Modeling Continuous Emotional Appraisals of Music Using System Identification

by

Mark David Korhonen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2004

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The goal of this project is to apply system identification techniques to model people's perception of emotion in music as a function of time. Emotional appraisals of six selections of classical music are measured from volunteers who continuously quantify emotion using the dimensions valence and arousal. Also, features that communicate emotion are extracted from the music as a function of time. By treating the features as inputs to a system and the emotional appraisals as outputs of that system, linear models of the emotional appraisals are created. The models are validated by predicting a listener's emotional appraisals of a musical selection (song) unfamiliar to the system. The results of this project show that system identification provides a means to improve previous models for individual songs by allowing them to generalize emotional appraisals for a genre of music. The average R^2 statistic of the best model structure in this project is 7.7% for valence and 75.1% for arousal, which is comparable to the R^2 statistics for models of individual songs.

Acknowledgements

The first people I would like to thank are my supervisors Prof. David A. Clausi and Prof. M. Ed Jernigan. They allowed me the freedom to work independently while providing advice as I needed it. If it wasn't for their confidence in me, I'm certain my research would not have progressed as well as it did.

I would like to thank my thesis readers Prof. P. Fieguth and Prof. S. Lipshitz for reviewing my thesis and providing suggestions to improve my work.

The members of the SIP (Signal and Image Processing) group deserve a special thank you as well as other students, staff and faculty in Systems Design Engineering at the University of Waterloo. The advice and support from these people have greatly helped my research progress.

I would like to thank all of the people who have volunteered for this study. I have greatly appreciated the interest the Laurier Centre for Music Therapy Research (LCMTR) at Wilfrid Laurier University has had in my research, as well as their help recruiting volunteers. Also, the Office of Research Ethics at the University of Waterloo deserves credit for their assistance in organizing my study.

I'd like to thank Emery Schubert for allowing me to use his 2DES software and for providing an excellent starting point for this research. Also, Maneesh Gupta deserves thanks for working with me on a project similar to this one several years ago.

Thanks are extended to Naxos for allowing the use of CD 8.550035-36 in this research. Also, I'd like to thank NSERC (Natural Sciences and Engineering Research Council of Canada) and the OGS (Ontario Graduate Scholarship) program for funding this project.

I'd especially like to thank my wife Nancy for her support while I spent countless hours researching and writing. She made this process a lot more pleasant.

Finally, I'd like to thank Max for always being willing to go for a walk when I needed a break.

This thesis is dedicated to my wife Nancy.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General

Music has the ability to communicate emotion and thus emotion can be perceived in music[24]. However, establishing exactly how music communicates emotion is a topic of much debate. This thesis aims to investigate how music communicates emotion by creating models of emotional appraisals of musical stimuli as described by listeners.

This investigation is accomplished by creating models using system identification techniques. The process of creating models is based on the analysis of emotional appraisals to a variety of musical stimuli. Once the models are created, it is possible to gain insight into the importance of features of the music by examining the model.

These models can also be used to predict a listener's emotional appraisals of several different songs of the same genre. Using a model to predict appraisals for multiple songs can generalize how musical features influence a listener's perception of emotion. Each model can be validated by comparing the model's predictions of emotional appraisals to true emotional appraisals.

System identification is used because it overcomes many of the difficulties associated with previous efforts to analyze continuous emotional appraisals[24]. By overcoming these difficulties, further analysis of continuous emotional appraisals should be more appealing.

1.2 Organization of Thesis

This thesis is organized to clearly evaluate the use of system identification for modeling continuous emotional appraisals of music. Any discussion that supplements the main body of the thesis can be found in the appendix.

Chapter 2 provides the background necessary for the thesis. Section 2.1 defines emotion and introduces methods of measuring emotional appraisals of music. Section 2.2 describes musical features and how they are measured.

Chapter 3 discusses general approaches that can be taken to model emotional appraisals of music. Section 3.1 is a literature review discussing how emotional appraisals of music have been modeled previously. Section 3.2 introduces system identification and how it can model emotional appraisals of music.

Chapter 4 discusses the specific approach taken to model emotional appraisals of music in this thesis. Section 4.1 summarizes the issues raised in the previous two chapters and lists objectives for overcoming these issues. The remainder of Chapter 4 discusses the methodology and evaluation techniques used to apply system identification to model emotional appraisals of music from musical features.

Chapter 5 provides the results of applying the methodology described in Chapter 4. Finally, Chapter 6 summarizes the results of Chapter 5, provides conclusions and recommendations for future development.

Chapter 2

Background

2.1 Emotion and Music

2.1.1 Emotional Appraisals and Responses

There is no unanimously agreed upon definition of emotion[7]. However, if we consider Bower's network theory of emotion, a working definition of emotion can be obtained[1]. According to Bower's network theory, every emotion is represented by a "node" in the brain that is associated with autonomic reactions and expressive behaviours for that emotion. When stimuli activate an emotion node above a threshold, the emotion node produces patterns of autonomic arousal and expressive behaviour commonly assigned to that emotion. From this theory, emotion can be defined either as *the nodes* or *a cluster* of behaviours and reactions connected to that emotion node[24].

When presented with emotional stimuli, a person may experience the autonomic reactions and expressive behaviours associated with an emotion. In this thesis, the term "emotional response" is used to indicate the person's experience of emotion. However, a person may simply recognize the emotion in stimuli without experiencing the reactions or behaviours associated with the emotion[7]. To recognize the emotion in stimuli, the stimuli are appraised and associated with particular emotion nodes. In this thesis, the process of recognizing emotions in stimuli is referred to as "perceiving emotion" and the term "emotional appraisal" is used to indicate the emotion perceived to be in the stimuli.

There are three advantages to study emotional appraisals over emotional responses. Firstly, emotional responses to the same stimuli can vary depending on many factors external to the stimuli. For example, if a person is in a positive mood, the emotional response they have to a stimulus could be quite different than the emotional response they would have to the same stimulus if they were in a very bad mood[7][26]. Also, a person can associate stimuli with a memory that causes an emotional response different from the emotion they appraise from the stimuli (e.g. if music appraised to be happy is associated with an unhappy event, the emotional response could be unhappiness even though the emotional appraisal is happy)[7][13]. With respect to music, emotional appraisals are more consistent than emotional responses[24].

The second advantage of studying emotional appraisals is based on our investigation of how emotion is communicated. If we consider the stimulus to be a medium for communicating emotions, it can be argued that emotional appraisals are more intuitive to investigate than emotional responses. For example, if facial expressions are considered to be a medium for communication, a primary goal of crying is to communicate that a person is unhappy, not necessarily to induce the experience of unhappiness in other people[9]. Investigating another person's emotional appraisal would involve determining whether they recognize unhappiness from the facial expression. However, investigating that person's emotional response would involve determining whether they are sad to see the unhappy facial expression. Because a goal of an unhappy person is to communicate their own emotion, it can be argued that investigating another person's emotional appraisal of the stimulus is more informative about the communication medium than investigating that person's emotional response.

The third advantage of studying emotional appraisals is logistical. There are several reliable methods to measure emotional appraisals, but measuring emotional responses can be prone to biases or involve many different, simultaneous measurements (e.g. physiological reactions, subjective feeling, motor expressive behaviour)[7].

The process of appraising stimuli is time-varying. Assuming a person is capable of sensing their surroundings and that the world is changing around them, stimuli perceived by a person will change with time. For this reason, stimuli are considered to be time-varying. As stimuli change, emotional appraisals of the stimuli are capable of changing as well. Therefore, we must consider emotional appraisals to be time-varying.

2.1.2 Measuring Emotional Appraisals

Measuring emotional appraisals of stimuli is accomplished by having the person report the emotions they perceive in the stimuli. This can be done in several different ways such as verbal descriptions, choosing emotional terms from a list, or rating how well several different emotional terms describe the appraisal[7][24]. These will be briefly described in the following paragraphs but are more comprehensively reviewed by Schubert[24].

Verbal descriptions of a person's emotional appraisal provide the most freedom in describing the emotions perceived in the stimuli. However, people describe their emotions using different words and different levels of detail[24]. Therefore, the verbal descriptions from different people can be difficult to compare to each other.

Emotional appraisals can also be described by having a person choose emotional terms from a checklist to describe the emotions perceived in the stimuli. The perceived emotions can be analyzed by determining the terms common to particular stimuli. However, the type of statistical analysis that can be applied to the checklist is usually limited[24].

The third method is an extension of the checklist approach. Instead of a checklist of emotional terms, the person is asked to rate the relevance of emotional terms on a numerical scale[24]. The emotional terms used need to be carefully selected to avoid ambiguity. Also, there should be a limited number of terms to make it feasible to measure how the appraisal of the stimuli changes with time. If these conditions are met, this technique of measuring emotional appraisals appears to be the most promising method to use in this thesis.

By rating emotional terms, emotions can be described using a vector. The dimension of the vector is the number of emotional terms and each component of the vector corresponds to the rating assigned to the corresponding emotional term. In other words, the emotional terms can be considered components or dimensions of emotion. Fischer et al. illustrate that splitting emotion into dimensions is consistent with Bower's network theory of emotion[6].

The next step is to decide what components (emotional terms) to use so that many emotions can be described using few dimensions. Dividing emotion into the components that form a basis to describe as many emotions as possible is appealing. However, the number of components and the type of components vary between studies (e.g. [5][23][24][30]). Results from multivariate analysis studies have "...suggested that many, perhaps most, emotions recognised in music may be represented in a two-dimensional space with valence (positive vs. negative feelings) and arousal (high–low) as principal axes ..."[7]. These are the dimensions suggested by Russell to describe emotion[21].

Figure 2.1 is an adapted version of Russell's figure showing how several different emo-

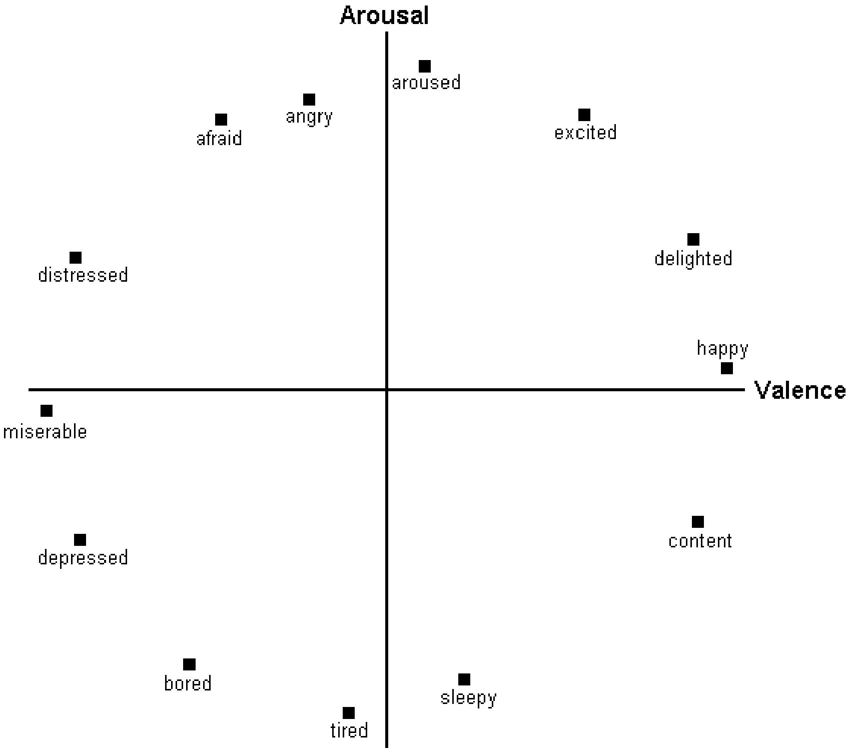


Figure 2.1: Possible descriptions of emotion using valence and arousal[22].

tions can be described using the dimensions valence and arousal[22]. Valence refers to the happiness or sadness of the emotion and arousal is the activeness or passiveness of the emotion[25]. A positive valence corresponds with positive emotions such as joy, happiness, relaxing and a negative valence corresponds with negative emotions such as fear, anger and sadness. In the other dimension, emotions such as anger, excitement and interest are more arousing than emotions such as sadness, relaxed or bored. Each component can be quantified by limiting the range of each dimension to $[-100\%, 100\%]$ and rating each component on this scale[25].

By describing emotion using the two dimensions of valence and arousal, a person can describe his/her emotional appraisal on a computer by using a joystick, mouse or similar input device. The person would use the input device to move a cursor around in the two-dimensional emotion space (2DES) and the cursor position would correspond to the emotional appraisal. By recording how the cursor position changes with time, the person can easily describe how his/her emotional appraisals change with time as the stimulus changes. *FEELTRACE*[3] and *EmotionSpace Lab*[25] are examples of software that are able to collect reliable time-varying emotional appraisals using a 2DES to emotionally appraise stimuli (e.g. words, faces, music and video). People appear to find this approach more intuitive than checklists or standard rating scales but the data are more difficult to analyze[13].

2.1.3 Perceiving Emotion in Music

Music cognition researchers often investigate how music evokes an experience and perception of emotion. Work has been done on analysing the role of particular features in music (e.g. [8][28]), modeling how musicians express emotion while they perform (e.g. [12]) and how listeners of music perceive emotion (e.g. [24]). A common goal shared by these studies is to determine the features in music that communicate emotion. In this thesis, the goal is to model how emotions are perceived from musical features so the focus will be on how listeners of music perceive emotion.

When people perceive emotion in music, there are some emotions that are reliably perceived and other emotions that are confused with different emotions[7]. The emotions that are reliably perceived (i.e. happiness, sadness), each appears to have a distinctive arousal and/or valence. Generally, the emotions that are confused (i.e. calm vs. sorrow, anger vs. fear) appear to have similar arousals and valences. This may mean that while emotion may consist of other components than arousal and valence, these two components may be the ones that are most clearly communicated through music. These reasons provide additional motivation for using the 2DES to emotionally appraise music.

2.2 Musical Features

To model a person's emotional appraisal of music, the music needs to be represented in a form suitable for modeling. The music is assumed to have certain *properties* or attributes that allow emotional appraisals to be distinguished. Measurements are then taken of the music that either directly or indirectly represent the properties[4]. *Musical features* are functions of the measurements that facilitate the modeling process and are used in the models. For example, measurements useful for modeling can be treated as features or new features can be calculated as functions of other features. Ideally, the musical features quantify and represent all of the properties about the music needed to model emotional appraisals[1].

Musical features can be considered either global or local. Global features are measured over an entire selection of music (e.g. dynamic range, genre, etc.). Global features can only be obtained by analysing the whole selection of music and are not time-varying. Local features are measured over small sections of time and re-measured many times over the selection of music (e.g. loudness, pitch, etc.). Since emotion is treated as a time-varying quantity in this thesis, only local features will be considered.

Schubert has performed a comprehensive review of studies that determine which musical properties cause listeners to perceive emotion[24]. The properties identified by Schubert are dynamics, mean pitch, pitch range, variation in pitch, melodic contour, register, mode, timbre, harmony, texture, tempo, articulation, note onset, vibrato, rhythm and metre. There are two different approaches to measure these properties for calculating musical features.

One approach is to measure, sometimes approximately, these properties using software

[1] If features are correlated, the feature vectors can be reduced in dimension using techniques such as *principal component analysis* or *independent component analysis*. Reducing the dimension of the feature vectors reduces the number of parameters in the model. For more information about these techniques, consult Hyvärinen et al.[11].

algorithms. Software such as MARSYAS and PsySound can measure some of these properties, as well as providing measurements based on the human auditory system that may indirectly measure other properties[2][29]. The advantage of this approach is that it is easy to measure many different features in a short period of time using standard techniques. The primary disadvantage to this approach is that the algorithms used to calculate the features vary in robustness, or may only work under certain assumptions, so care needs to be taken when using these features.

Another approach to measuring these properties is to have an expert analyze a transcribed version of music as they listen to the music. This approach can be used to estimate features such as tempo, dynamics and metre. The primary advantage of this approach is that it can be used to measure properties difficult to measure using software algorithms (e.g. tempo). The disadvantages include variations and bias due to subjectivity (e.g. beat detection can vary by a few milliseconds resulting in an artificially varying tempo) and it can be time-consuming.

Finally, some properties are difficult to quantify in a meaningful way (e.g. rhythm), or are unknown. Until these properties can be quantified, they cannot directly be included in a mathematical model. If these properties are necessary to model emotional appraisals, and if other features cannot indirectly represent these properties, the model will not perform as well as desired.

Chapter 3

Models of Emotional Appraisals

3.1 Review of Current Models

3.1.1 Introduction

Most research on emotional appraisals of music focus on determining the musical properties that communicate emotion, or focus on verifying these musical properties by composing or performing music to convey particular emotions. These studies must be considered when selecting the musical features used in the models, but they do not provide much insight into possible modeling techniques. Only a small amount of research into modeling emotional appraisals has been done. The following sections describe models used to estimate a listener's emotional appraisals to musical stimuli.

3.1.2 Mapping Musical Instruments

Suzuki and Hashimoto modeled emotional appraisals of tones played on different instruments[27]. Emotional appraisals were measured relatively by comparing the similarity between pairs of instruments. The goal was to map the *timbre* of an instrument to a two (or three) dimensional emotion space using these similarity measures.

Sounds for twenty-two different instruments were used in the experiments. A 1.5 second recording of each instrument playing one note was recorded digitally at 44100Hz. A spectrograph measuring the power spectrum of the audio data as a function of time was represented using a 128 dimensional vector. Principal Component Analysis (PCA)[11] was then applied to reduce the dimensionality of the vector to p.

Each subject was asked to listen to all possible pairs of ten of the instruments. While listening to the pairs of instruments, the subject was asked to evaluate the similarities of their emotional appraisals to the two instruments presented. The similarity was measured using a seven grade score from similar–1 to not similar–7.

A nonlinear mapping was then estimated to place the instruments in a space so that the similarity measures between pairs of instruments correspond to the euclidean distance between the instruments in that space. The nonlinear mapping was constructed using a three-layer perceptron neural network. The input to the nonlinear mapping is the p dimensional vector representing the audio data of the instrument and the output is the estimated location in the emotion space for that instrument. The nonlinear mapping is trained by setting the similarity between two instruments equal to the desired euclidean distance between the two instruments in the emotion space (and backpropagating the error).

Once the nonlinear mapping was determined, the remaining twelve instruments were mapped to the emotion space. The mapped instruments were in a location very similar to where they would be by calculating the mapping using a multidimensional scaling method.

This modeling framework is able to generalize emotional appraisals to any instrument. Unfortunately, the axes of the emotion space have no direct interpretation. Also, these models have only been tested on single notes from instruments so it is unclear how to map a sound of multiple notes from multiple instruments. The authors mention that pitch and loudness affect emotional appraisals but it is unclear if the location of the mapped instruments would change if they were all played at a different pitch and/or loudness. However, this model provides a promising method of representing the musical property *timbre* as a two (or three) dimensional vector of musical features.

3.1.3 Classifying Musical Selections

Li and Ogihara modeled emotional appraisals of 30 second selections of songs from four different genres of music[14]. Emotional appraisals were classified into 13 different emotion groups and 6 different emotion “supergroups”. The goal was to identify the emotion groups that could be associated with each selection of music.

Random thirty second selections of 499 different songs were used in this study. A 39 year old male listened to each selection of music and identified the emotion groups and supergroups that he felt should be associated with the selection of music.

Thirty features were extracted from the musical selections using MARSYAS and stored in a vector. The features selected were used to represent the musical properties of timbral texture, rhythmic content and pitch content.

Fifty percent of the musical selections were used to train a set of binary classifiers. There was one classifier for each emotion group/supergroup and each classifier identified whether or not the music should be associated with that emotion group. Support vector machines (SVMs) were used as the classifiers and the 30-dimensional feature vector for the music was the input to the classifier.

The remaining data were used to evaluate the performance of the set of binary classifiers. An information retrieval performance measure, the breakeven point, was calculated and found to be approximately 45% with 13 emotion groups and increased to approximately 50% with 6 emotion supergroups. The authors state that this performance is poor and suggest that performance could be improved by including genre/style information, using more data, improve the method of labeling the data, or using different features.

Although the performance is poor, this modeling framework is able to generalize emotional appraisals of musical selections from a variety of genres of music. Because only one person identified the emotion in the musical selections, this study cannot be used to generalize emotional appraisals of music for a population of listeners. This study assumes that the emotional appraisals are fairly constant over the 30 second selection of music, which may not always be true. Also, this study suggests that by properly labeling emotional appraisals, limiting the music selection to one genre of music and using many different features, a model with improved performance can be created.

3.1.4 Time Series Analysis

Schubert modeled emotional appraisals of four selections of classical music[24]. Emotional appraisals were measured using the dimensions valence and arousal in the 2DES as a function of time. The goal was to model the emotional appraisals of each song as a time series using musical features as input variables.

Four different selections of classical music were appraised by 67 different people using Schubert's *EmotionSpace Lab* software. The cursor position in the 2DES, corresponding to the listener's emotional appraisal of the music at a particular moment in time, was recorded every second. A mean emotional appraisal as a function of time was then calculated by averaging across participants at each second of music.

Local musical features were extracted from the four songs every second to represent the properties of loudness, pitch, tempo and texture. Loudness was represented using Densil Cabrera's algorithm to measure A-weighting decibels (dBA)[2]. Pitch was represented using the features power spectrum centroid and MIDI note number of the melody. Tempo was represented by having a musical expert estimate the instantaneous beats per minute (BPM) by examining the audio file and the score. Texture was represented using the number of instruments playing concurrently[1].

For each song, a first-order differenced OLS linear regression model (or, equivalently, a first-order FIR model[15]) with first-order autoregressive noise was fit to the arousal component of the emotional appraisal. Another model of the same architecture was fit to the valence component of the emotional appraisals. The lags used for each feature are those determined to be statistically significant from zero in the residual cross-correlation function.

Schubert demonstrated that combinations of musical features could explain 30-70% of first-differenced emotional appraisals using these models. Schubert was also able to infer some causal relationships between particular musical features and emotional appraisals.

This appears to be the first attempt to analyze time-varying emotional appraisals.

[1] For a list of rules Schubert used to determine the number of instruments playing from the score, see p. 261 of Schubert's thesis[24].

More sophisticated linear and nonlinear models can easily extend these models. With this approach, a model needs to be created for each song so one model cannot generalize to other songs.

3.2 Motivation for System Identification

The goal of this thesis is to model time-varying emotional appraisals. Ideally, the models should be able to generalize what the emotional appraisals should be for any song within a genre of music. The models reviewed are either able to generalize emotional appraisals or model time-varying emotional appraisals, but not both. System identification is a signal processing technique that can be used to achieve both of these goals[15].

To understand system identification, the terms *signal* and *system* need to be defined. A signal is a function of time (and/or other independent variables) that contains information about the nature of some phenomenon. A system responds to particular signals to produce observable signals. In other words, input signals interact in a system to generate observable, output signals. For a further discussion about signals and systems, see Oppenheim et al.[19] and/or Porat[20].

System identification is a technique to create mathematical models of a system given examples of its input and output signals. A traditional application of system identification is to model a system so that the output signals can be controlled[2]. The models are formulated to predict the output of the system to any given input signal. The models are usually parameterized so that a vector $\vec{\theta}$ completely characterizes the model. The goal of system identification is then to select the value of $\vec{\theta}$ so that the model best represents the observed data. The model is never accepted as the "true" description of the system but rather as a tool to describe the aspects of the system that are of interest to the user.

Typically, the models used in system identification assume that the output signals are caused by a deterministic function of the inputs, delayed versions of the inputs and a stochastic noise process. The noise can be auto-correlated, and if there is feedback from the output, the noise can also be correlated with the inputs.

Experiments are run by applying input signals to the system to record what output signals the system generates. The input and output data are typically split into *training data* and *testing data*. The training data are used to estimate $\vec{\theta}$. The testing data are used to validate the model to assess how the model relates to observed data, to prior knowledge, and to its intended use. If the model is not valid, then a different model is considered and a new $\vec{\theta}$ is considered.

There are several motivations for using system identification to model time-varying emotional appraisals:

[2] In control systems literature, the system to be controlled is commonly referred to as the "plant".

1. It is possible to model a listener's perception of emotion as a system by treating the m musical features as an input signal and the emotional appraisals as an output signal. The digitally sampled, m-dimensional input signal and two-dimensional output signal can easily be generated in a similar manner as done by Schubert[24]. The system to model represents the generation of emotional appraisals in the human brain from musical stimuli.

2. Performing time series analysis to examine the relationship between the inputs and the outputs is limited. Most time series analysis assumes that all signals are either stationary or homogeneous nonstationary stochastic processes. System identification extends time series analysis to pseudo-stationary signals, which consist of a deterministic component plus a stationary component. Also, system identification provides techniques to evaluate how well the system generalizes through evaluation of predicted output signals to arbitrary input signals. This generalization cannot be directly measured through standard time series analysis.

3. Splitting the emotional appraisal into a deterministic function of the inputs and a stochastic component is intuitive. The deterministic component may model the cognitive appraisal of the music and the stochastic component would model measurement error and components of emotional appraisals that are not represented in the model.

4. The goal of system identification is to predict the outputs of a system given the inputs to learn how they affect the outputs. In this study, if emotional appraisals of musical stimuli by a listener can be predicted, the model has successfully generalized the relationship between musical features and the emotional appraisals. If the models are successful, it is possible to examine the model to learn how musical features create the perception of emotion in the listener.

5. System identification literature addresses particular challenges that occur while creating models. For example, if features are measured at a different frequency than the emotional appraisals, resampling techniques from signal processing can be used. Also, emotional appraisals by different people for different songs can be combined using data fusion techniques. Similarly, if different people appraise the same songs (input signal), techniques exist for estimating properties of the stochastic component of the emotional appraisals (output signals)[15].

For the reasons discussed in the preceding paragraphs, system identification will be used to construct models for emotional appraisals of music. Chapter 4 discusses in detail the methodology used in this thesis to generate and evaluate models.

Chapter 4

Proposed Method

4.1 Issues

The goal of this thesis is to model the emotional appraisals of music made by a population of listeners. As discussed in the previous chapters, a model should meet the following criteria:

1. The measured emotional appraisals of the listeners need to be time-varying.

2. The musical features that are inputs to the model need to represent many musical properties that communicate emotion and also be time-varying.

3. The model needs to be estimated/trained using emotional appraisals to musical se-

lections representing a genre of music.

4. The model needs to accurately simulate emotional appraisals to any musical selection from the genre of music.

Currently, no model exists that satisfies all four of these criteria. For this reason, the primary research concept for this thesis is to show that system identification provides a means to create a valid model with all four of these properties. Several different models will be created and evaluated for comparison.

Evaluation of a model will be based on how well it meets all four of the criteria above. The following sections in this chapter will describe the methodology used to construct models meeting the first three criteria. To evaluate the final criterion, each model will be evaluated to measure how well it generalizes emotional appraisals. The evaluation methods will be discussed in this chapter and the results can found in Chapter 5.

The system identification process consists of six stages that can be performed iteratively. The first step is the design of the experiment to gather the input and output data needed to construct models. After the data are collected, the data are preprocessed to minimize problems in the identification procedures. Then, several model structures are selected to be evaluated and the criterion used to estimate the models is selected. Finally, the models are estimated and evaluated to determine how to improve the model. If the model is inadequate, then other model structures are considered. These steps are described in this chapter and the results of the iterative process of model selection, estimation and validation are described in Chapter 5.

4.2 Experiment Design

The models created using system identification are based entirely on the input and output data. This means that the set of input data should represent all of the inputs that we wish the system to model. Similarly, the outputs of the system should represent as many potential outputs as possible. Because the outputs are assumed to be a function of the inputs, the inputs selected for use in this procedure must be selected carefully to create valid models. The inputs determine which parts of the system are investigated during the experiment. The importance of selecting appropriate inputs becomes even more evident when considering the cost and time required to create a new set of input data after starting the analysis.

Ljung provides six guidelines for selecting the input signals[15]. First, to minimize the bias in the parameters of the model, the experiment needs to resemble the situation under which the model is to be used. Second, to minimize the variance of parameters used to describe the models, inputs and outputs should be chosen to make the predicted output sensitive with respect to each important parameter. Third, to have informative experiments the inputs need to be persistently exciting of a large order. Fourth, the inputs need to be capable of validating and invalidating the models. Fifth, if noise estimation or reduction is important and the inputs are independent of the noise, a periodic signal should be used as input. Finally, if the system is nonlinear, the best prior information available about the system should be used to select the inputs.

In each model, the input signals are the musical features and the output signals are the emotional appraisals. Therefore, selecting the inputs involves selecting the musical stimuli that will be appraised by people during a study. Because the system may be nonlinear, prior information from Schubert suggests that the musical stimuli should be real music as opposed to melodies or isolated sounds[24]. To be persistently exciting, measurable musical properties identified to communicate emotion (such as tempo, pitch, volume, articulation, timbre and harmony) need to vary regularly in the selected music throughout the duration of the experiment. The songs will be selected to represent a large operating range of the 2DES. Also, by exposing the same songs to different people, the inputs can be treated as stationary over the ensemble and thus be used for noise reduction. The assumption that the inputs are independent of the noise process needs to be evaluated.

The input and output signals will be sampled at discrete points in time so the sampling interval needs to be determined. Given a sampling interval of T seconds (s), the maximum frequency that can be represented is $\frac{1}{2T}$Hz. Schubert's *EmotionSpace Lab* software[24] samples emotional appraisals at 1Hz so this is the sampling rate used in this thesis. Thus, it is assumed that emotional appraisals contain information only at frequencies below 0.5Hz. It would be worthwhile to sample much faster in future studies and then resample the signal to a desired frequency to ensure all frequencies of interest are collected.

The number of input and output measurements to record is another design variable. To ensure that each listener is able to concentrate throughout the experiment, the duration of the session with each listener is limited to twenty minutes[17]. Thus, it is impractical to have each listener appraise a large number of pieces from the same genre. For the data to be maximally informative, the musical selections need to differ and vary considerably. This is accomplished by using as many songs as possible in a twenty minute period that have been slightly modified for duration[1](e.g. [8]).

To satisfy the third model criterion, the musical selections will be from the same genre of music. The pieces are selected from the Western Art musical style for ease of comparison with Schubert's work[24]. Since the total duration of the music is limited to twenty minutes, it is unlikely that the entire genre will be represented. However, this shortcoming is acceptable because the goal of this thesis is to show that system identification is capable of constructing valid models. If the goal was to create a model that is capable of representing a genre of music, then this shortcoming would need to be addressed.

[1] Specially composed pieces that cover a large range of inputs and outputs could also be appropriate.

Table 4.1: Musical selections used in this study.

