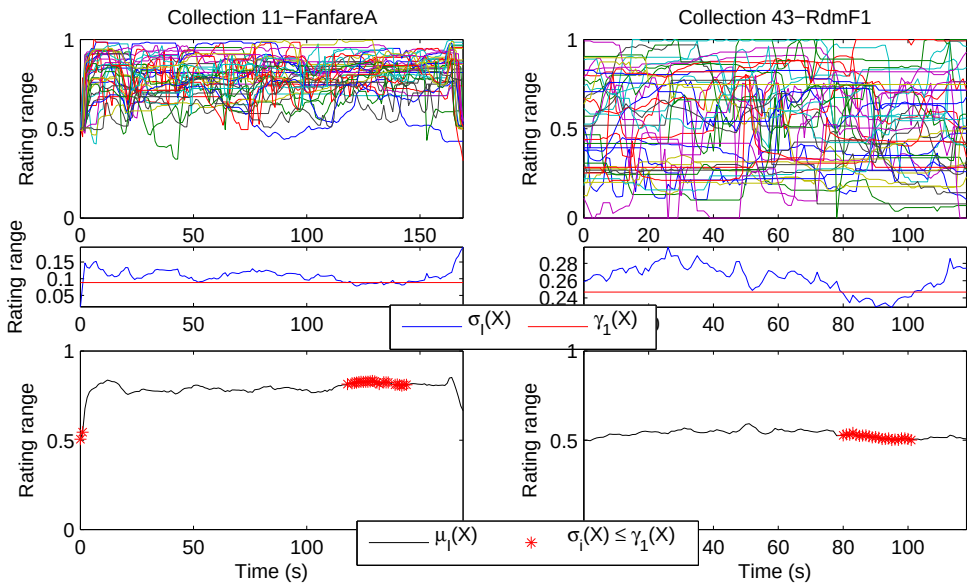

Figure 4–15: second-order standard deviation test as applied to two collections with contrasting degrees of inter-response variability: 11-FanfareA of the Kor data set and 43-RdmF1 from the Rdm set of unrelated response collections. Below the plots of all responses of each collection is the standard deviation time series with the threshold for $k = 1$ and at the bottom, the average time series with "significant" time points highlighted in red.

For analytic intuitions trained on discrete collections of data, the cross-sectional standard deviation seems to be a reasonable indication of how much the average value can be trusted. However, by setting the threshold in terms of the actual distribution of variance, the measure can only tell which time points are more "reliable" than others within a particular data collection. When considering different collections of responses, the same threshold definition could treat time points with the same cross-sectional distributions differently, depending on the behaviour of their respective collections.

Figure 4–15 shows the time points selected as "significant" from two collections using a threshold of $\mu(\sigma_{\mathbf{I}}(X)) - \sigma(\sigma_{\mathbf{I}}(X))$, with $k = 1$. The middle graphs are the standard deviation time series with the second-order standard deviation threshold in red, and the selected time points marked by red stars on the average response series in the bottom graph. Notice how the scale of the cross-sectional standard deviation time series differ between these two collections. It is no surprise that the cross-sectional standard deviation values of the unrelated response collection are, on average, more than twice that of the experimental collection (right). Despite this sizeable difference between the collections, a similar proportion of time points are selected as significant, i.e., as reliable, in each collection by this test. Thresholding $\sigma_{\mathbf{I}}(X)$ in terms of its own distribution makes this measure blind to the actual variability of a collection.

Testing the performance of this test on all of the behavioural collections, figure 4–16 shows the distribution statistics of the cross-sectional standard deviations of each collection above the graph reporting the precentage of time points found to be significant in each collection, according to different values of k. The longitudinal average of the cross-sectional standard deviation series varies within and between data sets, while the variance of these summary time series seem to be much more limited in range. Some researchers stopped publishing graphs of the cross-sectional standard deviation series because they found it "provided little information " [GMG04] with such small variance over time. The patterns of percentages found significant for each k do not correlate across collections with either the $\mu(\sigma_{\mathbf{I}}(X))$ or the $\sigma(\sigma_{\mathbf{I}}(X))$. The performance of the

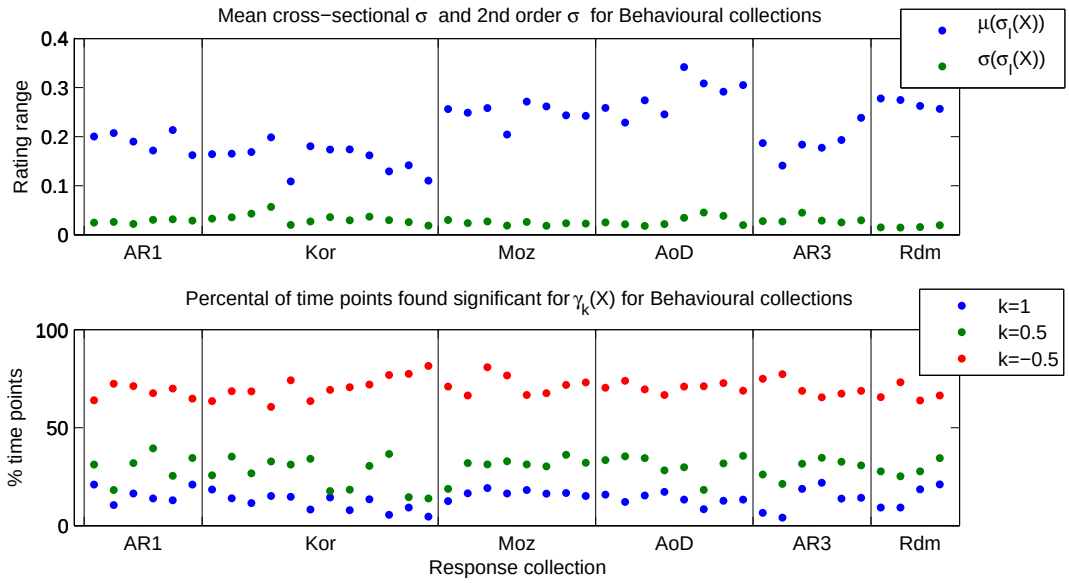

Figure 4–16: Distribution parameters of cross-sectional standard deviations and performance of second-order standard deviation significance test on all 44 behavioural response collections with three thresholds of standard deviation time series $\bar{\sigma}_X - k * \sigma(\sigma_{\mathbf{I}}(X))$ with k of three values used in publications.

test on the unrelated response collections also show that the moments found, whether many or few, can not be assumed to be indicative of inter-response coordination. While simple to calculate, this technique is evidently too arbitrary to make "significant" distinctions between time points in a collection of continuous responses.

4.2.2 An extreme event test

At the other end of the time-point differentiation problem is indentifying moments of extreme or deviant response behaviour. Rather than remove the unreliable time points from future analysis, the goal for this kind of "significant event" test is to identify those few moments over the course of the stimulus during

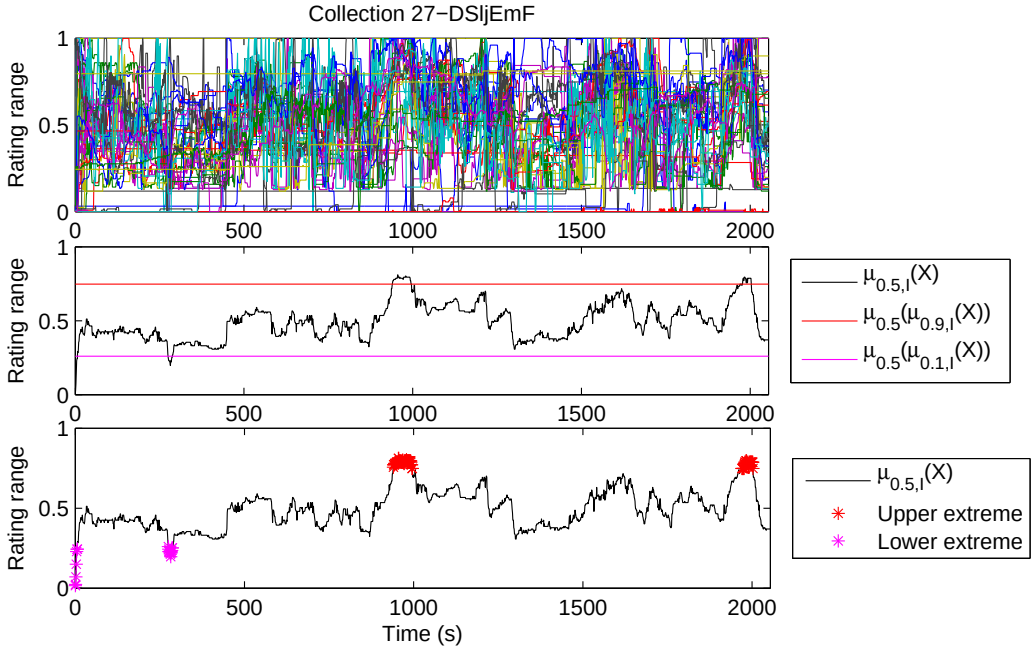

Figure 4–17: Upper and lower extreme event test for collection 27-DSljEmF of the AoD data set. Below the plot of all responses is the median response time series with thresholds for testing, and on the bottom, the median with points starred for passing the Wilcoxon signed-rank test against the longitudinal median.

which particularly strong responses are reported by/recorded from many or most listeners. Oliver Grewe et al., published an interesting technique for identifying moments of extreme responses using the Wilcoxon signed-rank test to evaluate the significance of these response events against the average, or rather median, cross-section of responses [GNKA07a]. Using non-parametric distribution criteria, a threshold is calculated for the median time series of a collection of responses. The cross-section at time points whose median values exceed this threshold are then tested against the collections longitudinal median distribution. If the test

shows the response values at that time point to have a significantly different distribution from the longitudinal median, this time point is labeled as an extreme event. Grewe et al., use the term affect event because this test was developed for continuous responses related to emotion [GNKA07a], however the same process may be applied to measures of other aspects of responses.

The process outlined in the *Emotion* article defined the threshold for cross-sectional medians as the median of the 90th percentile values of individual time series. Note that if the response are really variable in range, it could be that the median time series would never reach threshold. The authors did not explain the choice of 90th percentile, and lower bounds, such as those used in figure 4–18. Using the notation described in chapter 3 for the median and quartile values of a set, we define the longitudinal 90th percentile response values of collection X to be $\mu_{0.9,R}(X) = \{\mu_{0.9}(\mathbf{x_r})\}_{r=1}^{M}$. The threshold is then the median of this set, $\mu_{0.5}(\mu_{0.9,R}(X))$. A time point, t_i, would qualify to be tested for significance if $\mu_{0.5,i}(X) \geqslant \mu_{0.5}(\mu_{0.9,R}(X))$, or, more explicitly, if:

$$\frac{\|\{r \in \{1, 2, ..., M\} \mid x_{i,r} \geqslant \mu_{0.5}(\mu_{0.9,R}(X))\}\|}{M} \geqslant 1/2 \tag{4.15}$$

This thresholding of the median can be seen in figure 4–17. In this figure, the threshold has been calculated for the 90th percentile and the 10th percentile to catch both upper and lower extremes. Applying the threshold to the median time series ensures that only moments when the majority of responses are above the articulated threshold are considered for testing. Even though the threshold is based on the actual distribution of values within the collection, the median time

series many never reach these extreme values, as is the case for the lower bound on 19-K492L in figure 4–18.

