SPACIOUSNESS IN RECORDED MUSIC: HUMAN PERCEPTION, OBJECTIVE MEASUREMENT, AND MACHINE PREDICTION

Andy M. Sarroff

Submitted in partial fulfillment of the requirements for the
Master of Music in Music Technology
in the Department of Music and Performing Arts Professions
in The Steinhardt School
New York University
Advisor: Dr. Juan P. Bello
2009/05/05

ACKNOWLEDGMENTS

Deep gratitude goes to my advisor, Juan P. Bello, for his tireless dedication to his students, good research, and creative thinking. Our weekly group meetings were as valuable as anything that I have learned in the classroom. I thank Agnieszka Roginska for pointing me in the right direction at NYU. Her early guidance has had more impact on me than any other's at NYU. My peers, including Ernest Li, Jan Hagevold, Arefin Huq, Zeeshan Lakhani, Loreto Sanchez, Makafui Kwami, Adam Rokhsar, Aron Glennon, and Tae Min Cho, who have shown an enthusiasm for our community which I hope to keep close to me. The remaining core of the Music Tech program at NYU—Robert Rowe, Kenneth Peacock, Mary Farbood, and Panos Mavromatis—you have each contributed greatly to my academic success. Sarah Freidline and other close friends, thank you for helping me find the transition from recording studio to research. And finally, my parents, Alan and Eileen, and my sister, Amanda, who are the best family anyone could have.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Music making and music listening are cross-cultural human activities that pervade every aspect of daily life. We hear music intentionally, e.g. when we attend a music concert or turn on the radio; and we experience it passively when we are subjected to commercial advertisements, or the person sitting next to us has their headphones turned up. Musical activities are culturally ubiquitous—there is no record of modern human society without some form of music. And music has been with us for a long time—archeological evidence suggests that humans may have been building musical instruments as many as 45,000 years ago (Kunej & Turk, 2000).

Musical activity recruits neural resources inter-hemispherically and across the human brain, including centers for pleasure, movement, visuospatial tasks, emotion, pitch processing, and memory. We have distinct circuitry for managing non-speech, musical auditory stimuli. Some researchers find (uniquely to our species) a strong evolutionary basis for the the musical disposition that we find ourselves having (Levitin, 2006).

It is difficult to come up with a singular definition for music, but one popular explanation, by Edgard Varèse, is that it is "organized sound" (Clayson, 2002). If that is so, than we can understand music as an organization of basic perceptual components. When we compose, perform, and record music, we aim to execute explicit control over the organization and interpretation of its components. For these reasons, we must understand what they are, and how they are important to us.

Can we canonically organize music into distinct perceptual dimensions? If so, can stimuli that act upon our perceptual dimensions be qualitatively and quantitatively evaluated? And do people exhibit enough consistency in reaction that we can predict human perception? To answer the first question, Levitin (2002) identifies 8 separable musical attributes that we perceive—loudness, pitch, contour, duration, tempo, timbre, spatial location, and reverberation—which are often organized into higher level concepts of musical hearing, such as key. And to answer the second and third questions, we have methods of evaluating stimuli and predicting response to some, but not all, perceptual dimensions. For instance, (Suzuki & Takeshima, 2004) define an objective measurement of loudness based upon sound-pressure level and hearing experiments collected from 12 countries. Yet their "equal-loudness contours" only explain our perception of pure-tones, not complex musical stimuli. So it is with many higher-level concepts of musical hearing; as the stimulus get more complex (and more musical), a robust model of perception is increasingly difficult to build. Yet this does not negate the value of building such models.

To exploit Varése's definition further, music, in order to be interpreted as such and not noise, requires skilled execution of organization. Good musicians organize the basic perceptual components of music to form (usually enjoyable) higher-level impressions such as mood, color, emotive valence, and space. It is the job of recording engineers and music producers to faithfully transfer a musician's expression to a static medium. And furnished with expert knowledge of music theory, acoustics, and signal processing technology, they optimize this process to elicit and manipulate desirable musical impressions from listeners.

One such impression is auditory spaciousness—the concept of type and size of an actual or simulated space (Blauert & Lindemann, 1986). The perception of space is an important component to the way humans hear recorded music;

engineers and producers capture, manipulate, and add spatial cues to provide robust impressions of simulated acoustic spaces, whether intentionally natural or unnatural sounding. During recording, the character and extent of these spatial cues are controlled through means such as relative placement of microphones, performers, and reflecting surfaces. When mixing, engineers control the character and extent of spatial attributes through means such as source mixing, digital signal processing, and multichannel panning. The artful handling of these cues creates novel and enjoyable experiences for the listener. For instance, Västfjäll, Larsson, & Kleiner (2002) have shown that reverberation time in recordings influences emotional interpretation of music. The management and manipulation of recorded and synthesized spatial cues are a necessary and important step in music production.

Yet the concept of spaciousness in recorded music has not been treated explicitly in terms of the questions posed above. We do not know whether it is heard consistently by humans; we do not have an objective means of measuring spaciousness in recorded music; and, to the best of this author's knowledge, no study has attempted to predict perceptual response to spaciousness for music recordings. Here, an answer to these questions is attempted.

More specifically, this paper answers these questions from a Music Information Retrieval (MIR) perspective. MIR systems perform analyses upon symbolically-represented music or music recordings and retrieve human-relevant information about them. In its entirety, this work presents a complete system for retrieving a stream of perceptually meaningful information (spaciousness) from its digital recording. The paper will show that humans perceive the spaciousness of music recordings in a consistent fashion. It will present two new signal analysis techniques to measure spatial information in recorded music. And it will

demonstrate a means of mapping subjective experience to objective measurements of musical recordings.

The approach of the paper is outlined in Figure I.1 and is organized as follows: The next chapter (II) will provide detail on which dimensions of spaciousness have been studied previously, and how those works relate to this one. Based on those studies, the concept of spaciousness will be modeled as an aggregation of three nonorthogonal dimensions of perception. In Chapter III, a data set of musical recordings is built and a human subject study is executed to collect quantitative ratings on spaciousness for recorded music along the three dimensions. The results are examined for their consistency, individual correlation to demographic factors, and cross-correlation. Chapter IV proposes two objective measurements of digital signal for spaciousness. These are empirically validated in an experimental framework. Finally, machine learning is used to predict perceived spaciousness by mapping the subjective data collected in Chapter III to objective measurements, including the ones proposed in Chapter IV. Concluding remarks and future work are laid out in Chapter VI.

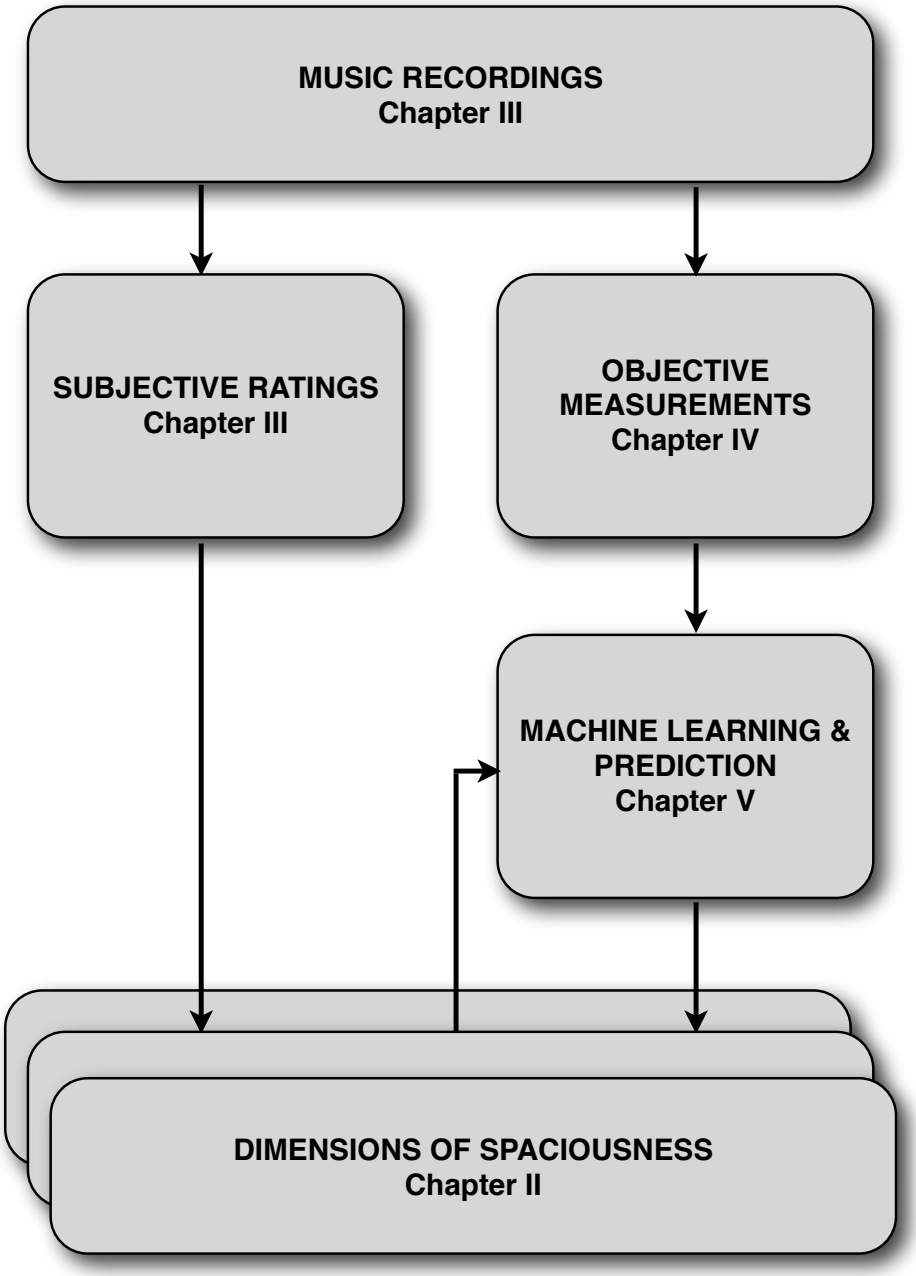


Figure I.1: Framework for predicting perceived spaciousness of music recordings.

CHAPTER II

SPACIOUSNESS

The spaciousness of musical recordings is not a well-defined concept. Casual conversation about a musical recording often leads to such comments as, "The lead singer sounded far away," or "That mix sounded really large." Yet, to my knowledge, there have been no empirical investigations into what perceptual attributes lead to such space-related comments for music recordings. When a person describes their listening experience in such a way, what exact electro-acoustic properties of the recorded signal bring about their response, and what are the specific perceptual components that inform such decisions? Because one of the goals of this paper is to answer the first question, a satisfactory answer to the second must be obtained. For the answer, this work turns to research in two related domains—natural acoustics and audio quality.

Natural Acoustics

In natural acoustics, researchers question what the physical properties are that lead some listening environments to sound better than others.
In 1967, Marshall determined that "spatial responsiveness," is a desirable property of concert halls. By analyzing echograms and architectural drawings of two dissimilar rooms, he concluded that good spatial responsiveness arises from well-distributed early reflections of the direct sources. After Marshall, spaciousness in music halls was parameterized by two distinct dimensions: Apparent Source Width (ASW) (Keet, 1968), and later, Listener

Envelopment (LEV) (M. Morimoto & Maekawa, 1989; M. Morimoto, Fujimori, & Maekawa, 1990). The first has consistently been attributed to early lateral reflections and the latter to the late arriving sound in an acoustic space. While the terms have been distinguished by different labels and varying definitions, they have more or less been used to describe the same distinct phenomena throughout. (For a brief overview of the development and semantic meanings of the terms ASW and LEV, I recommend Marshall & Barron, 2001.)

Despite minor differences in interpretation across studies, the perceptual dimensions of ASW and LEV can be defined thusly:

> Apparent source width (ASW) is the apparent auditory width of the sound field created by a performing entity as perceived by a listener in the audience area of a concert hall.
> . . .
> Listener envelopment (LEV) is the subjective impression by a listener that (s)he is enveloped by the sound field, a condition that is primarily related to the reverberant sound field. (Okano, Beranek, & Hidaka, 1998)

In natural acoustic environments, the relative positions of sound sources to each other, the relative positions of sound sources to a listener, the listener's and sources' relative positions to the surfaces of the listening environment, and the physical composition of the structures that form and fill the listening environment are each factors that contribute to ASW and LEV. Because ASW and LEV are experienced in a linear, time-invariant system (a "live" listening environment), the transfer function for various source-listener relationships can be captured and analyzed for spatial impression. There have been many such objective measurements for each. The inter-aural correlation function is usually used to measure ASW (Barron & Marshall, 1981; Okano et al., 1998; Vries, Hulsebos, & Baan, 2001; M. . Morimoto & Iida, 2005), for a refutal (Mason, Brookes, & Rumsey, 2005), and varying measurements of late arriving energy are used for

LEV (Bradley & Soulodre, 1995a,b; Furuya, Fujimoto, Young Ji, & Higa, 2001; Barron, 2001; Evjen, Bradley, & Norcross, 2001; Hanyu & Kimura, 2001; M. Morimoto, Jinya, & Nakagawa, 2007). ASW and LEV provide not only well-defined semantic meanings for perceived spaciousness in"live" listening environments, but a means of studying their relationship to measurable quantities in the physical world.

Audio Quality

Spaciousness has been a focal point of research for audio quality evaluation, especially for multi-channel sound reproduction systems. Such systems, like Surround Sound, create a virtual representation of spatial sound out of a discrete number of audio channels. Because the quality of these systems hinge on the believability and enjoyability of the display, researchers must have an empirical system for qualitative evaluation. Investigators must know the dimensions of spaciousness that are most important to human listeners for any meaningful evaluation of sound quality for spatial reproduction systems. Experiments with various attribute elicitation techniques have been reported, including Repertory Grid Technique and non-verbal techniques (Rumsey, 1998; Berg & Rumsey, 1999; Mason, Ford, Rumsey, & Bruyn, 2001; Ford, Rumsey, & Bruyn, 2001; Ford, Rumsey, & Nind, 2003b,a, 2005). And commonly elicited attributes have been analyzed with respect to preference of reproducing system, sound stimulus, and factor analysis (Berg & Rumsey, 1999, 2000, 2001; Zacharov & Koivuniemi, 2001; Rumsey, 2002; Berg & Rumsey, 2003; Guastavino & Katz, 2004; Choisel & Wickelmaier, 2007).