#	Alias	Title of Musical Selection	Composer	Times	Duration
1	Allegro	Piano Concerto No. 1 – Allegro Maestoso	Liszt	0:00 – 5:15	5:15
2	Aranjuez	Concierto de Aranjuez – Adagio	Rodrigo	7:05 – 9:45 & 5s silence	2:45
3	Fanfare	Fanfare for the Common Man	Copland	0:00 – 2:50	2:50
4	Moonlight	Moonlight Sonata – Adagio Sostenuto	Beethoven	0:00 – 0:22 & 3:08 – 5:19	2:33
5	Morning	Peer Gynt – Morning	Grieg	0:00 – 2:39 & 5s silence	2:44
6	Pizzicato	Pizzicato Polka	J. Strauss	0:00 – 2:31	2:31

Table 4.1 lists the musical selections from Naxos's "Discover the Classics" CD (8.550035-36) that are used in this study as well as the aliases used to refer to them. Portions of Aranjuez, Morning and Pizzicato were selected to allow comparison with models from Schubert's study. Allegro, Fanfare and Moonlight were selected from the same CD and assumed to contain musical features and emotional appraisals that are different from the other three songs. The duration of each musical selection is adjusted to be approximately 2min40s to equally weight each song in the models. The duration of Moonlight Sonata is reduced by removing 0:22 – 3:08 because the music is (almost) identical at 0:22 and 3:08. The entire duration of Allegro is used because initial testing showed that the emotional appraisals from this song span a broad range of the 2DES and thus may be more informative than the other songs. Also, to ensure that each listener has the same amount of time to finalize their appraisal at the end of each song, five seconds of silence are added to the end

of Aranjuez and Morning. Finally, these musical selections are burnt onto a CD for use with *EmotionSpace Lab*.

For the remainder of this thesis, these six musical selections will be refered to as *songs*. Although the term *song* is technically incorrect, usage of this term improves readability.

4.3 Data Sets

4.3.1 Musical Features

To use the musical selections as input signals in the model, the music needs to be represented by m time-varying musical features to satisfy the second model criterion. These m features are measured every second and treated as an m-dimensional vector, $\underline{u}_i(t)$, where t is the time in seconds when the features are calculated and i is the song number in Table 4.1. As mentioned in the background, the goal of selecting musical features is to quantify and represent all of the properties about the music needed to model emotional appraisals. The methodology used to achieve this goal is described in this section.

In constructing a model, it is important to model true relationships between inputs and outputs and avoid including false relationships. In general, increasing the number of features in a model increases the number of parameters in the model. For a fixed amount of training data, increasing the number of parameters in a model increases the significance of variance error or *overfitting*. Therefore, to reduce the effects of overfitting the model to the training data, smaller values of m are desirable. This conflicts with the goal to improve the representation of the musical selection by increasing m. In this thesis, overfitting is addressed by evaluating how well the model generalizes; many features will be used initially in the models and evaluation of the models will determine if there are too many features.

The eighteen musical features used in this thesis to achieve the second model criterion are summarized in Table 4.2. All musical features were local features extracted using PsySound, the FFT extractor from MARSYAS or extracted manually[2][29]. Features are extracted using established algorithms to minimize subjectivity in the features. PsySound is used because it extracts psychoacoustic features that represent many musical properties that communicate emotion. MARSYAS is used for feature extraction because it has successfully been used in music information retrieval applications.

The diffuse field was used for PsySound analysis because music is the auditory stimulus and the music may be interpreted as originating around the listener since they are wearing headphones[2]. The features extracted by MARSYAS were resampled from $\frac{44100}{512}$Hz to 1Hz using a polyphase, anti-aliasing filter to eliminate high frequency noise[20].

Features are selected to represent the musical properties that communicate emotion such as dynamics, mean pitch, variation in pitch, timbre, harmony, articulation, tempo, texture, vibrato, register, mode, note onset, melodic contour, pitch range, rhythm and

Table 4.2: Musical features used in this study.

Musical Property	**Alias**	**Musical Feature**	**Extraction Method**
Dynamics	LN	Loudness Level	PsySound
	NMax	Short Term Maximum Loudness	PsySound
Mean Pitch	Centroid	Power Spectrum Centroid	PsySound
	MeanCentroid	Mean STFT Centroid	MARSYAS FFT
Pitch Variation	MeanFlux	Mean STFT Flux	MARSYAS FFT
	StdFlux	Standard Deviation STFT Flux	MARSYAS FFT
	StdCentroid	Standard Deviation STFT Centroid	MARSYAS FFT
Timbre	TW	Timbral Width	PsySound
	S(Z&F)	Sharpness (Zwicker and Fastl)	PsySound
	MeanRolloff	Mean STFT Rolloff	MARSYAS FFT
	StdRolloff	Standard Deviation STFT Rolloff	MARSYAS FFT
Harmony	SDiss(H&K)	Spectral Dissonance (Hutchinson and Knopoff)	PsySound
	SDiss(S)	Spectral Dissonance (Sethares)	PsySound
	TDiss(H&K)	Tonal Dissonance (Hutchinson and Knopoff)	PsySound
	TDiss(S)	Tonal Dissonance (Sethares)	PsySound
	CTonal	Complex Tonalness	PsySound
Tempo	BPM	Beats per Minute	Schubert's method
Texture	Mult	Multiplicity	PsySound

metre. Seven of these properties are directly represented by features and six others may be indirectly represented by the same features. These features are described in detail in the following paragraphs.

Dynamics are represented using PsySound's loudness level (LN) and short term maximum loudness (NMax). The weighted and unweighted sound pressure levels calculated in PsySound for the songs, such as the A-Weighted sound pressure level used by Schubert[24], were found to be similar to the loudness level and thus were not included. The mean loudness was similar to the short term maximum loudness but appeared to have a smaller signal-to-noise ratio (SNR) and thus was not included. The two features selected for use are assumed to be adequate to represent musical dynamics.

Mean pitch is represented using two different power spectrum centroid calculations from PsySound and MARSYAS (Centroid, MeanCentroid). The mean is measured over one second windows.

Pitch variation is represented using statistics of the Short-Time Fourier Transform (STFT) measured by MARSYAS. It is assumed that calculating the standard deviation of the power spectrum centroid (StdCentroid), the mean of the STFT flux (MeanFlux) and the standard deviation of the STFT flux (StdFlux) over one second windows will represent pitch variation.

Timbre is represented primarily using PsySound's timbral width (TW) and one of

PsySound's sharpness measures (S(Z&F)). Zwicker and Fastl's algorithm for calculating sharpness is used because the values are similar to sharpness calculated using Aures's algorithm but the SNR of Zwicker and Fastl's algorithm appears to be higher[2]. Even though the mean and standard deviation of the STFT rolloff (MeanRolloff, StdRolloff) may not directly represent timbre, they are included because they have been used successfully in music information retrieval.

Harmony is represented using four different measures of dissonance (SDiss(H&K), SDiss(S), TDiss(H&K), TDiss(S)) as well as complex tonalness (CTonal). Each of the four dissonance measures calculated by PsySound are different so no decision could be made to omit any particular one. Also, the complex tonalness and pure tonalness calculated by PsySound are very similar for these musical selections so the complex tonalness is selected arbitrarily.

Tempo was calculated manually using the same method described by Schubert[24]. Some of the music varied in tempo considerably so the beats were manually detected to overcome shortcomings in beat-detection algorithms. To estimate the beats per minute (BPM) every second, linear interpolation was used between beats. The tempo of the silence at the end of each song was assumed to remain constant because the tempo of silence is meaningless.

Texture is represented using Parncutt's algorithm for calculating multiplicity (Mult).

Multiplicity is an estimate of the number of tones playing in the sound and is measured using PsySound.

There are other musical properties that are assumed to be implicitly represented using the features above. These properties are articulation, vibrato, register, mode, note onset and melodic contour. Articulation may be partially accounted for by Zwicker and Fastl's sharpness measure. Vibrato may be represented by the pitch variation features. Register may be represented by the mean pitch and timbre features. The mode may be accounted for by the harmony features. Note onset may be represented by the sharpness measures. Finally, melodic contour is assumed to be represented by the mean pitch features because the model is capable of subtracting a lagged version of the mean pitch to approximate the rate of change of pitch.

Finally, some musical properties have not been included in the models. Pitch range is a global feature and thus cannot be represented as an input signal. MARSYAS provides global features that may represent rhythm and metre but no features were found to represent rhythm and metre as input signals. The portion of the emotional appraisals influenced by these musical properties, and other unknown musical properties, are assumed to be accounted for by the stochastic component of the models.

All of these features will be calculated and m of the features will used in the models. The value of m will vary depending on which features are being investigated for use in particular models. The data collected from extracting features from the musical selections is described in Section 5.1.1. Graphs of the features can be found in Appendix B.1.

4.3.2 Emotional Appraisals

Emotional appraisals are measured using *EmotionSpace Lab*, which quantifies emotion using the dimensions valence and arousal[24]. The emotional appraisal data is collected at 1Hz as volunteers use *EmotionSpace Lab* to appraise the same six music selections using the 2DES. This method of measuring emotional appraisals satisfies the first model criterion because the emotional appraisals can change with time.

Each person who volunteers to emotionally appraise music goes through the same procedure. First, each volunteer reads an information letter describing the study as well as the purpose of the research. The study then begins by asking each volunteer five questions to record their gender, age and musical background using the questionnaire shown in Appendix A.1. At this point, the volunteer is asked to run *EmotionSpace Lab*.

EmotionSpace Lab is configured so that each participant goes through a tutorial to learn how to appraise emotion using the 2DES. During the tutorial, the emotional stimuli consist of the same words and faces used in Schubert's study and the volunteer appraises the stimuli using valence only, arousal only and both valence and arousal[24]. Then, after a sound check, each participant moves the mouse in the 2DES to emotionally appraise each of the musical selections in random order. At the completion of each song, the participant has the opportunity to rest until they are ready for the next song. After appraising all six songs, the participant has the opportunity to ask the researcher questions, and is given a feedback letter thanking them for participating in the study.

The data collected from the studies are described in Section 5.1.2. See Appendix B.2 for graphs of the emotional appraisals.

Finally, two possible approaches are considered for modeling the emotional appraisals for the sample population. The first approach is to estimate an emotional appraisal that represents the emotional appraisal of most people in the population. The second approach is to generate a model for each listener and then compare the parameters of each of the models to determine what parameter values are typical for the population. Because there is less computational effort required to use the first approach, and because this approach allows a direct comparison with the models generated by Schubert, it will be the approach taken in this thesis.

4.3.3 Preprocessing

To create an emotional appraisal representative of the sample population, the emotional appraisals are preprocessed in the following manner. First, explicitly deal with outliers and missing data. Second, to reduce the effects of noise, apply filters to the input and output signals. Finally, an emotional appraisal to represent the sample population is calculated for each song.

The first step is to remove occasional bursts/outliers and handle missing data. Outliers consist of signals that are non-representative of the rest of the data. For example, data segments that contain no information or have non-representative data can be considered outliers. Missing data are unknown signal values at particular points in time due to errors in the measurement process.

In this study, only output signals (emotional appraisals) have outliers and missing data because the algorithms used to calculate the input signals (musical features) are assumed to be robust. During data collection, some output data are missing due to the nature of *EmotionSpace Lab*. If the user moves the mouse outside of the 2DES axes, no coordinates are stored for each second that the cursor is outside of the box[24]. Also, at certain times the outputs can be considered outliers. For example, some people appraise emotion much differently from the majority of the people in the sample population; thus at these times, their appraisals can be considered outliers. Outliers will be identified according to the following rules and treated as missing data:

Rule 1 *If fewer than 10% of the emotional appraisals at a particular region in time are over two standard deviations away from the mean for the region, treat those appraisals as outliers.*

Rule 2 *If a person's arousal or valence appraisal is considered to be an outlier by Rule 1 for over 30% of the duration of the song, remove that component of the person's appraisal for the entire song.*

Rule 3 *If removing the emotional appraisals of the first song heard by the participants reduces the average variance by at least 5% for a component of a song's appraisal, remove that component of the appraisal for the participants who heard that song first.*

The motivation for Rule 1 is that a minority of the population may re-evaluate their emotional appraisal during that region and attempt to change it. If this occurs, then the person may not consider their own emotional appraisal to be accurate for that region. Rule 2 provides a criterion to remove appraisals detrimental to estimating an emotional appraisal representative of the population. Rule 3 provides a method to remove emotional appraisals collected while a participant was still learning how to appraise music using *EmotionSpace Lab*.

Once the outliers are removed, missing data need to be considered. There are several approaches that can be taken to process missing data. When the data are non-periodic, typically a time varying Kalman filter or the Expectation-Maximization algorithm is used[15]. However, since the data are treated as periodic in this study, the outliers and missing data will simply be omitted from the calculation of the population's emotional appraisal by assuming fewer emotional appraisals were recorded at those times.

The second step is to filter the input and output signals to reduce the effects of noise. To simplify analysis, assume that people respond to the same emotional stimuli at the same time with the same amplitude and that noise causes emotional appraisals to vary between listeners. This assumption may not be accurate but will be considered a starting point for identifying models to limit the scope of this thesis. Future studies may be concerned with addressing the possibility that different people have different reaction times to musical stimuli, may not respond to certain stimuli and may use various sized regions of the 2DES.

The signals can be temporally low-pass filtered to remove high-frequency disturbances in the data that are above the frequencies of interest to the system dynamics or high-pass filtered to remove drift, offset and low-frequency disturbance. Filtering the signals before fitting linear models can optimize the bias and MSE (mean squared error) in the frequency response[15]. It can be shown that filtering the input-output data is the same as filtering prediction errors and the same as changing the noise model[15].

To remove offsets in the data and to investigate relative changes in the signals over absolute values of the signals, the inputs and outputs will be treated as deviations from their means. In other words, because many frequencies are potentially of interest in the data, the high-pass filter that will be applied is a notch filter that removes the mean of the input and output signals (i.e. DC removal). Performing the first difference of the inputs and outputs is not considered because it over-emphasizes the high-frequencies[15]. Because it is unclear what the frequency response should be of a filter to remove high-frequency disturbances, the data are not low-pass filtered.

The third step is to calculate an emotional appraisal that is representative of the J people in the sample population. For the discussion that follows, define the following two-dimensional, time-varying vectors:

$\underline{\gamma}_{ij}(t)$ the emotional appraisal of person j to song i at time t, $j = 1, \ldots, J$

$\underline{Y}_i(t)$ a random vector describing the population's emotion appraisal of song i at time t

$\underline{y}_i(t)$ an emotional appraisal representative of the population for song i at time t

$\underline{\xi}_{ij}(t)$ the difference between the representative emotional appraisal and the emotional appraisal of person j to song i at time t; $\underline{\xi}_{ij}(t) = \underline{y}_i(t) - \underline{\gamma}_{ij}(t)$

The pdf (probability distribution function) of $\underline{Y}_i(t)$ is a function of musical features and emotional appraisals prior to time t. However, by considering the marginal pdf of the emotional appraisal as a function of time only, it is possible to calculate an emotional appraisal representative of the population at a particular time t by considering only the observed emotional appraisals at t. This is acceptable because the models that will be identified determine how the musical features and emotional appraisals affect $\underline{Y}_i(t)$.

The vector $\underline{\gamma}_{ij}(t)$ can be interpreted as the j^{th} observation of $\underline{Y}_i(t)$. To construct $\underline{y}_i(t)$, an emotional appraisal representative of the sample population, statistics such as the mean, median or mode of $\underline{Y}_i(t)$ can be used. Because the pdf of $\underline{Y}_i(t)$ is unknown, these statistics can be estimated from the observations $\underline{\gamma}_{ij}(t)$. Therefore, $\underline{y}_i(t)$ will be a function of $\underline{\gamma}_{i1}(t), \underline{\gamma}_{i2}(t), \ldots, \underline{\gamma}_{iJ}(t)$.

The sample mean of $\underline{\gamma}_{i1}(t), \underline{\gamma}_{i2}(t), \ldots, \underline{\gamma}_{iJ}(t)$ can be used to calculate $\underline{y}_i(t)$ at time t. For this to be valid, the mean of the emotional appraisals should represent the data for all values of t. In terms of model estimation, using the sample mean is equivalent to treating $\underline{\gamma}_{ij}(t)$ as a periodic signal that is repeated J times. This implies that $\underline{\xi}_{ij}(t)$ can be used for noise estimation and that noise is reduced in $\underline{y}_i(t)$.

The sample median can also be used to calculate $\underline{y}_i(t)$. The advantages of using the sample median of the emotional appraisals over the sample mean are that $\underline{y}_i(t)$ is not as sensitive to outliers and the median represents the data differently. Unfortunately, the median may not be as smooth as the mean as a function of t and it is not a linear function of the individual emotional appraisals so it is more difficult to use for noise estimation. However, the median may be more representative of the data and so it will be compared to the sample mean in Section 5.1.3.

It is possible to estimate the mode of $\underline{Y}_i(t)$ at each time t by estimating the pdf using a nonparametric pdf estimation technique such as Parzen Windows. To accurately estimate the pdf of $\underline{Y}_i(t)$, J needs to be large[4]. Unfortunately, not enough emotional appraisals are collected in this study to be able to estimate the pdf of $\underline{Y}_i(t)$ accurately, so the mode will not be used to calculate $\underline{y}_i(t)$[4].

The calculations and comparisons of the mean and median emotional appraisals can be found in Section 5.1.3. After the representative emotional appraisal $\underline{y}_i(t)$ is calculated, the mean is subtracted from it as discussed on p. 40.

4.3.4 Summary

The following summarizes the methodology used to generate the data sets. The results from applying this methodology can be found in Section 5.1.

1. Measure the eighteen musical features from the six songs to calculate $\underline{u}_i(t)$

   $i = 1, \ldots, 6$

2. Collect emotional appraisal data $\underline{\gamma}_{ij}(t)$ from J volunteers

   $i = 1, \ldots, 6$, $j = 1, \ldots, J$

3. Preprocess the data to remove outliers and calculate $\underline{y}_i(t)$

   $i = 1, \ldots, 6$

As part of preprocessing of the data, the sample mean and sample median emotional appraisals will be compared to determine which represents the population better. Also, the temporal means of $\underline{u}_i(t)$ and $\underline{y}_i(t)$ will be removed.

4.4 Model Identification

4.4.1 Model Structures

Once the data sets used for training are collected, the next step is to select the model structures to use. Because there are m inputs and two outputs, only multivariable model structures are considered. Two linear model structures are investigated in this thesis. Each model structure is parameterized using a d-dimensional vector $\underline{\theta}$ consisting of all of the parameters needed to describe the model. The nonparametric impulse response and frequency response models are also considered but only for validation as described in Section 4.4.3.

Only simulation models will be constructed in this thesis. A simulation model predicts the output based entirely on the input signal and delayed versions of the input. The alternative, a k-step ahead predictor, assumes that the "true" output is known k samples after the output was predicted. This implies that the k-step ahead predictor models can only be used when measured emotional appraisals are available. Because the fourth model criterion requires models to estimate emotional appraisals for musical selections with unknown "true" appraisals, k-step ahead predictor models are inappropriate.

The model structures considered can be linear or nonlinear. Linear model structures are simpler to estimate and analyze and nonlinear model structures are complex but more flexible. Generally, linear model structures should be considered first and nonlinear model structures should be considered only if linear model structures are inaccurate. To limit the scope of this thesis, only linear model structures will be considered[2].

The two linear models considered are the ARX (Auto-Regression with eXtra inputs) and State-Space model structures[16]. These models are the only linear models considered in this thesis to avoid difficulties estimating other multivariable linear structures. Also, limiting the discussion to these models allows usage of MATLAB's System Identification Toolbox[16].

Given m-dimensional input data $\underline{u}(t)$ and 2-dimensional output data $\underline{y}(t)$, the ARX model structure can be described using the following expression[3]:

$$\underline{y}(t)+A_1(\underline{\theta})\underline{y}(t-1)+\ldots+A_{n_a}(\underline{\theta})\underline{y}(t-n_a)=B_0(\underline{\theta})\underline{u}(t)+\ldots+B_{n_b}(\underline{\theta})\underline{u}(t-n_b)+\underline{e}(t) \quad (4.1)$$

where,

[2] Preliminary models considered suggest that nonlinearities exist if the means of the input/output data are not removed from each song. If one felt it important to include the means in the model, there would be more motivation to consider nonlinear model structures.

[3] If $\underline{u}(t)$ were treated as a white noise process, the ARX model would be equivalent to an ARMA model. However, $\underline{u}(t)$ is deterministic so $\underline{y}(t)$ is not an ARMA process.

$A_k(\underline{\theta})$ is a 2x2 matrix

$B_k(\underline{\theta})$ is a 2xm matrix

$\underline{e}(t)$ is a 2-dimensional white noise process with zero mean

n_a is the maximum number of auto-regressive terms in the model

n_b is the maximum number of lagged inputs in the model

$\underline{\theta}$ is a d dimensional vector containing all of the non-zero elements of $A_k(\underline{\theta})$ and $B_k(\underline{\theta})$

The matrices $A_k(\underline{\theta})$ and $B_k(\underline{\theta})$ are composed of zeroes and the parameters that need to be estimated from the input and output data.

By introducing the unit-shift operator q, $q^{-k}\underline{y}(t) = \underline{y}(t-k)$, it is possible to describe (4.1) using transfer function matrices:

$$A(q,\underline{\theta})\underline{y}(t) = B(q,\underline{\theta})\underline{u}(t) + \underline{e}(t) \tag{4.2}$$

$$A(q,\underline{\theta}) = I + A_1(\underline{\theta})q^{-1} + \ldots + A_{n_a}(\underline{\theta})q^{-n_a} \tag{4.3}$$

$$B(q,\underline{\theta}) = B_0(\underline{\theta})q^{0} + \ldots + B_{n_b}(\underline{\theta})q^{-n_b} \tag{4.4}$$

Using (4.2), the simulated output of the ARX model is described using the following equation:

$$\hat{\underline{y}}(t|\underline{\theta}) = \Big[A(q,\underline{\theta})\Big]^{-1} B(q,\underline{\theta})\underline{u}(t) \tag{4.5}$$

Given the same input and output data as in the ARX model structure, the state-space model structure can be described using the following expressions:

$$\underline{x}(t+1) = A(\underline{\theta})\underline{x}(t) + B(\underline{\theta})\underline{u}(t) + K(\underline{\theta})\underline{e}(t) \tag{4.6}$$

$$\underline{y}(t) = C(\underline{\theta})\underline{x}(t) + D(\underline{\theta})\underline{u}(t) + \underline{e}(t) \tag{4.7}$$

$\underline{x}(t)$ n-dimensional state-vector

$A(\underline{\theta})$ a nxn matrix representing the dynamics of the state-vector

$B(\underline{\theta})$ a nxm matrix describing how the inputs affect the state variables

$C(\underline{\theta})$ a 2xn matrix describing how the state-vector affects the outputs

$D(\underline{\theta})$ a 2xm matrix describing how the current inputs affect the current outputs

$K(\underline{\theta})$ a nx2 matrix used to model the noise in the state-vector

The initial state $\underline{x}(0)$ can be set to zero or estimated from the data by including it in $\underline{\theta}$. Also, all non-zero elements of the matrices are represented using $\underline{\theta}$.

When used in control systems, the state vector $\underline{x}(t)$ typically represents information with physical significance (e.g positions, velocities, voltages) so that the measured outputs are a known linear combination of the state variables[15]. Therefore, in this thesis, one could speculate that the state vector may represent the information needed to generate an emotional appraisal.

By combining (4.6) and (4.7), the simulated output of the state-space model is described using the following equation:

$$\underline{\hat{y}}(t|\underline{\theta}) = \Big[C(\underline{\theta})\big[qI - A(\underline{\theta})\big]^{-1}B(\underline{\theta}) + D(\underline{\theta})\Big]\underline{u}(t) \tag{4.8}$$

After selecting these structures, the number of parameters needs to be chosen. In the ARX model, this involves selecting n_a, n_b and the non-zero elements of matrices $A_k(\underline{\theta})$ and $B_k(\underline{\theta})$. From the work of Tillman and Bigand, it appears that less than six seconds of musical stimuli are needed to represent emotion so the maximum order considered will be five[28]. The parameters will then be modified using trial and error from insight gained from analyzing parameters of the models. In the state-space model, choosing the number of parameters involves selecting the order n of the system and the delays $n_{k,ij}; i = 1, \ldots, m; j = 1, 2$ from each of the m inputs to both of the outputs. The order will be selected by determining the n singular values of the extended observability matrix that are significantly larger than zero for a large order state-space system[15].

The delays for the inputs are chosen by using correlation analysis to estimate the nonparametric step responses[15]. Each delay is estimated by determining when the nonparametric step response for each input becomes significantly different from zero. Because of the number of inputs and the fact that the data are divided into experiments, only a subset of the inputs can be used in the correlation approach[16]. Therefore, the inputs will be randomly split into multiple subsets of 6, 9 and 12 inputs to generate estimates of the step responses. See Appendix D.1 for graphs of the nonparametric step responses.

The initial ARX model identified will be a fourth order ARX model with delays for each input. Each delay is estimated by examining the nonparametric step response for each input to determine when it is significantly different from zero. Variations of this model will be examined by changing the orders for each input and the auto-regression. The models will be compared using techniques described in Section 4.4.3.

Both the ARX model structure and the state-space model structure can be shown to be mathematically equivalent. Typically, a transfer-function model, such as ARX, is used when the form of each transfer function can be estimated and there is no a priori knowledge of the mathematical model. State-space models are used when the order of the transfer functions is unknown or if a priori knowledge can be expressed using the state vector. Because the form of each transfer function is unknown, several different orders of state-space models will be examined and then several ARX models structures will be constructed based on insights gained from evaluating the state-space models.

The inputs to the models will also be investigated. Based on results from the model validation, certain inputs will be removed in some models to see how the fit is affected. Another possibility would be to try adding nonlinear combinations of inputs, but this is not done because it is unclear what transformations may improve the identification.

Also, separate MISO (multiple input, single output) models will be created for the arousal and valence components for comparison with the MIMO (multiple input, multiple output) models. While MISO models are unlikely to provide a better fit for the outputs, estimating MISO models provides an efficient method to investigate how the orders for the input transfer functions affect the fit[16].

For details on what models are constructed, see Section 5.2. For more information about these linear model structures, see [15].

4.4.2 Model Estimation

Once the structure of the model is selected, the parameters of the model need to be estimated so that the model fits the input and output data. The estimation techniques used depend on factors such as the model structure, algorithmic complexity, optimization difficulties and the intended use of the model.

Before estimating the models, the data is divided into the *training set* for estimating the parameters and the *testing set* for validating the models. Initial model estimation will be done using the data from songs Allegro, Aranjuez, Fanfare, Moonlight, Pizzicato. These songs are all from the same genre of music and thus the third model criterion is satisfied. Models estimated using these songs will be validated by using the data from Morning, which was not used to train the model. Because Morning is a song unfamiliar to the system, it can be used to determine how well the model generalizes to other songs, and thus satisify the fourth model criterion. Evaluation of the most promising model structures will be done using cross-validation techniques to avoid biasing the models as described in Section 4.4.3.

For linear models there are three general approaches to estimating the parameters in a model[15]. First, there is PEM (the Prediction Error Method) which can be thought of as a generalization of least-squares because $\underline{\theta}$ is selected to minimize a function of the output error of the one-step ahead predictor[15]. The main design variables in PEM are the norm used to measure the error and the preprocessing filter. Another approach is the correlation approach which selects $\underline{\theta}$ so that the error at time t is uncorrelated with data prior to t. The multistep IV (Instrumental-Variable) implementation of the correlation approach as described by Ljung is a simple technique to estimate $\underline{\theta}$ where the only design variable is the linear regression structure[15]. The third approach is the subspace approach to estimating the matrices in state-space models using an estimate of the extended observability matrix. There are several design variables in the subspace approach as described by Ljung such as the maximum prediction horizon, the weighting matrices, the "post-multiplication matrix" R and the correlation vector.

Although PEM is the most computationally demanding of the three approaches, it will generate unbiased estimates of the parameters if the 'true' system is not in the model set[15]. Because it is unlikely that all of the variables of the 'true' system are included in this thesis, PEM will be used to estimate the models. In this thesis, the preprocessing filter used is described in Section 4.3.3. The determinant of the estimated error covariance, $\widehat{\Lambda}_N(\underline{\theta})$, will be used as the norm as shown in the following equations:

$$\hat{\underline{\theta}} = \arg\min_{\underline{\theta}} V_N(\underline{\theta}) \tag{4.9}$$

$$V_N(\underline{\theta}) = \left|\widehat{\Lambda}_N(\underline{\theta})\right| \tag{4.10}$$

$$\widehat{\Lambda}_N(\underline{\theta}) = \frac{1}{N-d}\sum_{t=1}^{N} \underline{\epsilon}(t|\underline{\theta})\underline{\epsilon}^T(t|\underline{\theta}) \tag{4.11}$$

$$\underline{\epsilon}(t|\underline{\theta}) = \underline{y}(t) - \underline{\hat{y}}_p(t|\underline{\theta}) \tag{4.12}$$

where,

$\underline{\hat{y}}_p(t|\underline{\theta})$ is the one-step ahead prediction for $\underline{y}(t)$ for the model structure

N is the total number of samples in the training set

d is the number of parameters in $\underline{\theta}$

The *loss function*, $V_N(\underline{\theta})$, is the determinant of the estimated noise (prediction-error) covariance. Minimizing the estimated noise covariance to solve for $\hat{\underline{\theta}}$ is equivalent to finding the maximum-likelihood estimate when the prediction-errors are jointly Gaussian[15].

Before estimating the parameters in the linear models, data fusion needs to be used to combine the input and output data from all of the songs. There are two possible approaches to combine the data from the songs. A model could be built for each song and then all of the models can be combined. Alternatively, the songs could be treated as one continuous musical selection but the initial conditions are reset at the beginning of each song. The latter approach is used because it is more efficient to only calculate one model and the model estimation will be better conditioned because each song has vastly different inputs[15].

See Section 5.2 for estimation results of these linear models.

4.4.3 Model Validation

After selecting the model structures and estimating the parameters in the models, the models need to be validated. Validating models involves assessing how they relate to observed data, to prior knowledge and their usage. This implies that several different tests will be performed to evaluate the models.

To assess how a model relates to observed data, the simulated emotional appraisals need to be compared to true emotional appraisals. In this thesis, this comparison will be done by evaluating the bias and variance of each model. Bias is the systematic difference between the simulated output and the true outputs and should ideally be close to zero. Variance can be thought of as uncertainties in model parameters and the output that is caused by having too many parameters, too much noise or too little data. Ideally, there should be little variance so the outputs can be predicted with some certainty; therefore, once all of the data has been measured, it is important to have a model with as few parameters as possible.