This test was initially applied to the rating response collections under the first-order difference transformation alongside physiological responses [GNKA07a]. To demonstrate its effectiveness on both the original rating values and differenced responses, figure 4–18 shows the results for two collections under less stringent thresholds based on the 80th and 20th percentiles. Even with this more inclusive criterion for time-points to qualify for testing against the longitudinal median, these events are not common. In collection 19-K492L, there are not time points at which the median of the differenced data is significantly non-zero. This kind of thresholding ensures that the points investigated always reflect the behaviour of the majority of responses, and if a majority never acts in concert, the collection fails to show any extreme events as defined by this test.

The fact that the Wilcoxon signed-rank test is non-parametric makes it all the more appropriate for evaluating these collections' cross-sections, as can be seen in figure 3–1. The Wilcoxon signed-rank test is a more powerful version of the non-parametric pair-difference test, the sign test. Given two sets of values, $X_i := \{x_{i,r}\}_{r=1}^{M}$ and $X_j := \{x_{j,r}\}_{r=1}^{M}$ paired by index r, a sign test evaluates the difference of each pair, $z_R := \{x_{j,r} - x_{i,r}\}_{r=1}^{M}$, with the null hypothesis that $\mu_{0.5}(z_R) = 0$. This hypothesis is tested by counting the number of positive z_r and number of negative z_r and the test statistic is the minimum of these two values. The sign-rank test adds another factor to this comparison of positive and negative paired differences by summing the ranks of the absolute value of

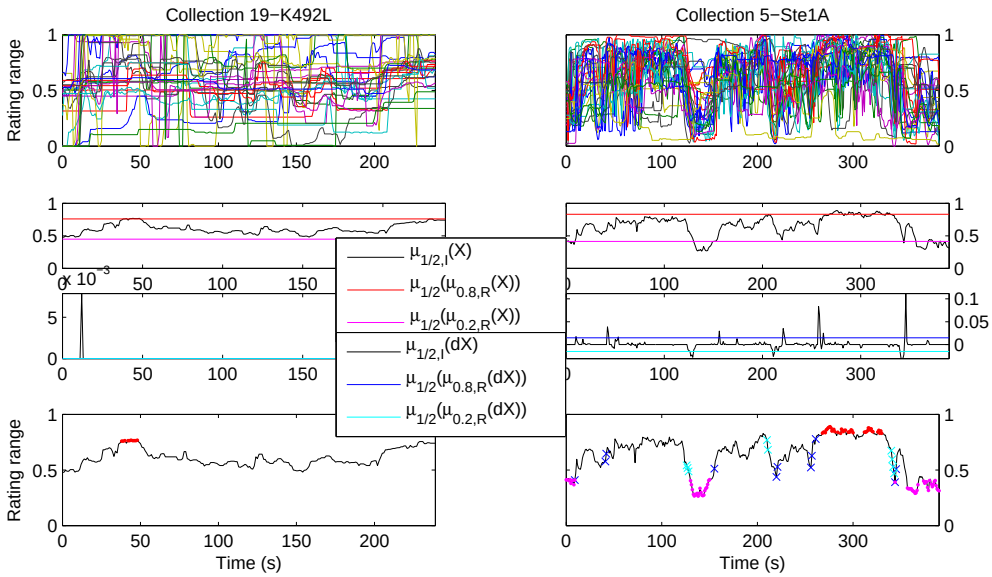

Figure 4–18: Example results of extreme event test for collection 19-K492L of the Moz data set and 5-Ste1A from the AR1 set. Below each collections plot of responses are the median time series for the collections and their first-order difference. The bottom graphs marks extreme events of either type with stars in the colour of the threshold on the cross-sectional median time series.

each difference z_r for positive and for negative differences. The minimum of these sums is the measure of deviation from the median null hypothesis [Wil45]. Given the number of pairs, or rather the number of non-zero pair-wise differences, the likelihood of the signed test statistic can be calculated with respect to their expected distribution under the null hypothesis. Grewe et al. used a significance level of 0.05 [GNKA07a].

Besides highlighting the differences in response ranges per collection, figure 4–19 shows the difference between the median longitudinal percentile criterion and a threshold based on the distribution of the cross-sectional medians (the latter

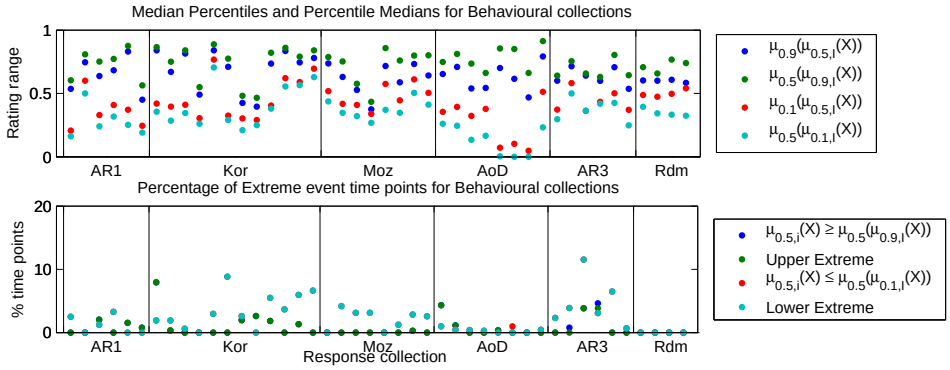

Figure 4–19: Comparison of median-based thresholds and proportion of time points measured as significant events across behavioural response collections.

being the non-parametric equivalent of second-order standard deviation threshold). The former thresholds are always more extreme, and often the difference is enough to prevent all time points from being tested. The median 90th percentile criterion yields no extreme events in the unrelated responses collections, but it also shuts out some of the experimental collections. The lower graph of figure 4–19 shows the percentage of time points exceeding the thresholds (high and low) as well as the number passing the Wilcoxon signed-rank test against the longitudinal medians of each collection. Most of the collections had some time points selected to be tested by the initial thresholding, at least on one of the extremes, but few tested both high and low. In only three collections did any of the selected time points fail to reject the null hypothesis of the signed-rank test. This suggests that more time points in these collections would be found "significant" by the same test. Rather than violate multiple test constraints further by testing all points, a lower threshold, as seen in figure 4–18, could catch more extreme events. Complimentary

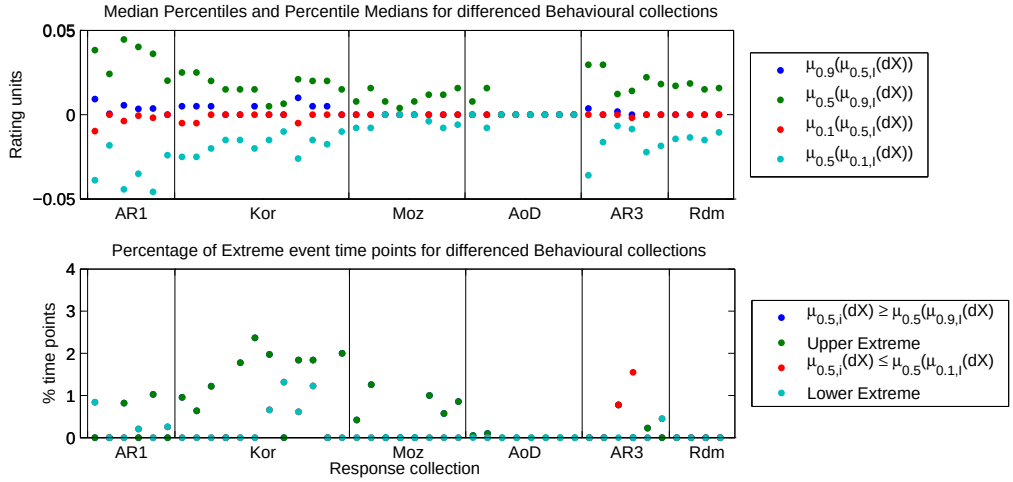

Figure 4–20: Comparison of median based thresholds and proportion of significant events measured across the first-order difference (1 Hz)of all behavioural response collections.

evaluation of these extreme events could inform the selection of a most practical percentile value or other thresholding criteria.

The same evaluation applied to the first-order difference of these data collections (sampled at 1Hz) shows similar results, plotted in figure 4–20. For these difference data, it is important to note the number of collections for which the 90th and 10th percentile of the median series are zero. As discussed in the context of activity analysis, it is very rare to have even half of the responses in a collection showing the same kind of change in ratings in the same short time interval. Those few points found to satisfy the cross-sectional median threshold almost never fail the significance test. The exceptions here are from data collections of very small numbers, containing less than 10 responses. More collections would have time points exceeding these thresholds if the difference transformation was on

a lower sampling frequency, but even with such adjustments, another kind of selection criterion may be needed to find moments of sufficiently drastic change. Another option might be to first sort responses within collections to get higher concentrations of similar patterns of response variation.

The purpose of this test is to find moments of coordinated extreme responses within the audience. By using the median time series and a threshold derived from longitudinal distributions and a statistically sound test of significant difference from median response behaviour, this test does a good job of picking out extreme time points. It is not clear, however, what might be missed in the process of applying such strict selection criteria. For both the ratings and the first-order differenced ratings, the cutoff of $\mu_{0.5}(\mu_{0.9,R}(X))$ does not seem to be the most efficient means of selecting time points of extreme responses as it renders the signed-rank test redundant in the vast majority of cases. Further study of this test and the implications of each step may yield a less stringent but more useful variation of Grewe et al.'s initial parameters.

4.2.3 Conclusion

There is a great deal of interest in the moment-by-moment details of continuous responses. The number of pages spent expounding on the details of the average time series' contour is but one sign of the need for reliable methods for systematically and skillfully detecting time points of interesting character. The methods presented in the section are just a beginning.

Assessing reliability or validity of responses, and in particular the average response time series, could ground a lot of speculation. However the test proposed

by Emery Schubert in 2007 is not the solution. Such a test would need to manage both the non-parametric distribution of cross-sections and the consequences of the serial nature of these data.

