Toward a Computational Model of Perceived Spaciousness in Recorded Music*

ANDY M. SARROFF,** *AES Associate Member,* **AND JUAN P. BELLO**
(sarroff@cs.dartmouth.edu) (jpbello@nyu.edu)

Music and Audio Research Laboratory (MARL), New York University, New York, NY 10012

A computational model of perceived spaciousness in recorded musical signals, inspired by research in music information retrieval (MIR), is presented and evaluated experimentally. First three dimensions of spaciousness are selected for computational modeling: width of source ensemble, extent of reverberation, and extent of immersion. Next human subject responses to stereophonic music along these dimensions were collected. Then audio features from this data set of music were extracted and finally three exemplar-based machine learning models for estimating the three dimensions of spaciousness were trained and tested. The worst predictor was found to perform at least 32% better than a baseline predictor. These results are important to the music and audio engineering communities, as computational models for spaciousness in music may present new, perceptually meaningful methods of analysis and processing for the spatial characteristics of recorded music.

0 INTRODUCTION

Spatial impression in music transmits important semantic information, such as genre, acoustic context, or emotional content. The artful handling of spatial cues in recorded music is in part responsible for creating an enjoyable listening experience. Because of this, music engineers allocate significant energy to controlling the character and extent of acoustically related spatial attributes and attributes of space in music by relative placement of microphones, performers, and reflecting surfaces. Spatial impression is further manipulated by source mixing, signal processing, and multichannel panning techniques. As such, perceived spaciousness is a crucial attribute of recorded music, and one that music engineers attend to with care. Yet spatial impression as it is related to the recorded musical signal has not been thoroughly explored in research literature.

Because spatial cues are closely managed by music engineers, two assumptions may be made. The first is that digital musical signals may exhibit measurable qualities that relate to spaciousness. The second is that human perception of spaciousness in recorded music may be consistent across subsets of human listeners. If these are true, then perceived spaciousness may be modeled computationally as a function of measurable properties of digital audio. Such a model would be useful for music engineers, musicians, and music listeners, allowing new interaction with music in perceptually meaningful ways. With an accurate model of spaciousness we may have access to new analytic tools and signal processing mechanisms for mixed music. For instance, extent of spaciousness may be displayed visually for multimodal music analysis. Or top-down controls for spaciousness may be designed that deliver novel ways to customize the music listening or making experience. In this paper we design a framework for such a model and evaluate it experimentally.

There are several challenges to modeling spaciousness. As yet there is no universally agreed upon definition for spaciousness in recorded music. And there is no predetermined basis for a comparison of spaciousness across recordings of music. We therefore lack the necessary data sets and tools to evaluate a model for spaciousness. Accordingly, in this paper we synthesize knowledge from related fields and provide a low-dimensional parameterization of spaciousness, comprised of the width of the source ensemble, the extent of reverberation, and the extent of immersion. We then collect a data set of human subject responses to auditory stimuli. Finally we train a computational model that links subjective responses to a set of audio features, which are inspired by research in music information retrieval (MIR). The model's predictive accuracy will be shown to be high, beating a baseline predictor by more than 32%.

Our study contributes a data set of over 2000 responses across 50 musical stimuli; a model for a low-dimensional parameterization of spaciousness; and a framework (Fig. 1) for modeling spaciousness using exemplar-based machine learning techniques. In this paper we model three dimensions of spaciousness for stereophonic music recordings. However, we believe the framework is generalizable to other multichannel formats. The organi-

*Manuscript received 2010 April 29; revised 2011 February 8.

**Now with the Computer Science Department, Dartmouth College, Hanover, NH 03755.

zation of this paper is as follows. In Section 1 we parameterize spaciousness in three dimensions by synthesizing knowledge from other fields. Then, in Section 2, we execute a human subject study to collect quantitative ratings of spaciousness along these three dimensions. In Section 3 we map the subjective data to objective digital signal measurements using machine learning techniques. We discuss the significance of the model and some of its limitations in Section 4, and we present concluding remarks and future work in Section 5.

1 PARAMETERIZING SPACIOUSNESS

The concept of "spaciousness" is not universally defined. Rather, the auditory, physical, and engineering sciences have treated spaciousness somewhat independently in order to answer questions most relevant to their domains. In acoustics, for example, spatial impression is important for optimizing room design. In engineering, sound reproducing systems are evaluated for the quality and integrity of their spatial displays. Our work builds a model of perceived spaciousness with respect to the digitally recorded music signal, rather than the listening environment or reproducing system. However, spatial impression has not previously been parameterized explicitly for the recorded signal and from an MIR perspective. We therefore seek knowledge from related research to choose a low-dimensional parameterization of spatial impression. The spatial parameters we choose are not complete. It is beyond the scope of this work to attempt to model the full set of perceived spatially related attributes. Rather, we somewhat arbitrarily choose attributes that describe various spatial relationships for listeners, and our choices are motivated by research in spatial perception by acousticians and engineers. In the following subsections we summarize the motivating work and our choices of spatial parameters.

1.1 Acoustics

In 1967 Marshall determined that "spatial responsiveness" is a desirable property of concert halls [1]. By analyzing echograms and architectural drawings of dissimilar rooms, he concluded that good spatial responsiveness arises from well-distributed early reflections of the direct sources. Since Marshall's findings, spaciousness in music halls has been parameterized by two distinct dimensions: apparent source width (ASW) [2] and, later, listener envelopment (LEV) [3], [4]. The first has consistently been attributed to early lateral reflections and the latter to the late arriving sound in an acoustic space. While the respective terms have been distinguished by varying labels and definitions, they have more or less been used to describe the same two phenomena throughout. (For an overview of the development and semantics of the terms ASW and LEV, we recommend [5].)