Attribute	**Description**
Naturalness	How similar to a natural (i.e. not reproduced through e.g. loudspeakers) listening experience the sound as a whole sounds.
Presence	The experience of being in the same acoustical environment as the sound source, e.g. to be in the same room.
Preference	If the sound as a whole pleases you. If you think the sound as a whole sounds good. Try to disregard the content of the programme, i.e. do not assess genre of music or content of speech.
Low frequency content	The level of low frequencies (the bass register).
Ensemble width	The perceived width/broadness of the ensemble, from its left flank to its right flank. The angle occupied by the ensemble. The meaning of "the ensemble" is all of the individual sound sources considered together. Does not necessarily indicate the known size of the source, e.g. one knows the size of a string quartet in reality, but the task to assess is how wide the sound from the string quartet is perceived. Disregard sounds coming from the sound source's environment, e.g. reverberation—only assess the width of the sound source.

Attribute	Description
Individual source width	The perceived width of an individual sound source (an instrument or a voice). The angle occupied by this source. Does not necessarily indicate the known size of such a source, e.g. one knows the size of a piano in reality, but the task is to assess how wide the sound from the piano is perceived. Disregard sounds coming from the sound source's environment, e.g. reverberation—only assess the width of the sound source.
Localisation	How easy it is to perceive a distinct location of the source—how easy it is to pinpoint the direction of the sound source. Its opposite is when the source's position is hard to determine—a blurred position.
Source distance	The perceived distance from the listener to the sound source.
Source envelopment	The extent to which the sound source envelops/surrounds/exists around you. The feeling of being surrounded by the sound source. If several sound sources occur in the sound excerpt: assess the sound source perceived to be the most enveloping. Disregard sounds coming from the sound source's environment, e.g. reverberation—only assess the sound source.
Room width	The width/angle occupied by the sounds coming from the sound source's reflections in the room (the reverberation). Disregard the direct sound from the sound source.

Attribute	Description
Room size	In cases where you perceive a room/hall, this denotes the relative size of that room.
Room sound level	The level of sounds generated in the room as a result of the sound source's action, e.g. reverberation—i.e. not extraneous disturbing sounds. Disregard the direct sound from the sound source.
Room envelopment	The extent to which the sound coming from the sound source's reflections in the room (the reverberation) envelops/surrounds/exists around you i.e. not the sound source itself. The feeling of being surrounded by the reflected sound.

Table II.1: Most common spatial attributes reported by Berg & Rumsey (2003).

Berg & Rumsey (2003) review the collective results of this research, and attributes that they have found to be the most important are reprinted in Table II.1. They note that evaluating reproduced sound quality necessitates higher demarcation of perceptual attributes than for live sound because spatial representations in reproduced sound are often intentionally fictional, not purposed to accurately depict the physical world. Their fundamental findings are that attributes referring to space are judged differently from those that deal with the sources; perception of room properties might be perceived in two dimensions—one which leads to a sense of being in the room, and another which deals with room characteristics, such as size; and spatial dimensionality can be globally categorized into dimensions of width, sensations of being present in the room, and distance to the source. They make the suggestion that the width

dimension observed in their studies might be similar to the ASW of natural acoustics, and that presence in the room might be similar to LEV.

Recorded Music

For research in natural acoustics and audio quality, there is an implicit need to understand how spaciousness affects the quality of their respective systems. An underlying similarity exists between these goals and the ones of this paper. But, importantly, the objectives of this paper diverge from those fields in that here the reproduced content is under evaluation, rather than the reproducing system.

This paper borrows from the literature of both fields for identifying salient spatial dimensions, and in doing so, focuses on three relations between listener and music—the source group relation, the environment relation, and the global relation. These embody 3 of the 4 basic categories that Rumsey (2002) declares in his "scene-based" paradigm for subjective evaluation of spatial quality (the last relates to individual sources). Specifically, the concept of spaciousness is modeled in this paper as an aggregated interaction between the "width of the source ensemble," the "extent of reverberation", and the "extent of immersion" that a listener perceives (Table II.2). At the outset, however, this paper makes no explicit assumptions about the orthogonality of these dimensions. They may be perceived in parallel, and perception of one may influence perception of the others.

The width of the source ensemble is a listener-source group relation. It describes the listener's perception of how widely the entire group of sources is representative in the sound field, irrespective of any room characteristics. This dimension lies closest to the "ensemble width" dimension in Table II.1 and is believed to be similar to ASW. The extent of reverberation is a listener-environment relationship, in which the listener perceives the overall

• The “width of the source ensemble” of a sound is how widely spread the ensemble of sound sources appears to be. • The “extent of reverberation” of a sound is the overall impression of how apparent the reverberant field is. • The “extent of immersion” is how much the sound appears to surround one’s head.

Table II.2: Definitions of learning concepts.

reverberation of the room. This is most closely related “room sound level” in Table II.1 and is believed to be one of the chief contributing factors to LEV (Okano et al., 1998). The last dimension considered, extent of immersion, is a global relation in which the listener perceives spaciousness as a macro assemblage of micro factors and can be considered a combination of “source envelopment” and “room envelopment.” The three dimensions have been chosen for their simplicity, overlapping treatment in natural acoustics and audio quality evaluations, and their monotonically increasing scene-based representation of source, environment, and global scene.

CHAPTER III

HUMAN PERCEPTION

In order to build a predictive function for spaciousness (Chapter V), a reliable "ground truth" for spaciousness is needed. As such a ground truth has not been previously established for evaluation of spaciousness in recorded music, this work necessitated the creation and annotation of one. This chapter explains how musical recordings were selected and segmented. It then describes two related experiments in which humans were asked to rate musical recordings for spaciousness. The results of the experiments are analyzed for statistical robustness.

Music Selection and Segmentation

All songs were selected from a single online music web site[1]. The web site is a free service that allows musicians to disseminate their work to the public in Mp3 format. As a large repository of free music, the web site allowed careful selection of appropriate recordings. Music was picked with the following criteria in mind: It should be representative of several genres; it should be unfamiliar, so as to avoid bias by recognition; it should represent the major parts of a song, i.e. verses, choruses, etc; the audio quality of the recordings should not be sub-par; and it should encompass widely varying degrees of spaciousness.

In order to satisfy the first criterion, songs were selected from and equally distributed across each of the popular genre categories on the site. These were: "Alt/Punk," "Classical," "Electronic-Dance," "Hip-Hop," "R&B/Soul," and

[1]Mp3 Music Downloads, `http://www.mp3.com/`

"Rock/Pop." The genre-label for each song had been selected by the artist who uploaded the song. There was therefor high variability in interpretation of genre across the songs. This was deemed as a positive side-effect, as it increased the broadness of the data set's genre representation.

None of the songs that were picked were commercially distributed on a large scale. Therefor they were each likely to be unfamiliar to most listeners. In order to satisfy the third criterion, a segment was chosen from each song so as to fall into either a "verse," "chorus," or "bridge" section. Sections were determined as verse if they contained novel lyrical content, and chorus sections were deemed such if they contained repeated lyrical content. Any section that did not contain lyrics or that encompassed a major shift in structure (e.g. a key change) was deemed a bridge. Twice as many bridge sections were included as verses and choruses so as to have a roughly equal number of lyrical and non-lyrical sections. No song segments were chosen from the beginnings or endings of songs.

The third and fourth criteria were satisfied by careful screening of each song amongst hundreds. If a song's audio quality was comparable to that of a commercially-distributed Mp3's, it was marked as appropriate for inclusion. The selection of songs that were chosen had varying degrees of source panning, from monophonic to very wide, and many levels of auditory spatial cues. Each song selection was segmented to be exactly seven seconds long, with a 50 ms fade-in and fade-out to avoid clicks. The duration was chosen, by informal evaluation, to be long enough to develop concrete impressions of spaciousness, yet short enough to prevent much temporal variation in spaciousness within the excerpt.

Experiment

Two experiments were conducted on the assembled database—one online and one in a laboratory. The experiments were similar in nature and goal; the first targeted a larger subject base, at the acknowledged cost of poorly controlled experimental conditions. The second optimized experimental control at the cost of subject pool size. The results of the second experiment were first used to substantiate the quality of the results from the first experiment and were thereafter combined with the results of the first experiment to finalize the annotated "ground truth" data set of music recordings. The materials and methods of each are explained below and followed by analysis.

Materials and Methods, Online Experiment

Subjects

Subjects were recruited by posting advertisements on nearly twenty online forums for musicians and music producers. Specific forums were targeted so as to recruit a high proportion of experienced listeners. The advertisement summarized the nature of the experiment and instructed interested parties to visit the experiment's web site. The experiment was approved by the New York University Committee on Activities Involving Human Subjects; by beginning the experiment, the participants acknowledged informed consent of the experiment.

There were 78 total participants across both studies. Their demographic data is summarized in Table III.1. Online participants, of which there were 58, varied in age from approximately 18 to 65 years of age and were distributed across 19 countries. They had varying degrees of experience regarding working or

studying in a music related field and were dispersed in the number of hours a day they spent listening to music.

Experimental Conditions

Before participants began the online experiment, they were informed that they were to use headphones. The first screen encountered (after entering some basic personal information) was a headphone calibration screen, where a series of simple tones were played to facilitate volume adjustment.

The next four screens were designed to train the participant for the experiment (see Appendix A for screen shots). First, a definition of the term "spatial attributes" was given. Next, participants were informed of which components in the sound field they were to listen for. Then, explicit definitions of the attributes they were to rate were given. For these screens, participants were asked to listen to a non-musical mixture of sources (a room of applause) in order to focus their hearing. This training phase was designed to give participants time to familiarize themselves with the concepts and focus their listening on a simple stimulus. The nonmusical recordings exhibited characteristics of the spatial dimensions but, to avoid pre-biasing their judgments of spaciousness, participants were not told how spacious the recordings were to be perceived. Finally, after training, a sample page with a real musical example was given.

Subjects were then asked to rate, on a bipolar 5-ordered Likert scale from "Less" to "Neutral" to "More," each of the dimensions for each test song. Participants were allowed easy access out of the experiment at any time via a button in the corner of the screen. An informational button activated a pop-up screen with the term definitions, in the case that a participant needed to be reminded. The experiment proceeded until all 50 song excerpts were played, or the

		Online	Laboratory
Gender	M	45	13
	F	13	7
Age Range	18-25	21	13
	26-35	23	7
	36-45	5	0
	46-55	5	0
	56-65	4	0
Country of Res.	US	40	20
	Non-US	18	0
Native English Speaker	Y	46	17
	N	12	3
Work in Music	Y	35	20
	N	23	0
Years in Musical Field	<5	6	7
	5-10	13	7
	11-20	4	5
	21-30	5	1
	31-40	6	0
	N/A	24	0
Hours Listening/Day	<.5	1	2
	.5-1	11	2
	1-2	14	8
	2-4	18	7
	4-8	11	1
	8-12	1	0
	>12	2	0
Usually Listen Through	Headphones	14	9
	Headphones & Speakers	28	6
	Speakers	16	5
Critical Listening Ability	1	N/A	1
	2	N/A	1
	3	N/A	3
	4	N/A	8
	5	N/A	7
TOTAL PARTICIPANTS		**58**	**20**

Table III.1: Demographics of subjects from the two experiments.

participant exited. The order of the songs was randomized so as to eliminate any order bias across participants. A web browser cookie-tracking mechanism prevented any subject with their browser cookies enabled from participating more than once.

Materials and Methods, Laboratory Experiment

Subjects

Subjects were recruited by posting advertisements on several email lists targeted to music technology and music performance university graduate and undergraduate students. The advertisement summarized the experiment and offered a small compensatory fee for completing the experiment. A total of 20 subjects were recruited for this experiment. The experiment was approved by the New York University Committee on Activities Involving Human Subjects; before beginning the experiment, signed consent forms were obtained.

The subject pool's demographics (see Table III.1) were rather homogenized compared to the online experiment. Participants were distributed over a smaller age range, they were all US residents, and they were each active workers in a music related field. These subjects were asked to rate their level of critical-listening ability on a scale of 1 to 5. Most subjects rated themselves highly, at 4 or 5.

Experimental Conditions

The experimental conditions were very similar to the ones in in the online experiment, with a few key differences. These participants were compensated; in order to receive their payment, they were required to rate all 50 song excerpts in

the data set. All participants took the test (at staggered times) in the same room using the same model of high-fidelity open back headphones, Sennheiser HD650. In addition, participants had the benefit of an experiment investigator on hand to precisely answer questions about the terms in the experiment. The average time it took for laboratory subjects to complete the experiment was roughly 30 minutes.

Post-Processing and Outlier Removal

The results of the two experiments were combined into one data set, providing 2,523 ratings over 50 songs and three dimensions of spaciousness. Ratings were transformed from a Likert space to a numerical space by assigning the 5-ordered response categories integer values of -3 to 3. Any rating for a song and dimension that exceeded three standard deviations was deemed an outlier and removed from the data set. Additionally, any participant that had outliers for more than one song in a dimension was removed entirely from the dimension. In total, 119, 140, and 128 ratings were removed from the width, reverberation, and immersion dimensions respectively. After outliers were removed, the ratings for each dimension were standardized to zero mean and unit variance. By doing so, the trends of the ratings for each dimension were preserved, while at the same time shifting them into a standardized space for easy cross-comparison. Figure III.1 shows the sorted mean value and standard deviation in response for each song for the three standardized dimensions. It can be seen that, after standardization, responses were skewed to the negative range, reflecting compensation for a larger quantity of positive responses. It is not clear if this is due to a tendency for subjects to rate selections more positively, or if this reflects the true nature of the distribution of spaciousness in the data set.