The model of extreme event finding laid out here is also a starting point for identifying isolated moments of particularly distinct responses. Given the results of activity analysis, study of simultaneous changes in response may require criteria that do not depend on the majority of responses showing the same activity, and further consideration must be given to selecting the most useful threshold, but the introduction of non-parametric criteria to this area of analysis is a very helpful step towards more reliable results.

Besides these two, there are many possible methods for differentiating time points or time intervals within collections of continuous responses. Activity analysis, for example, provides new basis for this process without depending on central tendency statistics. But before methods become established, definitions of what kind of response behaviour should count as significant need further development.

4.3 Clustering: grouping responses by profile and character

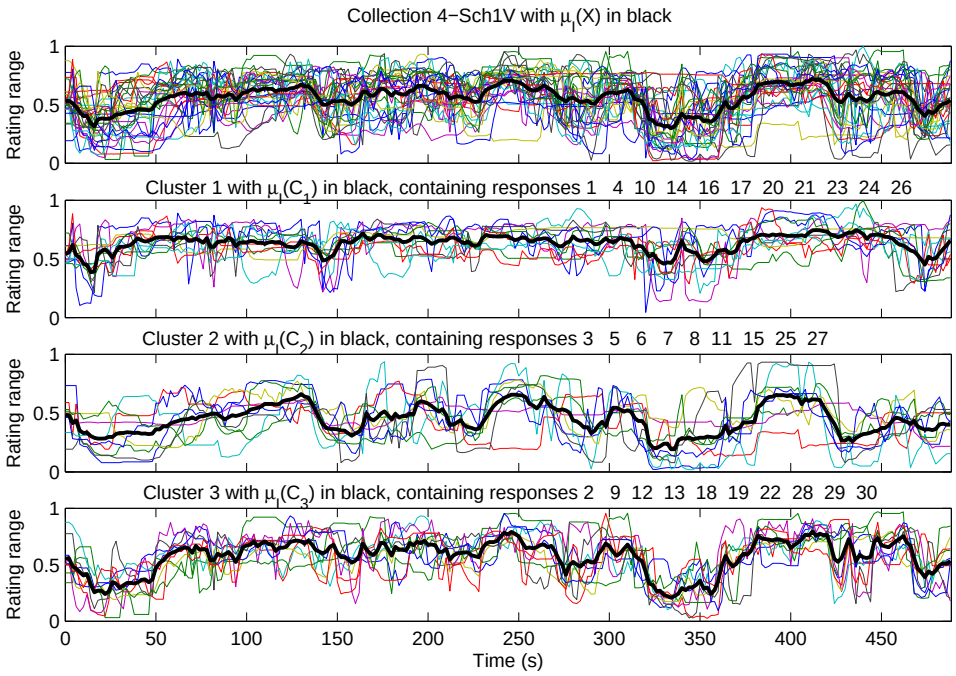

Figure 4–21: Clustering of responses in collection 4-Sch1V from the data set AR1. Using Euclidean distance and Ward linkage, three clusters are presented below the plot of all the responses, each with the responses' average time series.

The variation between participants responses to the same stimuli has often been noted, but despite this, most traditional analysis methods presume that there is a single ideal response underlying the responses collected to the same stimulus. Such singular representations of a collection's continuous responses mask the contradictory patterns of experience noticeable when studying individual responses. Currently, there are no reliable measures of how well a central tendency time series represents the individual responses in a collection. Given the detrimental effect of this diversity on standard summary statistics, it may be practical to consider the

possibility of multiple distinct temporal profiles of response within a given collection, as suggested in figure 4–21. In this section, simple hierarchical clustering is used to explore whether responses within collections can be separated into distinct clusters of related temporal profiles.

Botryology is admittedly a finicky science, if not a black art [vD00], and time series data add to it the same challenges that hinder other forms of time series analysis. There have been many creative time series clustering studies [L+05], but not all have the same purpose. The following goals have contributed to the selection clustering criteria:

1. Similar response values: Fundamentally, we want to know if different participants show similar responses to the stimulus at any given moment. The simplest evidence of this in collections of response ratings would be responses showing the same rating values at the same time.
2. Similar response behaviour: Participants may have similar experience but perform rating tasks differently. Differences such as varying degrees of sensitivity to the rating scale, faster or slower reaction times, or distinct interpretations of the task can hide similar experiences. To study these general behaviours, responses can be grouped by characteristics such as rating range use, or sudden versus gradual rating changes, whether or not these characteristics are separable from the gross rating value information.
3. Similar moments of response: Participants may have different responses while being affected by the same events in the time course of the stimulus. This would be expressed by changes in response around the same time, despite

possible differences in response values. Studying responses by contour or rating change activity would show groups of this type.

It is common place to normalize time series prior to clustering [KK03], by range, mean, or variance. Continuous rating data, however, are collected on a fixed, finite range, and normalization would interfere with the first goal of clustering similar response values. Some distance metrics are sensitive to the absolute values of data sets, while others are designed to measure relative values.

As mentioned above, continuous responses may show differences due to reaction times of participants: it takes a moment for subjects to realize their experience and another moment for them to express it. To reduce the effect of variable reaction times, the responses have been down-sampled to 1 or 0.5 Hz (depending on the technique). This lower sample rate is also useful for reducing the sensitivity of rating change timing for goal no. 3.

4.3.1 Hierarchical clustering of continuous responses

There are many ways to build hierarchical clusters. The process and result depend on choices of distance metrics and linkage criterion. For a first attempt, the standard Euclidean distance metric should suffice. Given two vectors of the same n-dimensional vector space, $\mathbf{x_r} := \{x_{r,i}\}_{i=1}^{i=n}$ and $\mathbf{x_s} := \{x_{s,i}\}_{i=1}^{i=n}$, the Euclidean distance between these vectors is defined as:

$$d(\mathbf{x_r}, \mathbf{x_s}) = \sqrt{\sum_{i=1}^{n}(x_{r,i} - x_{s,i})^2} \tag{4.16}$$

This metric measures the distance per time point between responses on the units of the response scale. This is different from a Pearson product-moment

correlation, a more popular similarity measure for these data, which is variance neutral. The Euclidean measure is effective at capturing the difference between responses as per the first clustering goal: grouping by rating value.

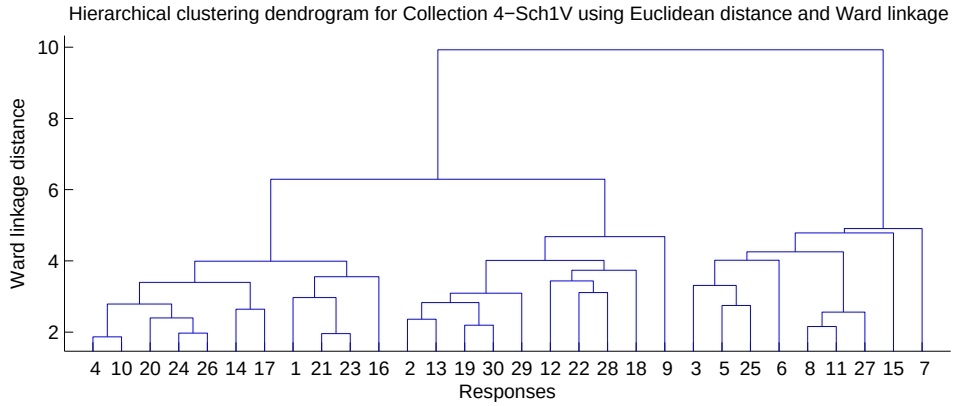

Figure 4–22: Hierarchical clustering of collection 4-Sch1V of the AR1 data set. The dendrogram is a bottom-up construction using Euclidean distance with Ward linkage criterion for cluster formation.

Bottom-up hierarchical clustering is a recursive process of joining the elements, or previously clustered subset of elements, which are closest to each other until the whole collection of elements have been joined into one cluster. After the first step of measuring the distance between all elements and joining the closest two, it is necessary to have some way of measuring the distance between subsets of elements of size greater than one. The method to assess that distance is called the linkage criterion and it is defined in part by the element distance measure. The y dimension on dendrograms such as figure 4–22 corresponds to the linkage distance measure: the height of each link joining two clusters represents the measured distance between the merging subsets.

A common measure of the distance between two sets of vectors is called the Single linkage criterion. For two subsets R and S with elements $\mathbf{x_r}$ and $\mathbf{x_s}$, the Single distance is the minimum distance measured between all possible element pairings between the sets. This linkage criterion makes it easy for sets to grow one element at a time, often resulting in staircase-like dendrograms. A more exclusive distance metric between sets is the opposite measure, called the Complete linkage distance. This takes, instead, the maximum distance between any two pairs of vectors:

$$D(R,S) = max\{d(\mathbf{x_r}, \mathbf{y_s}) \mid \forall\, \mathbf{x_r} \in R, \forall\, \mathbf{x_s} \in S\} \tag{4.17}$$

The Single and Complete distance measures are linkage criteria that can be applied to any well-defined element-wise distance metric, $d(\mathbf{x_r}, \mathbf{x_s})$. When the vector distance metric is Euclidean, there are other commonly used options that manage to compromise between these extremes. Clustering responses to minimise variance would satisfy goal No. 1, for which the Ward measure would be appropriate. According to Liao, et al.: "The Ward's minimum variance algorithm merges the two clusters that will result in the smallest increase in the value of the sum-of-squares variance. At each clustering step, all possible mergers of two clusters are tried. The sum-of-squares variance is computed for each and the one with the smallest value is selected." [L$^+$05] One method of calculating this measure evaluates the difference in the centroids of two vector sets weighted against the number of vectors composing the sets. If the centroid of a set R with N_R elements $\mathbf{x_r}$ is defined as the average of elements in the set $\mu_\mathbf{R}$, then the Ward

distance for two sets R and S would be given by:

$$D(R,S) = N_R N_S \frac{d(\mu_{\mathbf{R}}, \mu_{\mathbf{S}})}{N_R + N_S} \tag{4.18}$$

Hierarchical clustering is sensitive to the choice of linkage criterion. On these continuous response collections, the Single distance measure most often results in staircase dendrograms; making a long sequence of nested clusters rather than forming distinct subsets at any given level. The Complete distance measure, on the other hand, amplifies the distance between subsets and distinguishes clusters that may not be robustly distinct. The Ward metric is somewhere between the two, and is used in these first examples of clustering collections of continuous response.

Note that these measures are not designed for time series. When applied to serially sampled data, each sample point is treated as its own dimension. Insensitive to the visually obvious parallels behind spurious noise like the phase variation in reaction times, there are aspects of similarity and difference in these time series which will be missed by the above clustering approach. Despite this, alternative measures designed for time series rarely out-perform the Euclidean distance measure [KK03]. One successful technique to get around this temporal blind spot of the standard Euclidean measure is to apply controlled dynamic time warping [RK05], although that has not been attempted here.