Evaluating the bias of the model will be done using two measures. The MSE is used to evaluate the simulation errors and the squared multiple correlation coefficient[4] (R^2) is used to evaluate the percentage of the output variation that is explained by the model[15]. In each simulation, the initial value of the emotional appraisal is estimated as well, because subtracting the means from the signals results in an unknown initial appraisal. Ideally, the MSE should be as close to zero as possible and R^2 should be as close to one as possible. Because there are two outputs, these measures will be calculated separately for each of the outputs. The MSE for channel k (MSE_k) is related to the R^2 measure for channel k (R_k^2) by the following equations[15]:

$$MSE_k = \frac{1}{N}\sum_{t=1}^{N} |y_k(t) - \hat{y}_k(t|\underline{\theta})|^2 \qquad k = valence, arousal \tag{4.13}$$

$$R_k^2 = \left(1 - \frac{MSE_k}{\frac{1}{N}\sum_{t=1}^{N} |y_k(t)|^2}\right) * 100\% \qquad k = valence, arousal \tag{4.14}$$

If the R^2 measure for a channel is negative, the energy of the error is greater than the energy of the true emotional appraisals. This implies that the simulated emotional appraisal is extremely different than the true emotional appraisal. For reference, a constant simulated output results in the R^2 measure to equal zero.

It also possible to use a third method to measure bias in linear models. Comparisons

[4] The squared multiple correlation coefficient is sometimes referred to as the “fit”.

of the nonparametric step response and frequency response to the estimated step and frequency responses of the models should be similar if there is no bias[15]. However, because evaluating the fit of the simulated output is more important than comparing impulse and frequency responses, these comparisons will not be used to evaluate the bias.

Because it is important to have a model that has few parameters and little variance, two techniques will be used to evaluate the variance error. First, the variance of the parameters will be estimated to calculate 98% confidence intervals. This corresponds to $\pm 2.33\sigma$ since the estimated parameters are approximately normally distributed when N (the number of data samples) is large[15]. Parameters that reflect design decisions (such as model order or time delay) should be statistically significant from zero to be included in the model. For this reason, the percentage of the parameters that are statistically significant from zero will be calculated. Also, if the confidence intervals of many parameters are large, then this implies that there are too many parameters[15]. The covariance of the parameters are estimated using the following equations:

$$\widehat{P}_{\theta} = \left[\frac{1}{N} \sum_{t=1}^{N} \underline{\psi}(t, \underline{\hat{\theta}}) \widehat{\Lambda}_N(\underline{\hat{\theta}}) \underline{\psi}^T(t, \underline{\hat{\theta}}) \right]^{-1} \tag{4.15}$$

$$\underline{\psi}(t, \underline{\theta}) = -\frac{d\underline{\epsilon}(t|\underline{\theta})}{d\underline{\theta}} = \frac{d\underline{\hat{y}}(t|\underline{\theta})}{d\underline{\theta}} \tag{4.16}$$

where,

$\underline{\psi}(t, \underline{\theta})$ is a dx2 matrix representing the gradients/sensitivity of the simulated output with respect to each parameter at time t

The second measure used to analyze the variance of the model is the estimated variance of the output signals. Ideally, the variance of the output signals is small so that the output is known with some certainty. To analyze the variance of the output signals, 98% confidence intervals of the simulated output will be graphically compared to emotional appraisals. The maximum confidence interval size over each component of arousal and valence will be recorded.

Because the model structures are linear, the output is a linear function of $\hat{\underline{\theta}}$. This implies that $\hat{\underline{y}}(t|\hat{\underline{\theta}})$ can be expressed as follows:

$$\hat{\underline{y}}(t|\hat{\underline{\theta}}) = \underline{\psi}^T(t, \hat{\underline{\theta}})\hat{\underline{\theta}} + \underline{e}(t) \tag{4.17}$$

Since $\hat{\underline{\theta}}$ is assumed to be normally distributed, $\hat{\underline{y}}(t|\hat{\underline{\theta}})$ is normally distributed as well. The variance of $\hat{\underline{y}}(t|\hat{\underline{\theta}})$ can be calculated on the validation data using the following equation since $\underline{e}(t)$ will be uncorrelated with $\hat{\underline{\theta}}$.

$$Var\Big(\hat{y}(t|\hat{\underline{\theta}})\Big) = \underline{\psi}^T(t, \hat{\underline{\theta}})\widehat{P}_{\hat{\theta}}\underline{\psi}(t, \hat{\underline{\theta}}) + \widehat{\Lambda}_N(\hat{\underline{\theta}}) \tag{4.18}$$

To assess how a model relates to prior knowledge, assumptions made during the creation of the models need to be verified. To verify that the inputs are independent of the noise process, the cross-correlation function between each input and the model residuals will be examined to ensure no negative lags are significantly different than zero. The auto-correlation function (ACF) of the output residuals will also be calculated to ensure only the 0^{th} lag is significantly different than zero. This test will be done to ensure that the noise is white.

Finally, because each model will be used to simulate emotional appraisals of music to which the model has not been exposed, it is important to assess how well each model simulates them. If a model can accurately simulate emotional appraisals to any musical selection from the genre of music, the fourth model criterion will be satisfied. For this reason, cross-validation will be performed by using the MSE and R^2 measures with data that was not used to estimate the models. As mentioned in Section 4.4.2, for all of the models, the data for Morning will be used as validation data and the other five songs will be considered the training set.

To avoid biased results from using only Morning as testing data, the model structures with the highest R^2 measures will be evaluated further. These model structures will be evaluated using six-fold cross-validation, where each model structure will be estimated six times, using a different song for validation each time. As before, the songs in the training set will be treated as one continuous musical selection but the initial conditions will be reset at the beginning of each song. The R^2 measures for the six different songs will be combined using the following weighted average:

$$\overline{R}_k^2 = 1 - \frac{\sum_{i=1}^{6} N_i MSE_{ik}}{\sum_{i=1}^{6}\sum_{t=1}^{N_i} |y_{ik}(t)|^2} \qquad k = valence, arousal \tag{4.19}$$

where,

MSE_{ik} is the MSE measure for output channel k for song i

N_i is the number of input/output samples for song i

Once all of the model structures have been evaluated, a resultant model will be created for the best model structures. The resultant model will be estimated using all six of the songs. These models will be compared using Akaike's FPE (Final Prediction-Error Criterion) to assess the tradeoff between minimizing the MSE while minimizing the variance error by limiting the number of parameters in the model[15]. The expression to calculate the FPE is given by Ljung[15]:

$$FPE = \frac{(N+d)}{(N-d)} V_N(\underline{\theta}) \tag{4.20}$$

The Chapter 5 lists results for the model validation.

4.4.4 Summary

The following summarizes the methodology used to iteratively select model structures and to validate the models. The results from applying this methodology can be found in

Section 5.2.

For all of the model structures Morning is used as the validation data, and the other songs are used for estimation. Each model structure will be validated using the following measures:

1. MSE for each channel using the validation data – ideally should be close to zero
2. R^2 for each channel using the validation data – ideally should be close to 100%
3. Percentage of parameters that are statistically significant from zero – ideally should be close to 100%
4. Maximum output signal confidence intervals – ideally should be close to zero
5. Cross correlation of inputs with residuals – negative lags should be zero
6. Auto correlation of output residuals – all lags except 0 should be zero

Using the above validation measures, select model structures to estimate using PEM using the following methodology:

1. Compute non-parametric step response to estimate the delays use as a heuristic for removing inputs
2. Estimate state-space models – pick several orders and delays and evaluate the models

3. Estimate a fourth order ARX model – try several delays for each input and evaluate

4. Iteratively adjust the parameters for the models from the previous two steps and evaluate

   (a) Remove some inputs from the models

   (b) Model arousal and valence separately (make MISO models)

   (c) Try other orders and delays in the ARX models to see how the fit changes

   (d) Try any combination of these approaches

5. Select the best model structures from the previous steps for further evaluation

   (a) Perform six-fold cross-validation to calculate an average R^2 fit

   (b) Use all six songs as estimation data and calculate the FPE

   (c) Plot the simulated outputs from the best model structure

Chapter 5

Results

5.1 Data Collection

5.1.1 Musical Features

The musical features were calculated as described in Section 4.3.1. For the graphs of the features before the means are removed, see Appendix B.1. The graphs illustrate $\underline{u}_i(t)$

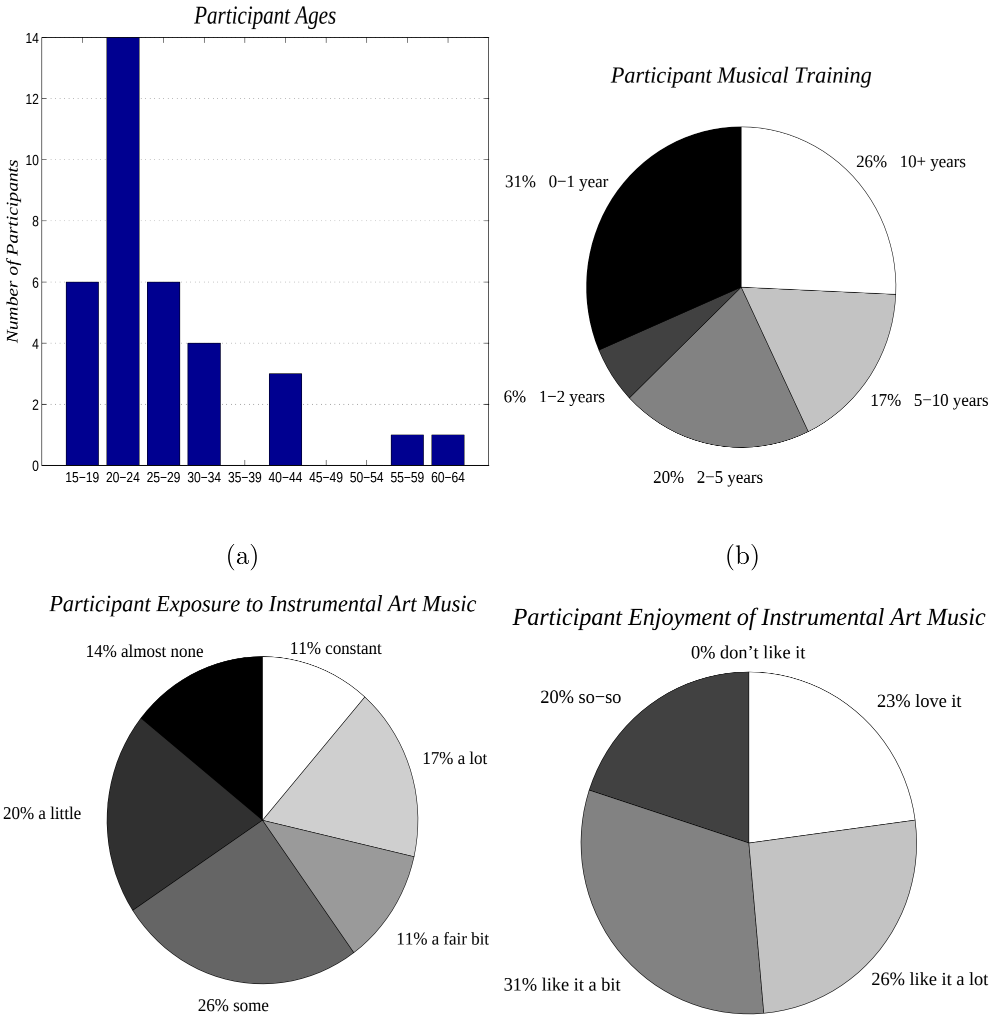
The number of samples output by the MARSYAS FFT feature extractor and the PsySound feature extractor were occasionally off by one because the algorithms are different. To ensure that all features had the same number of samples as the emotional appraisals, the last sample of each feature that was too short was duplicated. All of the songs had five seconds of silence at the end so the features are constant at the end.

5.1.2 Emotional Appraisals

Emotional appraisal data was collected from 35 volunteers – 21 male (60%) and 14 female (40%). Each participant was asked to fill out the questionnaire in Appendix A.1 to record information about their musical background. As shown in Figure 5.1(a), most of the participants were under the age of thirty-five. Figure 5.1(b) illustrates that the majority of the participants had some musical training but 31% of the participants had no training at all. According to Figure 5.1(c) and Figure 5.1(d), the participants had a broad range of exposure to classical music and enjoyed the music to various degrees, although nobody who took part in the experiment disliked classical music. There were more males and fewer musicians in this study than in Schubert's study[24].

Because it may take the participants some time to feel comfortable using *EmotionSpace Lab* to express their emotional appraisal, the emotional appraisal data may not be valid for the songs presented first. To verify that the songs heard by the participants were presented in different orders, the song orders are tabulated in Table 5.1. Morning and Pizzicato were presented first more often than other songs but this was not assumed to be significant. Section 5.1.3 discusses the preprocessing of the emotional appraisal data to address this issue.

The preprocessed emotional appraisals are shown in Appendix B.2. There is a large variance in the emotional appraisals relative to the scale so it appears that the SNR is



(a) (b)

(c) (d)

Figure 5.1: Participant statistics

Table 5.1: Song order heard by participants during the study.

Song	No. of Times Song Was Heard					
	1st	2nd	3rd	4th	5th	6th
Allegro	5	11	8	3	4	4
Aranjuez	4	5	6	6	8	6
Fanfare	4	7	7	1	7	9
Moonlight	5	5	5	7	7	6
Morning	8	6	5	9	3	4
Pizzicato	9	1	4	9	6	6

poor. The poor SNR of an individual appraisal provides additional motivation to use an emotional appraisal representing the population with a better SNR.

5.1.3 Preprocessing

As mentioned in Section 4.3.3, preprocessing the data consists of removing outliers and missing data, filtering the signals and then creating an emotional appraisal representative of the population.

Before preprocessing the data, the average standard deviation for arousal and valence over all the songs was 31.6%. Application of Rule 1 from Section 4.3.3 labelled 206 samples from 13 different people as outliers, corresponding to 0.53% of the emotional appraisal data to be removed. From Rule 2, the arousal appraisal of Pizzicato by one person was affected, removing an additional 0.39% of the data. From Rule 3, no samples were removed because removing the first song did not decrease the variance of any emotional appraisal significantly; thus the *EmotionSpace Lab* tutorial appears to effectively teach participants how to use the 2DES. In summary, removing 0.91% of the data reduced the variance of the emotional appraisals by 5% resulting in an average standard deviation over all the songs of 30.8%.

To filter the signals, the means were subtracted for all of the signals after creating the appraisal for the population[1]. Because it is unclear what the frequency response should be of a filter to remove high-frequency disturbances, no other filters were applied to the signals.

To generate the emotional appraisal representative of the sample population, the mean appraisal was compared to the median appraisal. As shown in Appendix B.2, the mean appraisal is similar to the median appraisal for most of the songs. The only appraisals where the mean appraisal is significantly different from the median appraisal are for Aranjuez-Valence (Figure B.14(a)), Aranjuez-Arousal (Figure B.14(b)) and Fanfare-Valence (Figure B.15(a)). For these appraisals, the marginal pdf appears to be either bimodal or

[1] The order of subtracting means and creating the appraisal only matters if a nonlinear method is used to create the appraisal. Therefore, the order would not matter when using the sample mean but it does matter for using the sample median.

skewed as many people appraised the music differently. Because the median is a more robust measure of centrality than the mean for bimodal and skewed distributions, the median emotional appraisal is used to represent the sample population[10]. The median emotional appraisal is described using the following equation:

$$\underline{y}_i(t) = \text{median}\Big(\underline{\gamma}_{i1}(t), \ldots, \underline{\gamma}_{i35}(t)\Big) - \underline{\mu}_i \qquad \forall i = 1, \ldots, 6; t = 1, \ldots, N_i \tag{5.1}$$

where,

median() is the sample median, ignoring outliers and missing data

$\underline{\mu}_i$ is the mean appraisal of song i used to ensure that $\sum_{t=1}^{N_i} \underline{y}_i(t) = [0, 0]^T$

5.2 Linear Models

5.2.1 Step Responses

Initially, nonparametric step responses between the inputs and outputs were estimated using correlation analysis. As mentioned in Section 4.4.1, the inputs are randomly split into 20 subsets of 6, 9 and 12 inputs to generate estimates of the step responses and are shown in Table D.1. Graphs of the estimated step responses can be found in Appendix D.1.

Table 5.2 summarizes the delay for each input-output pair estimated from the step responses. If an estimated delay could be one of two possible values, the lesser of the two values is used. The subsets sometimes varied considerably in their delays and slopes. Therefore, each input-output pair is also rated subjectively to describe the consistency of the estimates between subsets.

The consistency of the step response estimates is subjectively graded on a scale as follows:

A all estimates go in same direction, most are significantly different from zero at same locations

B most estimates go in same direction, most are significantly different from zero at same locations

C some estimates go in same direction, some are significantly different from zero at same locations, some are not significantly different from zero

D the estimates go in many different directions, many are not significantly different from zero

By arguing that inconsistent step response estimates for an input/output pair implies that that output is not a function of that input, it is possible to gain insight to determine which features may be worthwhile removing. Because of the subjectivity of the ratings, this approach to removing inputs can only be considered a heuristic. However, the input/output pairs in Table 5.3 will be removed in some model structures to determine how well the models perform without them.

Table 5.2: Input delays estimated from step response.

	Valence		Arousal	
Feature	Delay	Consistency Rating	Delay	Consistency Rating
LN	1	D	1	B
Centroid	0	A	1	D
NMax	1	B	0	A
S(Z&F)	1	D	1	B
TW	1	C	1	C
SDiss(H&K)	0	A	1	C
SDiss(S)	1	D	1	B
TDiss(H&K)	0	C	0	D
TDiss(S)	0	B	0	A
CTonal	1	B	1	B
Mult	1	A	0	A
MeanCentroid	2	D	1	B
MeanRolloff	1	C	1	C
MeanFlux	0	A	1	B
StdCentroid	1	D	1	C
StdRolloff	1	D	1	D
StdFlux	1	C	1	B
BPM	2	C	1	A

Table 5.3: Input/output pairs that will be removed in some model structures.

LN	Centroid	S(Z&F)	SDiss(S)	TDiss(H&K)	MeanCentroid	StdCentroid	StdRolloff	StdRolloff
Valence	Arousal	Valence	Valence	Arousal	Valence	Valence	Arousal	Valence

The step response estimates also provide some other suggestions for model structures. TDiss(H&K) and StdCentroid have a rating of C for one output and D for the other so perhaps these inputs can be removed altogether. Also, the TDiss(S) – Valence step response appears consistent around delay 0 but not for other delays so $n_k = 0$ and $n_b = 1$ will be tried for this input/output pair. A similar argument can be made to try $n_k = 1$ and $n_b = 1$ for MeanFlux – Arousal.

5.2.2 Investigated Model Structures

This section describes all of the model structures considered in this thesis. All of the models in this section use emotional appraisals for Morning as validation data and the other five songs as estimation data.

The first model structures considered are state-space models. To estimate the best order of the state-space models, the singular values of the extended observability matrix for a 6th order system are calculated and shown in Figure 5.2. This plot suggests that a second order state-space model is likely to be the most appropriate. For comparison,

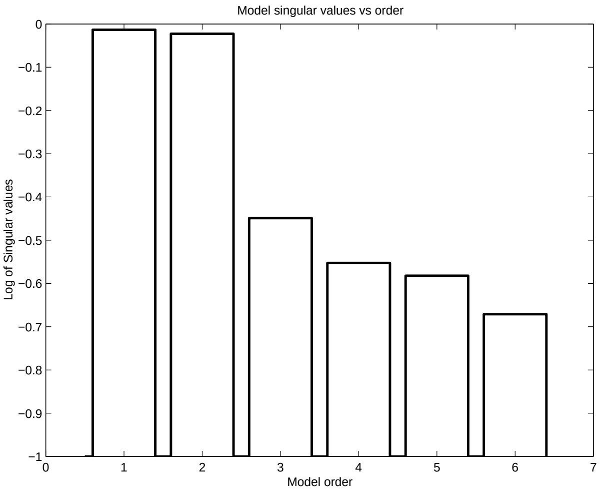


Figure 5.2: State-space model singular values vs. order

state-space models of orders 1 to 4 are created.

Each state-space model is labeled PSSn_p, where n is the order of the model and p is an index. PSSn_1 corresponds to a model where the delay is 1 for all inputs, PSSn_2 corresponds to a model where the delay is 0 for all inputs and PSSn_3 corresponds to a model where the delay is equal to the minimum value in Table 5.2 corresponding to each input.

The second model structure considered is the ARX model. Initially, the orders n_a, n_b in (4.1) are set equal to each other and vary from 1 to 4. To limit the number of parameters in the models, n_a and n_b are limited to be less than 5. Each ARX model is labeled ARX$n_a n_b$_p, where p is an index. ARX$n_a n_b$_1 corresponds to a model using delays estimated automatically from the step response using all of the inputs and ARXn_an_b_2 corresponds to a model using the delays estimated from the step response in the previous section shown in Table 5.2. The confidence intervals for ARXn_an_b_1 were found to be unacceptably large so only ARX44_1 is shown in Table 5.4 for comparison.

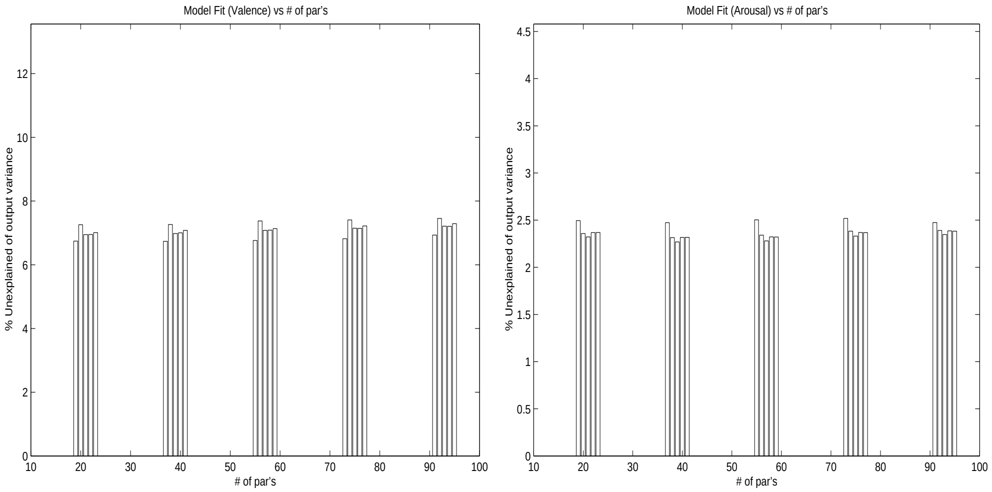
Table 5.4 summarizes the evaluation of the above state-space and ARX models. The MSE and R^2 values for valence and arousal are calculated as described in Section 4.4.3. "Conf. Int." is the maximum size of the 98% confidence interval of the simulated output. "No." list the number of parameters in the model structure and "Stat. Sig." lists the percentage of the parameters that do not include the value zero in their 98% confidence interval. See Appendix D.2 for more details about the model structures.

The crosscorrelation function and the autocorrelation function for residuals of all of the model structures will be discussed at the end of this section.

Next, MISO ARX models were analyzed to determine what values n_a and n_b should have for each of arousal and valence. 125 models were compared for each of arousal and valence, where n_a was allowed to vary between 1 and 5, n_b varied between 1 and 5 for all of the inputs and n_k varied between 0 and 4 for all of the inputs. The best fit for a model with a given number of parameters is shown in Figure 5.3. The results from this analysis can only be used heuristically because in the final models, n_b and n_k will have different values for different inputs.

Table 5.4: Summary of initial model structure comparison.

	Valence			Arousal			Parameters	
Model	*MSE*	R^2 (%)	Conf. Int.	*MSE*	R^2 (%)	Conf. Int.	No.	Stat. Sig.
PSS4_1	211	14.3	43.6	246	64.7	56.7	88	23.9%
PSS4_2	224	9.2	39.7	270	61.2	24.1	124	21.8%
PSS4_3	210	14.8	39.2	252	63.8	71.2	102	22.5%
PSS3_1	207	16.1	36.4	246	64.6	83.5	66	30.3%
PSS3_2	220	10.6	35.7	248	64.3	66.1	102	23.5%
PSS3_3	206	16.3	36.9	245	64.8	85.4	80	40.0%
PSS2_1	196	20.2	23.3	226	67.5	22.8	44	25.0%
PSS2_2	214	13.1	24.7	221	68.3	23.3	80	20.0%
PSS2_3	201	18.4	24.5	217	68.8	22.3	58	25.9%
PSS1_1	311	-26.1	**8.9**	235	66.3	17.6	**22**	40.9%
PSS1_2	265	-7.8	20.5	254	63.4	20.6	58	48.3%
PSS1_3	266	-7.9	17.9	255	63.3	17.7	36	**55.6%**
ARX44_1	220	10.6	1860	213	69.3	903	160	13.1%
ARX44_2	243	1.3	71.7	221	68.3	63.2	160	11.9%
ARX33_2	223	9.5	60.8	229	67.1	22.7	120	14.2%
ARX22_2	222	9.7	51.7	**202**	**70.9**	25.0	80	21.3%
ARX11_2	**179**	**27.2**	20.4	221	68.2	**16.3**	40	35.0%



(a) Valence

(b) Arousal

Figure 5.3: Goodness of fit vs. total number of parameters implied by n_a, n_b and n_k for MISO models.

In Figure 5.3(a) each group corresponds to the order of n_b and within each group, n_a increases from 1 to 5. Increasing n_b does not seem to improve the fit since each group has approximately the same fit. Therefore, n_b should be as small as possible for valence. The fit appears best when n_a is 1 or possibly 3. Therefore, n_a will be allowed to equal 1 or 3 for valence. Although not evident from this figure, this test implies that n_k should equal two or three but this conflicts with the delays estimated from the step responses. Therefore, n_k will be allowed to vary between 1 and 3.

Figure 5.3(b) illustrates the model fit for the arousal component and is organized identically to Figure 5.3(a). Again, increasing n_b does not seem to improve the fit so n_b should be as small as possible. The fit appears to be the best when n_a is 3 so values of n_a between 1 and 3 will be considered. Also, it appears that n_k should 0 or 1.

From the MISO model analysis, MIMO ARX models are created with diagonal matrices A_1, A_2 andA_3. The ARX models are labeled ARXAn_{a_a}Vn_{a_v}_p[2], where n_{a_a} is the order of n_a for arousal, n_{a_v} is the order of n_a for valence and p is an index. When $p = 1, 2, 3$, n_k equals 1 for all arousal inputs and equals p for all valence inputs. ARXAn_{a_a}Vn_{a_v}_4 uses the estimated delays from the step response for n_k. Table 5.5 summarizes the evaluation of these models and is structured in the same manner as Table 5.4.

Because the fit for valence in the ARXAn_{a_a}Vn_{a_v}_p models is poor, model structures with different combinations of orders and delays were iteratively estimated and evaluated. Using the estimated delays from Table 5.2 in the ARX models results in the best fit and smallest output confidence intervals so these delays will be used in other models. Of the models investigated so far, ARX11_2 has the most significant parameters, the largest R^2 values and smallest confidence intervals so it will be used as a starting point.

The next model structure considered is ARX11S_1, which is the same model structure as ARX11_2 but without inputs TDiss(H&K), StdCentroid and StdRolloff. Similarly, the models labelled ARXn1S_1 are the same as ARX11S_1 but n_a is a matrix where every element equals n. After evaluating these models, other model structures were constructed

[2] See Appendix D.2.2 for a detailed description of the models.

Table 5.5: Summary of investigative model structure comparison.

	Valence			Arousal			Parameters	
Model	MSE	R^2	Conf. Int.	MSE	R^2	Conf. Int.	No.	Stat. Sig.
ARXA3V1_1	223	9.2	119.3	221	68.2	14900	40	32.5%
ARXA3V1_2	257	-4.4	22.2	221	68.2	23.1	40	32.5%
ARXA3V1_3	280	-13.7	27700	233	66.5	20.9	76	15.8%
ARXA3V1_4	227	7.7	20.7	215	69.1	24.3	76	19.7%
ARXA3V3_1	266	-8.0	57.5	221	68.2	55.2	78	16.7%
ARXA3V3_2	239	2.9	37.7	221	68.2	61.0	78	17.9%
ARXA3V3_3	263	-7.0	5830	233	66.5	20.9	78	16.7%
ARXA3V3_4	249	-1.1	58.2	214	69.2	24.3	78	17.9%
ARX11S_1	185	24.8	**19.9**	227	67.3	15.9	34	44.1%
ARX21S_1	198	19.7	24.6	235	66.2	19.6	38	42.1%
ARX31S_1	212	13.8	25.0	247	64.5	21.9	39	33.3%
ARX11S_2	182	25.9	**19.9**	232	66.7	15.8	33	**48.5%**
ARX21S_2	197	19.8	24.7	236	66.1	19.4	36	44.4%
ARX31S_2	212	13.8	25.0	247	64.5	21.9	39	33.3%
ARX11S_3	191	22.4	20.0	219	68.5	16.0	37	37.8%
ARX11S_4	**179**	**27.5**	20.0	216	68.9	15.0	33	45.5%
ARX11S_5	221	10.1	20.3	207	70.3	16.3	49	36.7%
ARX11S_6	194	21.3	20.9	198	71.6	14.9	49	30.6%
ARX11S_7	214	13.0	25100	203	70.8	15100	31	32.3%
ARX11S_8	214	12.9	160.1	**194**	**72.1**	139.1	**27**	40.7%
ARX11S_9	189	23.4	**19.9**	202	71.0	**14.4**	41	41.5%

by removing parameters that were not statistically significant from zero and adding parameters to investigate increasing the order for particular inputs. For a detailed description about how models ARXn1S_p were constructed, consult Appendix D.2. Table 5.5 summarizes the results of the evaluation.

The crosscorrelation function for residuals of all of the model structures appeared to be statistically equivalent to zero for the majority of the lags. All of the significant lags in the crosscorrelation function are assumed to be due to chance. The autocorrelation function for residuals of the first order state-space models are statisically significant from zero for all lags so these models are not considered valid. All of the models appear to have a spurious autocorrelation at lag 10 in the residuals of the arousal component. This also occurred in Schubert's analysis and is assumed to occur due to chance[24]. The ACF of the residuals of some of the ARX11S_p models also appear to be marginally significant at lag 1, but this is assumed to be due to chance.

All of the model structures had many parameters that statistically are equivalent to zero. However, most of the models had relatively small output confidence intervals (i.e. most are less than 30) so the variance error is considered to be acceptable in these models.

5.2.3 Best Model Structures

The best model structures from the previous section are subject to further validation. Each of these models is six-fold cross-validated to calculate an average R^2 fit as described by (4.19). Also, all six songs are used as estimation data to calculate Akaike's FPE. Table 5.6 summarizes the results of this validation.