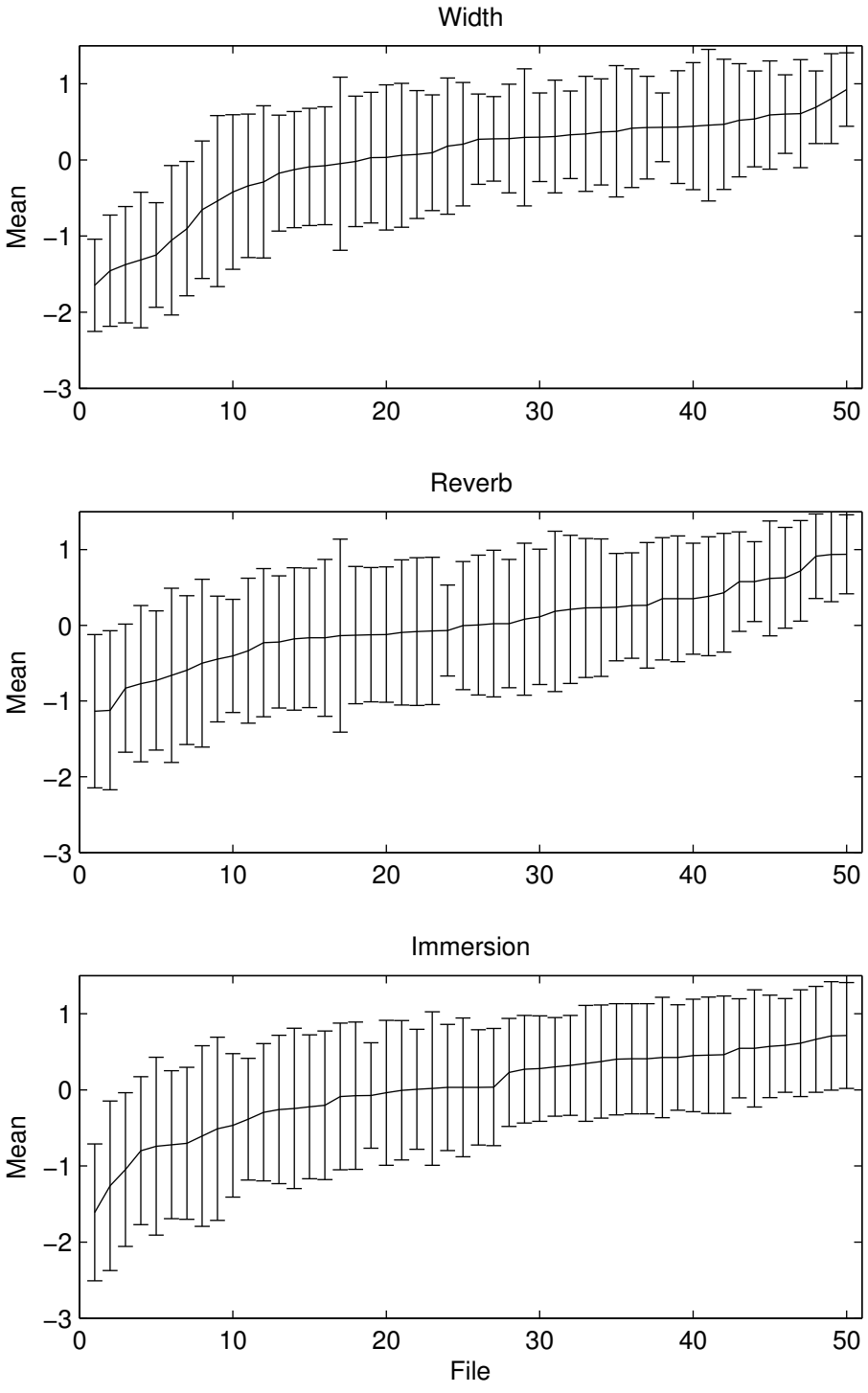


Figure III.1: The means and standard deviations of ratings for each song for each dimension of spaciousness. The songs are sorted by ascending mean response, and each dimension has been standardized for easy comparison.

Results

Pair-Wise T-Tests

A pair-wise T-test was computed for each song and each dimension to test the null hypothesis that the average ratings for the laboratory and online experiments share the same means. Since different experimental conditions were being compared, the p values were calculated assuming unequal variance, implementing Satterthwaite's approximation for standard error. The results are shown in Table III.2. The null hypothesis can be rejected at a 99% confidence level for only 2 songs, highlighted in grey.

Similar T-tests were conducted, per dimension, on the entire data set comparing three different demographics. The first was subjects who listen to more than 4 hours of music a day versus those who don't. The null hypothesis could not be rejected for any songs or dimensions. The second test was between subjects who work or study in a music-related field versus those who don't. In that test, there was a single song in the immersion dimension which was deemed to not share the same mean between populations. In the third test, those who usually listen to music through headphones were compared to those who usually listen to music through speakers. In this case, there were two instances of a rejected null hypotheses, both in the immersion dimension. These three tests were conducted at the 99% confidence level and with an equal variance assumption.

File	**Width**	**Rev**	**Imm**
bridge_10_Classical	0.852	0.858	0.923
bridge_11_Classical	0.655	0.889	0.897
bridge_11_ElecDance	0.592	0.108	0.494
bridge_11_RnBSoul	0.541	0.785	0.076

File	Width	Rev	Imm
bridge_12_Classical	0.678	0.247	0.137
bridge_12_HipHop	0.976	0.066	0.034
bridge_13_AltPunk	0.157	0.835	0.001
bridge_13_HipHop	0.363	0.606	0.067
bridge_14_Classical	0.718	0.188	0.365
bridge_15_Classical	0.837	0.373	0.699
bridge_16_ElecDance	0.413	0.425	0.064
bridge_18_Classical	0.409	0.877	0.707
bridge_18_ElecDance	0.379	0.102	0.550
bridge_1_ElecDance	0.150	0.194	0.846
bridge_22_Classical	0.833	0.150	0.540
bridge_2_RnBSoul	0.544	0.602	0.024
bridge_2_RockPop	0.018	0.233	0.411
bridge_3_AltPunk	0.534	0.165	0.307
bridge_3_Classical	0.685	0.256	0.751
bridge_5_RnBSoul	0.290	0.346	0.994
bridge_5_RockPop	0.123	0.247	0.294
bridge_6_HipHop	0.092	0.112	0.178
bridge_7_AltPunk	0.965	0.753	0.334
bridge_8_RockPop	0.441	0.227	0.144
bridge_9_RnBSoul	0.238	0.702	0.156
bridge_9_RockPop	0.771	0.523	0.032
chorus_10_AltPunk	0.536	0.818	0.439
chorus_11_HipHop	0.229	0.293	0.607
chorus_11_RockPop	0.833	0.462	0.845
chorus_12_AltPunk	0.849	0.375	0.121

File	Width	Rev	Imm
chorus_14_HipHop	0.176	0.427	0.485
chorus_20_ElecDance	0.933	0.605	0.119
chorus_2_HipHop	0.067	0.160	0.123
chorus_3_ElecDance	0.537	0.014	0.001
chorus_4_RockPop	0.665	0.242	0.181
chorus_6_AltPunk	0.852	0.787	0.219
chorus_7_RnBSoul	0.163	0.557	0.977
chorus_8_RnBSoul	0.321	0.576	0.034
verse_10_RnBSoul	0.734	0.269	0.362
verse_14_ElecDance	0.700	0.539	0.337
verse_15_ElecDance	0.190	0.406	0.143
verse_1_AltPunk	0.139	0.197	0.382
verse_1_HipHop	0.429	0.705	0.490
verse_3_RockPop	0.583	0.195	0.316
verse_4_AltPunk	0.876	0.538	0.313
verse_5_HipHop	0.144	0.257	0.735
verse_6_RnBSoul	0.222	0.381	0.981
verse_6_RockPop	0.353	0.513	0.305
verse_9_AltPunk	0.307	0.616	0.862
verse_9_ElecDance	0.832	0.481	0.045
Mean	0.499	0.426	0.389

Table III.2: p values calculated from pair-wise T-tests between online and laboratory experiments for each song and dimension. The null hypothesis is rejected at the 99% confidence level for two songs in the immersion dimension (highlighted in grey). The average of all T-tests for each dimension is shown at the bottom.

	Width	Rev	Imm
Laboratory	12.74	8.17	10.31
Online	17.79	7.39	9.49
All	29.66	14.59	18.74

Table III.3: F-values calculated for each dimension for each experiment and for both experiments.

	Width-Rev	Width-Imm	Rev-Imm
R	0.3186	0.8745	0.5679

Table III.4: Pearson's correlation coefficient *R* for averaged ratings between dimensions.

F-Statistic for Each Dimension

It was important to determine if the ratings between songs, for each dimension, were statistically different from each other. The F-test, which is the ratio of between-group variability to within-group variability was conducted on each dimension, the groups being the songs. A higher F-value indicates greater distance in ratings between songs. F-values were calculated independently for each experiment and for the data set comprising both experiments[2]. The results of the test are shown in Table III.3.

Correlation Between Dimensions

Finally, a measure of the cross-correlation in ratings between dimensions was needed. The subjective ratings were averaged for each song, and the Pearson's correlation coefficient *R* was calculated between dimensions. These coefficients are reported in Table III.4.

[2]The calculation of the F-value is dependent on the sample size. The F-value for the entire data set is therefor not meant to be compared directly to the F-values for the online and laboratory subsets.

Discussion

The inter-experiment T-test was important to determine if the the online experiment was robust compared to the laboratory experiment. It can be expected that the ratings in the online experiment would be less stable, as there was no way to control the experimental conditions for each participant. In fact, the average variance per song was consistently lower in the laboratory experiment. Only two instances out of 150 were rejected as sharing the same means between experiments. This is promising evidence that the full data set, including noisier data collected online, can be reliable for prediction of spaciousness. The additional T-tests were included to test if any specific variability would arise from demographic factors. It can be hypothesized that ratings from those who have more listening experience would be statistically different from those who have less. Again, the data set proves fairly robust with a statistical difference arising in only one instance (for a comparison between those who work and those who don't work in music). It may may be questioned whether subjects would rate songs consistently if presented the same song more than once. However, this analysis was deemed beyond the scope of the experiments' purpose. Additionally, enforcing multiple presentations of the same song would risk increased ear fatigue for the subjects.

One concern of the subjective experiments is whether the constraint of headphones would adversely affect the reliability of ratings. Headphone-listening can inhibit perceived externalization, a factor that might negatively affect perceived spaciousness. However, this paper aims to investigate the spaciousness of recorded music. In order to do so, any unrelated environmental acoustic factors of the listening environment must be eliminated from the experimental framework. If headphone-inhibited externalization affects perceived spaciousness, it can be

hypothesized that subjects that listen to music predominantly through headphones will be better-adapted to perceive differences in spaciousness. Therefor, T-tests were conducted on that population against participants that predominantly listen to music through speakers. The T-tests indicated only two instances, again for the immersion dimension, of a rejected null hypothesis. Collectively, the results of these T-tests indicate a robust data set for prediction tasks.

The F-statistics reported also indicate a robust data set. The p values of the group song means (not reported here) for each dimension indicated that they were statistically significant. The F-values, from which the p values are calculated, show that the width dimension has the greatest inter-song distance in rating variance, while the reverberation dimension has the least inter-song distance.

Finally, the R values of inter-dimensional correlation gives us some indication of whether the dimensions are perceived independently. Because width and immersion are highly correlated, it might be said that listeners perceive the two dimensions similarly. Or, conversely, it might be that production decisions that lead to wider mixes also lead to similar decisions to increase, in parallel, the extent of immersion. Similarly, the low width-reverberation correlation might reflect true orthogonality of dimensions, or it might be influenced by higher-level production choices.

CHAPTER IV

OBJECTIVE MEASUREMENT

Two independent mathematical models for two attributes of produced music that might correlate with the way humans perceive the spaciousness of recorded music are proposed here. Spaciousness is quantitatively modeled as a function of (1) the width of the source ensemble in a stereophonic field and (2) the level of overall reverberation in a musical sample. The models consider the stereophonic digital signal, rather than reproduction format or listening environment. The models are validated in a controlled experimental framework.

Source Width

This work is concerned with modeling components of music production that may be attributable to spatial perception for stereophonic music. As shown in Chapters II an III, music may be perceived as more or less spatial based upon the perceived wideness of sources. This model, using the azimuth discrimination strategy reported by Barry, Lawlor, & Coyle (2004) as its basis, blindly estimates through L-R magnitude scaling techniques how widely a mixture of sources is distributed within the stereo field. (The term azimuth is loosely used here to describe the virtual placement of a musical source in the horizontal plane by amplitude panning.) The source panning distribution model generates an azimuthal histogram of sources, and a musical sample's wideness of panning is estimated by calculating the full width half maximum value of a gaussian curve that is fit to the histogram.

As in Barry et al., it is assumed that the stereo signal is the weighted sum of J individual sources S_j, such that:

$$x_l(n) = \sum_{j=1}^{J} wl_j(n) S_j(n)$$

and (IV.1)

$$x_r(n) = \sum_{j=1}^{J} wr_j(n) S_j(n)$$

where x_l and x_r are the left and right signals, w_l and w_r are the left and right weighting coefficients, and n are discrete time samples. The source signal weight of J can also be represented as a left-right intensity ratio:

$$g_j = \frac{wl_j}{wr_j}$$

If g_j can be estimated for each source, then the wideness of panning can be estimated for the entire distribution of sources. To do this, phase cancellation is used to estimate panning intensity ratios for signal spectra. First, a set of arbitrary scaling coefficients is created:

$$g(i) = i \times \frac{1}{\beta}$$
$$i = \{0, 1, 2, \ldots, \beta\} \quad \text{(IV.2)}$$

where i is an azimuthal index, β is the azimuthal resolution for each channel, and both are integer numbers. Then, the magnitude spectrograms of the signals are calculated, $|X_l|$ and $|X_r|$, and arrays of frequency-azimuth planes, Az_l and Az_r are built. For every FFT frame m, $N/2$ frequency bins of each channel are scaled and

subtracted from the other channel by the scaling coefficients g:

$$\begin{aligned} Az_l^m(k,i) &= |X_r(k) - g(i) \cdot X_l(k)| \\ Az_r^m(k,i) &= |X_l(k) - g(\beta - i) \cdot X_r(k)| \end{aligned} \tag{IV.3}$$

where k is the frequency bin index, and N and M are the length of the FFT analysis window and number of FFT frames, respectively. The redundant azimuthal bin $Az_r^m(k,0)$ is discarded and the two arrays are concatenated to form array $Az^m(k,u)$ with azimuthal indices $u = [1,2,\ldots,(2\times\beta-1)]$.