The last step in generating clusters from an agglomerative hierarchical clustering process is to apply some criterion for cutting the dendrogram tree into distinct subsets of the full collection [TSK$^+$06]. One method for cluster selection depends on the linkage measure by selecting a threshold above which all links are

ignored. A similar tactic is to specify in advance how many clusters should come out of the process. This criterion uses an algorithm to find a threshold value at which slicing the tree will leave the "right" number of subsets. When something is known about the possible clusters, these methods are quick and simple means for selecting clusters.

With no precedent for clustering continuous response data, there is no experience to inform these criteria. Thus, it is more useful to work with a clustering criterion that measures the quality of each potential cluster. The inconsistency of a cluster is a measure of the highest link height against the nearest links the cluster contains. If one cluster is very tight and the next step up the hierarchy joins it with a very different cluster, the inconsistency measure will have a high value. The inconsistency measure uses the distribution of link heights of the present link and those immediately below it and can be expanded to include links deeper into the component clusters. The inconsistency clustering criterion results in a set of largest clusters which cover the full collection and in which all component links fall below the inconsistency threshold. The clusters in this section used a depth of 3 for the inconsistency measurements and the results of several thresholds were compared. Rather than present all clusterings found, below are some examples of typical clustering outcomes from the behavioural data collections.

Figures 4–22 and 4–24 are examples of dendrograms of agglomerative hierarchical clustering using the Euclidean metric and the Ward linkage criterion. The relatively large distances between higher levels of the hierarchy suggest that these responses separate into distinct clusters. In the example of the felt emotional

valence ratings to a recording of the Andante movement of Schumann's 3rd String Quartet, the dendrogram in figure 4–22 suggests three separate clusters, each with similar sizes and degrees of internal variation. An inconsistency threshold of 1.5 (depth 3) selects these three cluster as presented in the lower three graphs of figure 4–21. While these cluster do share some characteristics, their differences reflect the particularities of their subset. For example, the last two clusters are very similar between seconds 250 and 400, but the first and third are more similar between 100 and 250 seconds. All three show similar drops in ratings at the beginning of the stimulus, but the bottom two show much greater variation in mean rating values than the first cluster. Compared to the average rating for the entire collection, these cluster centroids some distinct "common" response behaviours.

Not all response collections group in comparable numbers. In figure 4–23, the colours of different inconsistency thresholds spread differently depending on the response collection. The second graphs shows the median within cluster cross-sectional variance against the same statistic for each entire collection. The complete sets have higher variance, but depending on the collection, their transition is smooth or there is a big gap from one thresholding criterion to the next. The last figure shows a similar trend, with the variance of the median being higher within the many clusters of small size than the complete or nearly complete collections. These graphs suggest that, at least in some cases, collections can be clustered to satisfy goal 1 and, for the purpose of finding ideal and distinct responses, many small clusters may be more interesting than a few big messy sets.

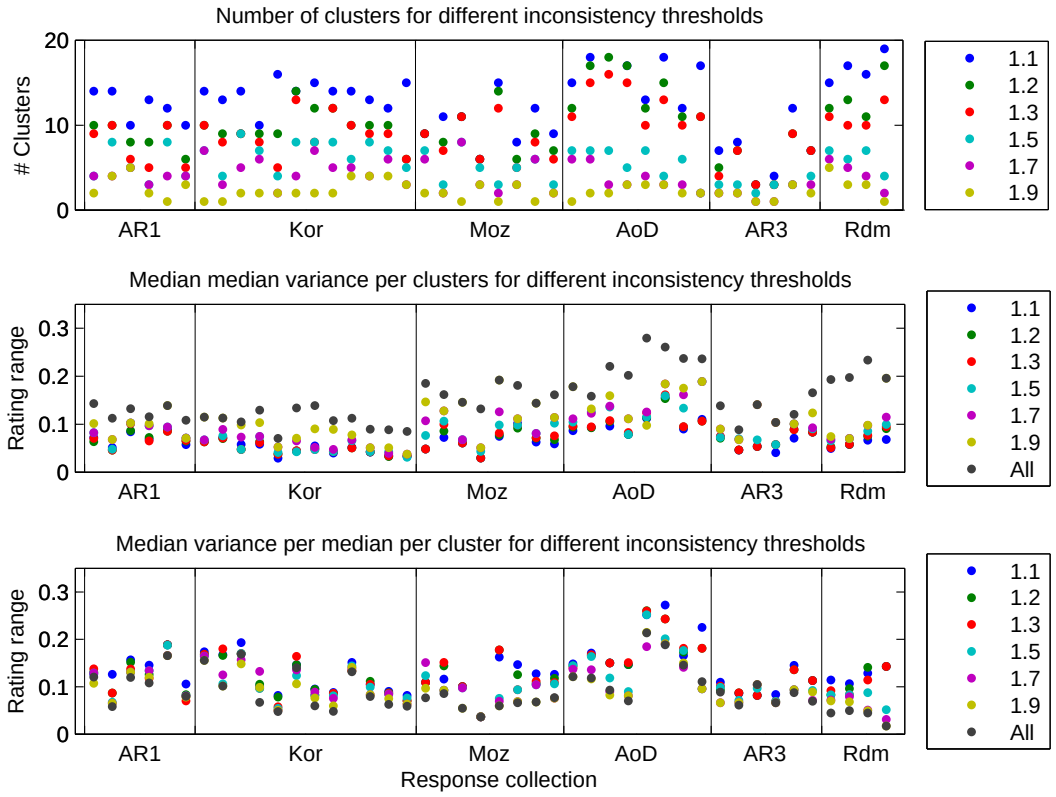

Figure 4–23: Hierarchical clustering summary of all behavioural collections with many degrees of inconsistency. The top graph presents the number of clusters resulting for each threshold; the second graph plots the median of the longitudinal median for the cross-sectional non-parametric half-interquartile evaluated per cluster; the third graph plots the median of the half-interquartile distance of the distribution of the median time series of each cluster.

The responses of 18-PizzicatoA do not cluster as quickly as the valence ratings of Schumann piece. Figure 4–24 shows the dendrogram generated by the algorithm. This clustering suggests a few thresholds for clean cuts, but the same inconsistency threshold of 1.5 with a depth of 3 results in seven distinct clusters within the collection. Figure 4–25 shows these clusters with their mean time series below the ratings of the complete collection.

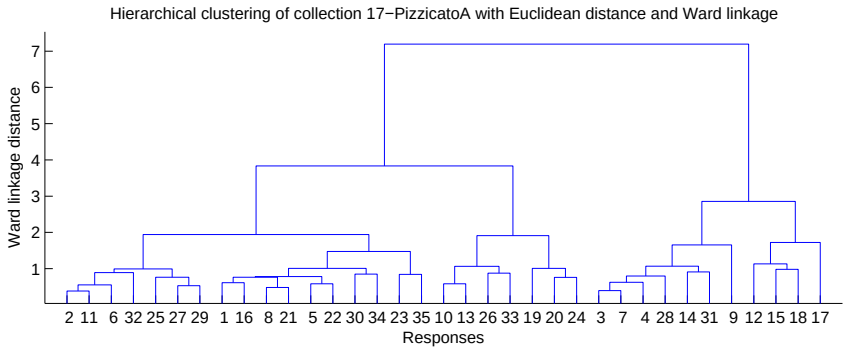

Figure 4–24: Hierarchical clustering of collection 17-PizzicatoA of the Kor data set. The dendrogram is a bottom-up construction using Euclidean distance with Ward linkage criterion for cluster formation.

In the clustering shown in figure 4–25, each subset is much less variable and more noticeably distinct that the three clusters in figure 4–21. The second-to-last of the seven clusters consists primarily of responses which go high quickly and stay there. The last cluster, in contrast, meanders around the midpoint of the range and shows similar points of change from the positive arousal to negative arousal range. While the relationship of the average time series within each cluster to that of the whole collection is visible, the differences in degree of change per moment, related to goal No. 2, are also reflected in these small subsets.

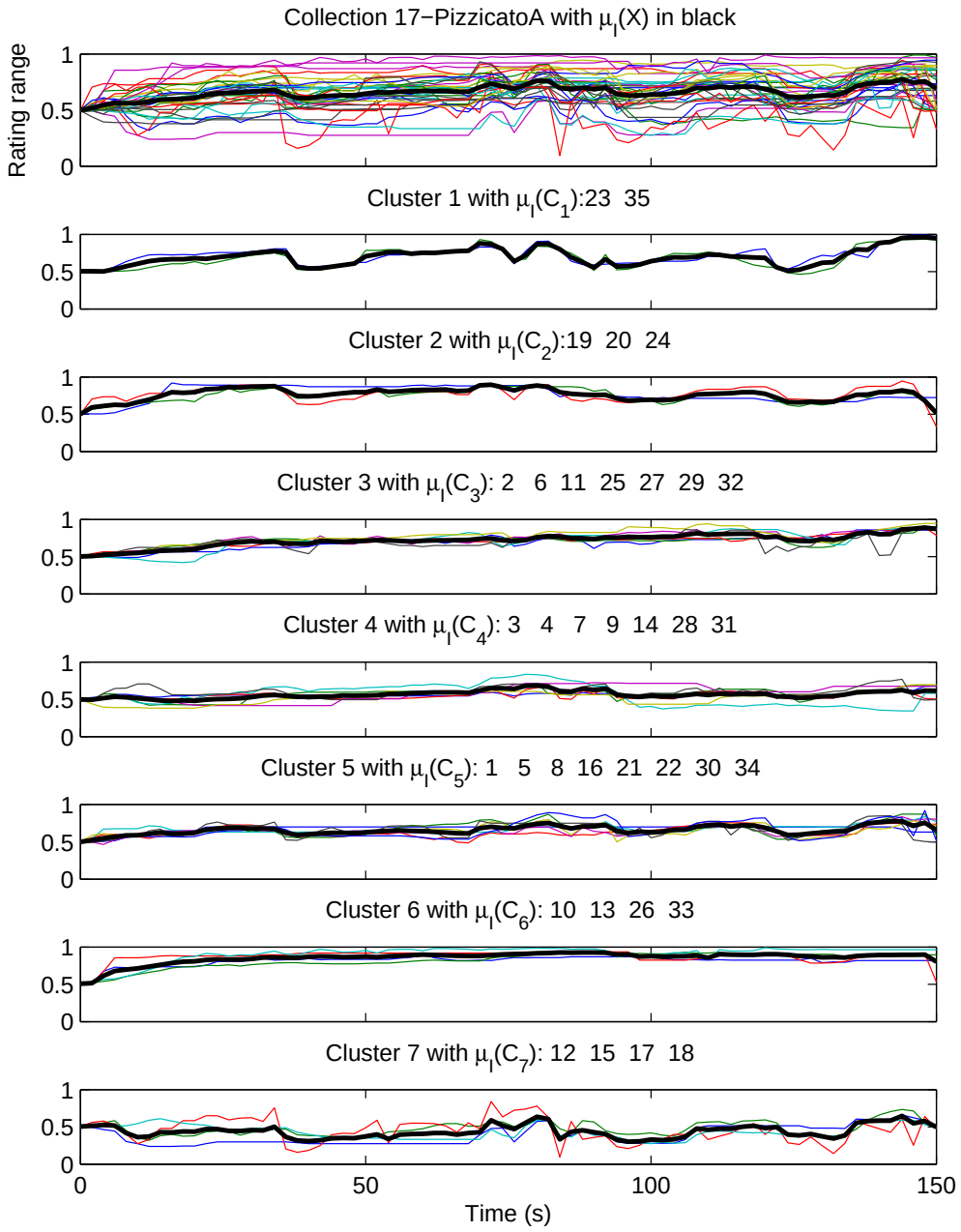

Figure 4–25: Clustering of responses in collection 17-PizzicatoA from the data set Kor. Using Euclidean distance and Ward linkage, seven clusters are presented below the plot of all the responses, each with the responses' average time series.