During six-fold cross-validation, there was a poor fit for Allegro – Valence, Aranjuez – Valence and Fanfare – Arousal. Perhaps these songs were too different from the songs in the training set to effectively generalize from them, or perhaps the representative emotional appraisal is poor. The second order ARX models had large confidence intervals when Aranjuez was not included in the training data.

The model structure that had highest R^2 values for six-fold cross-validated data was ARX11S_6. However, as shown in Table D.2, the residuals for 67% of the emotional appraisals are autocorrelated at lag 1. For this reason, it is not considered to be an acceptable model structure.

Model ARX21S_2 is considered to be the best linear model structure because it had the lowest FPE when using all six songs as estimation data. It has fewer parameters and has average R^2 values comparable to the ARX11S_6 model structure. The residuals on the validation data were only autocorrelated for songs that had generally poor fit. This model structure did have larger output confidence intervals than ARX11S_6, but were still

Table 5.6: Summary of best model structures

	Valence R^2 (%)						
Model	Alle	Aran	Fanf	Moon	Morn	Pizz	Avg.
PSS2_1	-7.1	-75.74	24.7	21.2	20.2	60.5	6.1
ARX11_2	-10.1	-164.1	33.7	30.0	27.2	65.5	2.6
ARX11S_1	-10.4	-145.3	42.3	27.1	24.8	65.0	6.6
ARX11S_2	-11.7	-142.3	42.6	26.3	25.9	63.3	6.4
ARX11S_3	-14.7	-139.3	43.1	24.6	22.4	56.2	4.9
ARX11S_4	-11.4	-143.4	45.1	25.7	27.5	66.0	7.7
ARX11S_6	2.4	-172.2	51.7	21.9	21.3	52.5	**11.4**
ARX11S_9	-0.9	-126.6	43.5	24.3	23.4	62.5	10.6
ARX21S_1	-0.9	-159.0	36.4	33.1	19.7	74.6	7.1
ARX21S_2	-0.9	-159.4	38.1	32.6	19.8	74.0	7.8

	Arousal R^2 (%)							Akaike's
Model	Alle	Aran	Fanf	Moon	Morn	Pizz	Avg.	FPE
PSS2_1	75.7	85.4	-243.7	7.2	67.5	59.6	66.5	142.0
ARX11_2	82.2	86.4	-298.2	6.3	68.2	68.1	68.8	158.1
ARX11S_1	81.4	88.8	-303.8	16.0	67.3	65.3	69.0	156.4
ARX11S_2	81.8	89.9	-198.1	22.4	66.7	66.1	72.2	156.8
ARX11S_3	82.3	90.9	-194.1	21.1	68.5	56.9	72.7	157.2
ARX11S_4	82.5	83.7	-106.8	14.7	68.9	68.2	73.2	157.1
ARX11S_6	84.4	91.6	-142.2	24.1	71.6	70.1	**76.2**	151.2
ARX11S_9	83.4	91.9	-137.1	25.5	71.0	69.8	76.0	152.1
ARX21S_1	86.0	90.1	-252.7	24.2	66.2	65.7	72.8	134.8
ARX21S_2	86.5	90.5	-169.8	25.6	66.1	65.5	75.1	**134.2**

reasonable (the confidence intervals will decrease as more data is included in the training set). See Appendix C for graphs showing the simulated outputs for ARX21S_2.

The ARX21S_2 model structure uses 15 of the musical features, 36 parameters and is described as follows:

$$\underline{y}(t)+A_1(\underline{\theta})\underline{y}(t-1)+A_2(\underline{\theta})\underline{y}(t-2) = B_0(\underline{\theta})\underline{u}(t)+B_1(\underline{\theta})\underline{u}(t-1)+B_2(\underline{\theta})\underline{u}(t-2)+\underline{e}(t) \quad (5.2)$$

$$\hat{\underline{y}}(t|\underline{\theta}) = [I + A_1(\underline{\theta})q^{-1} + A_2(\underline{\theta})q^{-2}]^{-1}[B_0(\underline{\theta}) + B_1(\underline{\theta})q^{-1} + B_2(\underline{\theta})q^{-2}]\underline{u}(t) \quad (5.3)$$

$$A_1(\underline{\theta}) = \begin{bmatrix} \theta_1 & \theta_2 \\ 0 & \theta_3 \end{bmatrix} \quad (5.4)$$

$$A_2(\underline{\theta}) = \begin{bmatrix} \theta_4 & \theta_5 \\ 0 & \theta_6 \end{bmatrix} \quad (5.5)$$

$$B_0(\underline{\theta}) = \begin{bmatrix} 0 & 0 & \theta_7 & 0 & 0 & 0 & 0 & \theta_8 & 0 & \theta_9 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{10} & 0 & 0 & 0 & \theta_{11} & 0 & \theta_{12} & 0 & 0 & 0 & 0 & \theta_{13} & 0 & 0 \end{bmatrix} \quad (5.6)$$

$$B_1(\underline{\theta}) = \begin{bmatrix} \theta_{14} & \theta_{15} & 0 & \theta_{16} & \theta_{17} & \theta_{18} & \theta_{19} & 0 & \theta_{20} & 0 & \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} & \theta_{25} \\ \theta_{26} & 0 & \theta_{27} & \theta_{28} & \theta_{29} & 0 & \theta_{30} & 0 & \theta_{31} & \theta_{32} & 0 & \theta_{33} & 0 & \theta_{34} & 0 \end{bmatrix} \quad (5.7)$$

$$B_2(\underline{\theta}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{35} & 0 & 0 & 0 & \theta_{36} \end{bmatrix} \quad (5.8)$$

where,

$\underline{y}(t)$ is a vector consisting of valence and arousal

$\underline{u}(t)$ is a vector consisting of the following features measured at time t: LN, Centroid, NMax, S(Z&F), TW, SDiss(H&K), SDiss(S), TDiss(S), CTonal, Mult, MeanCentroid, MeanRolloff, MeanFlux, StdFlux, BPM

Chapter 6

Conclusions

6.1 Summary

In this thesis, four criteria needed to be satisfied for a model of emotional appraisals of music to be valid. Through model construction, the first three criteria are met: the measured emotional appraisals of the listeners are time-varying, the musical features used in the model are time-varying and represent musical properties that communicate emotion, and the model is estimated using emotional appraisals to musical selections representing a genre of music.

To satisfy the fourth criterion, a model needs to accurately simulate emotional appraisals to any musical selection from the genre of music. Because the average R^2 statistic of the best model structure is 7.8% for valence and 75.1% for arousal, this criterion is met for arousal appraisals but not for valence appraisals. Although the models in this study did not completely satisfy this criterion, there is potential to improve the R^2 statistic for valence by considering other model structures and applying system identification techniques. Therefore, system identification provides a means to create a valid model of emotional appraisals of music.

6.2 Comparison with Other Research

It is difficult to directly compare the work in this thesis with other research because other models for time-varying emotional appraisals do not generalize to multiple songs. Comparing the R^2 values of models for individual songs with the R^2 values of the models in this study is possible, but it is expected that the models in this study fit more poorly than the models for individual songs.

In the study by Schubert, time series models of emotional appraisals were created for Pizzicato and longer versions of Morning and Aranjuez[24]. The R^2 values for the songs modeled in both of these studies are shown in Table 6.1 for comparison.

According to Table 6.1, it appears that models created in this thesis of emotional appraisals for Pizzicato are improvements over the models by Schubert. Improvements are

Table 6.1: Comparison with Schubert's models[24].

	Valence R^2 (%)		Arousal R^2 (%)	
Song	Schubert	Korhonen	Schubert	Korhonen
Pizzicato	38	74	36	65
Morning	40	20	67	66
Aranjuez	33	-159	57	90

probably due to the inclusion of thirteen more features in this study than in Schubert's study[24]. Similarly, the models for Morning – Arousal and Aranjuez – Arousal appear to be equal to or better than Schubert's models. Therefore, it appears that these models for the genre of classical music perform similarly to the models created by Schubert for individual songs.

The models of Morning – Valence and Aranjuez – Valence in this study have lower R^2 values than Schubert's models. There are several possible reasons for these lower values. First, shorter versions of these songs are used in this study so the R^2 values can only be used subjectively. Second, the R^2 values in this study are calculated using the testing set and the R^2 values for Schubert's models are calculated using the training set so lower R^2 values are expected in this study. Third, the R^2 values in this study are calculated using data filtered differently than in Schubert's study so different frequencies of the emotional appraisals are emphasized in this study. Because of the differences in these studies, definite conclusions about the model fit cannot be made by comparing the R^2 values.

However, despite the differences in the two studies, one can conclude that principles of system identification afford mathematical models of continuous emotional appraisals that generalize to a genre of music. By applying the systematic method used in system identification for designing experiments, selecting model structures and validating the models, valid models can be constructed to lead to an improved understanding of how musical features cause emotions to be perceived.

6.3 Future Work

The results from this thesis bring up two important issues. First, a method for improving the simulations of Allegro – Valence, Aranjuez – Valence and Fanfare – Arousal should be investigated. Second, a model structure should be found where most of the parameters are significantly different from zero while still having a small MSE and output confidence intervals. Overcoming these two issues should improve the validity of the models.

The remaining suggestions for future enhancement of the models can be divided into categories. First, more variables can be incorporated in the models. Second, other algorithms to preprocess data can be considered. Third, alternative model structures should be investigated. Fourth, several aspects of the methodology can be enhanced. Fifth, applications of the models should be investigated. Finally, possible alternatives to system identification can be considered.

Additional variables could be incorporated in the models through the use of parameters. For example, it is possible that emotional appraisals made by a person could be affected by factors such as their musical training, familiarity with the musical selections, mood or culture. Also, other musical features representing properties, such as rhythm or tempo variance, could be incorporated into the models. Finally, feature extraction methods other than those described in this thesis could be used.

There are three other algorithms that may be used to preprocess the data. First, people may respond to the same musical stimuli at different times, don't respond to certain stimuli or respond with different amplitudes. One possible method to overcome this problem is to normalize, rescale, and filter the emotional appraisals and then perform a time-alignment algorithm such as dynamic programming. Second, instead of trying to generate an emotional appraisal representative of the population, it is possible to create a model for each person and then analyze the parameter vectors. This approach may provide insight into whether a population can be represented by one model, or whether it needs to be modeled by several. Third, designing an appropriate low-pass filter to remove high-frequency disturbances should be investigated.

Several different model structures could be considered in addition to the ones used in this thesis. Other second and third order ARX model structures may result in a better fit than the model structures found in this study. Other linear model structures such as the ARMAX or Box-Jenkins models, or nonlinear model structures such as the two-layer artificial neural network could improve the fit of the simulated appraisals[15][18].

There are many ways that the methodology can be enhanced. First, incorporating nonlinear transformations of the features may improve the fit of linear model structures[16]. Second, features such as tempo and pitch variation could be measured more accurately to improve the SNR of the inputs. Third, an alternative interface to the 2DES in *EmotionSpace Lab* could be investigated to see if the noise in the emotional appraisals can be reduced. Fourth, the emotional appraisals should be sampled more frequently as it is straightforward to resample back to 1Hz if desired. Fifth, the validation routine used to evaluate all of the initial models should be improved to avoid arbitrary selection of estimation data without increasing the computation time. Sixth, more songs, and/or different genres of music should be included in the training set. Finally, it may be desirable to incorporate stochastic models of music into the models.

The resultant models could be used for further analysis. From the parameterization of the autoregressive components of the best models, it appears that only 2-3 seconds of musical stimulus is needed to perceive emotion in music, and that arousal is not a function of valence but valence is a function of arousal. This claim should be investigated further. Another area for investigation could be to investigate the significance of the state vector in the state-space models; one could investigate if the state variables have any neurological, physiological or other meaning. Also, the models could be used to determine which features cause people to perceive emotion and how they do so.

Finally, an alternative approach to modeling emotional appraisals could be considered. For example, a state machine (i.e. Markov model) may be able to model the noise in the emotional appraisal data.

References

[1] G. H. Bower. Mood and memory. *American Psychologist*, 36:129–148, 1981.

[2] Densil Cabrera. *PsySound2: Psychoacoustical Software for Macintosh PPC*, July 2000.

[3] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *Speech and Emotion, ISCA Tutorial and Research Workshop (ITRW)*, pages 19–24, Newcastle, Northern Ireland, UK, September 2000.

[4] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification.* John Wiley & Sons, Inc., 2nd edition, 2001.

[5] P. Ekman and W.V. Friesen. What emotion categories or dimensions can observers judge from facial behaviour? In P. Ekman, editor, *Emotions in the Human Face.* London: Cambridge University Press, 1982.

[6] K. W. Fischer, P. R. Shaver, and P. Carnochan. How emotions develop and how they organise development. *Cognition and Emotion*, 4:81–127, 1990.

[7] Alf Gabrielsson. Perceived emotion and felt emotion: Same or different? *Musicae Scientiae*, Special Issue 2001-2002, 2002. ISSN 1029-8649.

[8] Kate Hevner. Experimental studies of the elements of expression in music. *American Journal of Psychology*, 48:246–268, 1936.

[9] Gernot Horstmann. What do facial expressions convey: Feeling states, behavioral intentions, or action requests. *Emotion*, 3(2):150–166, 2003.

[10] Peter J. Huber. *Robust Statistics*. John Wiley & Sons, Inc., 1981.

[11] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, Inc., 2001.

[12] Patrik N. Juslin, Anders Friberg, and Roberto Bresin. Toward a computational model of expression in music performance: The GERM model. *Musicae Scientiae*, Special Issue 2001-2002, 2002. ISSN 1029-8649.

[13] Matthew Montague Lavy. *Emotion and the Experience of Listening to Music: A Framework for Empirical Research*. PhD thesis, Jesus College, Cambridge, 2001.

[14] T. Li and M. Ogihara. Detecting emotion in music. In *Proceedings of the Fifth International Symposium on Music Information Retrieval*, pages 239–240, 2003.

[15] Lennart Ljung. *System Identification: Theory for the User.* Prentice-Hall, Inc., 2nd edition, 1999.

[16] Lennart Ljung. *System Identification Toolbox: User's Guide.* The Mathworks Inc., 3 Apple Hill Drive, Natick, MA 0160-2098, 5th edition, March 2001.

[17] C. K. Madsen, R. V. Brittin, and D. A. Capperella-Sheldon. An empirical investigation of the aesthetic response to music. *Journal of Research in Music Education*, 41:57–69, 1993.

[18] M. Nørgaard, O. Ravn, N. K. Poulsen, and L.K. Hansen. *Neural Networks for Modelling and Control of Dynamic Systems.* Springer-Verlag, London, UK, 2000.

[19] Alan V. Oppenheim, Alan S. Willsky, and S. Hamid Nawab. *Signals and Systems.* Prentice Hall, 2nd edition, 1997.

[20] Boaz Porat. *A Course in Digital Signal Processing.* John Wiley & Sons, Inc., 1997.

[21] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[22] J. A. Russell. Measures of emotion. In R. Plutchik and H. Kellerman, editors, *Emotion: Theory Research and Experience*, volume 4, pages 81–111. New York: Academic Press, 1989.

[23] J. A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 1977.

[24] Emery Schubert. *Measurement and Time Series Analysis of Emotion in Music*. PhD thesis, University of New South Wales, 1999.

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

[26] J. A. Sloboda. Empirical studies of emotional response to music. In M. Jones and S. Holleran, editors, *Cognitive Bases of Musical Communication*. Washington: American Psychological Association, 1992.

[27] Kenji Suzuki and Shuji Hashimoto. Modeling of emotional sound space using neural networks. In *AIMI Intl. Workshop on Kansei - Technology of Emotion*, pages 116–121, Italy, October 1997.

[28] B. Tillman and E. Bigand. Does formal musical structure affect perceptions of musical expression? *Psychology of Music*, 24:3–17, 1996.

[29] G. Tzanetakis and P. Cook. MARSYAS: A framework for audio analysis. *Organized Sound*, 4(3):169–175, 2000.

[30] L. Wedin. Multidimensional study of perceptual-emotional qualities in music. *Scandinavian Journal of Psychology*, 13:241–257, 1972.

Appendix A

Study

A.1 Questionnaire

This questionnaire consists of questions identical to those asked in Schubert's study for ease of comparison[24]. These questions are asked to determine if this study is performed by people from many different age groups and many different musical backgrounds. If any person did not feel comfortable answering any of these questions, they did not need to answer them. The date and time of the study was collected with the questionnaire to identify each person without recording their name.

Each participant was asked to circle their answer to the following questions:

1. Gender:

Male   Female

2. What age group are you in?

15 – 19   20 – 24   25 – 29   30 – 34   35 – 39

40 – 44   45 – 49   50 – 54   55 – 59   60 – 64   65+

3. How many years of training do you have on a musical instrument (or in singing)?

0 – 1 years   1 – 2 years   2 – 5 years   5 – 10 years   10+ years

4. How much exposure do you have to Western instrumental art/classical music?

almost none   a little   some   a fair bit   a lot   constant

5. How much do you enjoy listening to Western instrumental art/classical music?

don't like it   so-so   like it a bit   like it a lot   love it

Appendix B

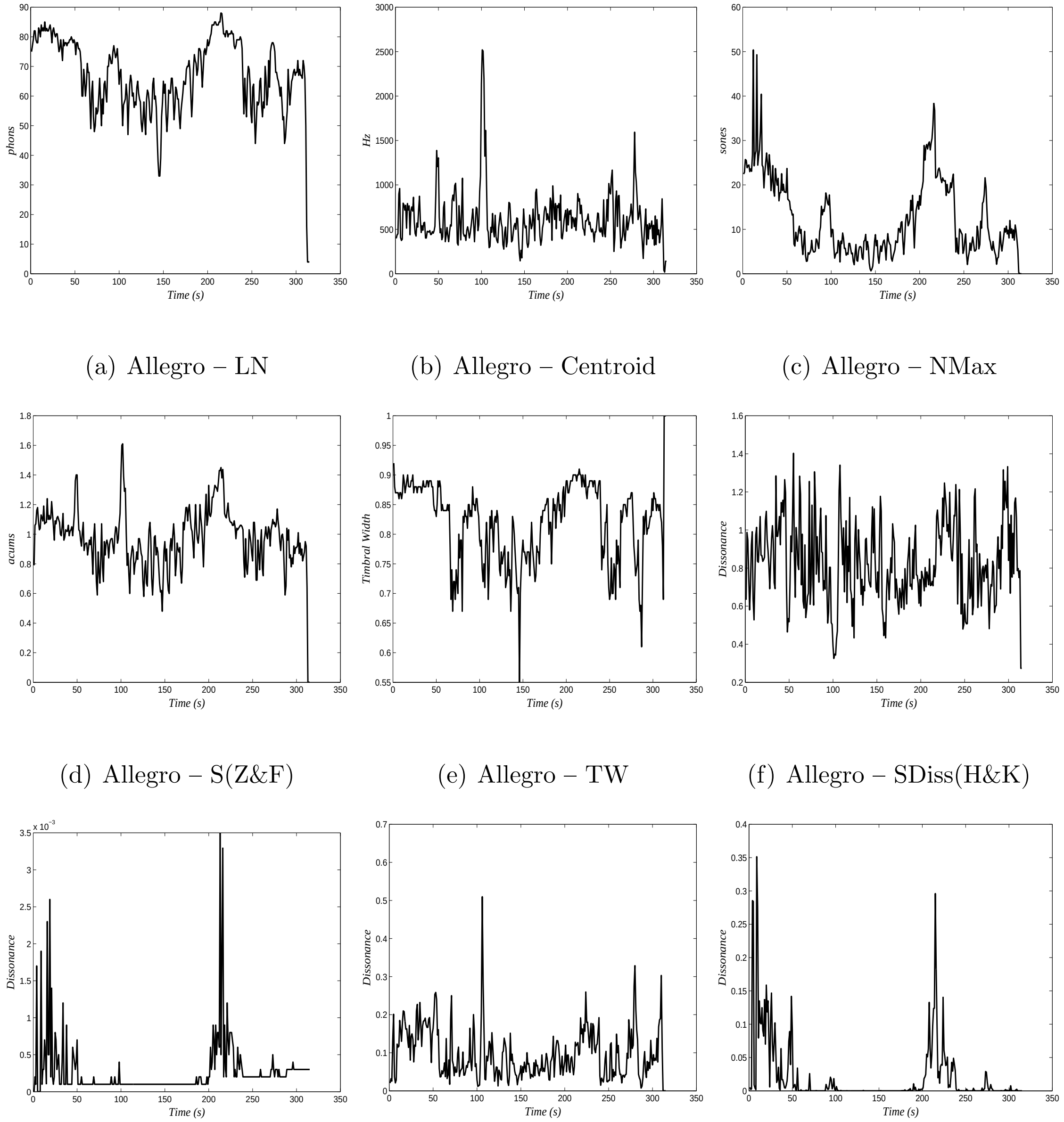
Raw Data

B.1 Musical Features

The following figures show the musical features calculated for the six songs. See Table 4.1 for information describing what musical selection to which the aliases refer and Table 4.2 for information about what musical property each feature represents.

Each graph represents $u_{i,k}(t)$, the measurement of feature number k for song i as a function of time. All of the features for song i can be combined as follows to create $\underline{u}_i(t)$:

$$\underline{u}_i(t) = \begin{bmatrix} u_{i,1}(t) \\ \vdots \\ u_{i,18}(t) \end{bmatrix} \tag{B.1}$$



(a) Allegro – LN

(b) Allegro – Centroid

(c) Allegro – NMax

(d) Allegro – S(Z&F)

(e) Allegro – TW

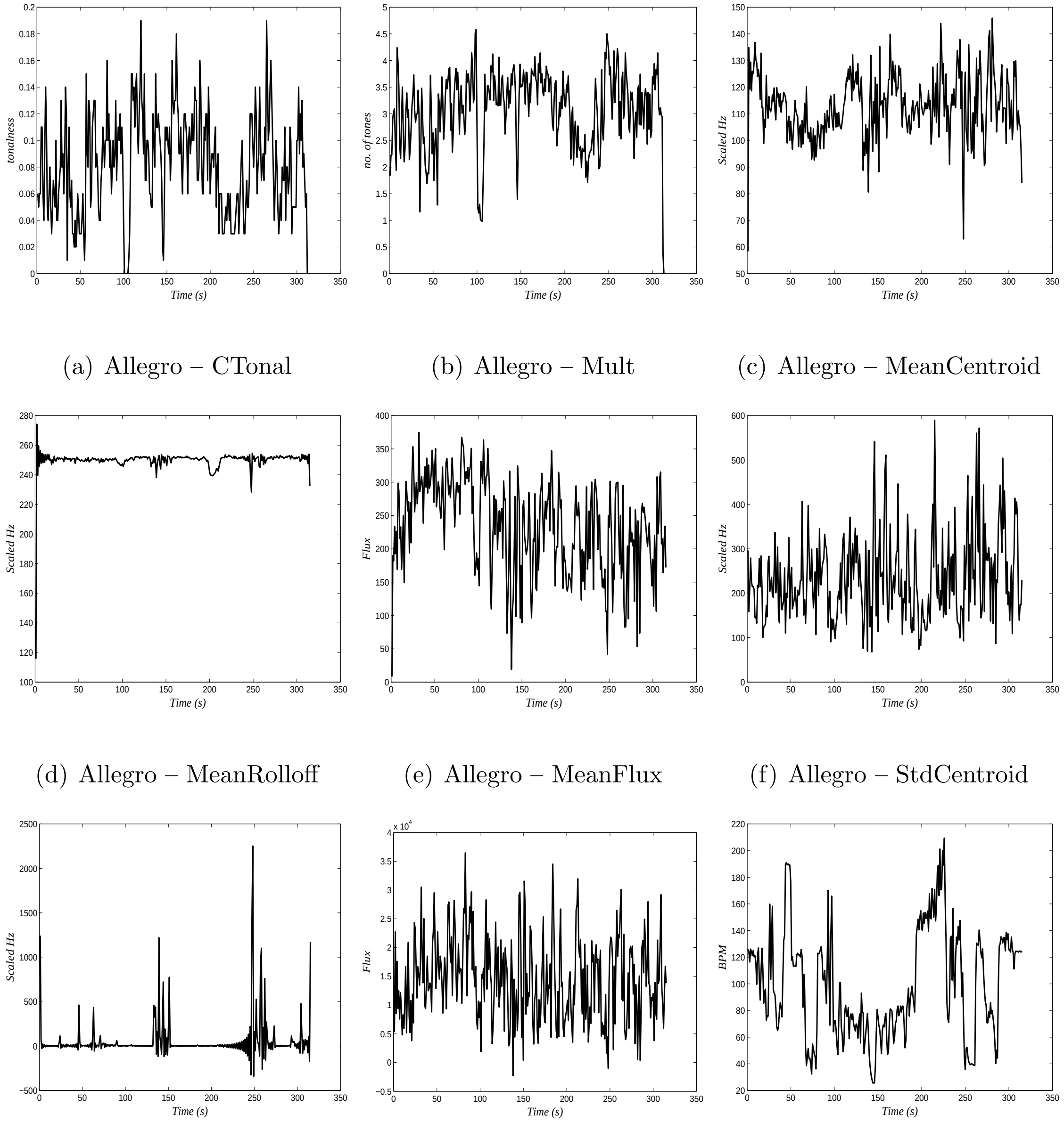
(f) Allegro – SDiss(H&K)

(g) Allegro – SDiss(S)

(h) Allegro – TDiss(H&K)

(i) Allegro – TDiss(S)

Figure B.1: The first nine features of Allegro.



(a) Allegro – CTonal

(b) Allegro – Mult

(c) Allegro – MeanCentroid

(d) Allegro – MeanRolloff

(e) Allegro – MeanFlux

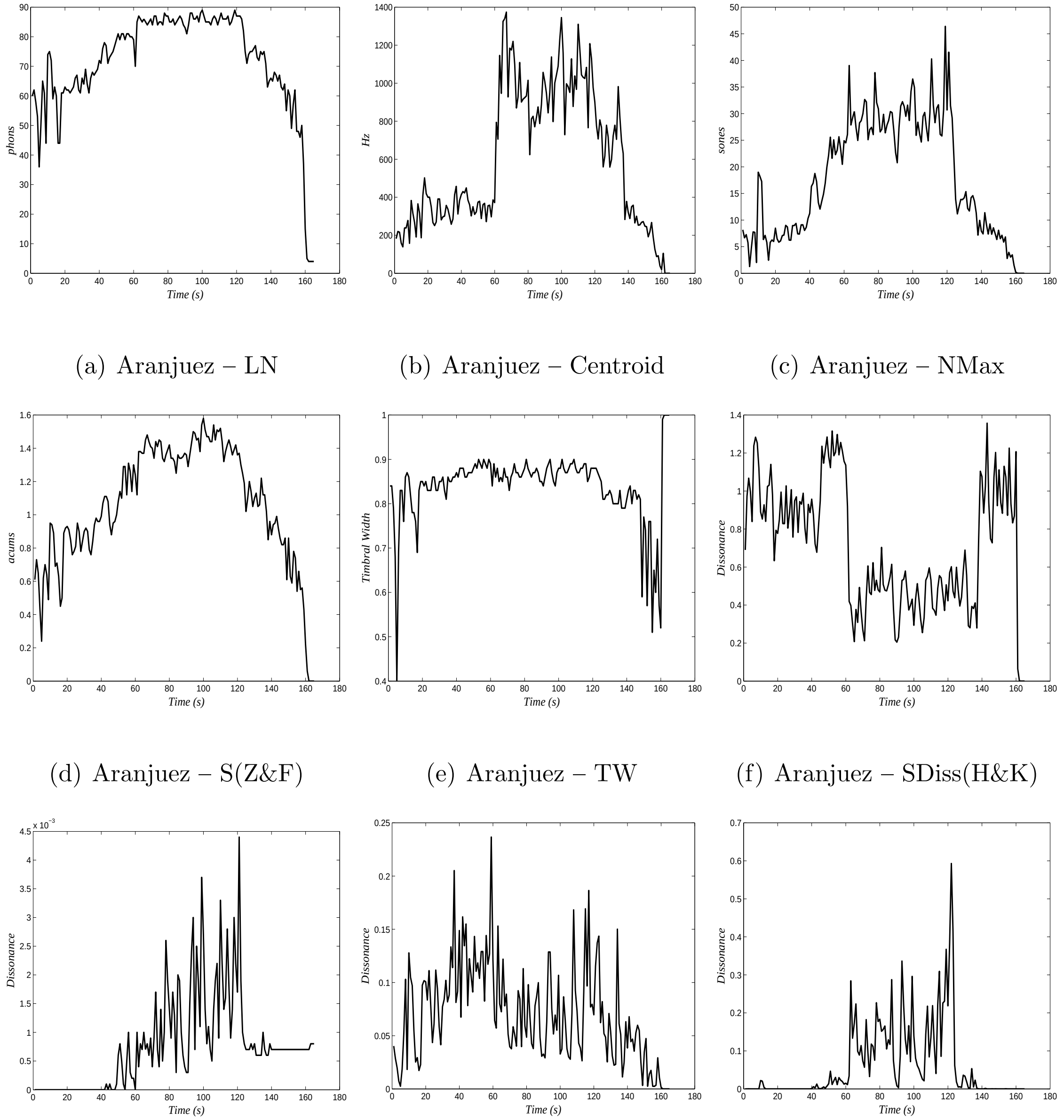
(f) Allegro – StdCentroid

(g) Allegro – StdRolloff

(h) Allegro – StdFlux

(i) Allegro – BPM

Figure B.2: The last nine features of Allegro.