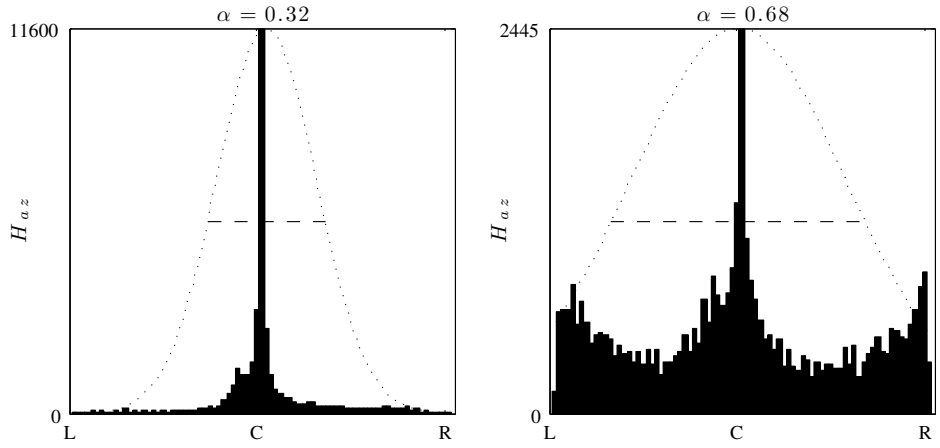
Only the maximal bins are of interest, so Az is filtered as follows:

$$\hat{A}z^m(k) = \begin{cases} max(Az^m(k)) - min(Az^m(k)) & if\ Az^m(k) = max(Az^m(k)) \\ 0 & otherwise \end{cases} \tag{IV.4}$$

From here, an azimuthal histogram of the analysis signal is built by summing the azimuthal bin values across all frames and all frequencies and weighting them by their indices:

$$H_{\hat{A}z}(u) = u\left(\sum_{m=0}^{M-1}\sum_{k=0}^{N/2-1}\hat{A}z^m(k,u)\right) \tag{IV.5}$$

Figure IV.1 shows azimuthal histograms for center-panned and a wide-panned distributions of sources, along with their estimated distributions. As can be seen, the azimuthal histograms tend to approximate normal distributions. When sources are more focused toward the center of the stereo field, the distribution exhibits less standard deviation. When sources are wider panned, the standard deviation is higher. The width of a statistical distribution with a single peak can be simply characterized by its Full Width Half Maximum (FWHM) value, or the distance between two half-maximal points in the distribution. The extent of source panning is estimated by calculating the FWHM of the data as if it



(a) Center-panned (b) Wide-panned

Figure IV.1: Source width estimation for center- and wide-panned guitars amongst a mixture of sources. Frame histograms have been fit with a gaussian curve and their Full Width Half Maxima are calculated to estimate α. Note: Y axes are not the same scale.

were a normal distribution and normalizing it by the total azimuthal resolution:

$$\alpha = \frac{\mu(H_{\hat{A}z}) \pm \sigma(H_{\hat{A}z})\sqrt{2\ ln\ 2}}{2 \times \beta - 1} \tag{IV.6}$$

Inspection of Figure IV.1 reveals that the gaussian fit for the left figure is wider than for the right, indicating a wider ditribution of sources.

Reverberation

In this section, a model for the blind estimation of the total reverberation of a musical sample is proposed. Reverberated musical sounds might be less linearly predictable than non-reverberated sounds, as uncorrelated signal causes spectral whitening in the temporal and frequency domains. As such, the residual of a linear predictor is used as the engine for the estimations. Linear prediction has been used previously in related applications such as blind de-reverberation (Gillespie,

Malvar, & Florencio, 2001) and source separation of speech (Kokkinakis, Zarzoso, & Nandi, 2003).

The model begins by mono-summing the input audio signal. If x_l and x_r are the left and right channels, then $x = (x_l + x_r)/2$. Then, p linear prediction coefficients are generated on non-overlapping blocks of audio and an excitation signal is filtered with the linear prediction coefficients:

$$\hat{x}^{m_{lpa}}(n) = y(n) - a_1 y(n-1) - \cdots - a_p y(n-p) \tag{IV.7}$$

where m_{lpa} is the linear prediction analysis frame index, n is a discrete time sample, a_i are the linear prediction coefficients ($i \in [0, p]$), and y is an excitation signal. The residual is calculated from the linear predictor and the frames are concatenated:

$$e(n) = x(n) - \hat{x}(n) \tag{IV.8}$$

As can be seen in the top graphs of Figures IV.2 and IV.3, the spectrum of the residual has plenty of high-frequency energy. The envelope of the residual is characterized as:

$$\hat{e}^{m_{env}} = \frac{\sum_{n=0}^{N_{env}-1} |e^{m_{env}}(n)|}{2N_{env}} \tag{IV.9}$$

where m_{env} is the envelope frame index and N_{env} is the size of the analysis window of the residual. As the smoothing window effectively down-samples the data, it is up-sampled with an interpolating filter by a factor of η to facilitate further processing. The up-sampled residual envelope is then transformed into the frequency domain and its log magnitude power is calculated so that $\hat{E}^{m_{fft}} = 20 \cdot \log(|\hat{E}^{m_{fft}}|)$, where m_{fft} is the FFT frame index. The middle graphs of Figures IV.2 and IV.3 show that the high frequency spectra of the envelopes of the residual for the non reverberated signal contain more power than for the reverberated signal. In order to characterize this feature, an arbitrary power

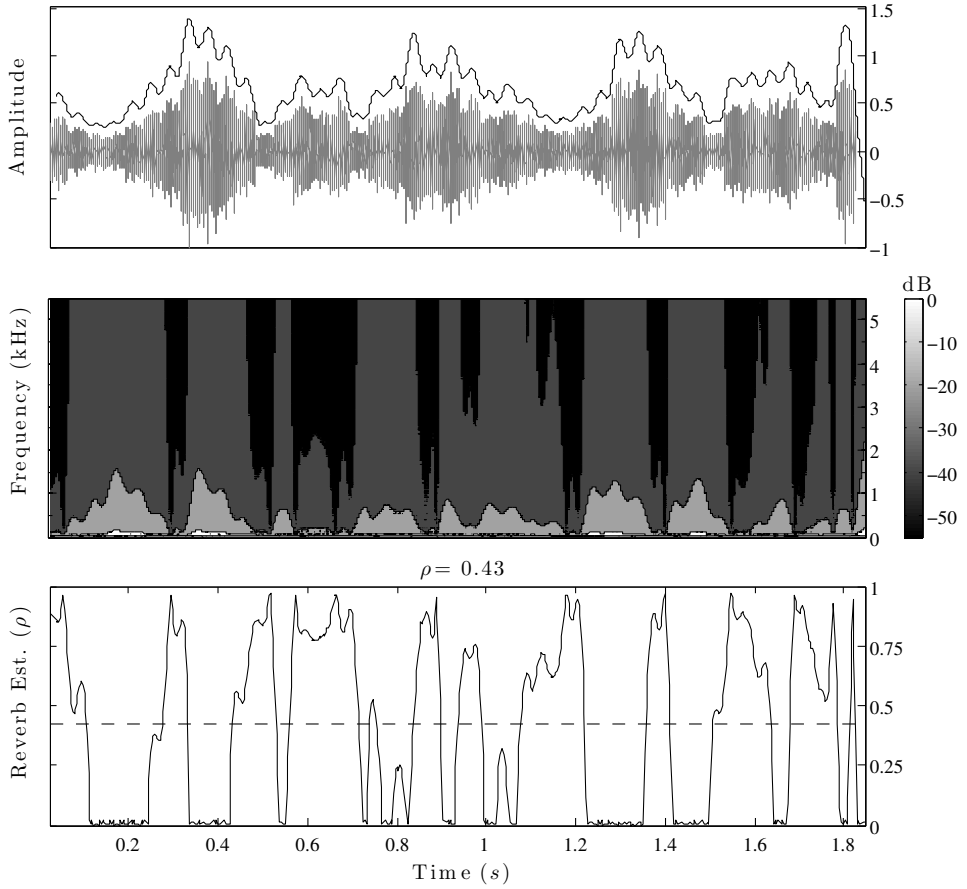


Figure IV.2: Comparison graphs for a non reverberated signal. Top: Linear predictor residual and its envelope. Middle: Frequency transform of the residual envelope. Bottom: Normalized maximum frequencies below power threshold γ and their mean, ρ.

threshold γ is decided upon. For each FFT frame of $\hat{E}$, the highest frequency bin index which contains approximately $\gamma\, dB$ of power is found. The mean of the resulting curve is calculated:

$$\rho = \frac{\sum_{m_{fft}=0}^{M_{fft}-1} max(\hat{E}_{fft}^{m}(n) \leq \gamma)}{M_{fft}} \tag{IV.10}$$

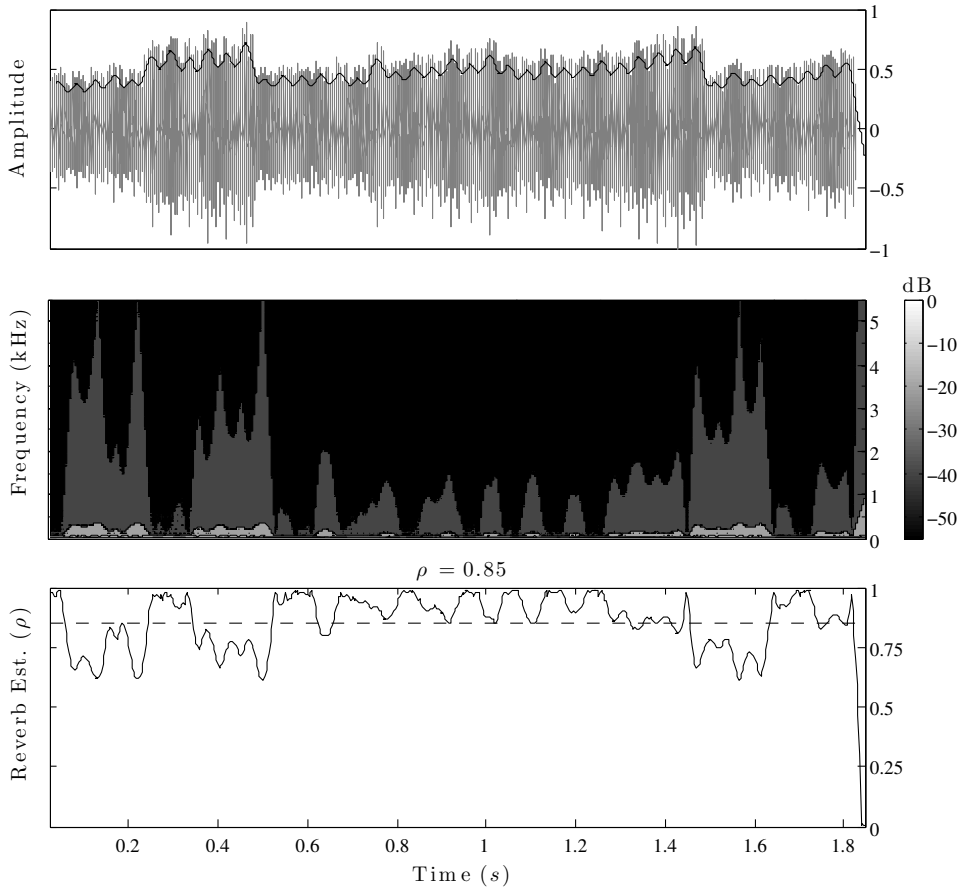


Figure IV.3: Comparison graphs for a reverberated signal. Top: Linear predictor residual and its envelope. Middle: Frequency transform of the residual envelope. Bottom: Normalized maximum frequencies below power threshold γ and their mean, ρ.

A normalization constant ν derived from the signal sampling rate (fs), the hop size of the envelope follower (N_{ehop}), and η is created:

$$\nu = \frac{fs}{2 \times N_{ehop} \times \eta} \tag{IV.11}$$

Finally, the output is normalized and subtracted from 1 so that an increasing estimator value indicates an increasing amount of reverberation:

$$\rho = 1 - \rho / \nu \tag{IV.12}$$

The bottom graphs of Figures IV.2 and IV.3 show ρ, the reverberation estimation for an analysis frame. Again, the figures represent two similar music clips. In the first, the guitars have no artificial reverberation added. In the second, artificial reverberation with a wet mix setting of -10 dB has been added to the guitars. It can be seen that the estimated reverberation is higher for the second.

Experiment

The models presented in the previous two sections were tested independently in controlled experiments. The estimators were each tested on multiple data sets; and each data set was tested under two conditions. The data sets and experimental methods are explained below, followed by results and a discussion.

Materials and Methods

Data Sets

Each data set consisted of mixed control and test tracks of musical audio. Data Set 1 was the chorus of a pop song, approximately 13 s in length. The instrumentation consisted of drums, bass, percussion, male vocals, electric guitar, and acoustic guitar. Data Set 2 was the chorus of a hip hop song, approximately 22 s in length. Its instrumentation consisted of kick drum, snare drum, percussion, bass, piano, synthetic horns, and assorted samples and sound effects. The last data set, Data Set 3 (approximately 13 s), was an electronica excerpt. Its tracks were comprised of several percussive loops, synthetic bass, several synthesizer pads, a synthesizer lead, and some effects tracks.