These clusters show that rating time series can show very similar responses despite individual differences while being distinct from most other responses in the collection. The clusters show a matching of task behaviour as well as response profile. With larger response collections, robust distinct clusters may emerge more clearly.

4.3.2 Hierarchical clustering on first-order differenced responses

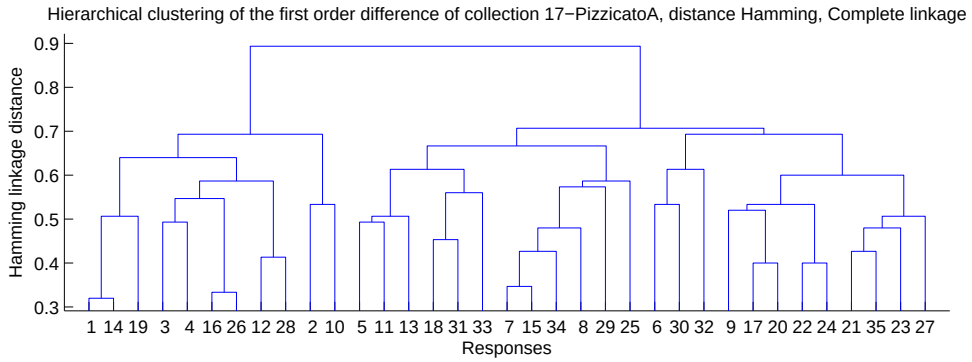

Figure 4–26: Hierarchical clustering of the sign first-order difference of the collection 17-PizzicatoA of the Kor data set. The dendrogram is a bottom-up construction using Hamming distance with Complete linkage criterion for cluster formation.

The first clustering approach tried to address the first two clustering goals, but within each cluster, differences in the "when" of ratings are suppressed in favour of the "how". Clustering the first-order difference of response collections would capture similarities in contour and other aspects of rating behaviour. To remove the bias of expressiveness, the first-order difference of each downsampled response is simplified to the sign-difference series of 0's 1's and -1's. For each sample, the sign-difference series specifies whether that participant's rating

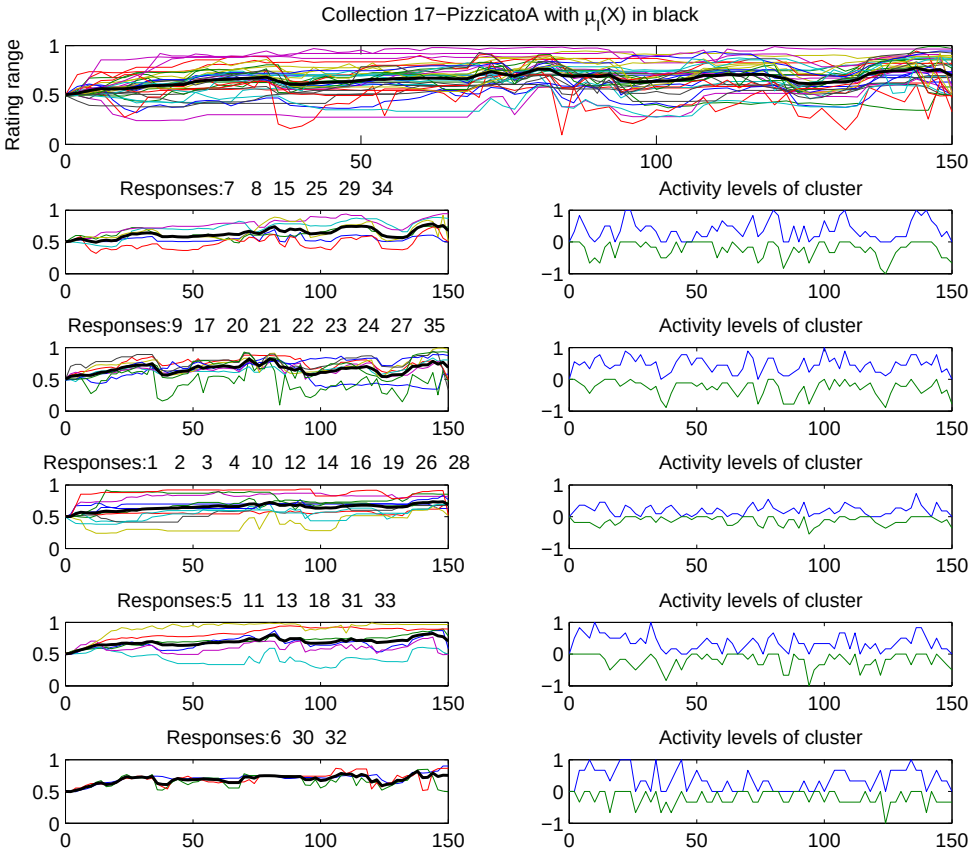

Figure 4–27: Clustering of the signed first-order differences (0.5 Hz) of responses in collection 17-PizzicatoA from the data set Kor. Using Hamming distance and Complete linkage, five clusters are presented below the plot of all the responses, each with the responses' average time series. To the right of each cluster plot is the rating increases (blue) and decreases (green, negative) activity-level time series to show the common activity within each cluster.

increased, decreased or stayed the same over the two second window centred on the associated time point.

For these series, it is also reasonable use another kind of distance measure. The purpose is to compare when response ratings change and when they stay constant. For this, a useful measure is the Hamming distance, a measure of the ratio of dimensions, or in this case, time points, in which the two vector elements do not match. Because the Hamming distance is not Euclidean, these hierarchical clusterings were performed using the Complete linkage criterion to sort responses into distinct clusters.

The clustering outcomes of the sign-difference data were similar in number and distinctness to the clustering using the Euclidean/Ward combination on the untransformed ratings. But one difference in the clustering behaviour is the speed of the collapse to smaller numbers of clusters. Using these distance measures and response representations, some collections require much higher allowances of inconsistency to move from many small clusters to larger subsets. To compare this clustering's performance with the last, consider again the collection of perceived emotional arousal ratings to the Strauss' Pizzicato Polka. The y axis in Figure 4–26 reflects the use of a different distance metric and linking criterion, and the hierarchical clustering differs from the initial pairings to higher clusters of the dendrogram in Figure 4–22. Still, the resulting clusters show behaviours of interest to goal No. 3, gathering responses which share moments of stillness and rating change. Figure 4–27 presents the five selected clusters, each beside their respective activity-level time series in Figure 4–27, as explained in section 4.1.

The five clusters selected by the inconsistency threshold of 1.5 are not very tight around the responses' mean time series, however the incidence of complete agreement in simultaneous increase and decreases show how these clusters differ from each other. The cluster rating change activity-levels is a counting of the ratio of responses changing ratings. The blue lines specify the proportion of responses in the cluster increasing ratings in that time window, while the green line traces the negative of the proportion of responses decreasing in each two-second interval. The first cluster then shows six time points at which all response move together with values of 1 and -1. The strongest contrast between clusters may be between the third and all other subsets, as this cluster brings together the responses which change the least often—clustering by stillness rather than by common moments of active response. Another type of difference can be seen between the last cluster and the rest. This set of three ratings share a calm in the middle of the piece by not changing ratings while the other clusters show a mix of strong and weak activity-levels. Instances of participants changing ratings together may be caused by similar sensitivities to the specific aspects of the stimulus.

Both hierarchical approaches expose patterns of rating behaviour while the clusters for specific collections might differ greatly depending on the clustering criteria. The activity profiles and cluster average time series suggest that separable clusters may still share similar response patterns to subsections of stimuli. It may be that participants do not fall into one or three ideal listening tracks but instead skip between listening approaches as their attention shifts.

4.3.3 Conclusions

Cluster analysis has a distinguished history of leaving researchers with more questions than answers. This introductory effort to cluster responses suggests new possibilities for continuous response analysis. Responses to specific stimuli may contain a lot more significant information than might be assumed when looking at the mess of complete response collections. Without considering the different ways in which participants can respond, it will be hard to make sense of the weak trends which are presented in summaries of the collection.

Time series clustering may also be greatly improved with the use of dynamic time warping. The responses here vary too widely in contour for simple techniques to be successful, but controlled warping may be the key to cleaning up these responses into clearer distinct subsets. At the very least, these clustering efforts once again show that these responses do diverge greatly from the simple average and, with clustering, those differences finally show themselves to be repeatable rather than accidental.

4.4 Conclusion

Collections of responses have a lot more to share than is made available through the techniques discussed in chapter 3. This set of novel approaches present new relevant and robust information about listeners' responses to music.

Given the results of section 4.1 on activity analysis, researchers using continuous response data should realize that testing for coordination should always precede analyses of summary time series like the cross-sectional average. And while responses in the collections studied here do not seem to express a single united

experience per stimulus, this diversity need not be dismissed as noise. Rather than lessening the information to be found in the collections, contradictory responses may yet show themselves to be repeatable and legitimate expressions of distinct experiences of the music.

Coordination plays an important role in identifying when responses more or less agree and when the music really moves listeners' responses. The exploration of extreme events and activity-level time series suggest an event representation of stimuli may be more fruitful for modelling responses than the current popular continuous feature approaches.

These techniques are admittedly just beginnings of new directions for continuous response analysis. However, they already promise to be useful tools, not only because they stand on firmer statistical and numerical ground than some of the more popular techniques in the literature.

CHAPTER 5
Conclusions

From the many analyses included in the previous chapters, a number of conclusions and many more questions arise. Besides discussing some outcomes of this study of methods, this last chapter addresses trends and characteristics of these collections as exposed by the different data treatments.