(a) Aranjuez – LN

(b) Aranjuez – Centroid

(c) Aranjuez – NMax

(d) Aranjuez – S(Z&F)

(e) Aranjuez – TW

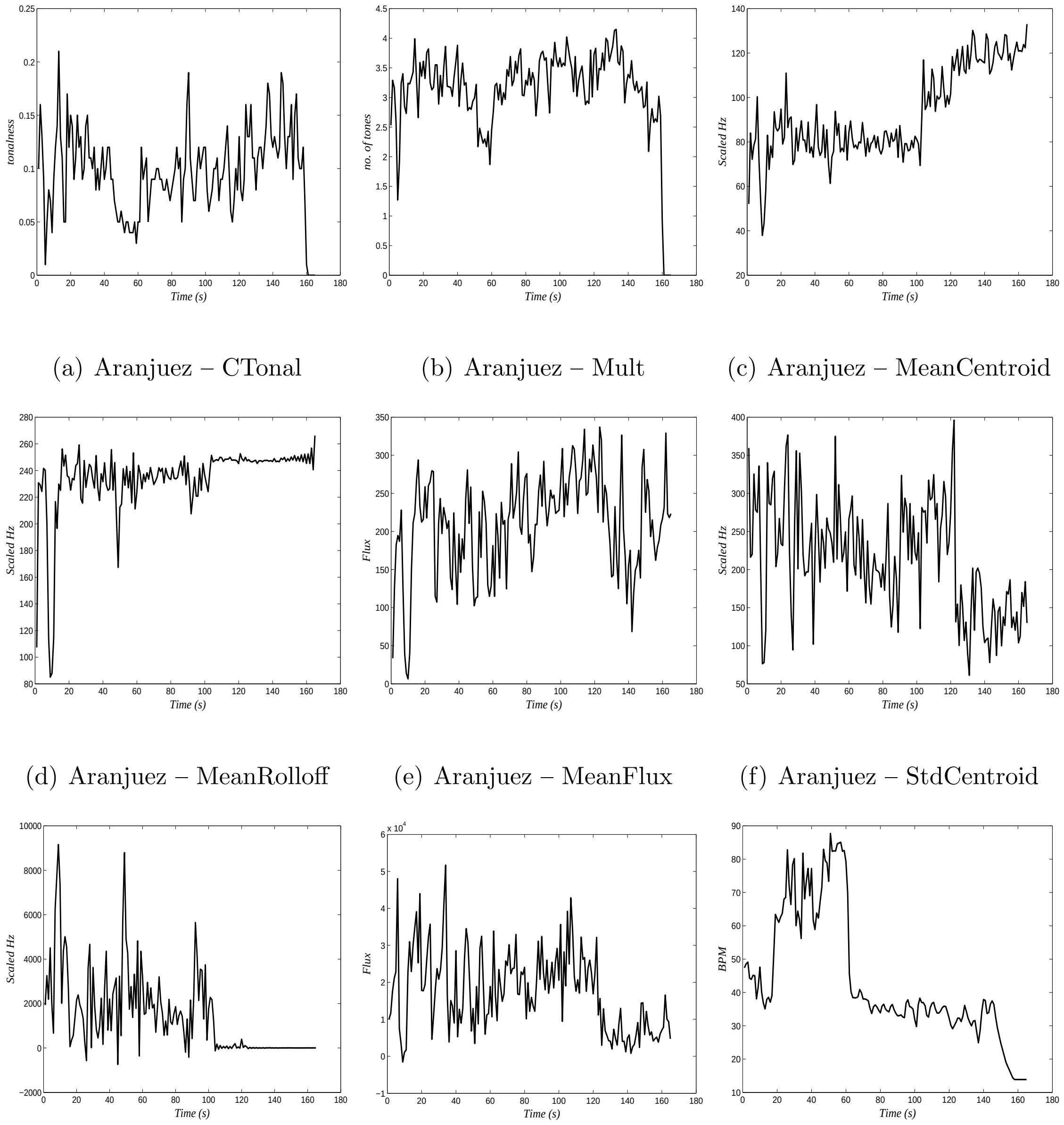
(f) Aranjuez – SDiss(H&K)

(g) Aranjuez – SDiss(S)

(h) Aranjuez – TDiss(H&K)

(i) Aranjuez – TDiss(S)

Figure B.3: The first nine features of Aranjuez.



(a) Aranjuez – CTonal

(b) Aranjuez – Mult

(c) Aranjuez – MeanCentroid

(d) Aranjuez – MeanRolloff

(e) Aranjuez – MeanFlux

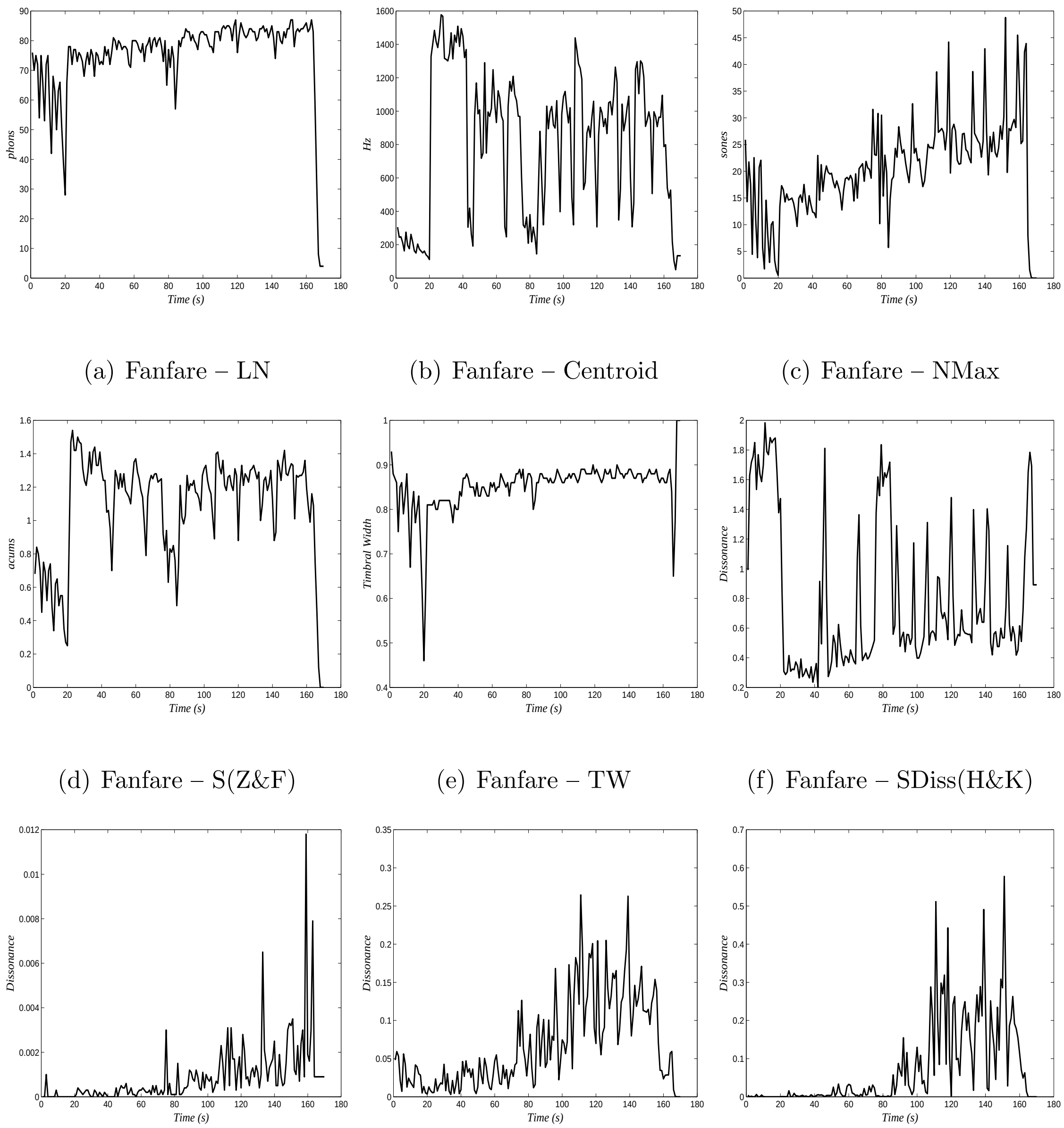
(f) Aranjuez – StdCentroid

(g) Aranjuez – StdRolloff

(h) Aranjuez – StdFlux

(i) Aranjuez – BPM

Figure B.4: The last nine features of Aranjuez.



(a) Fanfare – LN

(b) Fanfare – Centroid

(c) Fanfare – NMax

(d) Fanfare – S(Z&F)

(e) Fanfare – TW

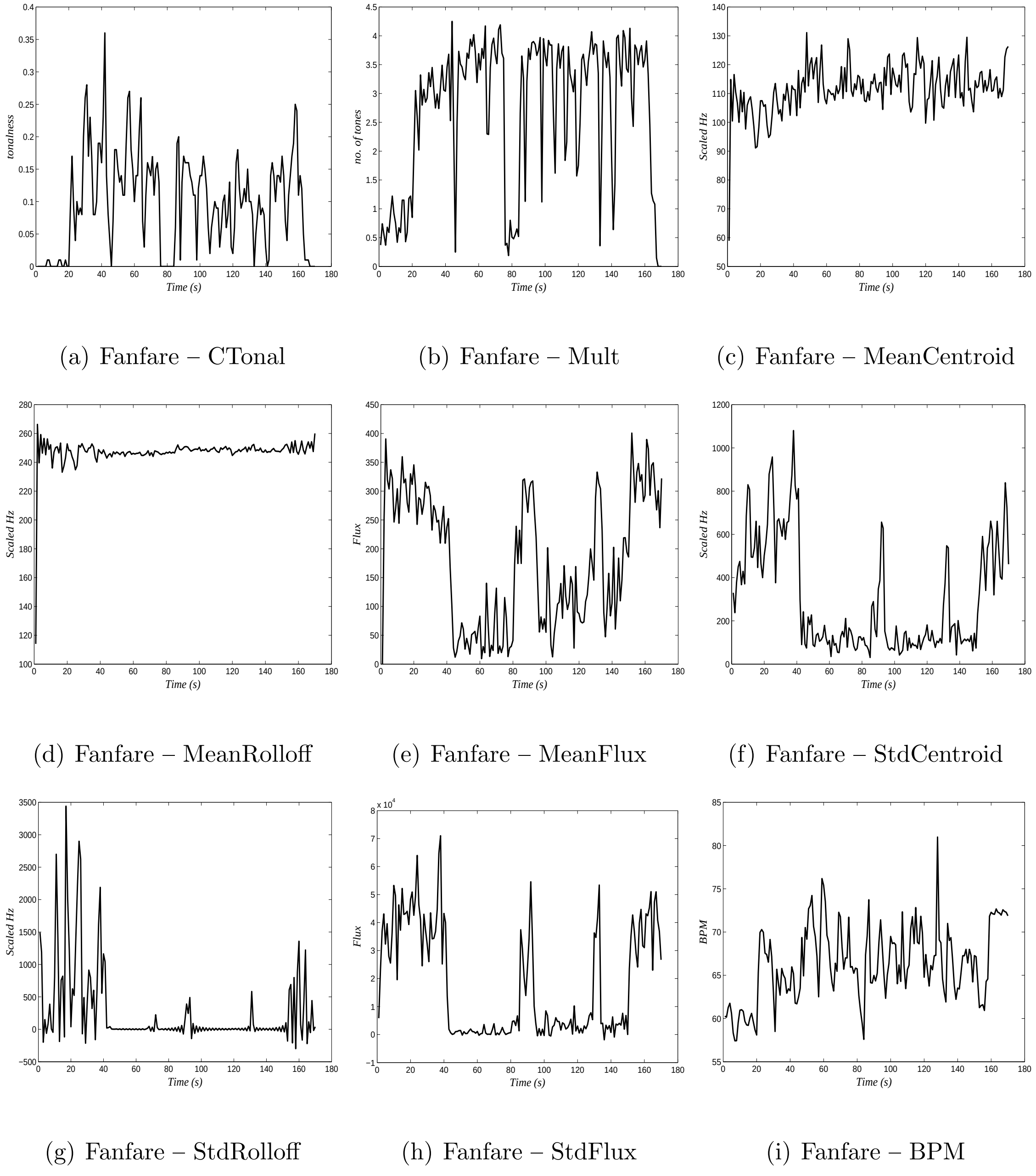
(f) Fanfare – SDiss(H&K)

(g) Fanfare – SDiss(S)

(h) Fanfare – TDiss(H&K)

(i) Fanfare – TDiss(S)

Figure B.5: The first nine features of Fanfare.



(a) Fanfare – CTonal

(b) Fanfare – Mult

(c) Fanfare – MeanCentroid

(d) Fanfare – MeanRolloff

(e) Fanfare – MeanFlux

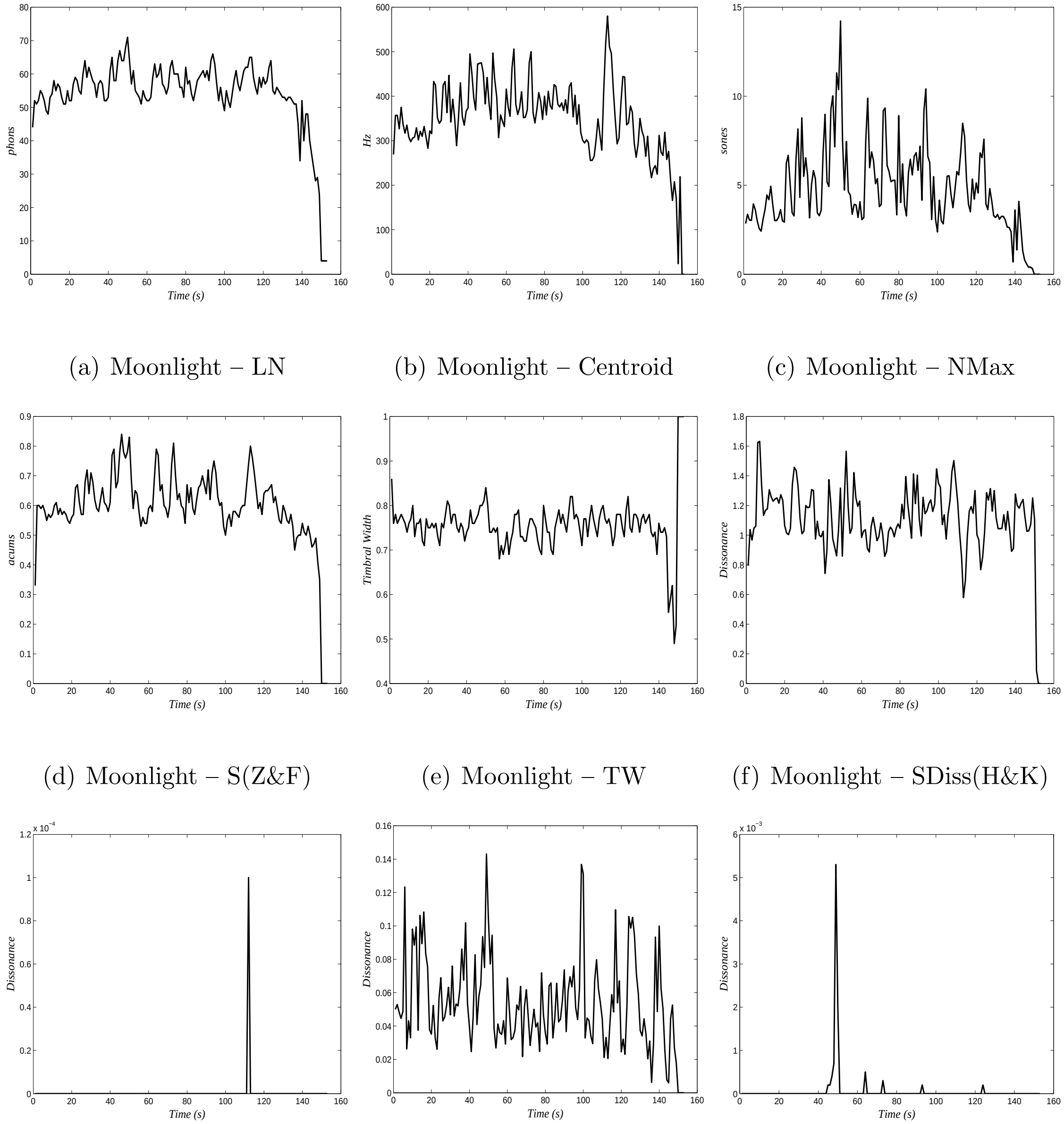
(f) Fanfare – StdCentroid

(g) Fanfare – StdRolloff

(h) Fanfare – StdFlux

(i) Fanfare – BPM

Figure B.6: The last nine features of Fanfare.



(a) Moonlight – LN

(b) Moonlight – Centroid

(c) Moonlight – NMax

(d) Moonlight – S(Z&F)

(e) Moonlight – TW

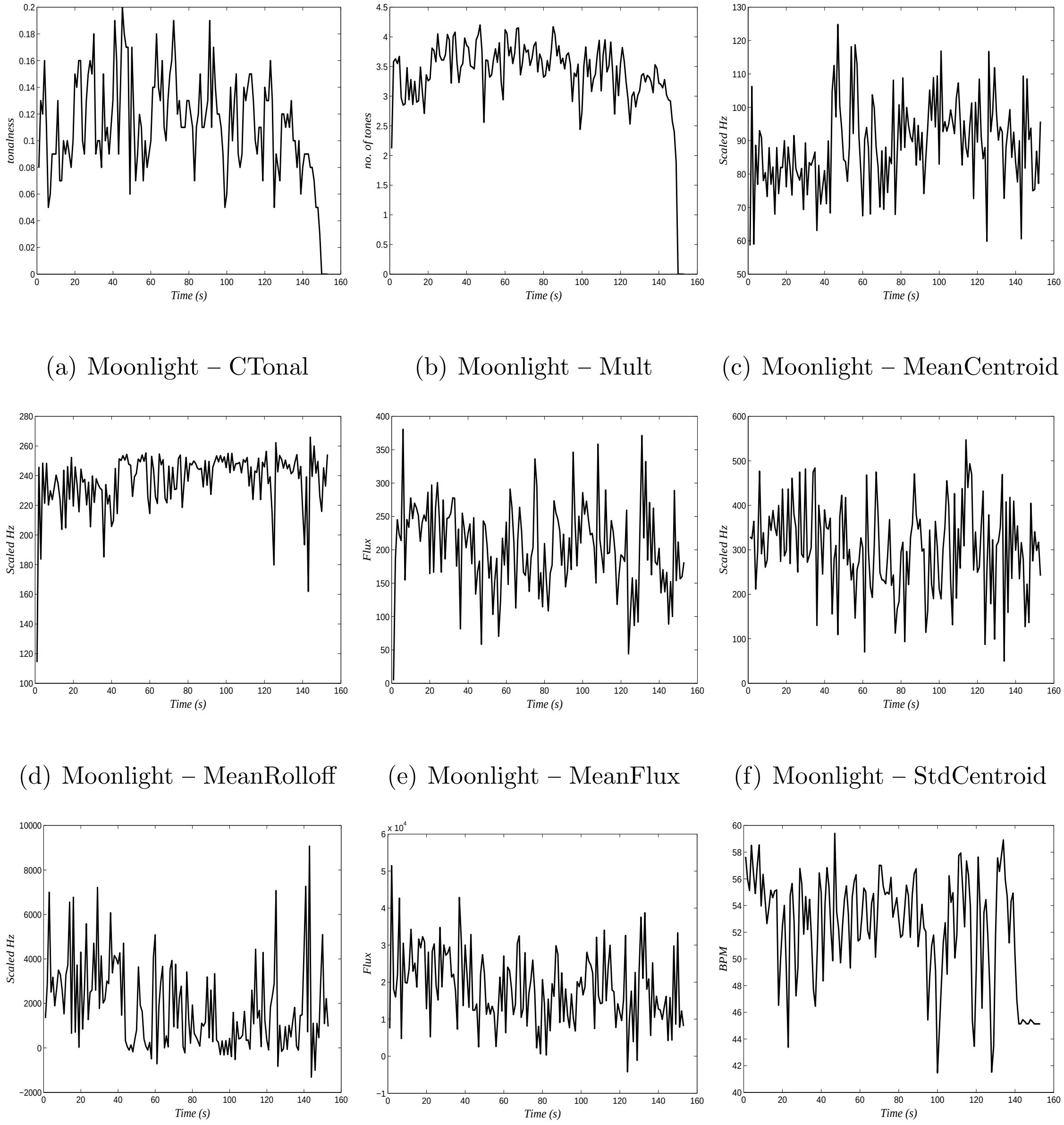
(f) Moonlight – SDiss(H&K)

(g) Moonlight – SDiss(S)

(h) Moonlight – TDiss(H&K)

(i) Moonlight – TDiss(S)

Figure B.7: The first nine features of Moonlight.

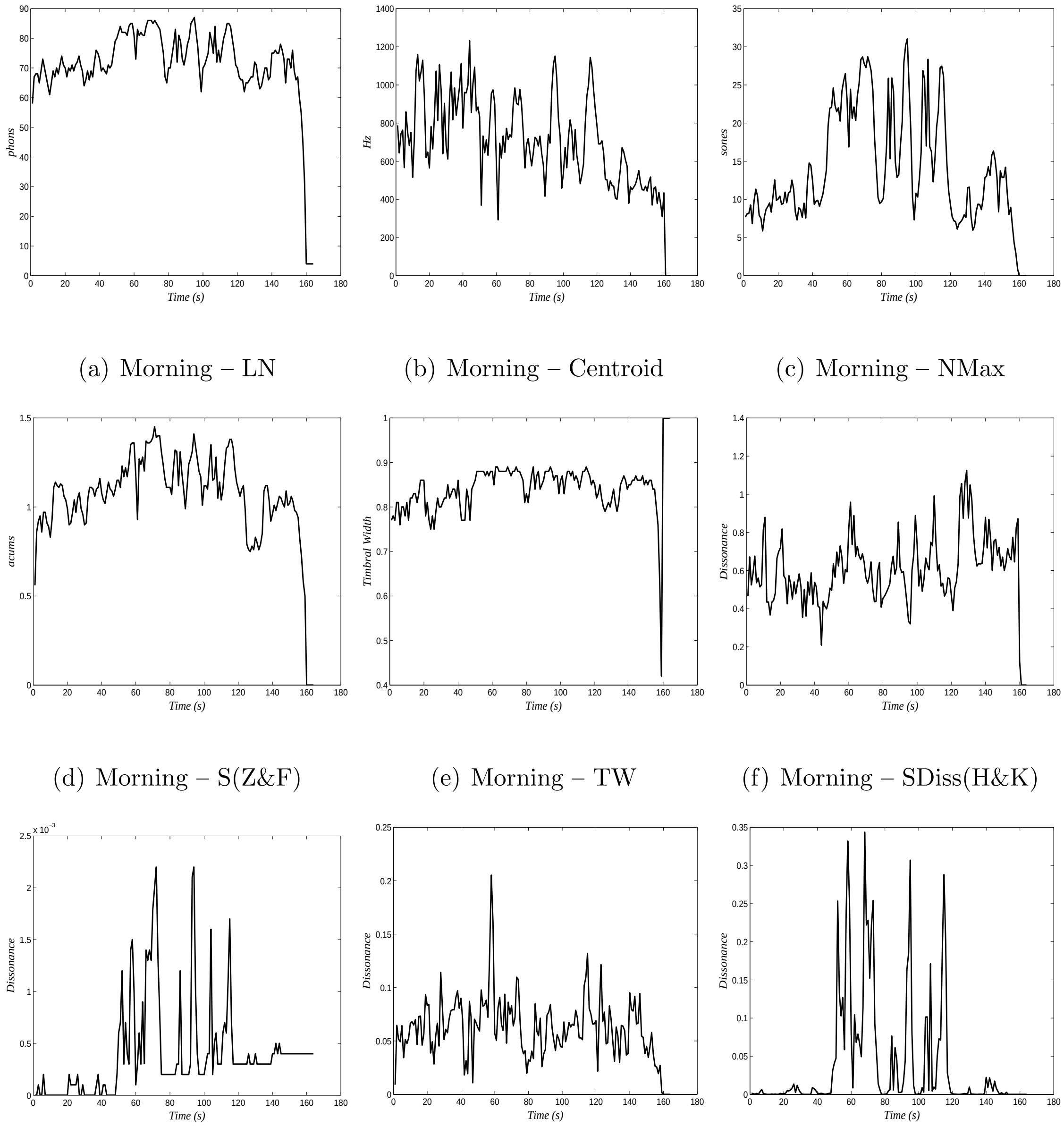


(a) Moonlight – CTonal (b) Moonlight – Mult (c) Moonlight – MeanCentroid

(d) Moonlight – MeanRolloff (e) Moonlight – MeanFlux (f) Moonlight – StdCentroid

(g) Moonlight – StdRolloff (h) Moonlight – StdFlux (i) Moonlight – BPM

Figure B.8: The last nine features of Moonlight.



(a) Morning – LN

(b) Morning – Centroid

(c) Morning – NMax

(d) Morning – S(Z&F)

(e) Morning – TW

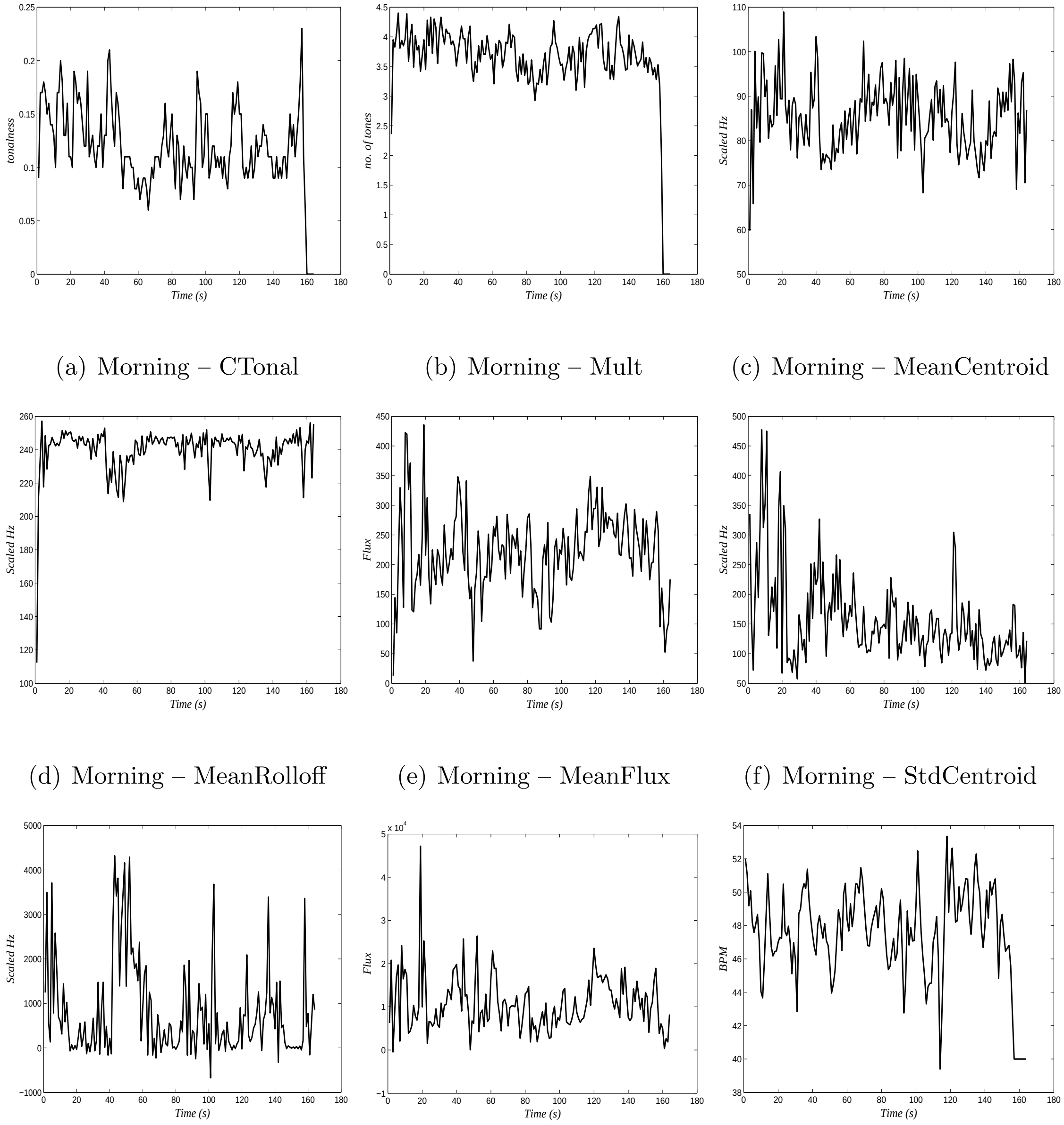
(f) Morning – SDiss(H&K)

(g) Morning – SDiss(S)

(h) Morning – TDiss(H&K)

(i) Morning – TDiss(S)

Figure B.9: The first nine features of Morning.



(a) Morning – CTonal

(b) Morning – Mult

(c) Morning – MeanCentroid

(d) Morning – MeanRolloff

(e) Morning – MeanFlux

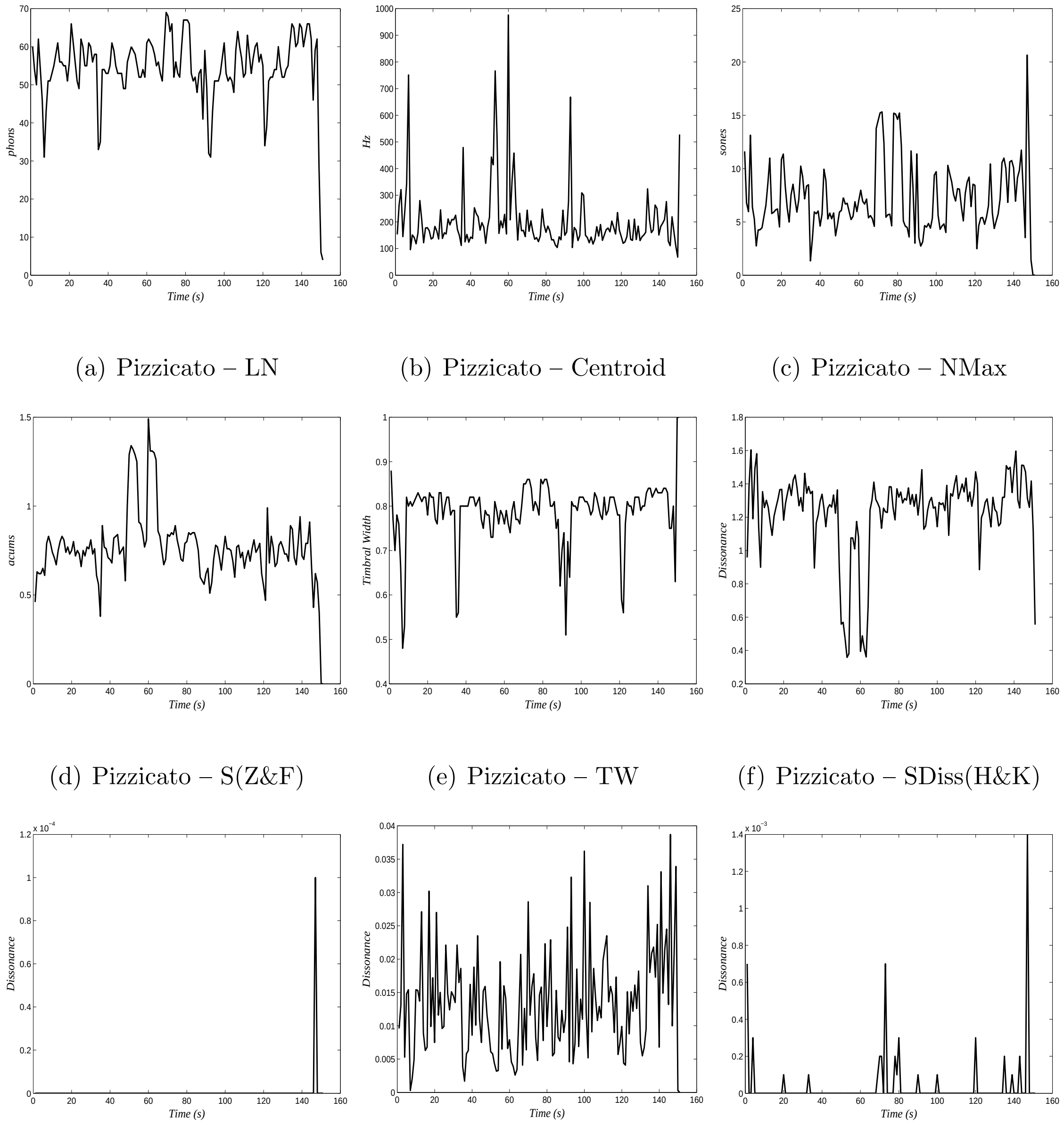
(f) Morning – StdCentroid

(g) Morning – StdRolloff

(h) Morning – StdFlux

(i) Morning – BPM

Figure B.10: The last nine features of Morning.