Each of the audio tracks for each data set were categorized as either "test tracks" or "control tracks." In the first experimental condition, the test tracks for Data Sets 1, 2, and 3 were acoustic guitar and electric guitar; doubled male lead vocal; and synthetic bass, respectively. In the second condition, the test tracks of Data Sets 1, 2, and 3 were acoustic guitar and electric guitar; snare drum; and lead synth pad, respectively.

Digital Audio Workstation (DAW)

The experimental conditions were implemented on a popular consumer-brand DAW. The workstation had virtual pan-pots for controlling the placement of sound sources. Panning values reported below reflect the MIDI numbers assigned to the virtual pan pots. For example, a MIDI value of "64" represents a center panned channel, and "127" a hard right panned channel.

Reverberation was implemented with a virtual insert on the DAW. A popular consumer-brand reverberation software plugin was used on a “warm space” setting with a reverb decay time of approximately 3 s and a pre-delay of approximately 16 ms.

Methods

In the first condition, two test tracks were iteratively panned from opposite outermost to center positions. The panning positions of the control tracks remained static in all iterations. The control tracks of all data sets were mostly, but not entirely, center-panned.

In the second condition, the wet mix control of the reverb plugin was iteratively lowered in 6 dB decrements on one or two test tracks. The reverb type remained constant through all iterations and for all data sets. The dry mix remained constant in all iterations. Reverb was monophonic in this experiment. (This would not affect results, as the estimator mono-sums the input signal.) Some control tracks were reverberant, either from the acoustic environment they were recorded in, or from preprocessing on mix stems. However, the extent of reverberation on the control tracks remained constant in all iterations. The lead synth pad in Data Set 3 had been preprocessed with synthetic reverberation; the track was tested, however, under the same conditions as the other test tracks.

All experiments were conducted with the parameters described in Tables IV.1 and IV.2 on 2-second windows of stereophonic music with a 50% overlap.

VARIABLE	SYMBOL	VALUE
sample rate	fs	44,100 Hz
FFT length	N	2048 samples
FFT overlap		hanning
FFT overlap		50%
channel azimuthal resolution	β	20

Table IV.1: Variable symbols and values used for source width estimation α.

VARIABLE	SYMBOL	VALUE
sample rate	fs	44,100 Hz
linear prediction frame size	N	2048 samples
linear prediction window		boxcar
linear prediction overlap		0%
number of linear prediction coefficients	p	20
excitation signal	y	white noise
envelope follower frame size	N_{env}	$N/2$
envelope follower window		hanning
envelope follower overlap		50%
up-sample factor	η	$N/16$
FFT length	N_{fft}	N
FFT window		hamming
FFT overlap		87.5%
power threshold	γ	-35 dB

Table IV.2: Variable symbols and values used for reverberation estimation ρ.

Results

Figure IV.4 shows the results of the source width estimator on the the three data sets. All data sets show decreasing estimations for decreasing panning widths. Additionally, the estimations are consistent with each other in the temporal domain. The estimations show relative values across sets that were consistent with the relative mixing intensities of the test tracks amongst the control tracks. Note that in Data Set 3, the range of estimation values is highly compressed relative to the other data sets. (The Y axis of the figure has been expanded to improve resolution.)

The results of the reverberation estimator are depicted in Figure IV.5. All data sets show decreasing estimations for decreasing reverberation. However, the estimator loses its ability to detect changes in reverberation at different levels for different data sets. For each data set, the figure shows the last iteration at which the estimator clearly predicted a change in reverberation level. For Data Set 1, this was at a wet mix level of -34 dB. For Data Set 2, it was -28 dB, and for data Set 3 -22 dB. The estimator's predictions in the temporal domain do not respond linearly with decreasing reverberation. For instance, at about 12 s in Data Set 2, a decrease in reverberation is estimated at -28 dB, but a slight increase is estimated at -22 dB. Test Set 3 performed worse than other test sets, detecting considerably less change in reverberation level than the other test sets.

Discussion

The temporal consistency of the source width estimator can be expected, as a change in intensity ratio at sample *n* should not affect intensity ratios in later frames. Likewise, it is possible to explain the lack of temporal consistency for the reverb estimation. Decreasing the wet mix parameter of a reverb with 3 s of reverb decay would probably affect the following analysis frames.

The "compression" of panning width estimation noted for Data Set 3 is probably due to the spectral characteristics of the test track, which was a bass. An instrument with fewer high frequency components would not be well represented in the linear time-frequency histogram that the estimator uses. There was a wide-panned hi hat loop in Data Set 3 that stops playing towards the end of the section. This is reflected in the graph, as the estimator slopes downward after approximately 8 s. The estimator was thus highly dependent on instrumentation with stronger high-frequency spectra. It might be appropriate to weight the

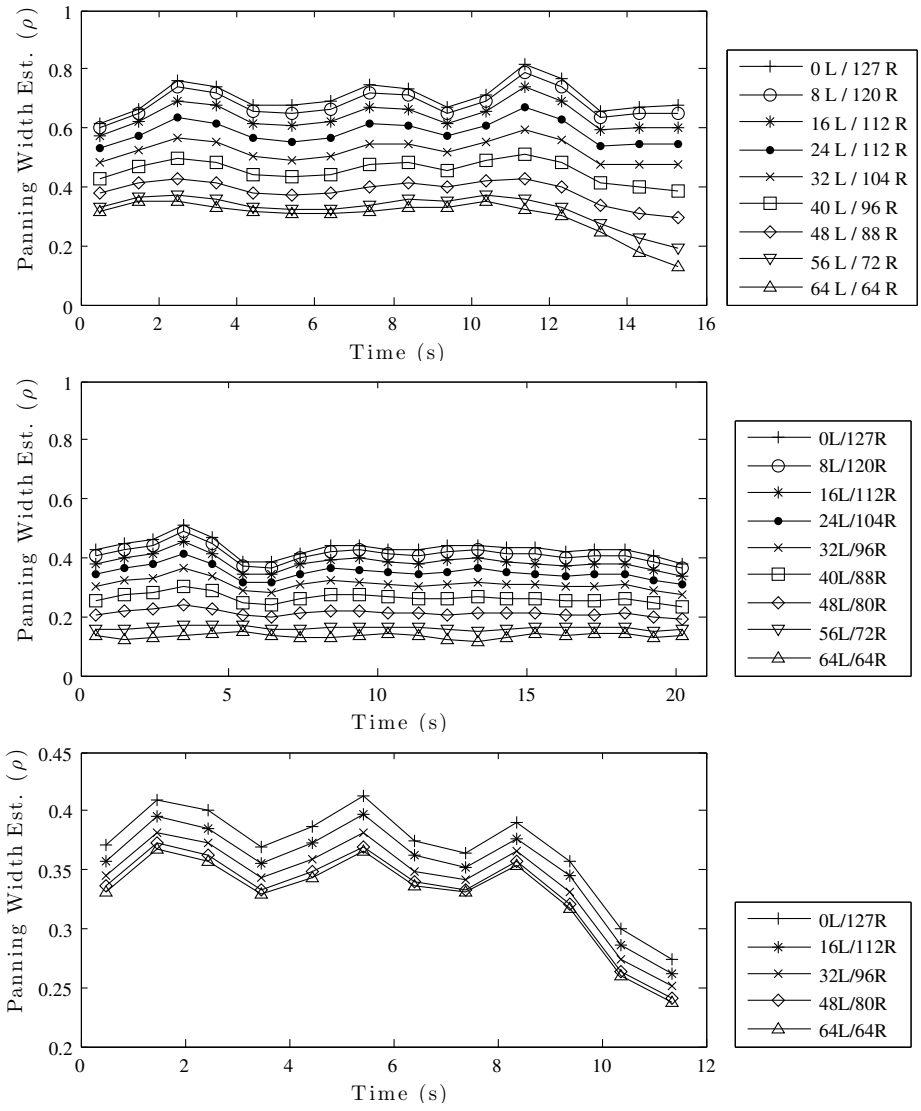


Figure IV.4: Source width estimation of three experimental data sets. Top: Data Set 1; Middle: Data Set 2; Bottom: Data Set 3. Note: Bottom graph is not to same scale as others.

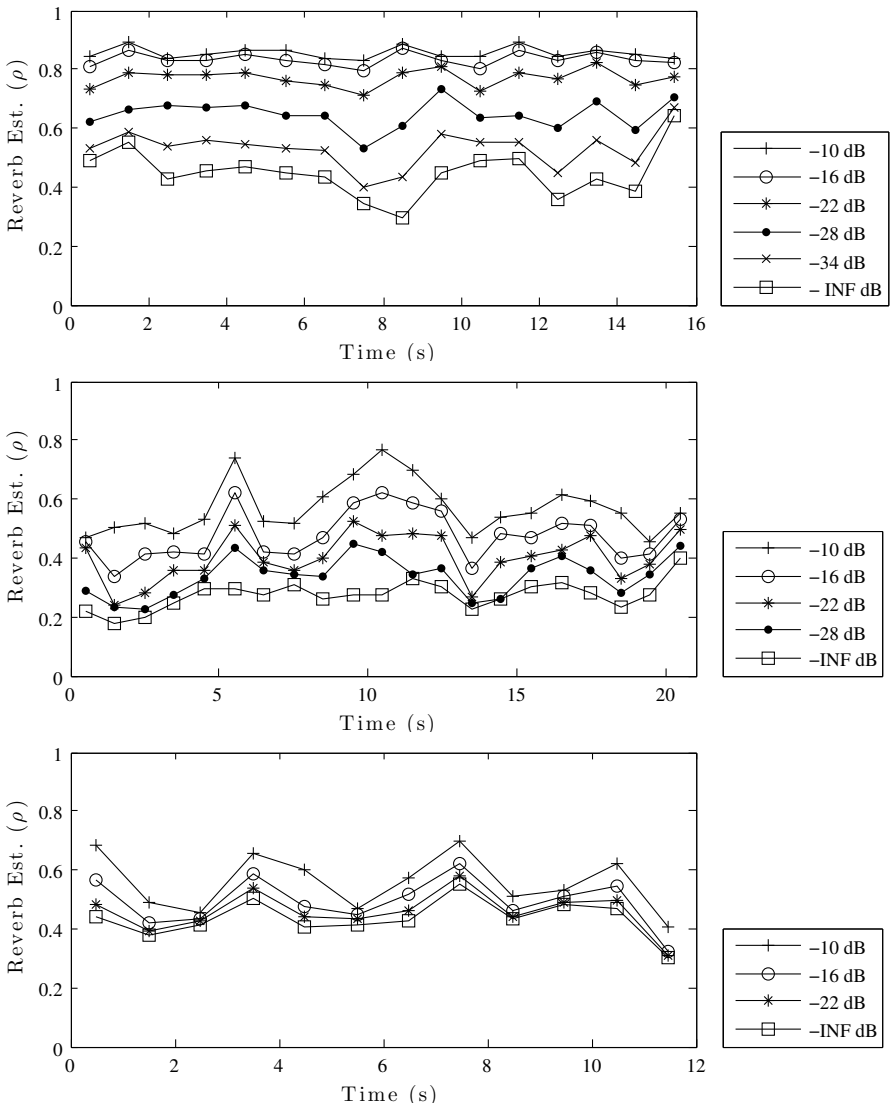


Figure IV.5: Reverberation estimation of three experimental data sets. Top: Data Set 1; Middle: Data Set 2; Bottom: Data Set 3.

frequency component of the time-frequency histogram logarithmically, so that low frequency components are more accurately represented.

Although the reverberation estimator ceased to detect changes in reverberation at different wet mix levels for different data sets, informal subjective listening tests revealed that reverberation was less perceivable in those data sets. For instance, the test and control tracks of Data Set 3 had been preprocessed with more reverberation than any of the other data sets, making additional reverberation more difficult to distinguish. In general, Data Sets 1 to 3 were increasingly dense in instrumentation and fluctuations of loudness. Despite absolute wet mix values across all data sets, reverberation was perceived less in denser sets. Further investigation needs to be done on the relationship between perception of reverberation and these other parameters.

It is important to note that the test conditions for reverberation estimation excluded multiple types of reverberation. The spectral and temporal characteristics of reverberation can vary wildly across many reverberation types. Different reverberations would almost certainly affect the results of these experiments. Further investigation needs to be done on the dependencies of the model upon the spectral and temporal characteristics of reverberation.

CHAPTER V
MACHINE PREDICTION

This chapter details the formulation of a mapping function between the ratings of the perceived spatial attributes obtained in Chapter III and objective measurements of digital audio, including the ones explained in Chapter IV. Since, to my knowledge, there are no extant objective measurements of recorded music for the concept of "spaciousness," the function must be newly created by machine learning. With the exception of listener experience, perceived attributes discussed in literature are consistently related to sound sources or their environment, rather than personal properties like gender. These are universal in nature and therefor support a model which maps spaciousness to objective measurements of the recorded signal. In the following sections, the components of the machine learning algorithm are discussed, followed by the results of an experiment which tests its validity.

Design of Machine Learning Function

A block diagram for building the objective-to-subjective mapping function is shown in Figure V.1. At the beginning is a large feature space that objectively describes the music recordings. At the end is a support vector machine that needs optimization to accurately predict subjective ratings. In between, a correlation-based feature selection and subset voting scheme are used to narrow down the feature space. Then, a grid search for the best parameterization of the

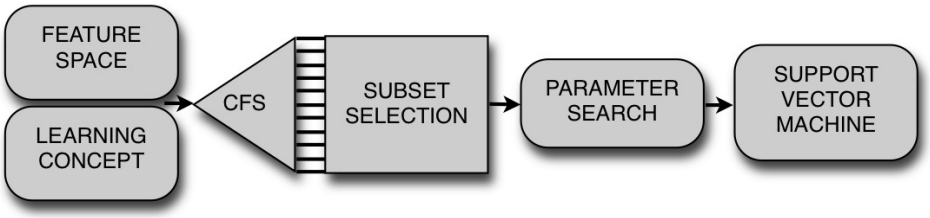


Figure V.1: Block diagram for building and optimizing the mapping function.

support vector regression function is conducted. Each stage is described in detail below.