5.1 Traditional analyses

The traditional analyses discussed were the most broadly employed techniques, and their popularity has been encouraged by their familiarity and simplicity. Each of the three categories of analyses described in chapter 3 had been scrutinized in the literature before, and the repeated applications and analyses here should give more weight to the concerns expressed by other researchers. The principle conclusions are as follows:

1. It cannot be assumed that these continuous responses data are distributed normally, either in the cross-sections of collections, or in the longitudinal distributions of values in each response. Using the median and the interquartile distance as alternatives to the arithmetic mean and standard deviation allows for more flexible and more explicit descriptors of these distributions, though these measures on cross-sections are still sensitive to individual responses dynamic behaviour.

2. Central tendency statistics, even non-parametric ones, should be treated with skepticism, as there can be broad disagreement between responses to the same stimuli. Beyond issues of skewness and kurtosis, the data being summarized may be bimodal, in which case, neither the mean nor the median effectively represent the distribution. Reporting dispersion measures, particularly the more sensitive interquartile range, is one way of acknowledging this disagreement.
3. Correlations, Pearson or otherwise, may not be measuring relevant differences between responses. And while they may be interpretable for some applications to continuous response data, the standard estimates of the significance of these measures of covariance do not apply to time series. Likelihood estimates that use the number of sample points directly as the degrees of freedom depend on the assumption that each datum is sampled independently. Though complicated alternatives may be found, the Student T estimate with $N - 2$ degrees of freedom is meaningless for serially sampled data sets.

Traditional analyses can pull significant conclusions from these collections, particularly when using response-wise statistics, but the results are large-scale descriptors of the responses. Numerically, these analyses do not take full advantage of the diversity of responses, or the temporal character of these collection.

The average response time series does not, as a rule, represent the average listener's continuous response. While this result is not surprising to those who often listen to music, this distinction has not be widely acknowledged in the

analyses and interpretations of these collections of responses. Though the central tendency time series does not necessarily represent the majority of responses, it may still be a meaningful summary of the collection. Even if music does not affect individuals deterministically, there appears to be significant consistency in how it effects groups of people. This collective experience of music can be studied explicitly through the dimensional reductions of collections of continuous responses.

5.2 Novel analyses

As explained in chapter 4, the novel techniques discussed here employ and express the inter-response variability and the temporal profiles of continuous responses. Some of the novel techniques did not get the same critical treatment as the often published traditional techniques as this would require a less bias eye than that of their designer. With this caveat in mind, there are some conclusions to be drawn from their application.

1. The novel analyses showed that responses are not generally unified. The rarity of a majority of responses being active within the two second time frames of the activity analysis is corroborated by the rarity of extreme events in the first-order differenced collections. The disagreement in rating responses was also noted in the rarity of extreme events on the untransformed collection, much less than the hypothetical maxima of $10 - 20\%$ of time points, and by the high inconsistency thresholds required to aggregate most collections into only a few clusters.

2. Despite the variablity of responses, many of the collections generally showed significant coordinated activity. The results of the activity coordination tests varied between collection of responses from the same participants, suggesting that differences in significant coordination are due to the coordinating power of stimuli. Cluster analysis also suggest that participants responses can resemble the responses of others, though the interpretation of these clusters and their differences requires further exploration.
3. It is possible and relevant to ask when responses show more agreement or when responses show coordinated extreme responses. Rather than lead the analysis of responses by that of the stimulus, as is the case for interrupted time series and regression analysis, it is possible to have the collection of responses direct the analysis of the stimulus by quantifying *when* and how the listening experience changes significantly.

There is robust information in these response collections which is not caught by traditional analyses. Activity analysis, for example, offers an alternative time series summary of collections for which the contribution of individual responses to the summary is clear at each time point. These techniques may be promising, however, they also need the establishment of reliable data-determined standards of application, standards as have yet to be fixed for the traditional analysis methods as well. Studies comparing the results of analysis techniques on multiple data sets are necessary for the development of sound methodologies.

5.3 Trends in response collections

Without discussing each collection in details, there are interesting trends in the results of the analyses which should be explored in the future. By comparing multiple data collections, differences are noticeable between groups of collections defined by common participants, by responses measures, by collection devices, and by stimulus characteristics. The significance of these factors have yet to be properly evaluated.

Different sets of collection have different degrees variability in response ratings, though within sets, collections also differ in range use and variability. The collections with the smallest cross-sectional dispersion were in those data sets collected using two dimensional GUI emotion rating system, both of perceived and experienced emotion. This difference is visible via clustering, activity analysis, and the traditional longitudinal summaries of collections.

The differences between collections in all of these analyses also show the importance of the response measure and the stimulus. For many stimuli of the collections, emotional arousal ratings were more active and variable than emotional valence ratings. The size of the collections also affect some of the analyses: the small size some collections in AR3 and the opposite distinction of some collection in AoD interfered with the interpretation of other factors across all collections.

Experimental collections vs unrelated collections

The unrelated response collections were quite useful as litmus tests for these analyses, setting bounds on unreasonable results for variability, coordination and other aspects of summary analyses. It would be useful to articulate in more detail

the differences between related and unrelated collections. Some of the experimental collections used here were not always measurably different from the artificially constructed collections, and while activity analysis suggests one explanation of this, there may be specific reasons for their lack of coordination.

Some of the less coordinated data collections, particularly in the Kor data set, did not show strong disagreement between responses. Instead, according to the average longitudinal standard deviation of responses, it seems that the stimuli for these collections did not provoke a wide use of the rating range. It could be that for stimuli which are fairly uniform in expression over their duration, response changes are not determined by the common stimulus. In such cases, the temporal dynamics may be too subtle to catch consistent effects in collections of only 35 participants.

Experienced measures vs perceived measures

In these data sets, there was only one with ratings of perceived emotion, Kor, and this set did contrast in some ways with the collections of experienced emotion ratings. It is easy to assume that ratings of the emotion perceived in the music would agree more than ratings of felt emotions. For most of the sets using experience measures, namely Moz and AoD, they showed less activity and more inter-response variability. However, AR1, also composed of felt emotion ratings, showed coordination and variability similar to that of the Kor data collections.

Analyses of rating values show many collections in the Kor data set with "high" inter-response agreement, a higher proportion than is seen in the the experience emotion data sets AR1, Moz, AoD (first 4 collections), and AR1. This

contrast is much stronger in analyses which depend on the original rating values than in the results of those techniques employing the first-order difference of the collections.

The first 20 seconds

For most collections, inter-response correlations did not change much as a consequence of removing the first 25 seconds of responses. However, for the few that it did, it seems possible that the effect may be worsened by the response collection device. According to extreme points analyses, the lowers extremes are often pulled out of collections like those of the Moz data set when the first few samples are removed. If they are artifacts of the rating task, rather that representative of participants responses, these time points should be excluded from further analysis.

5.4 Future work

The analyses of analyses presented here are beginnings; all of the novel techniques need further development. Conversations between researchers working on continuous responses to music should determine the direction taken for cluster analysis, event detection and activity analysis. Cluster analysis, for example, may be more useful for assess subsections of responses, which could then be interpreted like decision trees [TMCV06].

The distinct clusters of responses found in many collections, along with the many negative pairwise correlations between responses, are reminders that we should not expect music to produce a singular sequence experiences. Along with admitting inter-response variability, individual responses need to be explored

directly. Given the sparse step-like character of many individual rating responses, and the advent of event analyses, we may realize that a different approach to modelling responses is necessary. If participants seem to respond to cues rather than reflect gradual shifts in alignment with stimulus characteristics, it would be useful to quantifying stimuli and responses by looking at when these sharp changes are recorded.

From the beginning of research into continuous responses to music, there has been an interest in capturing *when* participants responsed and *how* their responses differ [Fra56] [Nie87] [Slo91] [CS92]. However, traditional analyses have mostly aggregated response to form a (possibly false) single ideal response for direct comparison to the stimulus. Publications of studies using continuous responses to music have reported results using ill-fitting statistical tools because of a lack of reliable and accessible methodologies suited to the questions at hand. Without sufficient implementation and comparison, many novel and promising approaches have sat waiting to be discovered again. This first comparison of analysis techniques, as applied to many collections of responses, is a starting point for developing accurate tools tailored to the nature of these data and the questions music cognition researchers hope to answer.

Appendix

Table 5–1: This table describes the collections in the Data set AR1, Audience Response system 1, recorded in March 2009. Participants were members of the university community who listened to the stimuli all together in concert like rowed seating. Their responses were recorded on iPod Touch devices with either a one dimensional slider GUI or a two-dimensional cartesian GUI.

1 - Arc1A	Il bianco e dolce cigno Recording (King Singers) J. Arcadelt	Emotional Arousal Experienced 1 of 1 or 2 dim	30 responses 10 Hz 119.7 seconds
2 - Arc1V	Il bianco e dolce cigno Recording (King Singers) J. Arcadelt	Emotional Valence Experienced 1 of 1 or 2 dim	30 responses 10 Hz 119.7 seconds
3 - Sch1A	Mvmt 1, String Qt No. 3, Op. 48 Recording (St. Laurence Quartet) R. Schumann	Emotional Valence Experienced 1 of 1 or 2 dim	30 responses 10 Hz 488.6 seconds
4 - Sch1V	Mvmt 1, String Qt No. 3, Op. 48 Recording (St. Laurence Quartet) R. Schumann	Emotional Valence Experienced 1 of 1 or 2 dim	30 responses 10 Hz 488.6 seconds
5 - Ste1A	Everybody to the Power of One Recording (d. andrew stewart) d. andrew stewart	Emotional Arousal Experienced 1 of 1 or 2 dim	30 responses 10 Hz 390.4 seconds
6 - Ste1V	Everybody to the Power of One Recording (d. andrew stewart) d. andrew stewart	Emotional Valence Experienced 1 of 1 or 2 dim	30 responses 10 Hz 390.4 seconds

Table 5–2: This table describes the collections in the Data set Kor, responses collected by Mark Korhonen for his 2004 thesis. Participants were divers in terms of musical expertise. Each listened to stimuli and responded using the EmotionSpace Lab, a 2D Emotion Space interface via computer screen and mouse. All musical stimuli were taken from the Naxos CD Discover the Classics, Vol. 1, and some stimuli were edited excerpts of the tracks of these compilation CDs.