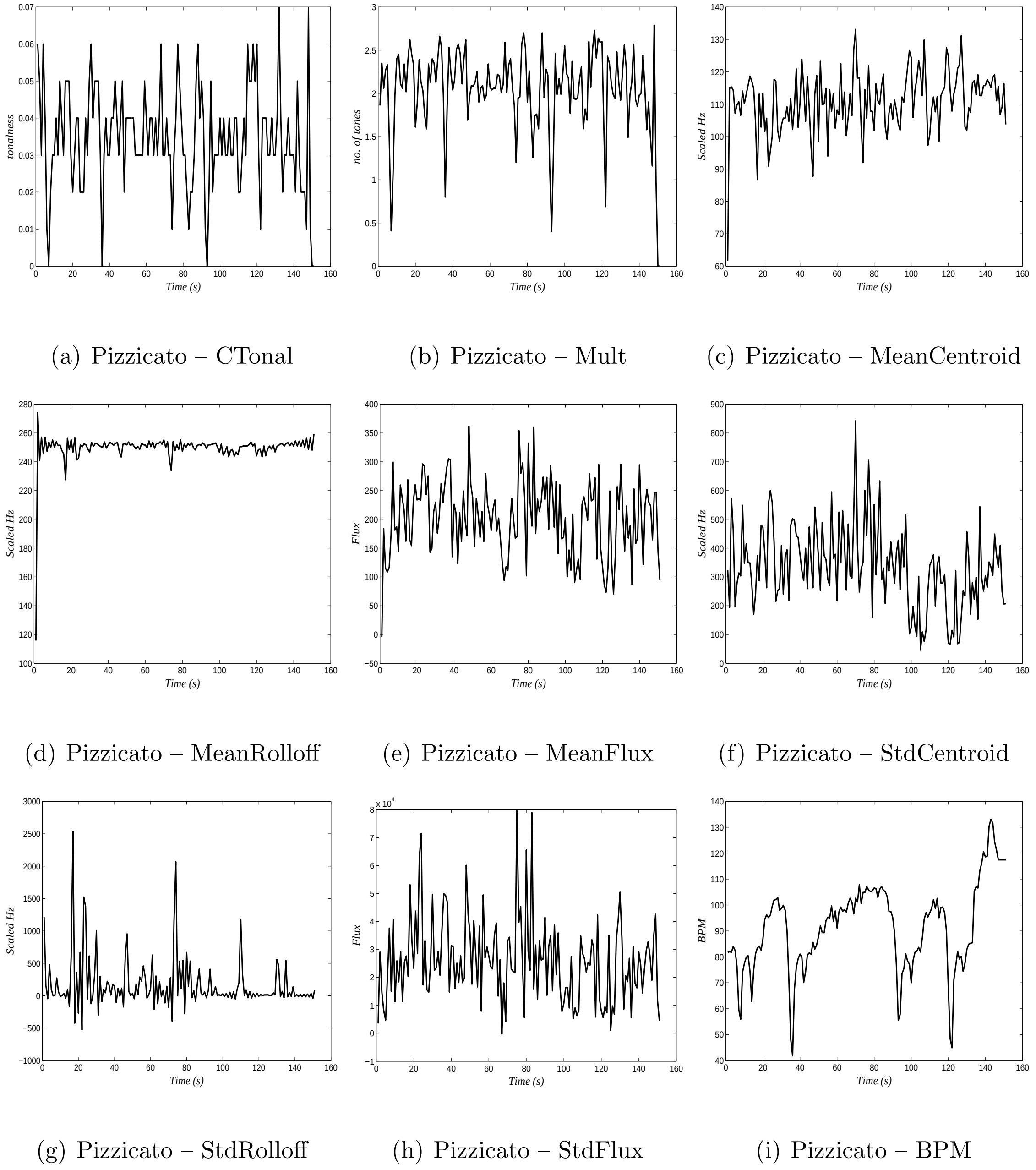


(a) Pizzicato – LN (b) Pizzicato – Centroid (c) Pizzicato – NMax

(d) Pizzicato – S(Z&F) (e) Pizzicato – TW (f) Pizzicato – SDiss(H&K)

(g) Pizzicato – SDiss(S) (h) Pizzicato – TDiss(H&K) (i) Pizzicato – TDiss(S)

Figure B.11: The first nine features of Pizzicato.



(a) Pizzicato – CTonal

(b) Pizzicato – Mult

(c) Pizzicato – MeanCentroid

(d) Pizzicato – MeanRolloff

(e) Pizzicato – MeanFlux

(f) Pizzicato – StdCentroid

(g) Pizzicato – StdRolloff

(h) Pizzicato – StdFlux

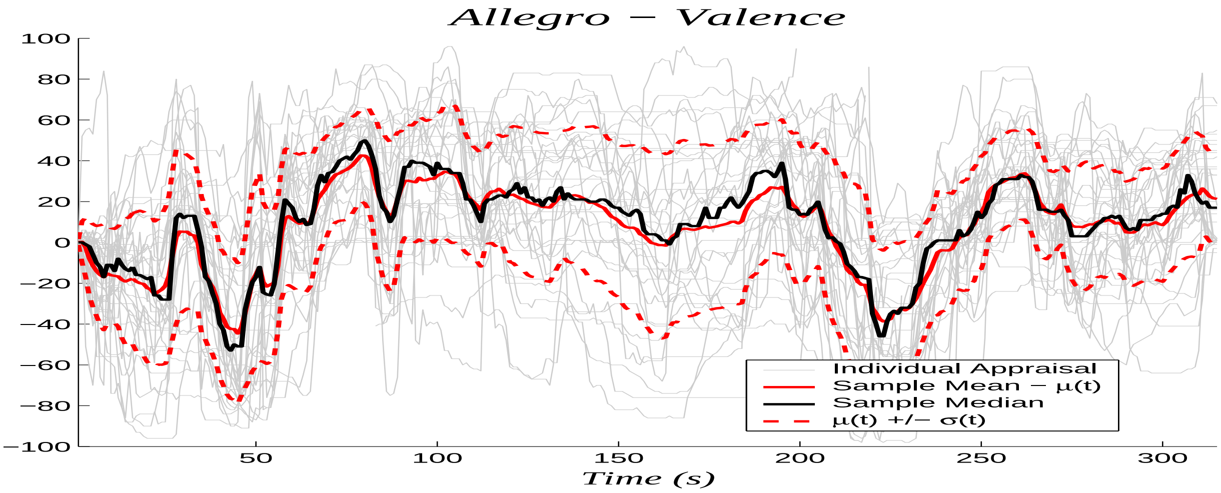
(i) Pizzicato – BPM

Figure B.12: The last nine features of Pizzicato.

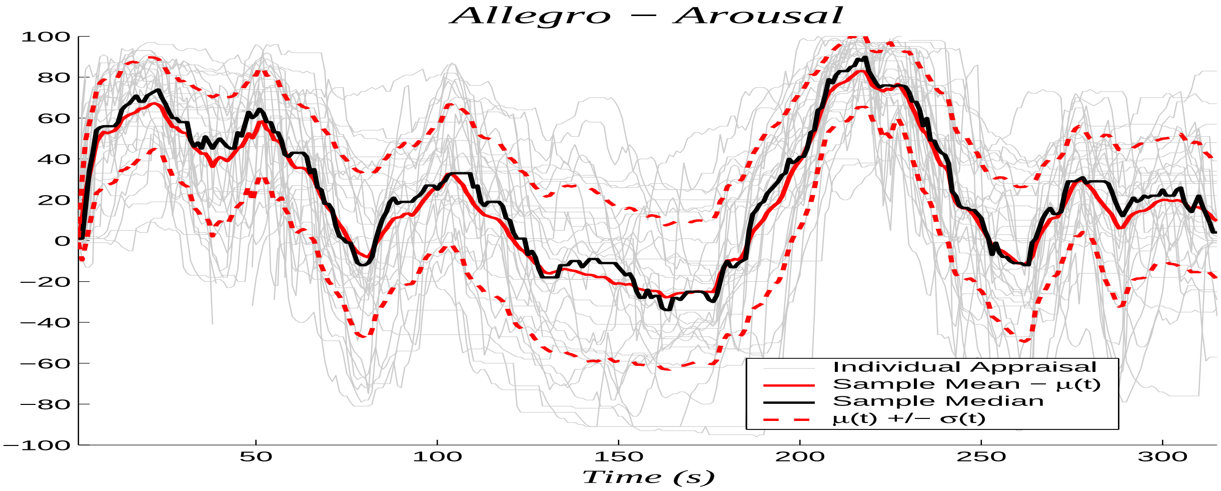
B.2 Emotional Appraisals

The following figures show the emotional appraisals gathered from the 35 volunteers with outliers removed. The median, mean and standard deviation of the emotional appraisals for each song are plotted as well. See Table 4.1 for information describing the musical selections to which the aliases refer.

The preprocessed emotional appraisal $\underline{\gamma}_{ij}(t)$ of participant j to song i is labeled as "Individual Appraisal" in the following graphs. The mean emotional appraisal for a song is labeled as $\mu(t)$ and the standard deviation of the appraisal at time t is labeled $\sigma(t)$.

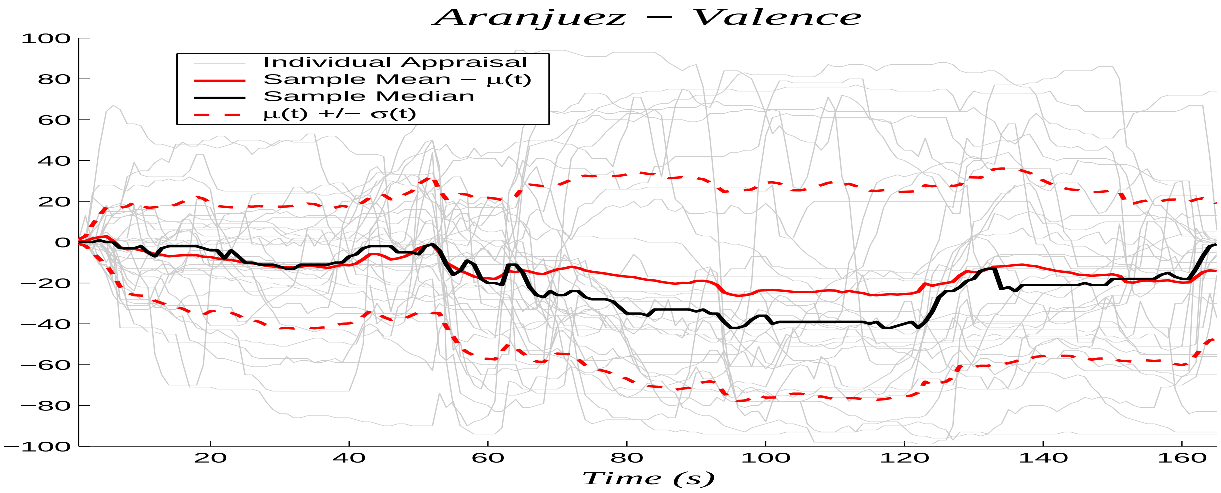


(a) Valence

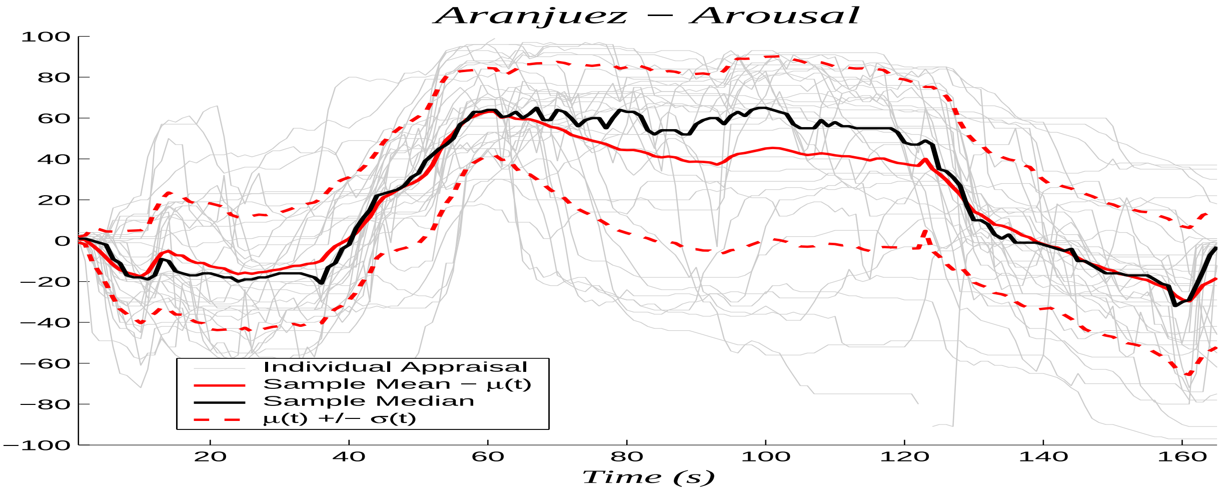


(b) Arousal

Figure B.13: Emotional appraisal of Allegro.

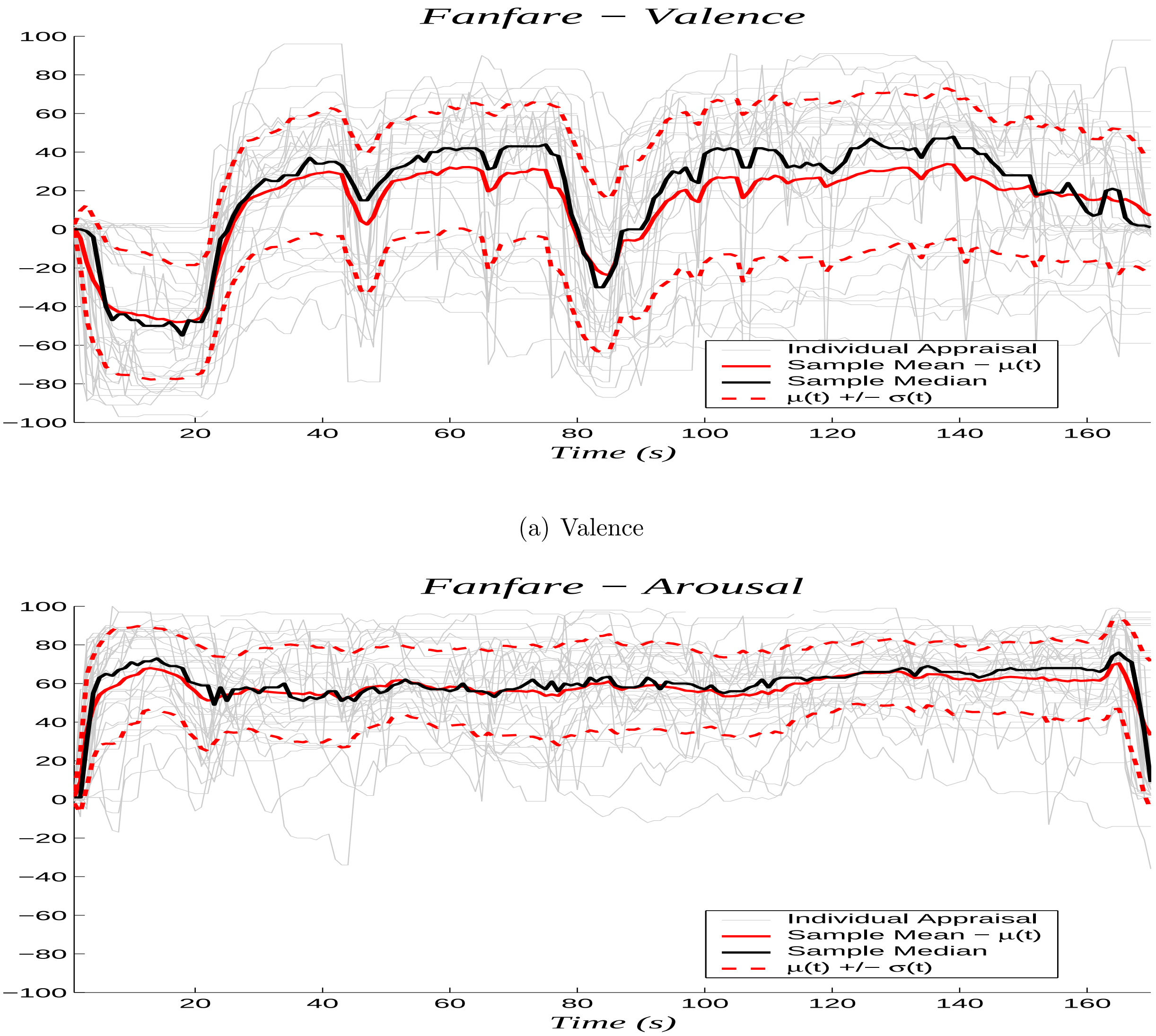


(a) Valence



(b) Arousal

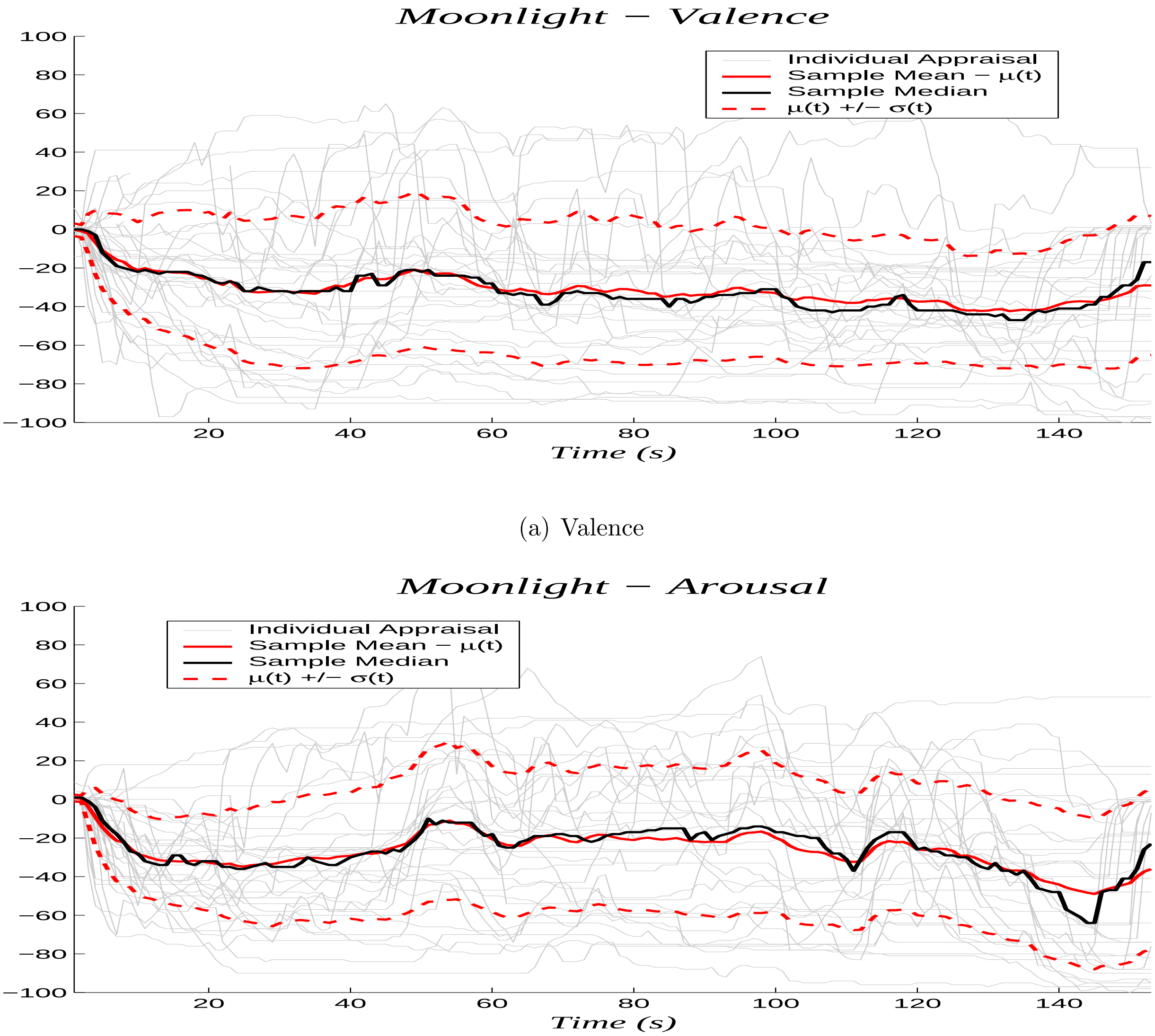
Figure B.14: Emotional appraisal of Aranjuez.



(a) Valence

(b) Arousal

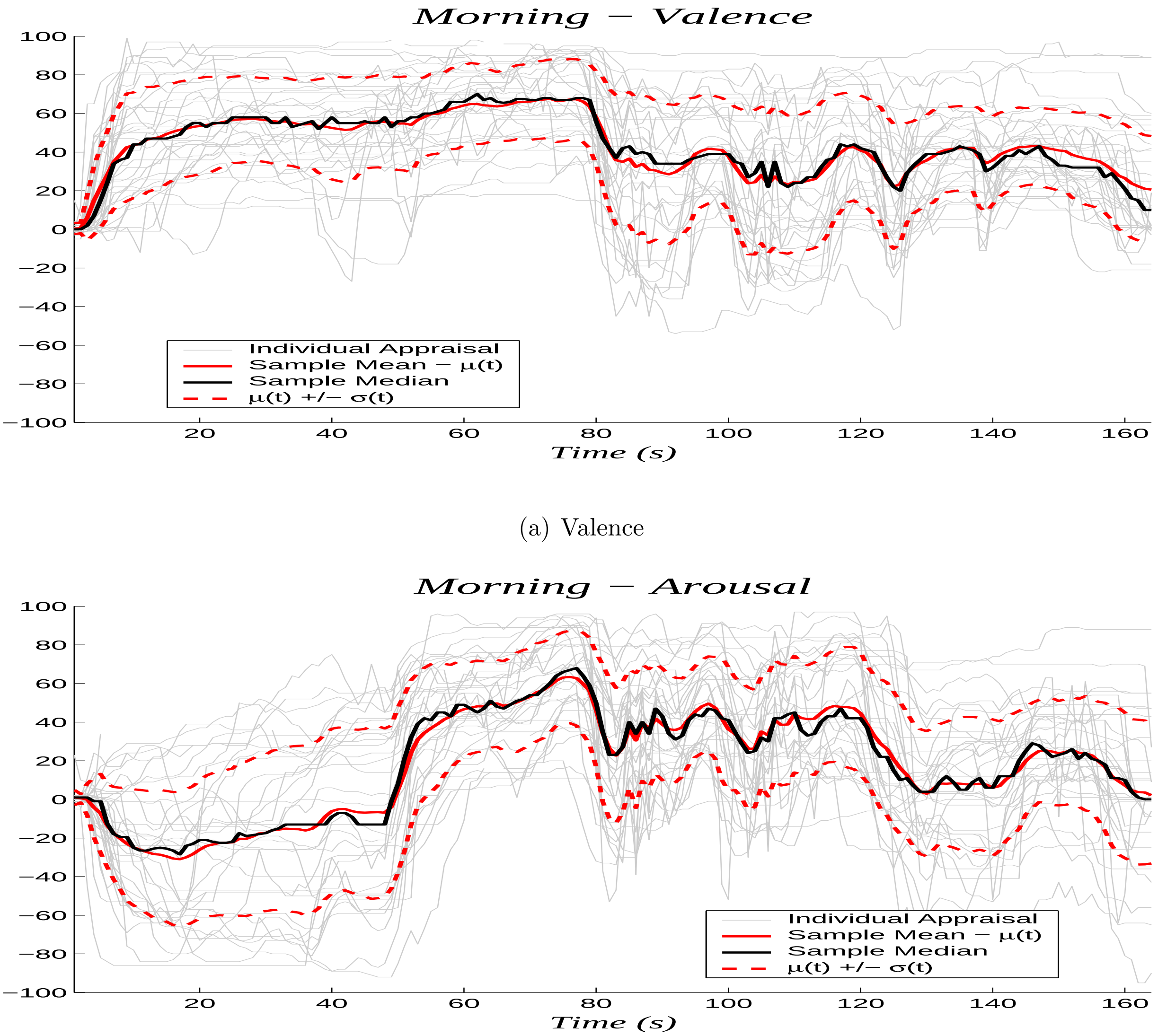
Figure B.15: Emotional appraisal of Fanfare.



(a) Valence

(b) Arousal

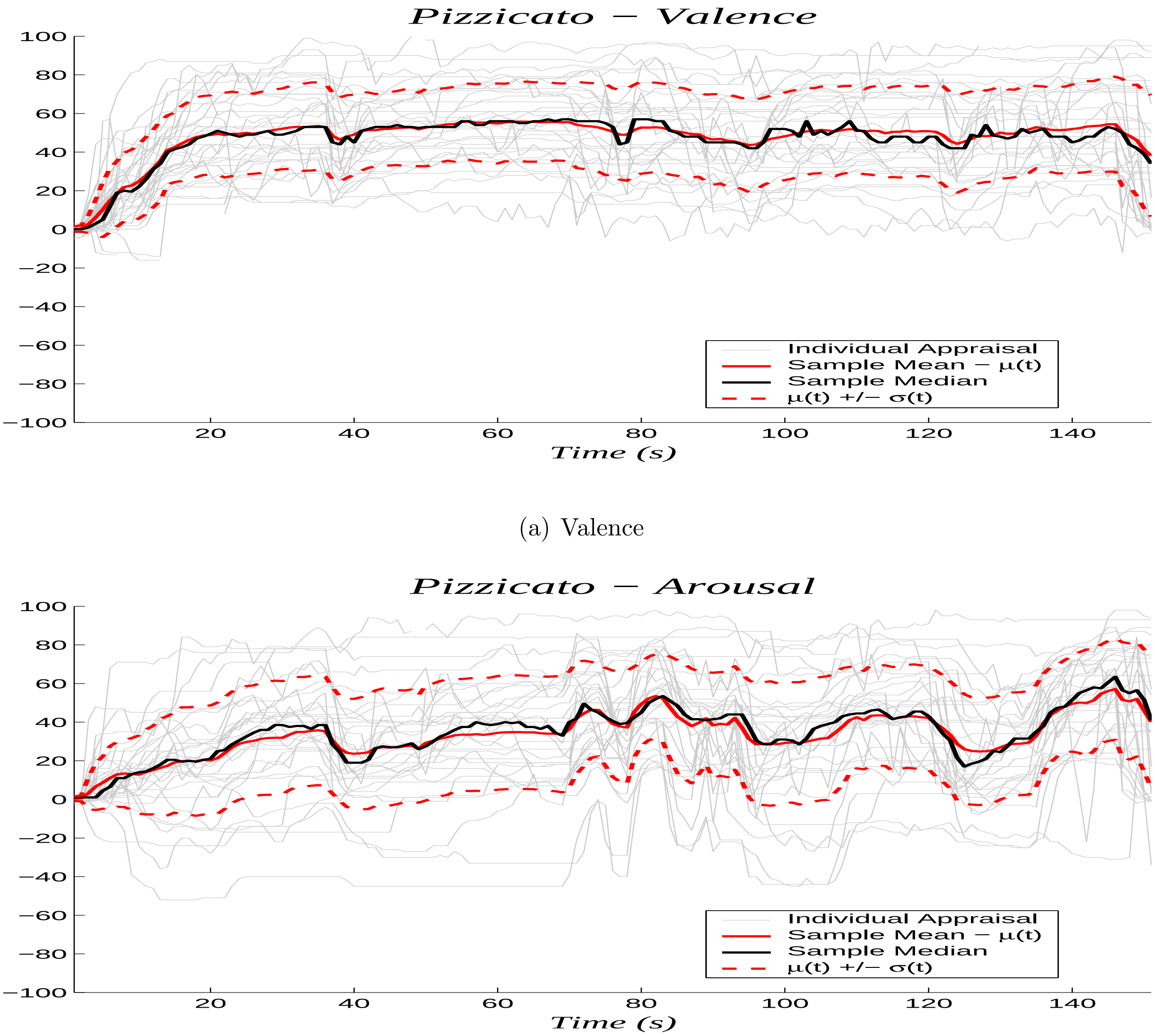
Figure B.16: Emotional appraisal of Moonlight.



(a) Valence

(b) Arousal

Figure B.17: Emotional appraisal of Morning.