Feature Generation

Features are descriptors of the audio signal obtained by signal filtering and analysis. By reducing an audio file to a set of audio features, one hopes to extract the most meaningful properties of the audio signal for the task at hand. For this project, a verbose set of attributes was batch-generated on the left-right difference signal of the data set using the MIR Toolbox (Lartillot, Toiviainen, & Eerola, 2008) and the two objective measurements reported in Chapter IV. The batch-generated features include many that are widely used, like MFCCs, Spectral Centroid, and Spectral Flatness. None of the features in the MIR Toolbox are intended to extract spatial features of a musical signal, like the ones presented in this paper. However, they are all initially included as it is unknown what characteristics of a signal might lead to perceived spaciousness.

For most features, the recording was frame-decomposed and feature extraction was performed on each frame. Some features, such as Fluctuation, were calculated on the entire segment. The frame-level features were summarized by their mean and standard deviation. Additionally, their periodicity was estimated by autocorrelation and period frequency, amplitude, and entropy was calculated. The

Category	Feature
Dynamics	RMS energy
Rhythm	Fluctuation Peak Position*, Fluctuation Peak Magnitude*, Fluctuation Spectral Centroid*, Tempo, Tempo Envelope Autocorrelation Peak Position, Tempo Envelope Autocorrelation Peak Magnitude, Attack Time, Attack Time Onset Curve Peak Position*, Attack Time Onset Peak Magnitude*, Attack Slope, Attack Slope Onset Curve Peak Position*, Attack Slope Onset Curve Peak Magnitude*
Timbre	Zero-Cross Rate, Spectral Centroid, Brightness, Spectral Spread, Spectral Skewness, Spectral Kurtosis, Roll-Off (95% threshold), Roll-Off (85% threshold), Spectral Entropy, Spectral Flatness, Roughness, Roughness Spectrum Peak Position, Roughness Spectrum Peak Magnitude, Spectral Irregularity, Irregularity Spectrum Peak Position, Irregularity Peak Magnitude, Inharmonicity, MFCCs, Δ MFCCs, $\Delta\Delta$ MFCCs, Low Energy*, Low Energy RMS, Spectral Flux
Pitch	Salient Pitch, Chromagram Peak Position, Chromagram Peak Magnitude, Chromagram Centroid, Key Clarity, Mode, Harmonic Change Detection
Spatial	Wideness Estimation*, Reverberation Estimation*
Summary	Mean, Standard Deviation, Slope, Period Frequency, Period Amplitude, Period Entropy

Table V.1: List of audio features and their categories. Features with an asterisk (*) only had their mean calculated.

size of the final feature space extracted from the recordings was 430 dimensions. The entire set of features, which can be sub-divided into categories of Dynamics, Rhythm, Timbre, Pitch, and Spatial, is listed in Table V.1.

Pre-Processing

The feature space was normalized to the range $[0, 1]$ and transformed into a principal components space. The non-principal components that accounted for the 5% least variance in the data set were discarded, and the data set was transformed back to its original symbolic attribute space. This transformation and reduction of

data by principal components analysis is an often-used means of performing data cleanup on a feature space (Witten & Frank, 2005).

Feature Selection

For each target concept, Correlation-Based Feature Selection (CFS) was performed with a greedy step-wise forward search heuristic. CFS chooses attributes that are well correlated to the learning target, yet exhibit low intercorrelation with each other. CFS has been shown to be good for filtering out irrelevant or redundant features (Hall, 1999).

However, supervised attribute selection can over-fit attributes to their learning concept when the same data set is used for training and testing (Miller, 2002). To minimize subset selection bias, a percentile-based voting scheme with 10×10-fold cross-validated attribute subset selection was performed. Multiple cross-validation (CV) is a robust way of estimating the predictive power of a machine when only a small data set is available. As each fold generated a different feature set, some features were selected more often than others. For each run, features were placed in a percentile bin based upon how many times that feature had been selected. Up to 11 new data sets with monotonically increasing feature spaces were generated in this way.

Each feature space was then used to learn a non-optimized support vector regression algorithm for each dimension. The subset that performed the best for each learning concept was voted as the final subset for further system optimization and training.

Regression

For each concept, a support vector regression model was implemented with the the Sequential Minimal Optimization (SMO) algorithm in Smola & Schölkopf, 2004. Support vector machines have shown to generalize well to a number of classification and regression tasks. Support vector machines implement a trade-off between function error and function flatness. An error threshold ξ is selected below which instance errors will be invisible to the loss function. A complexity constant C preserves the flatness of the function and prevents it from over-fitting the data. The higher the value of C, the more influence errors outside of ξ have upon the function. A kernel function generalizes the model to nonlinear fits. The SMO algorithm is a means of improving computational efficiency when analyzing large data sets. The data sets that were used in this work were relatively small, rendering SMO irrelevant to discussion.

The support vector model in this work employed a polynomial kernel, $K(x,y) = (< x,y > +1)^p$, chosen as the best in an informal kernel search. Support vector machines perform, to some extent, similarly well independent of kernel type if the kernel's parameters are well-chosen (Scholkopf & Smola, 2001). In the case of a polynomial kernel, the only parameter to choose is the polynomial exponent, p. An exhaustive grid search for the optimal values of the support vector machine complexity C and its kernel exponent p was conducted after the optimal feature space had been selected. The value of ξ was set at 1×10^{-3} for the entirety of this study.

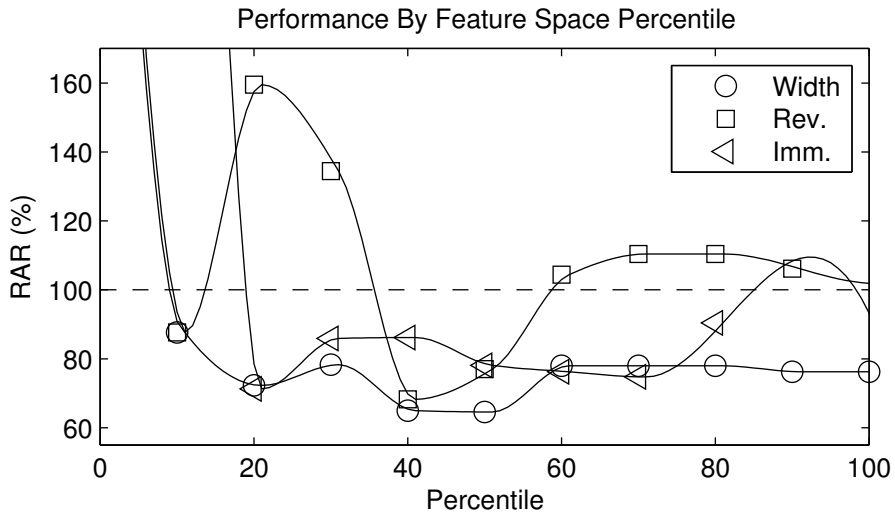


Figure V.2: Performance of non-optimized machine on monotonically decreasing feature spaces.

Experiment

Materials and Methods

Data Set

The averaged responses per song from Chapter III were used to train and test the learning algorithm after it was pre-processed as described above.

Computing Environment

All learning and training exercises were conducted on a Mac dual-core 2.4 GHz computer with 4 GB of memory on the Unix operating system. Machine training and testing was conducted in Weka, an open-source computing environment for machine learning (Witten & Frank, 2005).

Methods

For each dimension of spaciousness, the best feature space was found by using Multiple CV as described above. Then a systematic search for the support vector parameterization that yielded the lowest error for each concept was conducted. Success was evaluated by relative absolute error (RAE, explained in "Results"). The model that yielded the lowest RAE was retained and tested a final time, using Multiple CV, to obtain final results.

Results

The relative absolute error (RAE) was the primary error metric used to evaluate success. However, several secondary metrics were incorporated in evaluations, as well. RAE is the sum of all the errors normalized by the sum of the errors of a baseline predictor. The baseline predictor, Zero-R, picks the mean value of the test fold for every instance. An error of 0% would denote perfect prediction and 100% would indicate prediction no better than chance.

The final test results are depicted in Table V.2. The mean absolute error (MAE), which is dependent upon scale, was no more than 0.11 for any of the predictors. The average MAE for the Zero-R predictor is shown for comparison at the bottom of the table. All predictors had a correlation coefficient R of 0.73 or higher to the actual values. An R value of 0.0 would denote a complete lack of correlation between the predicted and actual values. The predictor for wideness of source ensemble performed the poorest, but was still well above chance. By all measurements of accuracy, the predictor for extent of reverberation performed the best. Its coefficient of determination (R^2) indicates that the function accounted for 62% of the variance in the test set.

	Width	**Rev.**	**Imm.**
RAE(%)	62.63	67.20	64.36
MAE	0.11	0.10	0.11
R	0.73	0.79	0.76
R^2	0.53	0.62	0.58
MAE (Zero-R)	*0.19*	*0.17*	*0.18*

Table V.2: The final mean absolute error (MAE), relative absolute error (RAE), correlation coefficient (R), and coefficient of determination (R^2) of the learning machines. The MAE for a baseline regression function, Zero-R, is given for comparison. All results are averaged from Multiple CV.

Discussion

The predictive capability of each of the mapping functions was much better than chance, as indicated by the RAE. The accuracies of the models suggest that objective measurements of digital audio can be successfully mapped to new dimensions of music perception. It is informative, however, to inspect the performance of the intermediate stages of model design. Figure V.2 shows the results of testing for the best feature space percentile. All predictors show two local minima: Width at the 20^{th} and 50^{th} percentiles; reverberation at the 10^{th} and 40^{th} percentiles; and immersion at the 20^{th} and 70^{th} percentiles. This indicates that there might have been more than one optimal feature subset percentile to use. In every case, the percentile that yielded the lowest RAE for the algorithm was chosen, without testing all local minima. The steepness of the error curves between the 0 and 10^{th} percentiles shows that simply using the entire feature set without any feature selection would greatly inhibit the performance of the support vector algorithm.

A summary of the final feature subset percentile used for learning each concept is shown in Table V.3. While most features are probably not individually useful, the correct combination of features is. Features that were selected for more

Concept (Percentile)	Features
Width (50 %)	**Tempo Envelope Autocorrelation Peak Magnitude Period Frequency**, **Spectral Flatness Period Amplitude**, **Wideness Estimation Mean**, **Reverb Estimation Mean**, Δ **MFCC Slope 5**, ΔΔ **MFCC Mean 11**
Rev. (40 %)	MFCC Mean 3, MFCC Period Entropy 3, MFCC Slope 3, ΔΔ MFCC Period Amplitude 13, Key Clarity Slope, Chromagram Peak Magnitude Period Frequency, Harmonic Change Detection Function Period Amplitude, Spectral Flux Period Amplitude, Pitch Period Amplitude, Δ MFCC Slope 10, Δ MFCC Period Frequency 10, Δ MFCC Slope 13
Imm. (20 %)	MFCC Period Entropy 6, Spectral Centroid Period Entropy, **Tempo Envelope Autocorrelation Peak Magnitude Period Frequency**, **Spectral Flatness Period Amplitude**, Spectral Kurtosis Standard Deviation, **Wideness Estimation Mean**, **Reverb Estimation Mean**, Mode Period Entropy, Pitch Period Frequency, Δ MFCC Slope 7, Δ **MFCC Slope 5**, Δ MFCC Slope 11, Δ MFCC Mean 11, ΔΔ**MFCC Mean 11**

Table V.3: Selected feature spaces after running on non-optimized machine. Features in boldface were picked for more than one learning concept.

than one learning concept are shown in boldface. Notably, the spatial estimators for wideness and reverberation were automatically chosen for the tasks of predicting source ensemble wideness and extent of immersion, but not for estimation of reverberation. This may denote a non-optimized parameterization of the reverberation measurement. The width and immersion dimensions shared the most features in common; this is understandable, as these dimensions shared the highest correlation among annotations (as reported in Chapter III). This may indicate that the dimensions are highly similar, that subjects assumed them to be the same, or that there exists a song-selection bias in the data set. Selected features for all three concepts were largely from the Timbre category. It is interesting that the reverberation predictor picked three features from the Pitch category. There are no obvious explanations for this behavior, and it merits further investigation.

The error surfaces for parameterizations of each of the machines is shown in Figure V.3. These surfaces show the RAE for each value in the grid search for optimal C and p values. It can be seen that the surfaces are not flat and that a globally optimal parameterization can be found for each. Yet they depict few local minima and are relatively smooth, suggesting that other parameter choices in between the grid marks would not have significantly improved results. It is worth noting that the flattest error surface, that for extent of reverberation, is also the one that performed the best, indicating robustness against parameter choices.