7 - AllegroA	Allegro - Piano Concerto No. 1 Recording (C. Oliver Dohnanyi) F. Liszt	Emotional Arousal Perceived 1 of 2 dim	35 responses 1 Hz 315 seconds
8 - AllegroV	Allegro - Piano Concerto No. 1 Recording (C. Oliver Dohnanyi) F. Liszt	Emotional Valence Perceived 1 of 2 dim	35 responses 1 Hz 315 seconds
9 - AranjuezA	Adagio - Concierto de Aranjuez Recording (s. Norbert Kraft) J. Rodrigo	Emotional Arousal Perceived 1 of 2 dim	35 responses 1 Hz 165 seconds
10 - AranjuezV	Adagio - Concierto de Aranjuez Recording (s. Norbert Kraft) J. Rodrigo	Emotional Valence Perceived 1 of 2 dim	35 responses 1 Hz 165 seconds
11 - FanfareA	Fanfare for the Common Man Recording (c. S. Gunzenhauser) A. Copland	Emotional Arousal Perceived 1 of 2 dim	35 responses 1 Hz 170 seconds
12 - FanfareV	Fanfare for the Common Man Recording (c. S. Gunzenhauser) A. Copland	Emotional Valence Perceived 1 of 2 dim	35 responses 1 Hz 170 seconds
13 - MoonlightA	Adagio - Moonlight Sonata Recording (s. Jeno Jando) L. van Beethoven	Emotional Arousal Perceived 1 of 2 dim	35 responses 1 Hz 153 seconds
14 - MoonlightV	Adagio - Moonlight Sonata Recording (s. Jeno Jando) L. van Beethoven	Emotional Valence Perceived 1 of 2 dim	35 responses 1 Hz 153 seconds
15 - MorningA	Morning - Peer Gynt Suite No. 1 Recording (c. Jerzy Maksymiuk) E. Grieg	Emotional Arousal Perceived 1 of 2 dim	35 responses 1 Hz 164 seconds
16 - MorningV	Morning - Peer Gynt Suite No. 1 Recording (c. Jerzy Maksymiuk) E. Grieg	Emotional Valence Perceived 1 of 2 dim	35 responses 1 Hz 164 seconds
17 - PizzicatoA	Pizzicato Polka Recording (c. Ondrej Lenard) Johann Strauss II, Josef Strauss	Emotional Arousal Perceived 1 of 2 dim	35 responses 1 Hz 164 seconds
18 - PizzicatoV	Pizzicato Polka Recording (c. Ondrej Lenard) Johann Strauss II, Josef Strauss	Emotional Valence Perceived 1 of 2 dim	35 responses 1 Hz 164 seconds

Table 5–3: This table describes the collections in the data set Moz, a collection of responses to pieces by W. A. Mozart, recorded during two concerts in 2006. Participants were diverse in age and musical expertise and recorded their responses on handheld slider potentiometers wired to their seats in the concert halls. All stimuli were performed by the Boston Symphony Orchestra under the direction of Maestro Lockhart.

19 - K492L	Overture - Marriage of Figaro, K492 Live (BSO) W. A. Mozart	Emotional Intensity Experienced 1 of 1 dim	30 responses 2 Hz 239.5 seconds
20 - K492R	Overture - Marriage of Figaro, K492 Recording (BSO) W. A. Mozart	Emotional Intensity Experienced 1 of 1 dim	23 responses 2 Hz 239.5 seconds
21 - K16L	Rondo - Symphony No. 1, K16 Live (BSO) W. A. Mozart	Emotional Intensity Experienced 1 of 1 dim	30 responses 2 Hz 128 seconds
22 - K16R	Rondo - Symphony No. 1, K16 Recording (BSO) W. A. Mozart	Emotional Intensity Experienced 1 of 1 dim	22 responses 2 Hz 128 seconds
23 - K622L	Adagio - Cl. Concerto in A, K622 Live (BSO) W. A. Mozart	Emotional Intensity Experienced 1 of 1 dim	31 responses 2 Hz 401.5 seconds
24 - K622R	Adagio - Cl. Concerto in A, K622 Recording (BSO) W. A. Mozart	Emotional Intensity Experienced 1 of 1 dim	22 responses 2 Hz 401.5 seconds
25 - K551L	Finale - Symphony No. 41, K551 Live (BSO) W. A. Mozart	Emotional Intensity Experienced 1 of 1 dim	31 responses 2 Hz 350.5 seconds
26 - K551R	Finale - Symphony No. 41, K551 Recording (BSO) W. A. Mozart	Emotional Intensity Experienced 1 of 1 dim	22 responses 2 Hz 350.5 seconds

Table 5–4: This table describes the collections in the data set AoD, of the Angel of Death project, recorded in two concerts in 2002, one in Paris, France and another in La Jolla, California. Participants were diverse in age and musical expertise and recorded their responses on handheld slider potentiometers wired to their seats in the concert halls.

27 - DSljEmF	Angel of Death, D-S version Live (G. Cheng, SONOR Ensemble) R. Reynolds	Emotional Force Experienced 1 of 1 dim	41 responses 2 Hz 2054.5 seconds
28 - SDljEmF	Angel of Death, S-D version Live (G. Cheng, SONOR Ensemble) R. Reynolds	Emotional Force Experienced 1 of 1 dim	51 responses 2 Hz 2033.5 seconds
29 - DSpaEmF	Angel of Death, D-S version Live (JM Cottet,Court Circuit) R. Reynolds	Force emotionnelle Experienced 1 of 1 dim	54 responses 2 Hz 2067.5 seconds
30 - SDpaEmF	Angel of Death, S-D version Live (JM Cottet,Court Circuit) R. Reynolds	Force emotionnelle Experienced 1 of 1 dim	41 responses 2 Hz 2067.5 seconds
31 - DSljFam	Angel of Death, D-S version Live (G. Cheng, SONOR Ensemble) R. Reynolds	Familiarity Experienced 1 of 1 dim	34 responses 2 Hz 2054.5 seconds
32 - SDljFam	Angel of Death, S-D version Live (G. Cheng, SONOR Ensemble) R. Reynolds	Familiarity Experienced 1 of 1 dim	43 responses 2 Hz 2033.5 seconds
33 - DSpaFam	Angel of Death, D-S version Live (JM Cottet,Court Circuit) R. Reynolds	Ressemblance Experienced 1 of 1 dim	36 responses 2 Hz 2067.5 seconds
34 - SDpaFam	Angel of Death, S-D version Live (JM Cottet,Court Circuit) R. Reynolds	Ressemblance Experienced 1 of 1 dim	40 responses 2 Hz 2067.5 seconds

Table 5–5: The collections in the data set AR3, Audience Response system 3, recorded in March 2009. Participants were members of the public and music theorists in town for a conference who attended this concert. Their responses were recorded on iPod Touch devices with a two dimensional cartesian GUI.

35 - Arc2A	Il bianco e dolce cigno Live (Orpheus Singers) J. Arcadelt	Emotional Arousal Experienced 1 of 1 or 2 dim	17 responses 10 Hz 128.4 seconds
36 - Arc2V	Il bianco e dolce cigno Live (Orpheus Singers) J. Arcadelt	Emotional Valence Experienced 1 of 1 or 2 dim	17 responses 10 Hz 128.4 seconds
37 - Sch2A	Andante-Allegro, String Qt No. 3, Op. 48 Live (Student Quartet) R. Schumann	Emotional Valence Experienced 1 of 1 or 2 dim	8 responses 10 Hz 130 seconds
38 - Sch2V	Andante-Allegro, String Qt No. 3, Op. 48 Live (Student Quartet) R. Schumann	Emotional Valence Experienced 1 of 1 or 2 dim	8 responses 10 Hz 130 seconds
39 - Ste2A	Everybody to the Power of One Live improvisation (d. andrew stewart) d. andrew stewart	Emotional Arousal Experienced 1 of 1 or 2 dim	30 responses 10 Hz 446.1 seconds
40 - Ste2V	Everybody to the Power of One Live improvisation (d. andrew stewart) d. andrew stewart	Emotional Valence Experienced 1 of 1 or 2 dim	30 responses 10 Hz 446.1 seconds

Table 5–6: The artificial "unrelated" collections constructed from responses selected from each of the 40 previous experimental collections.

41 - RdmI1	First 2 minutes of random responses Mixed Mixed	Mixed Mixed 1 of 1 or 2 dimensions	40 responses 1 Hz 119 seconds
42 - RdmI2	First 2 minutes of random responses Mixed Mixed	Mixed Mixed 1 of 1 or 2 dimensions	40 responses 1 Hz 119 seconds
43 - RdmF1	Last 2 minutes of random responses Mixed Mixed	Mixed Mixed 1 of 1 or 2 dimensions	40 responses 1 Hz 119 seconds
44 - RdmF2	Last 2 minutes of random responses Mixed Mixed	Mixed Mixed 1 of 1 or 2 dimensions	40 responses 1 Hz 119 seconds

Glossary of Math

This is a glossary of math symbols to clarify the particularities of the notation used here.

$:=$ such as $X := \{\mathbf{x_r}\}$, this modification of the equal sign, $=$, specifies that the equation gives the definition of the term X and thus taken to be a fact rather that a relationship to be proven.

$\{\}$ also found in $X := \{\mathbf{x_r}\}$, curly brackets mark a set of elements, a collection. In this case, X refers to the set rather than any element in the set. The elements in a countable set can be indexed (though they need not be), and a quick notation for this is with super- and sub-scripts such as $\mathbf{x_r} := \{x_{r,i}\}_{i=1}^{N}$. In this case, $\mathbf{x_r}$ is a set of N elements and the index of $x_{r,i} \in \mathbf{x_r}$ is significant. Note that the collections of time series are then sets of sets of data.

$\in$ as used above, $\in$ means the symbol to the left is an "element of the set" symbolized to the right.

$\lfloor y \rfloor$ as used in $\lfloor N/k \rfloor$. These partial square brackets indicate that the resulting numbers is the greatest integer less than the value in the brackets.

$\|A\|$ the size of a set, i.e., the number of elements in the set.

$\forall$ for all elements in the set.

References

[BBL+09] J.P. Bachorik, M. Bangert, P. Loui, K. Larke, J. Berger, R. Rowe, and G. Schlaug. Emotion in motion: Investigating the time-course of emotional judgments of musical stimuli. *Music Perception*, 26(4):355–364, 2009.

[BS95] R.V. Brittin and Deborah A. Sheldon. Comparing continuous versus static measurements in music listeners' preferences. *Journal of Research in Music Education*, 43(1):36–46, 1995.

[CC09] E. Coutinho and A. Cangelosi. The use of spatio-temporal connectionist models in psychological studies of musical emotions. *Music Perception*, 27(1):1–15, 2009.

[CJSK+10] H. Chapin, K. Jantzen, JA Scott Kelso, F. Steinberg, E. Large, and A. Rodriguez-Fornells. Dynamic emotional and neural responses to music depend on performance expression and listener experience. *PloS one*, 5(12):169–200, 2010.

[CL54] H. Chernoff and EL Lehmann. The use of maximum likelihood estimates in χ 2 tests for goodness of fit. *The Annals of Mathematical Statistics*, 25(3):579–586, 1954.