(a) Valence

(b) Arousal

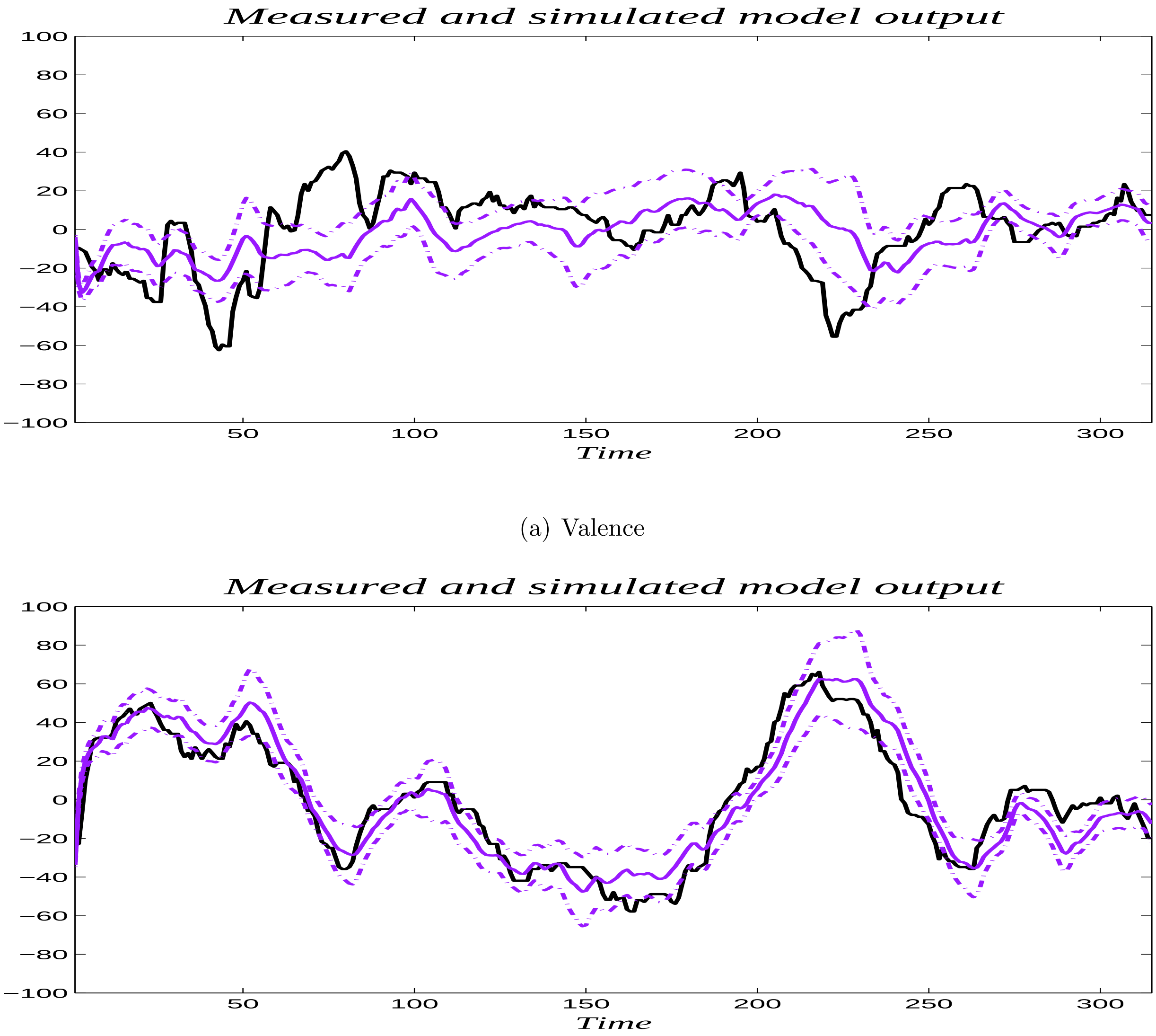
Figure B.18: Emotional appraisal of Pizzicato.

Appendix C

Model Outputs

The following figures show the simulated emotional appraisals of ARX21S_2 during six-fold cross-validation. This means that the model simulated the emotional appraisal of a song that was not in its training set.

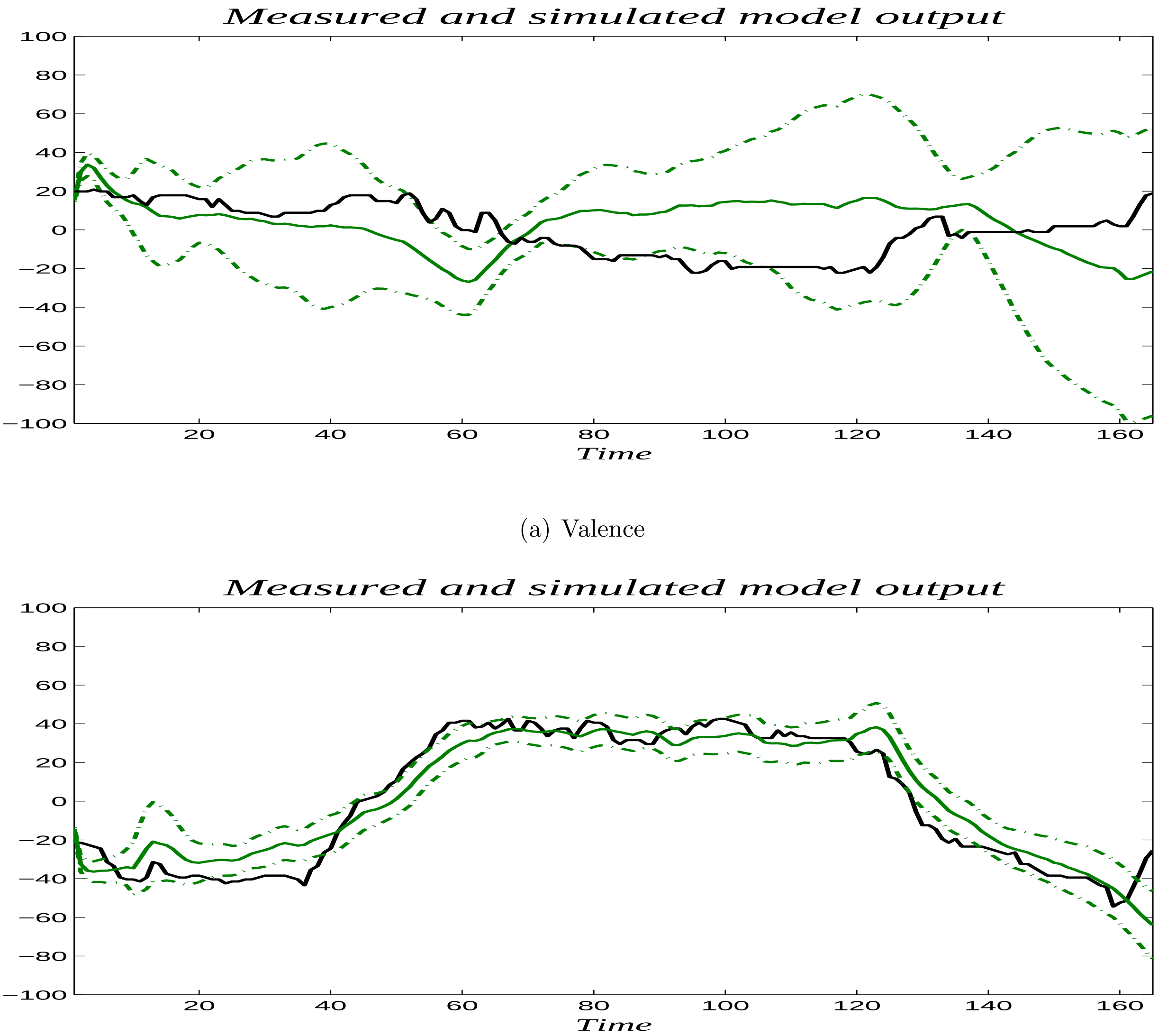
The solid black line is the measured median emotional appraisal. The solid, coloured line is the simulated output and the dotted lines show the 98% confidence interval of the simulated output. Each song is in a different colour to emphasize that six different training sets were used to generate these simulations.



(a) Valence

(b) Arousal

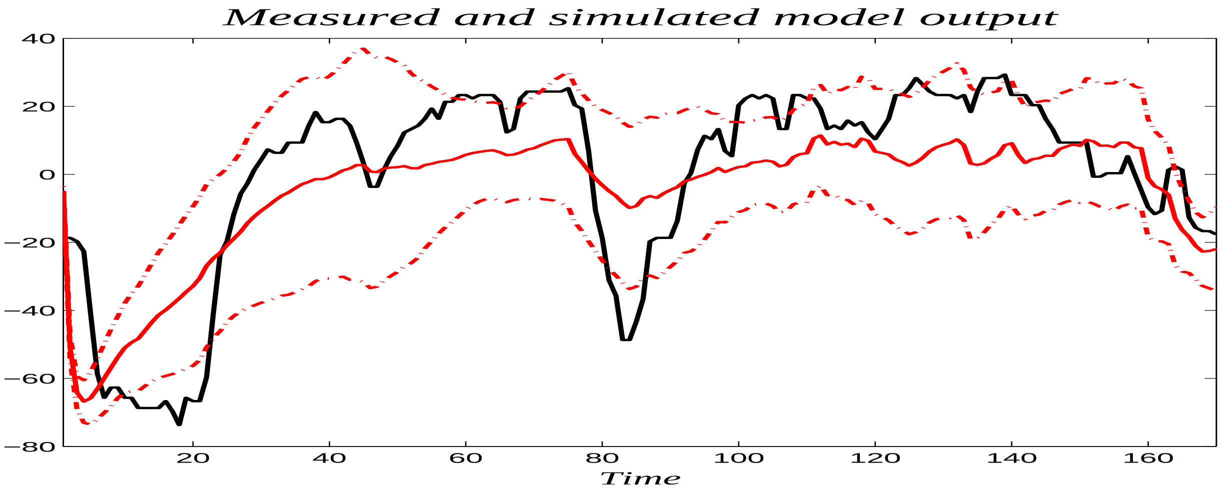
Figure C.1: Simulated emotional appraisal of Allegro.



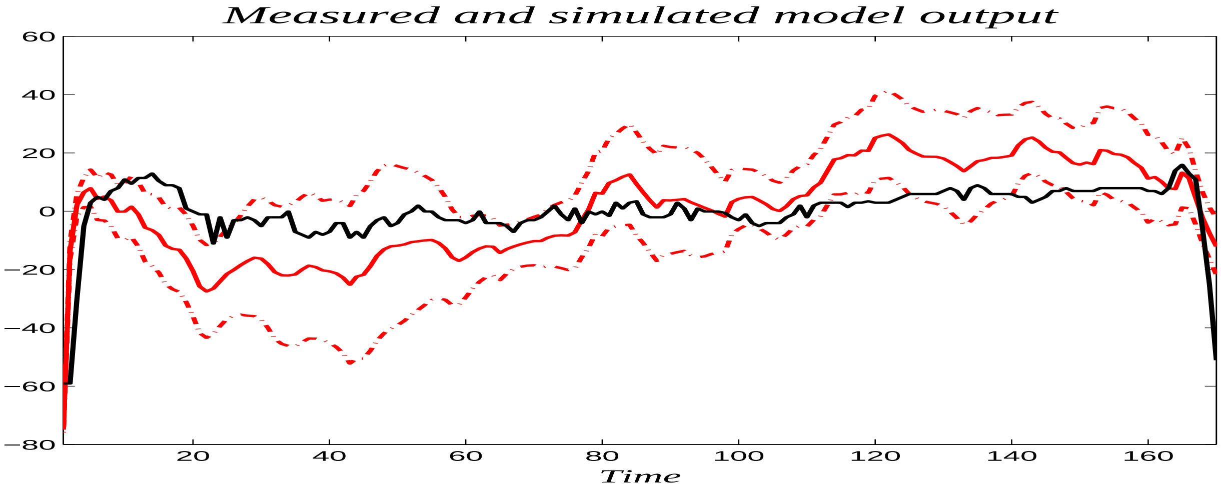
(a) Valence

(b) Arousal

Figure C.2: Simulated emotional appraisal of Aranjuez.

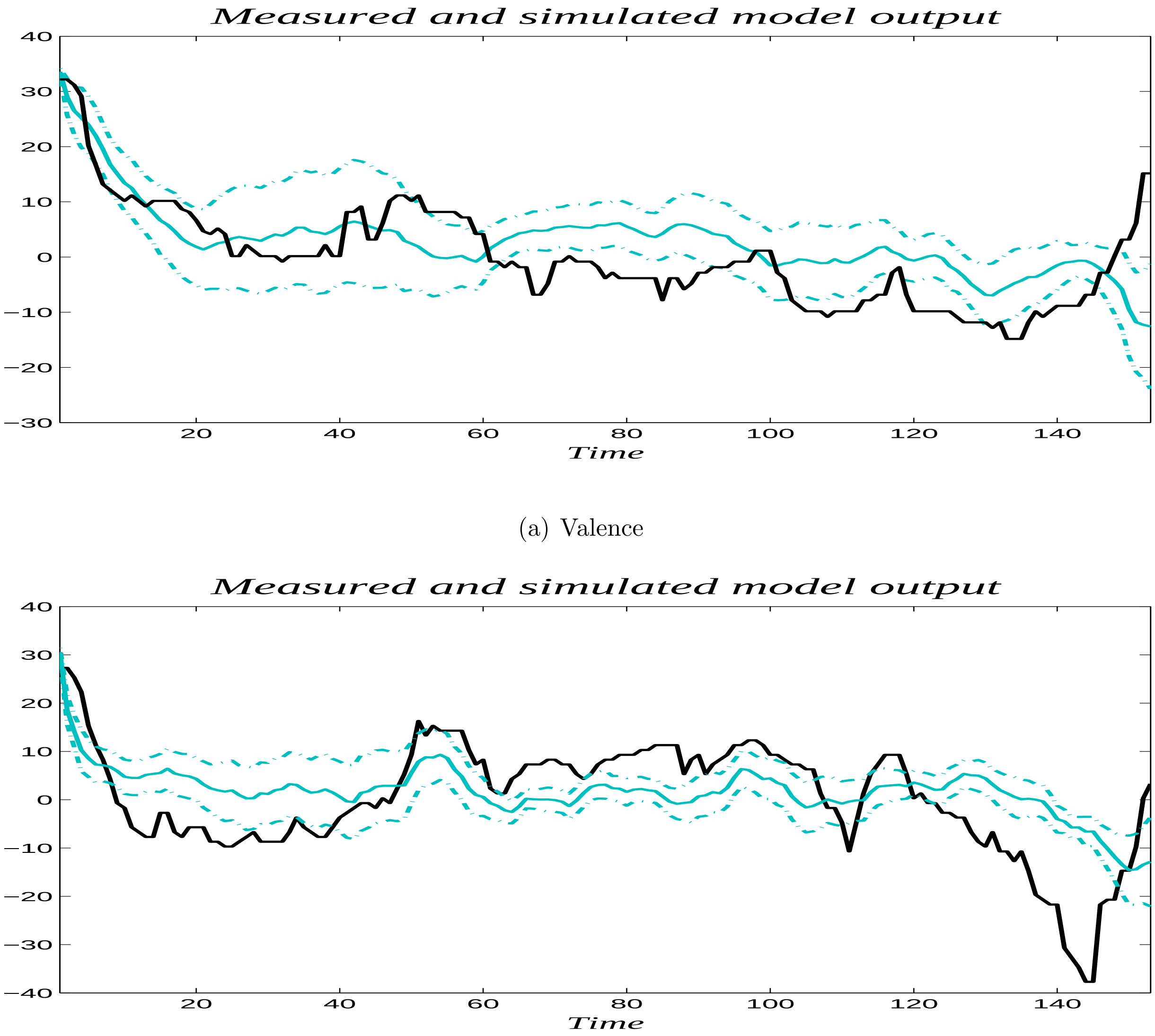


(a) Valence



(b) Arousal

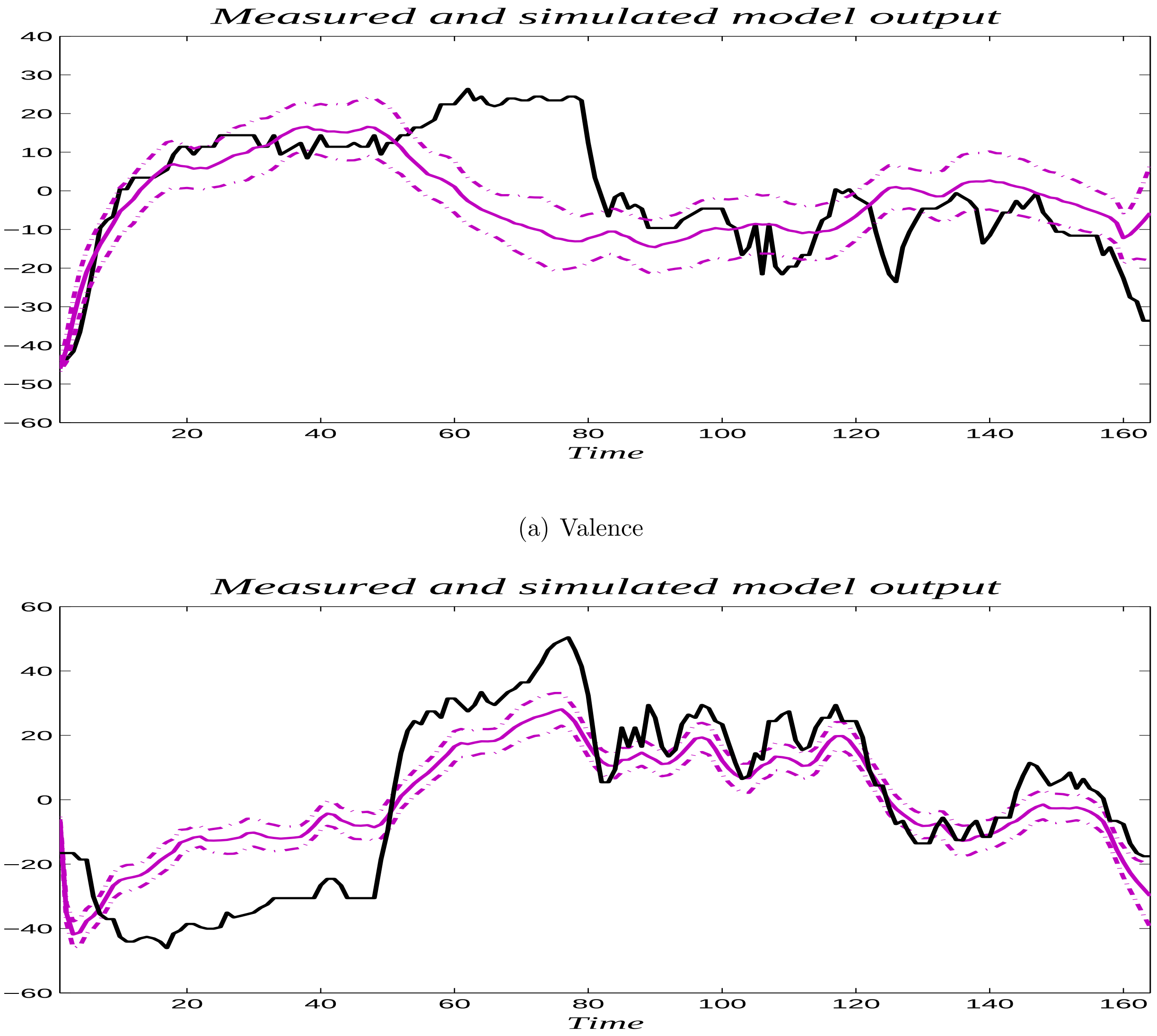
Figure C.3: Simulated emotional appraisal of Fanfare.



(a) Valence

(b) Arousal

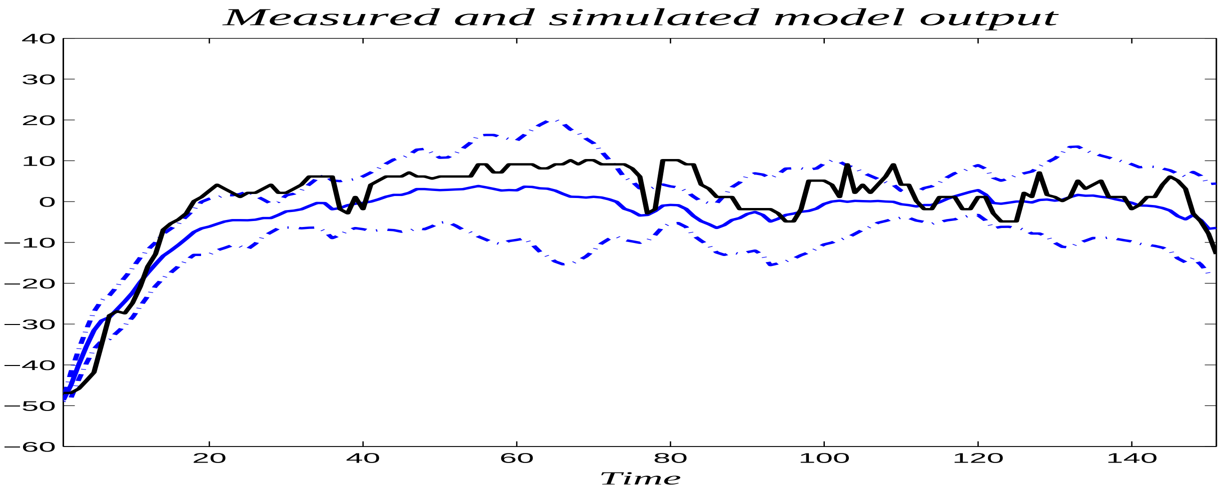
Figure C.4: Simulated emotional appraisal of Moonlight.



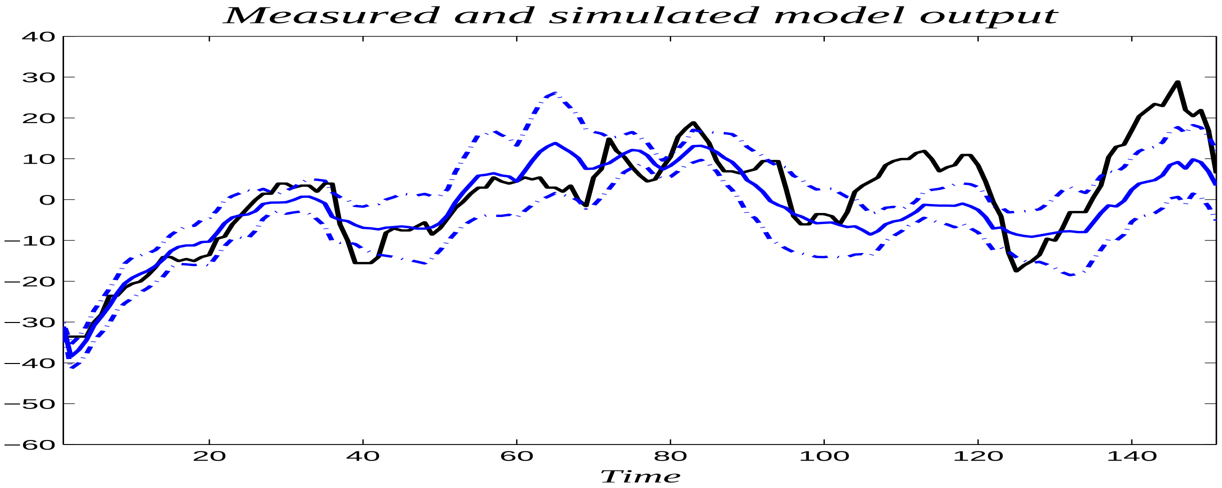
(a) Valence

(b) Arousal

Figure C.5: Simulated emotional appraisal of Morning.



(a) Valence



(b) Arousal

Figure C.6: Simulated emotional appraisal of Pizzicato.

Appendix D

Model Analysis

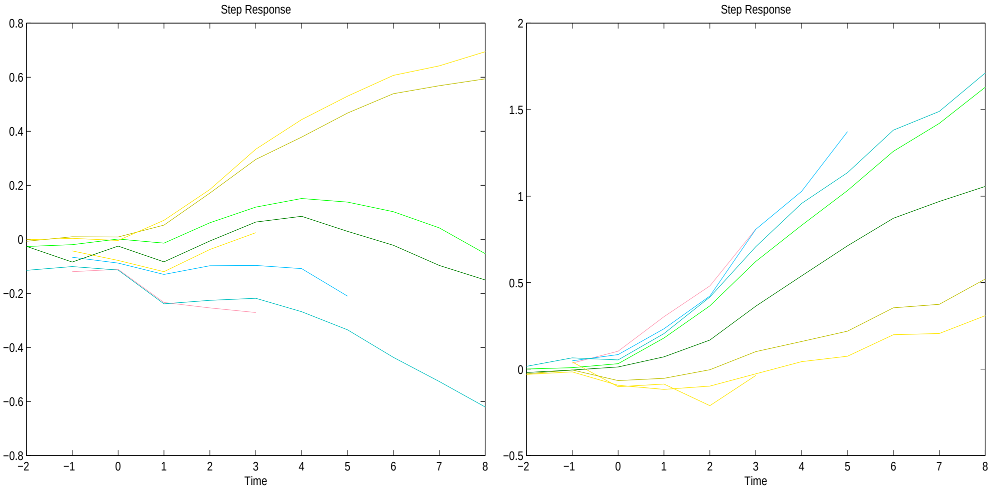
D.1 Step Response Estimates

The more inputs that are included in the estimated step response, the shorter the duration of the step response that can be reliably estimated using correlation analysis[16]. To overcome the short duration of the step response when 18 inputs are included, each input is included in eight random subsets of the inputs which are used to estimate the step response. Ideally, if an input signal affects an output signal, the time that the step response becomes significantly different from zero should not vary when different sets of inputs are included in the models. The sign of the step response should be consistent as well. If the estimated step response for an input signal has inconsistent delays or an inconsistent sign, one could

Table D.1: Input subsets used for estimating step response

Subset	Features Included
1	NMax, SDiss(H&K), TDiss(H&K), MeanFlux, StdCentroid, StdFlux
2	LN, Centroid, S(Z&F), SDiss(S), MeanCentroid, BPM
3	TW, TDiss(S), CTonal, Mult, MeanRolloff, StdRolloff
4	LN, S(Z&F), TDiss(H&K), CTonal, MeanFlux, StdRolloff
5	Centroid, NMax, TW, SDiss(S), MeanCentroid, StdFlux
6	SDiss(H&K), TDiss(S), Mult, MeanRolloff, StdCentroid, BPM
7	LN, NMax, SDiss(H&K), SDiss(S), TDiss(H&K), MeanRolloff
8	Centroid, TW, TDiss(S), Mult, MeanCentroid, MeanFlux
9	S(Z&F), CTonal, StdCentroid, StdRolloff, StdFlux, BPM
10	Centroid, S(Z&F), SDiss(S), TDiss(H&K), Mult, BPM
11	TW, TDiss(S), CTonal, MeanRolloff, StdCentroid, StdRolloff
12	LN, NMax, SDiss(H&K), MeanCentroid, MeanFlux, StdFlux
13	LN, SDiss(H&K), TDiss(H&K), MeanRolloff, MeanFlux, BPM
14	NMax, TW, SDiss(S), Mult, MeanCentroid, StdRolloff
15	Centroid, S(Z&F), TDiss(S), CTonal, StdCentroid, StdFlux
16	LN, Centroid, TW, SDiss(H&K), TDiss(S), CTonal, MeanRolloff, MeanFlux, StdFlux
17	NMax, S(Z&F), SDiss(S), TDiss(H&K), Mult, MeanCentroid, StdCentroid, StdRolloff, BPM
18	NMax, S(Z&F), TW, SDiss(H&K), SDiss(S), TDiss(H&K), CTonal, Mult, MeanCentroid, StdCentroid, StdRolloff, StdFlux
19	LN, Centroid, S(Z&F), TW, SDiss(H&K), TDiss(S), Mult, MeanCentroid, MeanRolloff, MeanFlux, StdCentroid, BPM
20	LN, Centroid, NMax, SDiss(S), TDiss(H&K), TDiss(S), CTonal, MeanRolloff, MeanFlux, StdRolloff, StdFlux, BPM

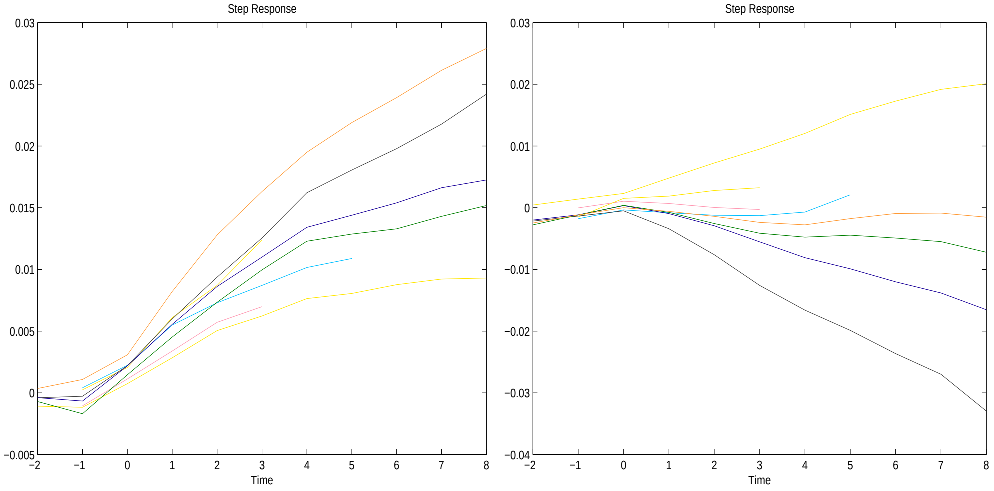
argue that the input is not explaining consistent patterns in the output signals and thus should not be included in the models.



(a) Valence

(b) Arousal

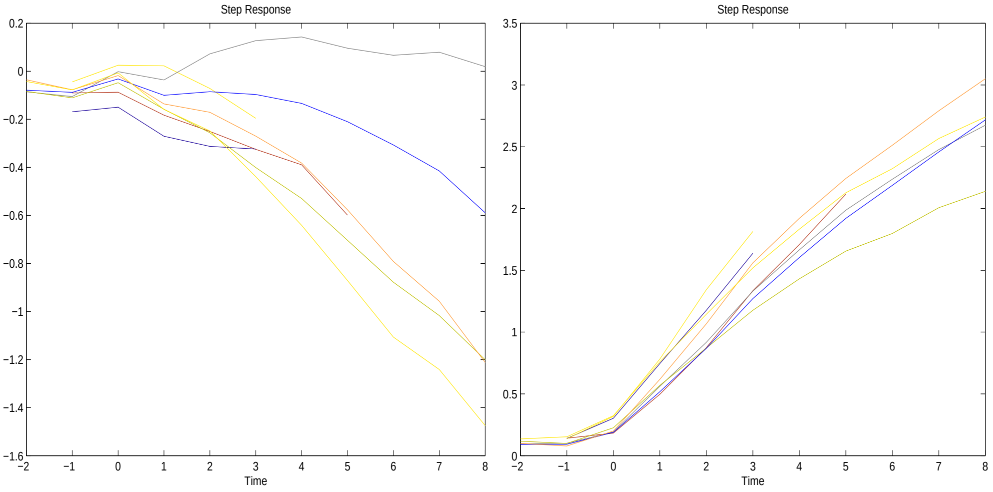
Figure D.1: Estimated step response for LN.