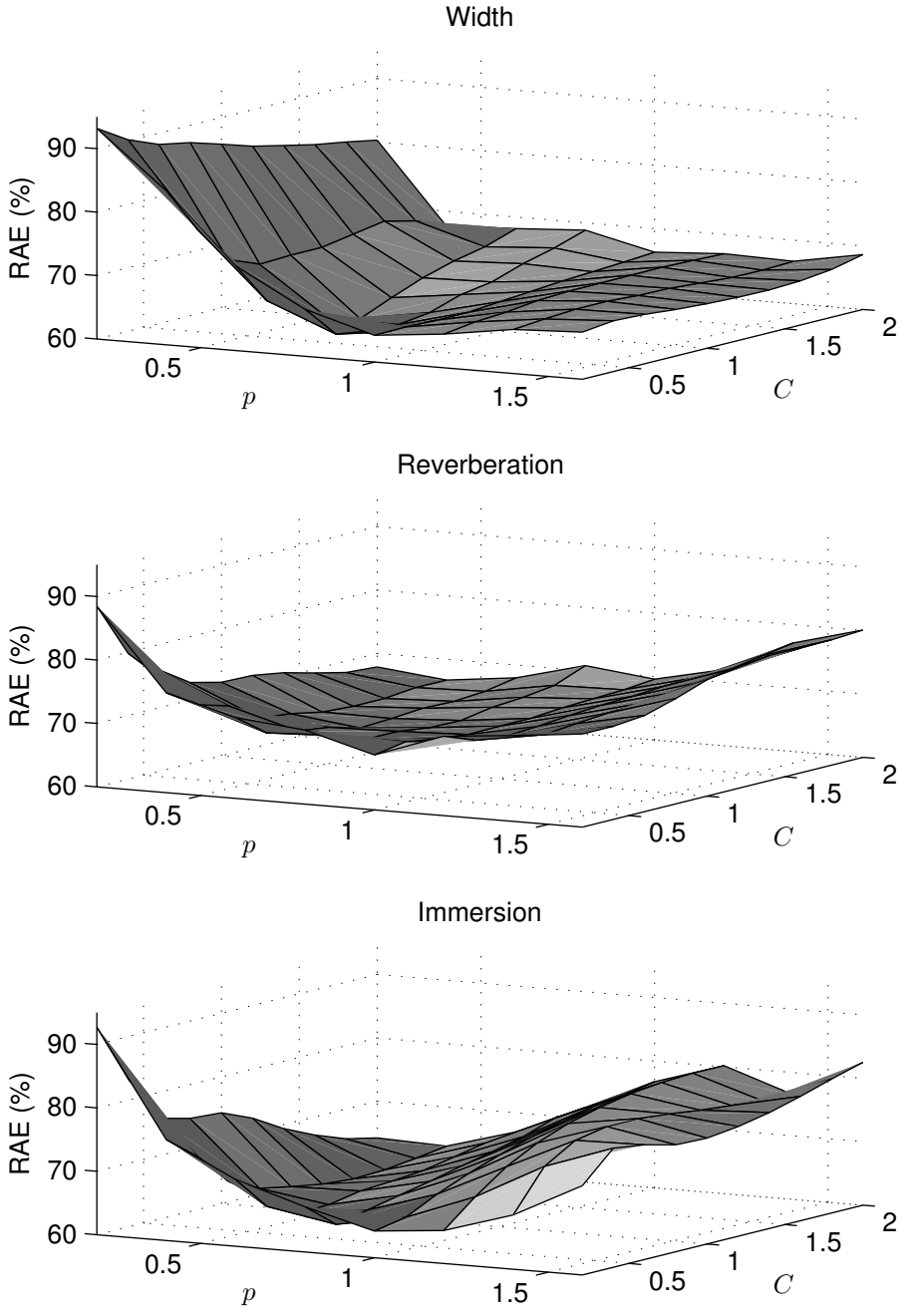


Figure V.3: Relative absolute error surface for machine parameter grid search of kernel exponent p and machine complexity C.

CHAPTER VI
CONCLUSIONS

This work presents a complete model for spaciousness in recorded music. First, the concept of spaciousness was discussed in context of previous work in other music-related fields. It was found that the spaciousness of a music recording could be parameterized by the width of its source ensemble, its extent of reverberation, and its extent of immersion—three dimensions which represent listener-source, listener-environment, and listener-global scene relationships, respectively. By doing so, each of these perceptual attributes could be studied independently, and in tandem.

A newly annotated set of music recordings was generated along the three dimensions of spaciousness. The annotations were compiled in two human subject studies. The first was conducted on a large population, at the acknowledged cost of experimental control. The second was conducted on a smaller population with increased experimental control. The results of the second test were used to validate the first. It was found, through pair-wise T-tests, that the first study was robust enough to include with the second to compile a complete set of annotations. Additionally, inter-popuation and inter-song T-tests showed that the data set was robust against demographic variations and that the set of musical recordings were statistically different from each other in ratings. It was concluded that the data set would be sufficient for accurate machine prediction.

Two new objective measurements were proposed for measuring spatial attributes of a recorded musical signal. The measurements predict the width of the

source ensemble and the extent of reverberation in a musical signal, respectively. Both algorithms were successfully validated in controlled experiments.

Lastly, a function was built to map the data set of music annotations to a large set of signal descriptors, including the two novel spatial descriptors introduced in this paper. Automatic feature selection was used in conjunction with exemplar-based support vector regression to build a mathematical model of spaciousness. The model was evaluated against the data set by Multiple CV and found to predict spaciousness at levels much better than chance.

This paper therefor concludes that perceived spaciousness of musical recordings can be effectively modeled and predicted along an arbitrary numerical continuum. These findings are significant because spatial impression is an important factor in the enjoyment of recorded music. Recording and mixing engineers stimulate attention to music by manipulating spatial cues. Novel spatial stimuli are often a major trait separating produced recorded music from strict documentation of a recorded performance, especially in the popular genres. By parameterizing an important perceived attribute of music and mapping it to measurable quantities of digital audio, a meaningful way of accessing and manipulating music is provided. By implementing a complete model of spaciousness for recorded music, musicians have another means of executing organization of sound. If we follow Varése's definition of music, we may argue that organizational capacity over sound is the single most important instrument of composition a musician can exercise.

Future work in several areas will improve the efficacy of this model. First, a larger data set, inclusive of more songs and human subjects will improve the model. A second human subject study in which humans evaluate the machine predicted values of spaciousness will bolster the model's validity.

The width estimator will benefit from a new frequency weighting which de-emphasizes the influence of higher frequency spectra. Further investigation into the performance and parameterization of the reverberation estimator for different types of reverbs is also warranted.

Lastly, this work examined one machine learning algorithm, support vector regression. Future work will evaluate the performance of other machine learning types, such as linear regression or support vector regression with different kernel functions.

REFERENCES

Barron, M. (2001). Late lateral energy fractions and the envelopment question in concert halls. *Applied Acoustics*, *62*(2), 185–202.

Barron, M., & Marshall, A. H. (1981). Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. *Journal of Sound and Vibration*, *77*(2), 211–232.

Barry, D., Lawlor, B., & Coyle, E. (2004, October 5-8). Sound source separation: azimuth discrimination and resynthesis. In *7th int. conference on digital audio effects (DAFX'04)*, Naples, Italy.

Berg, J., & Rumsey, F. (1999, May 8–11). Identification of perceived spatial attributes of recordings by repertory grid technique and other methods. In *106th AES convention*, Munich, Germany.

Berg, J., & Rumsey, F. (2000, September 22-25). Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction. In *109th AES convention*, Los Angeles.

Berg, J., & Rumsey, F. (2001, June 21–24). Verification and correlation of attributes used for describing the spatial quality of reproduced sound. In *AES 19th international conference: Surround sound – techniques, technology and perception*, Schloss Elmau, Germany.

Berg, J., & Rumsey, F. (2003). Systematic evaluation of perceived spatial quality. In *Proceedings of AES 24th international conference on multichannel audio*, Banff, Alberta, Canada.

Blauert, J., & Lindemann, W. (1986, Aug). Auditory spaciousness—some further psychoacoustic analyses. *Journal of the Acoustical Society of America*, *80*(2), 533–542.

Bradley, J. S., & Soulodre, G. A. (1995a, Apr). The influence of late arriving energy on spatial impression. *Journal of the Acoustical Society of America*, *97*(4), 2263–2271.

Bradley, J. S., & Soulodre, G. A. (1995b, Nov). Objective measures of listener envelopment. *Journal of the Acoustical Society of America*, *98*(5), 2590–2597.

Choisel, S., & Wickelmaier, F. (2007, Jan). Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA*, *121*(1), 388–400.

Clayson, A. (2002). *Edgard Varèse*. London: Sanctuary.

Evjen, P., Bradley, J. S., & Norcross, S. G. (2001). The effect of late reflections from above and behind on listener envelopment. *Applied Acoustics*, *62*(2), 137–153.

Ford, N., Rumsey, F., & Bruyn, B. de. (2001, May). *Graphical elicitation techniques for subjective assessment of the spatial attributes of loudspeaker reproduction – a pilot investigation.* (Presented at 110th AES Convention, Amsterdam, 12–15 May, Paper 5388)

Ford, N., Rumsey, F., & Nind, T. (2003a, Oct). *Creating a universal graphical assessment language for describing and evaluating spatial attributes of reproduced audio events.* (Presented at 115th AES Convention, New York, 10-13 October)

Ford, N., Rumsey, F., & Nind, T. (2003b, June 26-28). Evaluating spatial attributes of reproduced audio events using a graphical assessment language – understanding differences in listener depictions. In *AES 24th international conference*, Banff.

Ford, N., Rumsey, F., & Nind, T. (2005, May 28-31). Communicating listeners' auditory spatial experiences: a method for developing a descriptive language. In *118th convention of the audio engineering society*, Barcelona, Spain.

Furuya, H., Fujimoto, K., Young Ji, C., & Higa, N. (2001). Arrival direction of late sound and listener envelopment. *Applied Acoustics*, *62*(2), 125–136.

Gillespie, B. W., Malvar, H. S., & Florencio, D. A. F. (2001). *Speech dereverberation via maximum-kurtosis subband adaptive filtering.*

Guastavino, C., & Katz, B. F. G. (2004, Aug). Perceptual evaluation of multi-dimensional spatial audio reproduction. *JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA*, *116*(2), 1105–1115.

Hall, M. (1999). *Correlation-based feature selection for machine learning*. Phd thesis, University of Waikato, Department of Computer Science, Hamilton, New Zealand.

Hanyu, T., & Kimura, S. (2001, Feb). A new objective measure for evaluation of listener envelopment focusing on the spatial balance of reflections. *Applied Acoustics*, *62*(2), 155–184.

Keet, W. (1968). The influence of early lateral reflections on the spatial impression. In *Reports of the sixth international congress on acoustics*, Tokyo.

Kokkinakis, K., Zarzoso, V., & Nandi, A. (2003, April). Blind separation of acoustic mixtures based on linear prediction analysis. In *4th international symposium on independent component analysis and blind signal separation (ICA2003)*, Nara, Japan.

Kunej, D., & Turk, I. (2000). New perspectives on the beginnings of music: Archeological and musicological analysis of a middle paleolithic bone "flute". In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (chap. 15). Cambridge, Mass.: MIT Press.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). Mirtoolbox [Computer program and manual]. Internet web site. Retrieved 5/1/2009, from `http://www.jyu.fi/music/coe/materials/mirtoolbox`

Levitin, D. J. (2002). *Foundations of cognitive psychology: core readings*. Cambridge, Mass.: MIT Press.

Levitin, D. J. (2006). *This is your brain on music: the science of a human obsession*. New York, N.Y.: Dutton.

Marshall, A. H. (1967). A note on the importance of room cross-section in concert halls. *Journal of Sound and Vibration*, *5*(1), 100–112.

Marshall, A. H., & Barron, M. (2001). Spatial responsiveness in concert halls and the origins of spatial impression. *Applied Acoustics*, *62*(2), 91–108.

Mason, R., Brookes, T., & Rumsey, F. (2005). The effect of various source signal properties on measurements of the interaural crosscorrelation coefficient. *Acoustical Science and Technology*, *26*(2), 102-113.

Mason, R., Ford, N., Rumsey, F., & Bruyn, B. de. (2001). Verbal and non-verbal elicitation techniques in the subjective assessment of spatial sound reproduction. *Journal of the Audio Engineering Society*, *49*(5).

Miller, A. J. (2002). *Subset selection in regression*. Boca Raton: Chapman & Hall/CRC.

Morimoto, M., Fujimori, H., & Maekawa, Z. (1990). Discrimination between auditory source width and envelopment. *J Acoust Soc Jpn*, *46*, 449–457. (in Japanese)

Morimoto, M. ., & Iida, K. . (2005). Appropriate frequency bandwidth in measuring interaural cross-correlation as a physical measure of auditory source width. *Acoustical Science and Technology*, *26*(2), 179–184.

Morimoto, M., Jinya, M., & Nakagawa, K. (2007, Sep). Effects of frequency characteristics of reverberation time on listener envelopment. *Journal of the Acoustical Society of America*, *122*(3), 1611–1615.

Morimoto, M., & Maekawa, Z. (1989). Auditory spaciousness and envelopment. In *Proceedings of 13th ICA*.

Okano, T., Beranek, L. L., & Hidaka, T. (1998, Jul). Relations among interaural cross-correlation coefficient ($IACC_E$), lateral fraction (LFE), and apparent source width (ASW) in concert halls. *Journal of the Acoustical Society of America*, *104*(1), 255–265.

Rumsey, F. (1998). Subjective assessment of the spatial attributes of reproduced sound. In *AES 15th international conference: Audio, acoustics and small space*, Copenhagen, Denmark.

Rumsey, F. (2002). Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society*, *50*(9), 651-666.

Scholkopf, B., & Smola, J., Alexander. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT Press.

Smola, J., Alex, & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, *116*(2), 918-933.

Västfjäll, D., Larsson, P., & Kleiner, M. (2002). Emotion and auditory virtual environments: Affect-based judgments of music reproduced with virtual reverberation times. *CyberPsychology & Behavior*, *5*(1), 19-32.