[CS92] Deborah A. Capperella-Sheldon. Self-perception of aesthetic experience among musicians and non-musicians in response to wind band music. *Journal of Band Research*, 28:57–71, 1992.

[DC01] R.A. Duke and E.J. Colprit. Summarizing listener perceptions over time. *Journal of Research in Music Education*, 49(4):330, 2001.

[DM88] E.J. Dudewicz and S. Mishra. *Modern mathematical statistics*. John Wiley & Sons, Inc. New York, NY, USA, 1988.

[DMR06] S. Dubnov, S. McAdams, and R. Reynolds. Structural and affective aspects of music from statistical audio signal analysis. *Journal*

of the American Society for Information Science and Technology, 57(11):1526–1536, 2006.

[Fra56] Robert Francès. Recherches électro-polygraphiques sur la perception de la musique. *L'année psychologique*, 56(2):373–396, 1956.

[Fre95] W.E. Fredrickson. A comparison of perceived musical tension and aesthetic response. *Psychology of Music*, 23(1):81–87, 1995.

[Fre99a] W.E. Fredrickson. Effect of Musical Performance on Perception of Tension in Gustav Hoist's First Suite in E-flat. *Journal of Research in Music Education*, 47(1):44, 1999.

[Fre99b] D. Frego. Effects of aural and visual conditions on response to perceived artistic tension in music and dance. *Journal of Research in Music Education*, 47(1):31, 1999.

[FS09] M.M. Farbood and B. Schoner. Determining Feature Relevance in Subject Responses to Musical Stimuli. In *Mathematics and Computation in Music*, volume 38 of *Communications in computer and information science*, pages 115–129. Springer, 2009.

[GMG04] John M. Geringer, Clifford K. Madsen, and Dianne Gregory. A fifteen-year history of the continuous response digital interface: Issues relating to validity and reliability. *Bulletin of the Council for Research in Music Education*, (160):1–15, 2004.

[GNKA07a] O. Grewe, F. Nagel, R. Kopiez, and E. Altenmüller. Emotions over time: Synchronicity and development of subjective, physiological, and facial affective reactions to music. *Emotion*, 7(4):774–788, 2007.

[GNKA07b] O. Grewe, F. Nagel, R. Kopiez, and E. Altenmüller. Listening to music as a re-creative process: Physiological, psychological, and psychoacoustical correlates of chills and strong emotions. *Music Perception*, 24(3):297–314, 2007.

[Gre94] D. Gregory. Analysis of listening preferences of high school and college musicians. *Journal of Research in Music Education*, 42(4):331, 1994.

[Gre95] Dianne Gregory. The continuous response digital interface: an analysis of reliability measures. *Psychomusicology*, 14:197–208, 1995.

[IM99] M. Iwanaga and Y. Moroki. Subjective and physiological responses to music stimuli controlled over activity and preference. *Journal of Music Therapy*, 36:26–38, 1999.

[Iye11] S. Iyengar. *The Sourcebook for Political Communication Research*, chapter Experimental Designs for Political Communication Research, pages 129–148. Routledge, 2011.

[KCJ05] M.D. Korhonen, D.A. Clausi, and M.E. Jernigan. Modeling emotional content of music using system identification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(3):588–599, 2005.

[KK03] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

[Kor04] M. D. Korhonen. Modeling continuous emotional appraisals of music using system identification. Masters of applied science in systems design engineering, University of Waterloo, Waterloo, Ontario, Canada, 2004.

[Kru96] Carol L. Krumhansl. A perceptual analysis of mozart's piano sonata k. 282: Segmentation, tension, and musical ideas. *Music Perception*, 13(3):401–432, 1996.

[Kru97] C.L. Krumhansl. An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology*, 51(4):336–353, 1997.

[L+05] W. Liao et al. Clustering of time series data–a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[LC10] D. Lottridge and M. Chignell. Emotional Majority Agreement a psychometric property of affective self-report instruments. In *Science and Technology for Humanity (TIC-STH), 2009 IEEE Toronto International Conference*, pages 795–800. IEEE, 2010.

[LNVR07] D.J. Levitin, R.L. Nuzzo, B.W. Vines, and JO Ramsay. Introduction to functional data analysis. *Canadian Psychology*, 48(3):135, 2007.

[LTE+08] G. Luck, P. Troiviainen, J. Erkkilä, O. Lartillot, K. Riikkilä, A. Mäkelä, K. Pyhäluoto, H. Raine, L. Varkila, and J. Värri. Modelling the relationships between emotional responses to, and musical content of, music therapy improvisations. *Psychology of Music*, 36(1):25–45, 2008.

[Lyc98] J.A. Lychner. An empirical study concerning terminology relating to aesthetic response to music. *Journal of Research in Music Education*, 46(2):303, 1998.

[Mad97] C.K. Madsen. Focus of attention and aesthetic response. *Journal of Research in Music Education*, 45(1):80–89, 1997.

[Mar07] E.H. Margulis. Silences in music are musical not silent: An exploratory study of context effects on the experience of musical pauses. *Music Perception*, 24(5):485–506, 2007.

[MF93] Clifford K. Madsen and Willian E. Fredrickson. The experience of musical tension: A replication of nielsen's research using the continuous response digital interface. *Journal of Music Therapy*, 30(1):45–57, 1993.

[MG90] C.K. Madsen and J.M. Geringer. Differential patterns of music listening: Focus of attention of musicians versus nonmusicians. *Bulletin of the Council for Research in Music Education*, 1:45–57, 1990.

[MGF97] Clifford K. Madsen, John M. Geringer, and Willian E. Fredricksonrickson. Focus of attention to musical elements in Haydn's Symphony No. 104. *Bulletin of the Council for Research in Music Education*, (133):57–63, 1997.

[MR09] I.B. Mauss and M.D. Robinson. Measures of emotion: A review. *Cognition & emotion*, 23(2):209, 2009.

[MVV+04] S. McAdams, B.W. Vines, S. Vieillard, B.K. Smith, and R. Reynolds. Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. *Music Perception*, 22(2):297–350, 2004.

[Nie87] Frede V. Nielsen. Musical 'tension' and related concepts. In TA Sebeok and J. Umiker-Seboek, editors, *The semiotic web '86: An international yearbook*, pages 491–513. Mouton de Gruyter, Berlin, 1987.

[PK87] C. Palmer and C.L. Krumhansl. Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Attention, Perception, & Psychophysics*, 41(6):505–518, 1987.

[Ric04] N.S. Rickard. Intense emotional responses to music: A test of the physiological arousal hypothesis. *Psychology of Music*, 32(4):371, 2004.

[RK05] C.A. Ratanamahatana and E. Keogh. Three myths about dynamic time warping data mining. In *Proceedings of SIAM International Conference on Data Mining (SDM'05)*. Citeseer, 2005.

[RN88] J.L. Rodgers and W.A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[RT04] J.R. Rossiter and J. Thornton. Fear-pattern analysis supports the fear-drive model for antispeeding road-safety TV ads. *Psychology and Marketing*, 21(11):945–960, 2004.

[Say89] L.W. Sayrs. *Pooled time series analysis*, volume 70. Sage Publications, Inc, 1989.

[SBL+09] V.N. Salimpoor, M. Benovoy, G. Longo, J.R. Cooperstock, and R.J. Zatorre. The rewarding aspects of music listening are related to degree of emotional arousal. *PloS one*, 4(10):29–49, 2009.

[Sch99a] Emery Schubert. *Measurement and time series analysis of emotion in music*. PhD thesis, University of New South Wales, 1999.

[Sch99b] Emery Schubert. Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, 51(3):154–165, 1999.

[Sch02] Emery Schubert. Correlation analysis of continuous emotional response: Correcting for the effects of serial correlation. *Musicae Scientiae*, Special Issue 2001-2002:213–236, 2002.

[Sch04] E. Schubert. Modeling perceived emotion with continuous musical features. *Music Perception*, 21(4):561–585, 2004.

[Sch07] E. Schubert. When is an event in a time-series significant? *Proceedings of ICoMCS December*, page 135, 2007.

[Sch10] E. Schubert. Continuous self-report methods. In P. Juslin and J. A. Sloboda, editors, *Handbook of music and emotion: theory, research, applications*, Series in Affective Science, pages 223–253. Oxford University Press, 2010.

[SD04] E. Schubert and W.T.M. Dunsmuir. Introduction to Interrupted Time Series Analysis of Emotion in Music: The case of arousal, valence and points of rest. In *8th International Conference of Music Perception and Cognition. August*, pages 3–7, 2004.

[SKS06] N. Steinbeis, S. Koelsch, and J.A. Sloboda. The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. *Journal of Cognitive Neuroscience*, 18(8):1380–1393, 2006.

[Slo91] John A. Sloboda. Music structure and emotional response. *Psychology of Music*, 19:110–120, 1991.

[SSM+09] C.J. Stevens, E. Schubert, R.H. Morris, M. Frear, J. Chen, S. Healey, C. Schoknecht, and S. Hansen. Cognition and the temporal arts: Investigating audience response to dance using PDAs that record continuous data during live performance. *International Journal of Human-Computer Studies*, 67(9):800–813, 2009.

[TMCV06] R. Timmers, M. Marolt, A. Camurri, and G. Volpe. Listeners' emotional engagement with performances of a Scriabin étude: an explorative case study. *Psychology of Music*, 34:281–510, 2006.

[TSK+06] P.N. Tan, M. Steinbach, V. Kumar, et al. *Introduction to data mining*, chapter 8, Cluster Analysis: Basic Concepts and Algorithms, pages 487–568. Pearson Addison Wesley, 2006.

[TWV07] S. Thompson, A. Williamon, and E. Valentine. Time-Dependent Characteristics of Performance Evaluation. *Music Perception*, 25(1):13–29, 2007.

[vD00] Stijn van Dongen. *Graph Clustering by Flow Simulation*. Computer science, Universiteit Utrecht, May 2000.

[VKWL06] B.W. Vines, C.L. Krumhansl, M.M. Wanderley, and D.J. Levitin. Cross-modal interactions in the perception of musical performance. *Cognition*, 101(1):80–113, 2006.

[VVS93] D. Vaitl, W. Vehrs, and S. Sternagel. Prompts-leitmotif-emotion: Play it again, richard wagner. In N. Birbaumer and A. Ohman, editors, *The Structure of Emotion: Psychophysiological, Cognitive, and Clinical Aspects*, pages 169–189. Hogrefe & Huber, Toronto, 1993.

[Wil45] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[WM08] Y.J. Wang and M.S. Minor. Validity, reliability, and applicability of psychophysiological techniques in marketing research. *Psychology and Marketing*, 25(2):197–232, 2008.