(a) Valence

(b) Arousal

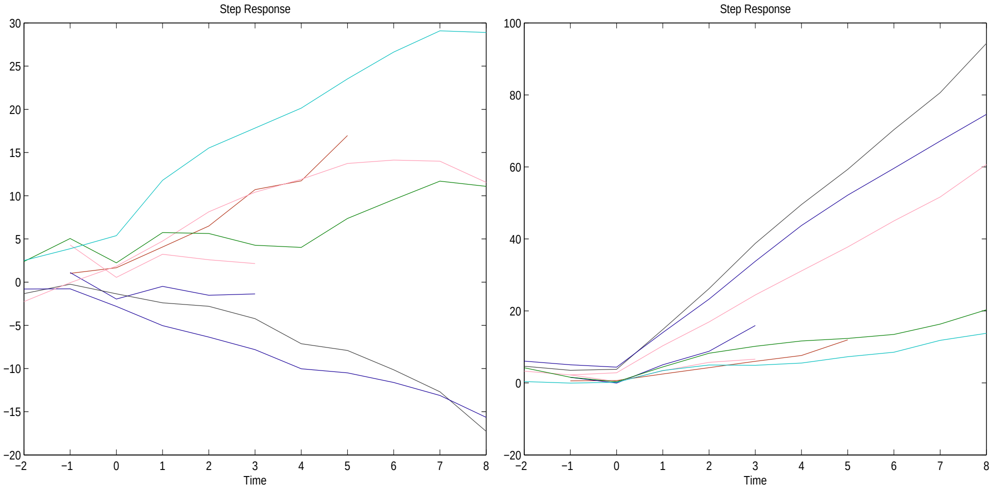
Figure D.2: Estimated step response for Centroid.



(a) Valence

(b) Arousal

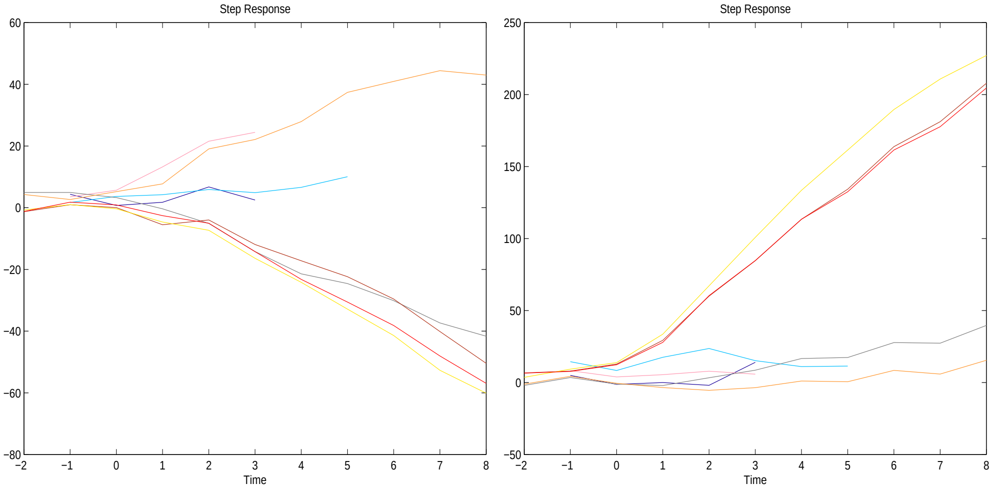
Figure D.3: Estimated step response for NMax.



(a) Valence

(b) Arousal

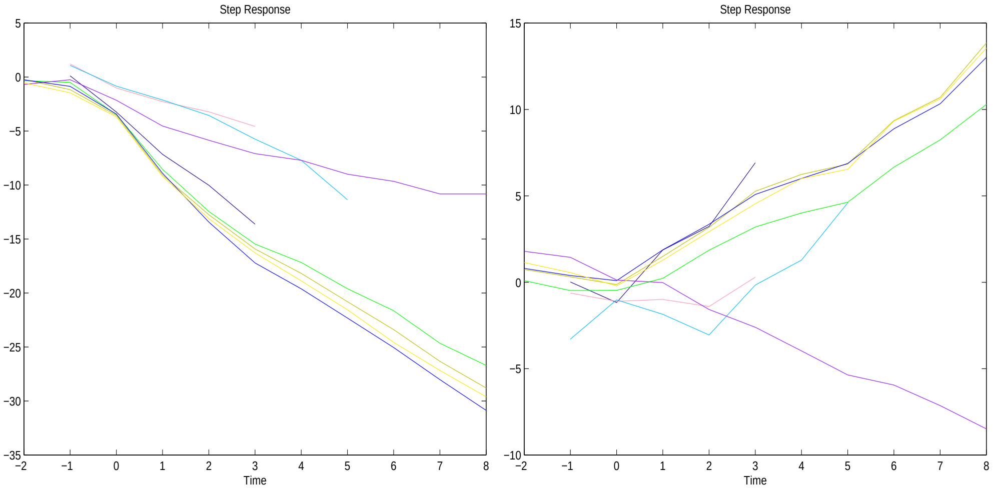
Figure D.4: Estimated step response for S(Z&F).



(a) Valence

(b) Arousal

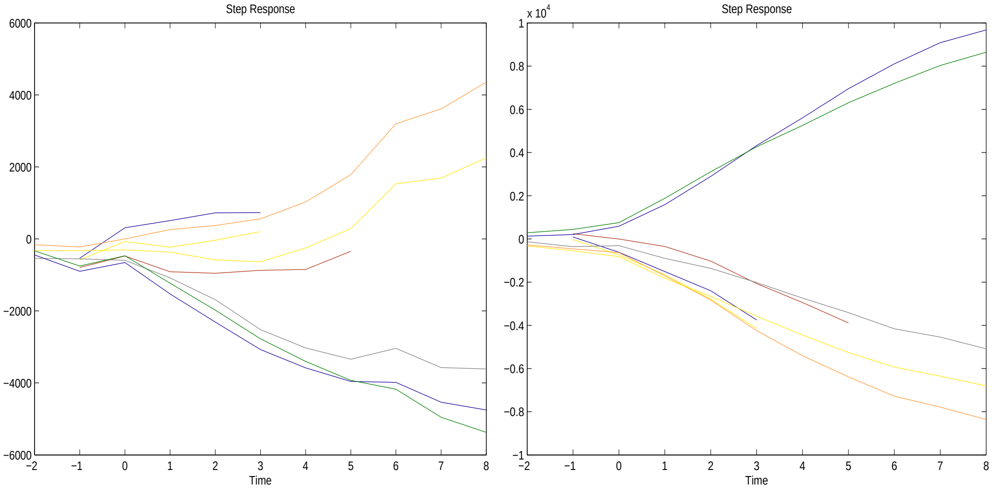
Figure D.5: Estimated step response for TW.



(a) Valence

(b) Arousal

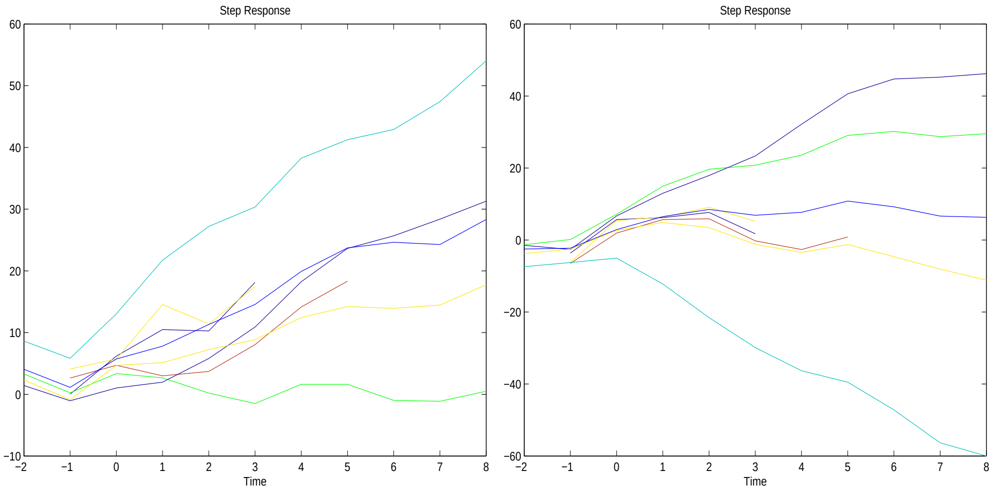
Figure D.6: Estimated step response for SDiss(H&K).



(a) Valence

(b) Arousal

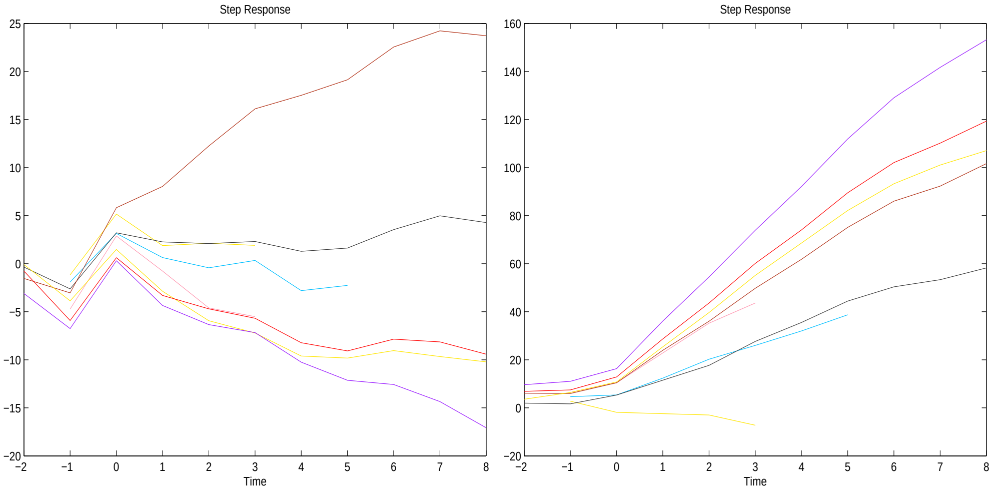
Figure D.7: Estimated step response for SDiss(S).



(a) Valence

(b) Arousal

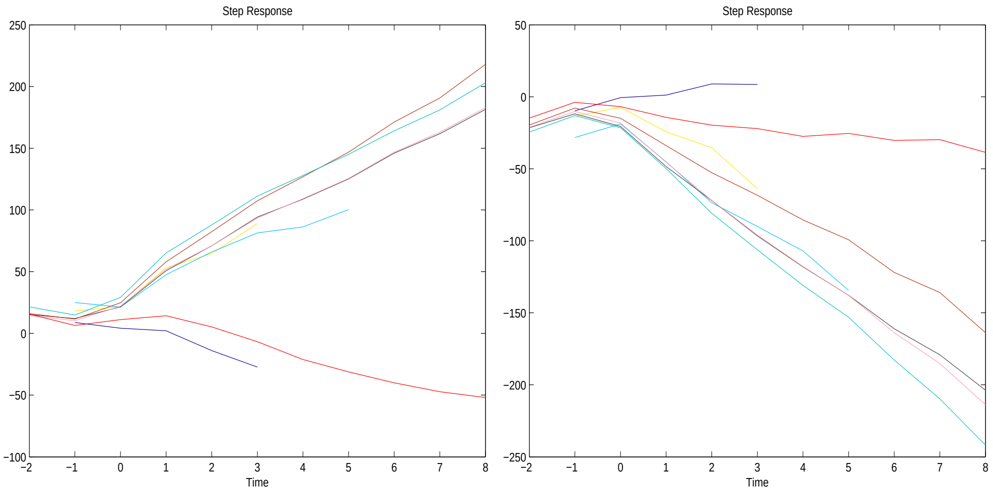
Figure D.8: Estimated step response for TDiss(H&K).



(a) Valence

(b) Arousal

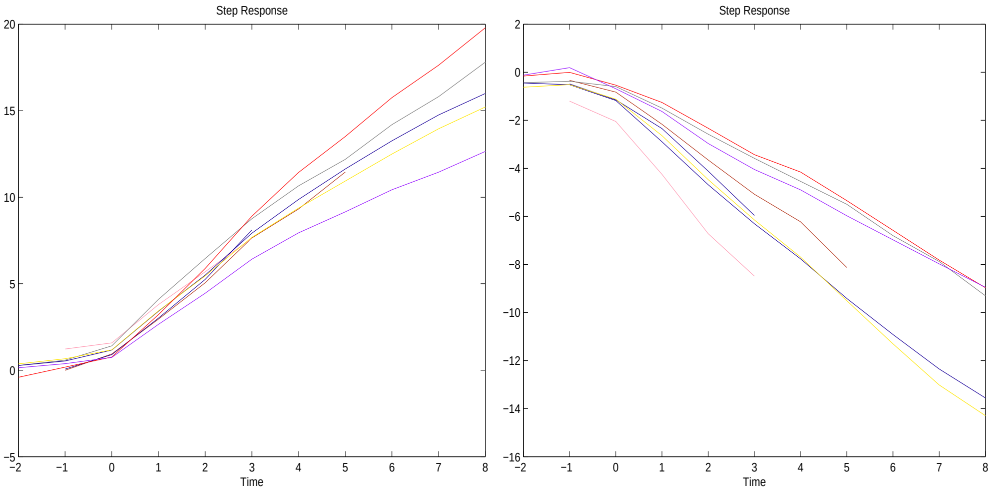
Figure D.9: Estimated step response for TDiss(S).



(a) Valence

(b) Arousal

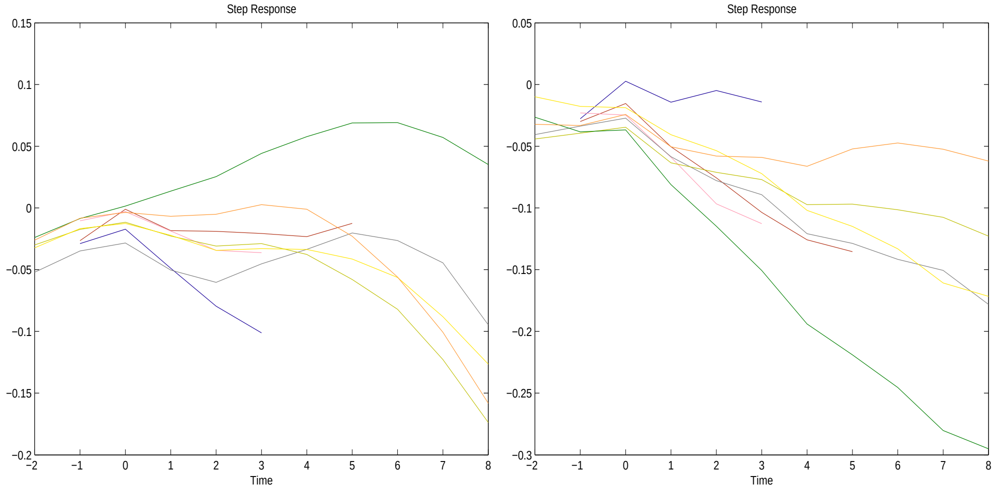
Figure D.10: Estimated step response for CTonal.



(a) Valence

(b) Arousal

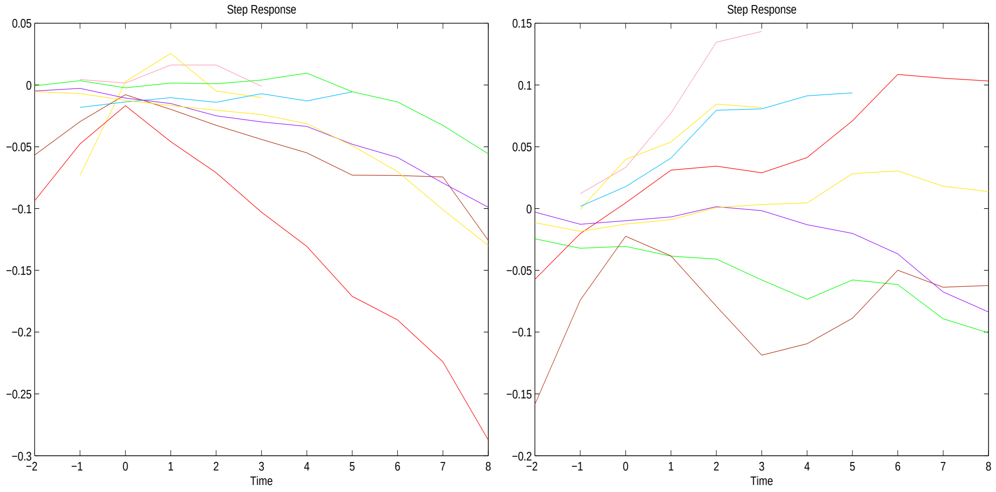
Figure D.11: Estimated step response for Mult.



(a) Valence

(b) Arousal

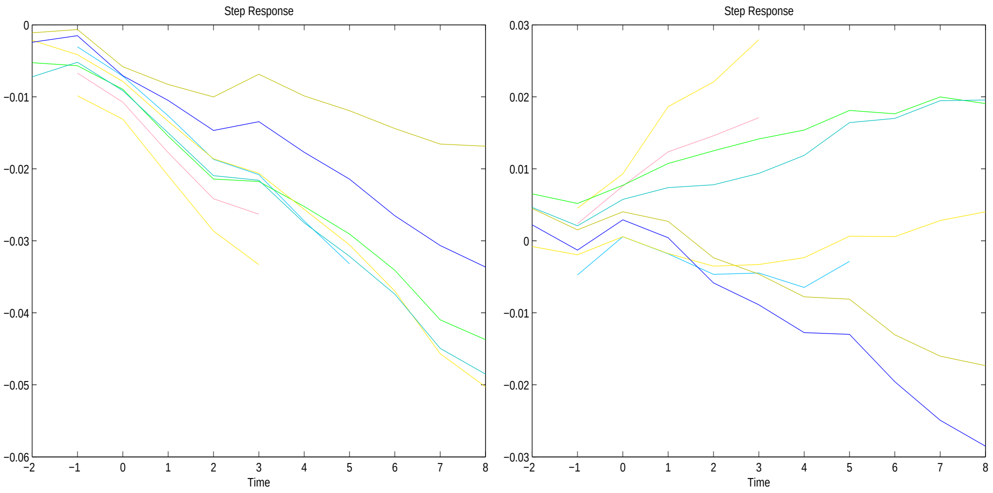
Figure D.12: Estimated step response for MeanCentroid.



(a) Valence

(b) Arousal

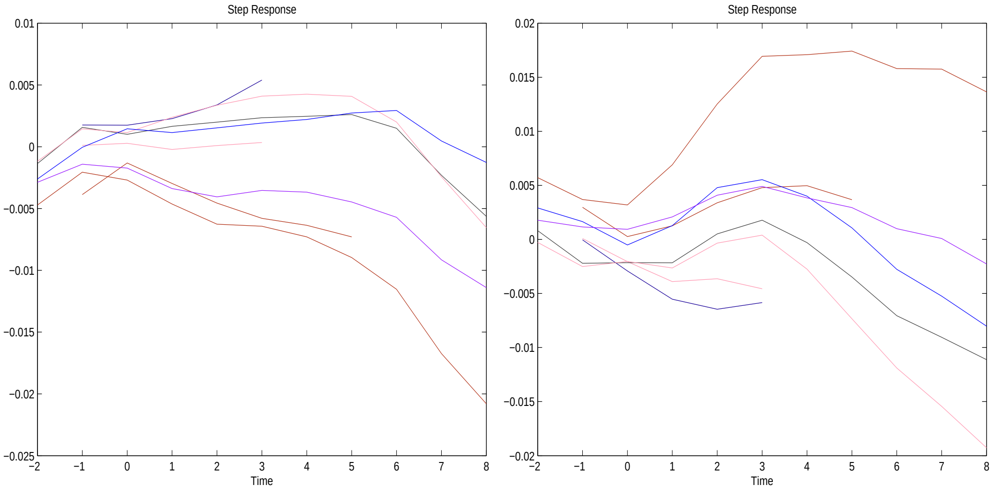
Figure D.13: Estimated step response for MeanRolloff.



(a) Valence

(b) Arousal

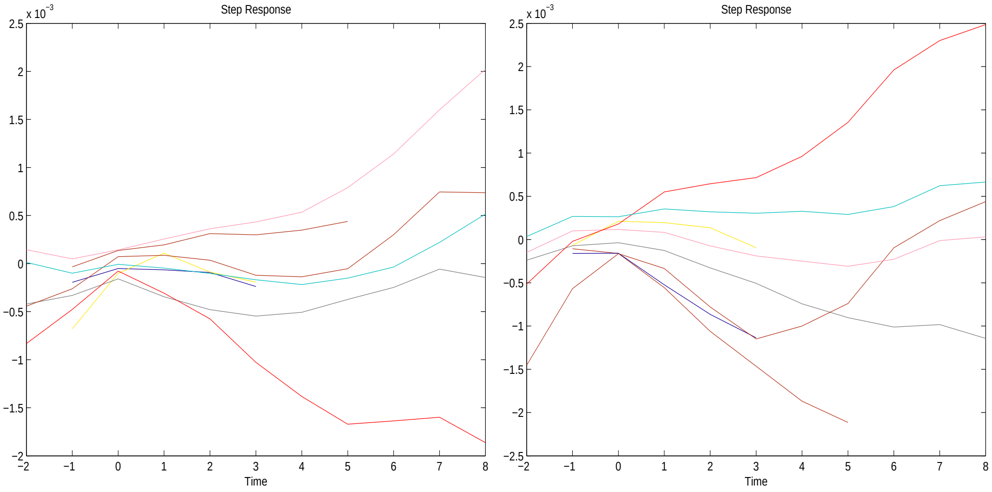
Figure D.14: Estimated step response for MeanFlux.



(a) Valence

(b) Arousal

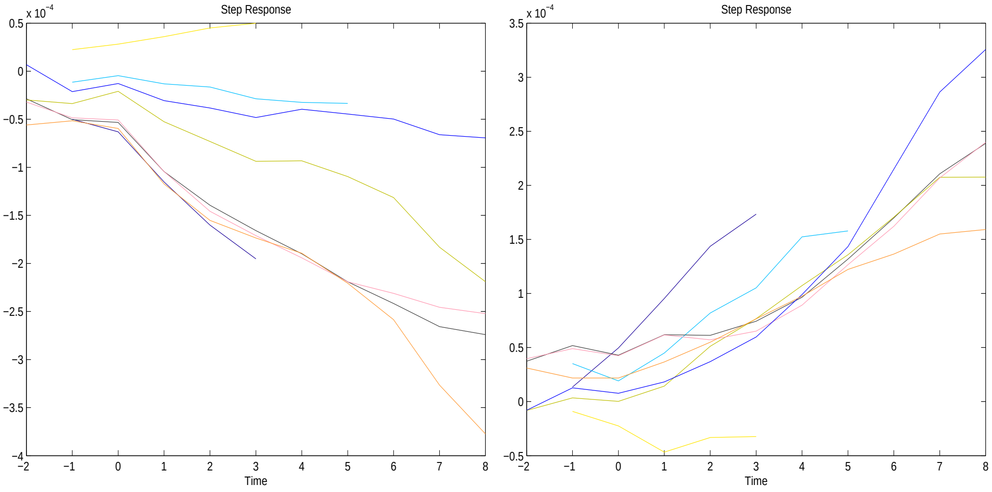
Figure D.15: Estimated step response for StdCentroid.



(a) Valence

(b) Arousal

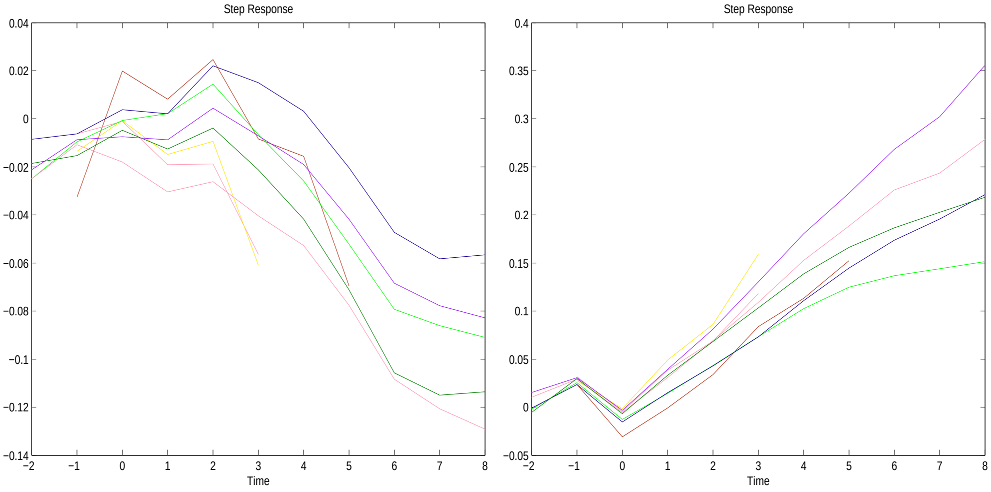
Figure D.16: Estimated step response for StdRolloff.



(a) Valence

(b) Arousal

Figure D.17: Estimated step response for StdFlux.



(a) Valence

(b) Arousal

Figure D.18: Estimated step response for BPM.

D.2 Model Structures

For all of the models that follow, the inputs and outputs are described using the following variables:

$\underline{y}(t)$ is a vector consisting of valence and arousal

$\underline{u}(t)$ is a vector consisting of the following features measured at time t: LN, Centroid, NMax, S(Z&F), TW, SDiss(H&K), SDiss(S), TDiss(S), CTonal, Mult, MeanCentroid, MeanRolloff, MeanFlux, StdFlux, BPM

All design variables are displayed in the format used by the System Identification Toolbox[16].

D.2.1 State-Space Models

PSS1_1: Quickstart First Order

$$n = 1$$

$$\text{Delays} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

PSS1_2: First Order with modified delay vector

$$n = 1$$

$$\text{Delays} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

PSS1_3: First Order with modified delay vector

$$n = 1$$

$$\text{Delays} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The delay vector is the minimum value of each row in Table 5.2.

PSS2_1: Quickstart Second Order

$$n = 2$$

$$\text{Delays} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

PSS2_2: Second Order with modified delay vector

$$n = 2$$

$$\text{Delays} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

PSS2_3: Second Order with modified delay vector

$$n = 2$$

$$\text{Delays} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The delay vector is the minimum value of each row in Table 5.2.

PSS3_1: Quickstart Third Order

$$n = 3$$

$$\text{Delays} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

PSS3_2: Third Order with modified delay vector

$$n = 3$$

$$\text{Delays} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

PSS3_3: Third Order with modified delay vector

$$n = 3$$

$$\text{Delays} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The delay vector is the minimum value of each row in Table 5.2.

PSS4_1: Quickstart Fourth Order

$$n = 4$$

$$\text{Delays} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

PSS4_2: Fourth Order with modified delay vector

$$n = 4$$

$$\text{Delays} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

PSS4_3: Fourth Order with modified delay vector

$$n = 4$$

$$\text{Delays} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The delay vector is the minimum value of each row in Table 5.2.

D.2.2 ARX Models

ARX44_1: Quickstart fourth order

$$na = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$$

$$nb = \begin{bmatrix} 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \end{bmatrix}$$

$$nk = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

ARX44_2: Fourth order with delays estimated from step responses

$$na = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$$

$$nb = \begin{bmatrix} 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX33_2: Third order with delays estimated from step responses

$$na = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX22_2: Second order with delays estimated from step responses

$$na = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

$$nb = \begin{bmatrix} 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX11_2: First order with delays estimated from step responses

$$na = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARXA3V1_1:

$$na = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

ARXA3V1_2:

$$na = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

ARXA3V1_3:

$$na = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

ARXA3V1_4:

$$na = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARXA3V3_1:

$$na = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

ARXA3V3_2:

$$na = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

ARXA3V3_3:

$$na = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

ARXA3V3_4:

$$na = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX11S_1: Remove inputs TDiss(H&K), StdCentroid and StdRolloff, then fit ARX11

$$na = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX21S_1: Remove inputs TDiss(H&K), StdCentroid and StdRolloff, then fit ARX21

$$na = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX31S_1: Remove inputs TDiss(H&K), StdCentroid and StdRolloff, then fit ARX31

$$na = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX11S_2: Remove three inputs, and dependence of arousal on valence

$$na = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX21S_2: Remove three inputs, and dependence of arousal on valence

$$na = \begin{bmatrix} 2 & 2 \\ 0 & 2 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX31S_2: Remove three inputs, and dependence of arousal on valence

$$na = \begin{bmatrix} 3 & 3 \\ 0 & 3 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

At this stage, the 1st order autoregression models appear to be better than the 2nd and 3rd order autoregression models. Therefore, only 1st order models will be considered.

ARX11S_3:

$$na = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 2 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 2 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 2 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX11S_4:

$$na = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}$$

ARX11S_5: Adjust parameters to see which ones are statistically significant

$$na = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 0 & 2 & 0 & 2 & 1 & 2 & 2 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 2 \\ 2 & 0 & 2 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 2 \end{bmatrix}$$

$$nk = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

ARX11S_6: Adjust parameters to see which ones are statistically significant

$$na = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 0 & 3 & 0 & 2 & 0 & 3 & 2 & 0 & 2 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 2 \\ 2 & 0 & 3 & 2 & 2 & 0 & 2 & 0 & 0 & 0 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 3 \end{bmatrix}$$

$$nk = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

ARX11S_7: Try to remove parameters statistically equivalent to zero

$$na = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 0 & 2 & 0 & 2 & 1 & 2 & 2 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 2 & 2 & 1 & 0 & 2 & 0 & 0 & 0 & 2 & 1 & 1 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

$$nk = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The confidence intervals are very large. SDissS parameters are no longer statistically significant from zero.

ARX11S_8: Same as ARX11S_7 but without SDissS

$$na = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 0 & 2 & 0 & 2 & 1 & 2 & 0 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 2 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 1 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

$$nk = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

ARX11S_9: To overcome the large output confidence intervals of the previous models, try combining the arousal from ARX11S_6 with the valence of ARX11S_4

$$na = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$nb = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 2 & 0 & 3 & 2 & 2 & 0 & 2 & 0 & 0 & 0 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 3 \end{bmatrix}$$

$$nk = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

D.2.3 Final Model Residual Evaluation

Table D.2 compares the autocorrelation function of the residuals for ARX21S_2 and ARX11S_6. Ideally, no lags are significantly different than zero (lag 0 is not included because it is constant).

Table D.2: ACF residuals of best models

Appraisal	ARX11S_6 Significant Lags	ARX21S_2 Significant Lags
Allegro - Arousal	1,2	—
Allegro - Valence	1,2	1,2
Aranjuez - Arousal	1	—
Aranjuez - Valence	1,2,3,4,5,6,7,8,12,13	1,2,4,5,6,7,12
Fanfare - Arousal	1	—
Fanfare - Valence	1	—
Moonlight - Arousal	1	—
Moonlight - Valence	—	—
Morning - Arousal	1,10	5,10
Morning - Valence	—	—
Pizzicato - Arousal	—	—
Pizzicato - Valence	—	—