Vries, D. de, Hulsebos, E. M., & Baan, J. (2001, Aug). Spatial fluctuations in measures for spaciousness. *Journal of the Acoustical Society of America*, *110*(2), 947–954.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (2nd ed ed.). Amsterdam: Morgan Kaufman. Retrieved May 3, 2009, from `http://www.cs.waikato.ac.nz/ml/weka/` (Computer software and manual)

Zacharov, N., & Koivuniemi, K. (2001, July 29–Augu 1). Audio descriptive analysis mapping of spatial sound displays [inproceedings]. In *Proceedings of the 2001 international conference on auditory display.* Espoo, Finland: ICAD: International Conference on Auditory Display. (Espoo, Finland)

APPEDIX A

HUMAN SUBJECT STUDY INTERFACE

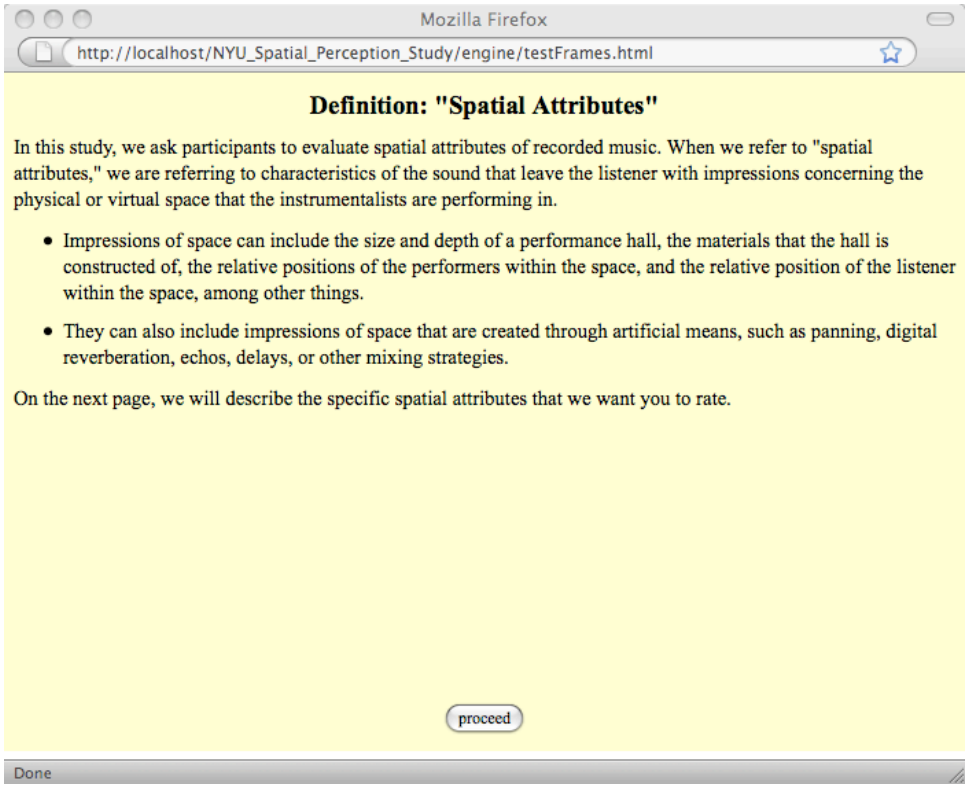


Figure A.1: Definitions for “spatial attributes.”

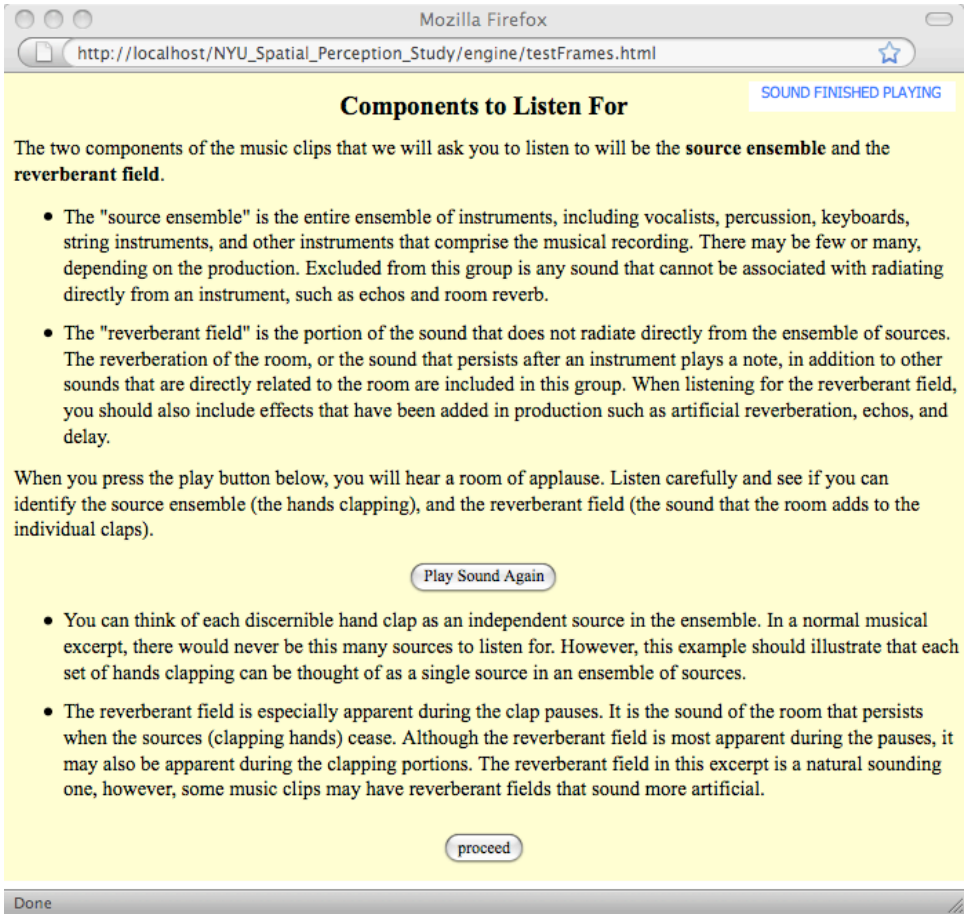


Figure A.2: Instructions on components to listen for.

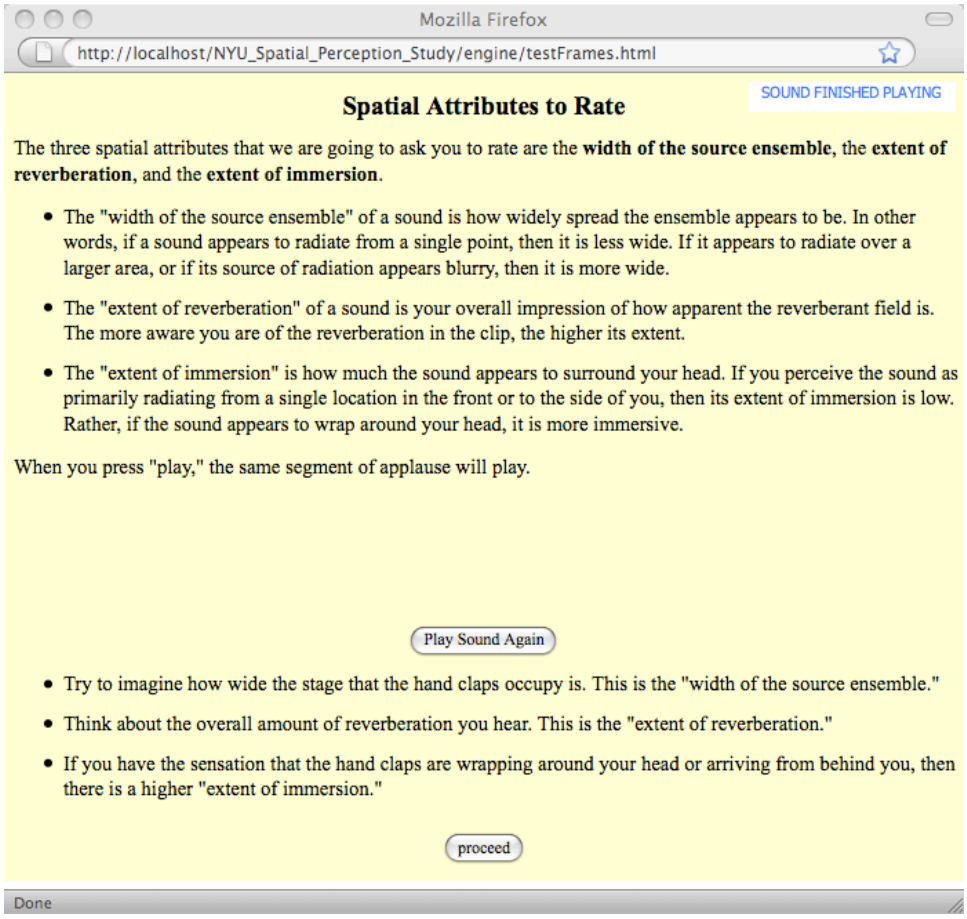


Figure A.3: Instructions on how to rate spatial attributes.

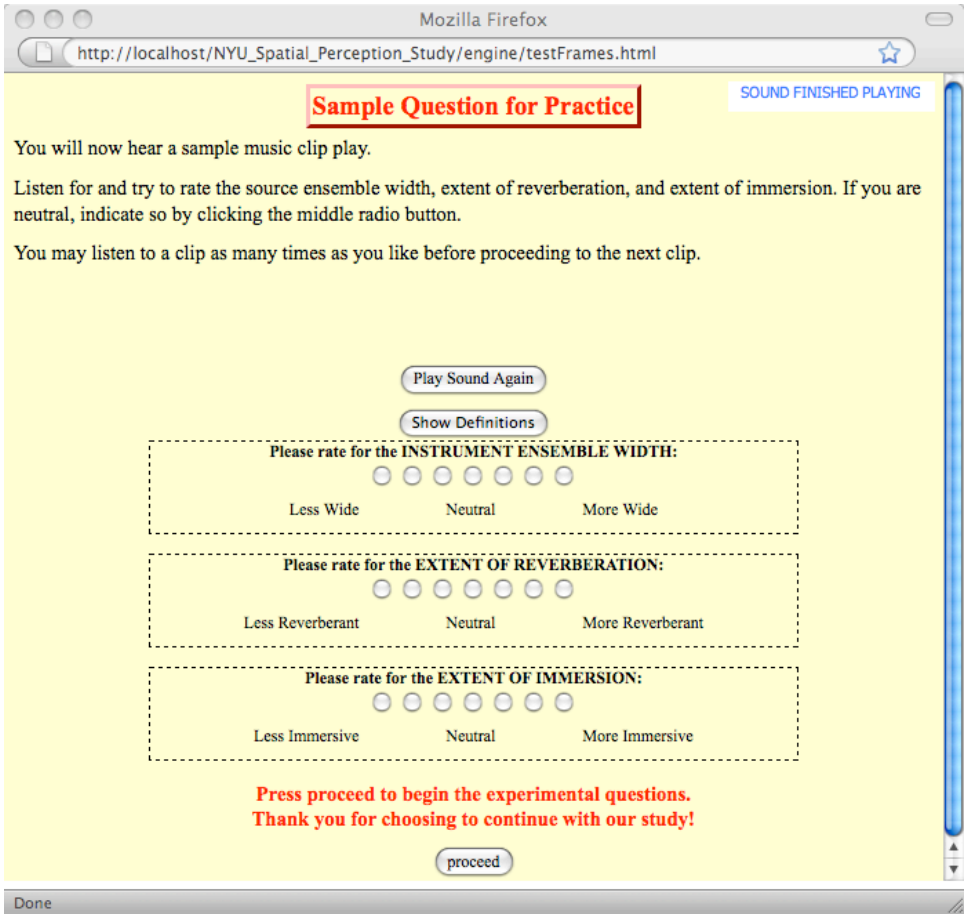


Figure A.4: Practice question.

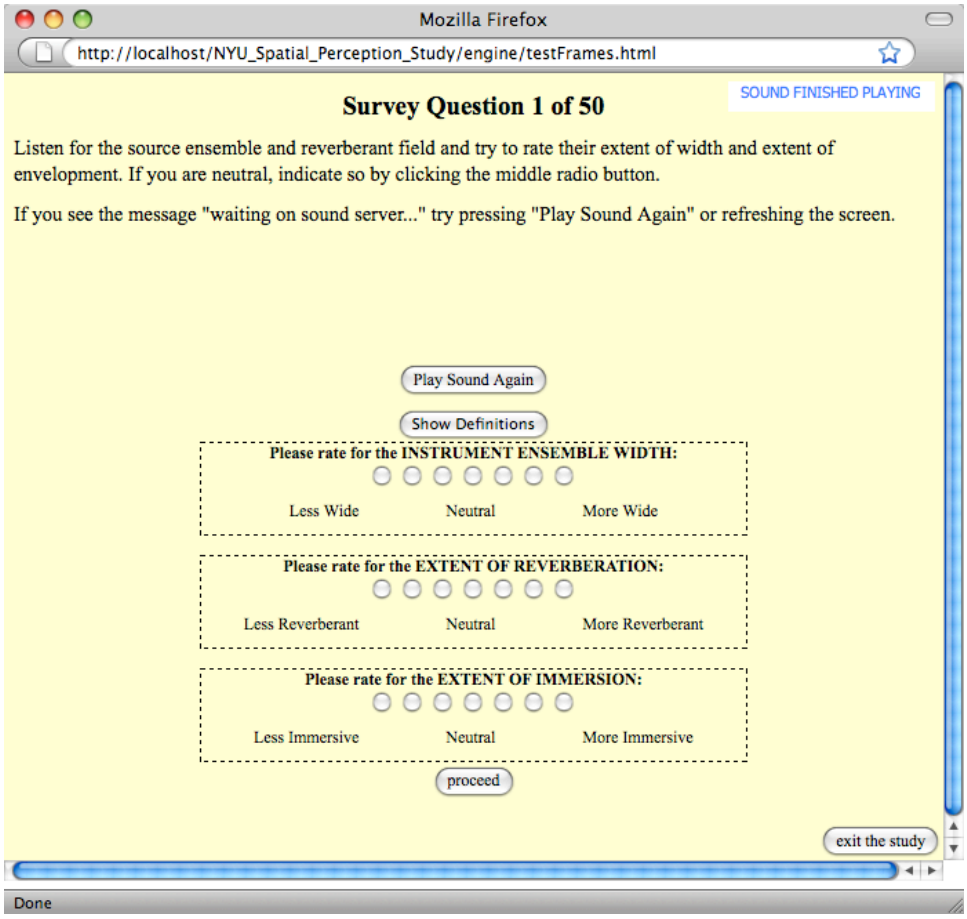


Figure A.5: Experimental question.