Despite minor differences in interpretation across studies, the perceptual dimensions of ASW and LEV can be defined as follows [6]:

> 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.

ASW includes sensations of broadness, blurriness, and ambiguity in localization. LEV, on the other hand, imparts sensations of fullness and surrounding.

ASW has been attributed to the early arriving lateral energy as early as 1971, when Barron determined that the relative level of the lateral reflections arriving within 80 ms of the direct signal contributed to spatial impression. While he did not use the term "apparent source width," he described an apparent broadening of the source [7]. Bradley and Soulodre were the first to show systematically that ASW and LEV are separably perceived components of spatial impression, and that perception of LEV can impact perception of ASW [8]. They show that energy arriving later than 80 ms after a direct sound produces a different spatial impression than ASW, and they describe the sensation as "listener envelopment." Bradley and Soulodre indicate that ASW and LEV may arise as distinct perceptual components because early arriving sound is temporally and spatially fused with the direct sound by the auditory system, whereas late arriving sound is not. The phenomenon of time-based sound integration was demonstrated on speech signals by Haas [9]. Bradley and Soulodre's findings are significant to concert hall design. Rather than optimize the level of early lateral reflections, as was popular practice at the time, the perceptual phenomenon of LEV indicated that the late arriving energy must be considered carefully when designing concert spaces.

In natural acoustic environments the relative positions of

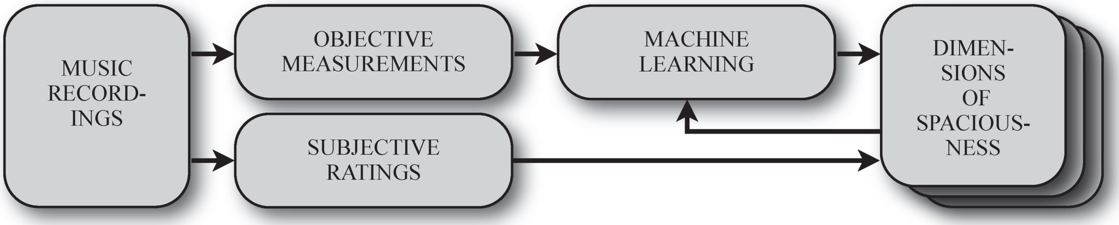

Fig. 1. Framework for predicting perceived spaciousness of music recordings.

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 normally experienced in a linear, time-invariant system, 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. Methods using the lateral energy or interaural cross-correlation function have been employed for measurements of ASW [6], [10], [11], and varying measurements of late arriving energy are used for LEV [8], [12]–[14]. ASW and LEV provide not only well-defined meanings for perceived spaciousness in "live" listening environments, but a means of studying their relationship to measurable quantities in the physical world.

1.2 Sound Reproduction

Sound reproduction systems such as Surround Sound create a virtual representation of spatial sound by utilizing a discrete number of audio channels. It is often necessary to evaluate the quality of spatial reproduction systems. For this, researchers must know which attributes of spaciousness are most important to human listeners. Nonverbal [15], [16] and verbal [17] elicitation techniques have been studied at the University of Surrey and elsewhere, with Berg and Rumsey reporting extensively on verbal assessments of spatial audio quality [18]–[20].

A systematic method for evaluating spatial quality is proposed in [21], in which Berg and Rumsey list 13 perceived attributes of spaciousness that they have found to be significant for evaluating audio quality. The attributes, which were validated in two separate experiments involving unique stimuli and subjects [19], [22], are suggested to be discriminative enough for other subjective evaluations of spatial quality. These attributes are duplicated in Table 1. Berg and Rumsey note that some spatial attributes are perceived in both audio reproduction systems and concert halls (such as those dealing with envelopment), while others are perceived only in the former. This leads to a higher parameterization of spaciousness when evaluating audio reproduction quality, with attributes such as naturalness arising. Berg and Rumsey have shown in [18] that attributes may be grouped into one of three classes: descriptive, emotional/evaluative, or naturalness.

In [21] Berg and Rumsey have offered some important insight, which we find helpful to this work. First, attributes that deal with the room tend to be judged differently than those that deal with the sources. Second, perceived room attributes can be categorized as those that deal with the physical properties of the room, such as room sound level (reverberation), and those that deal with the listener's feeling of being present in the room. Finally attributes seem to be perceived mainly in the dimensions of "source width," "distance to the source," and "sensation of presence in the room." Rumsey notes in [23] that attributes of space are often confused in the literature with spatial attributes, but are qualitatively different. The first class of attributes describe the physical characteristics of a space, such as the reverberation level of a concert hall. Spatial attributes, however, refer to those characteristics that impart sensations of dimensionality, size, and space of reproduced sources, groups of sources, or environments. In the following subsection we discuss how we borrow these concepts to construct our model parameterization.

1.3 Recorded Music

As far as we know there has not been a thorough attempt in the literature to parameterize perceived spaciousness for recorded musical signals. While researchers in acoustics and audio reproduction quality must decide on appropriate parameterization for spaciousness, their tasks are inherently different from ours. Researchers in acoustics are dealing with the physical environment. Recorded music often exhibits attributes of spaciousness that could never be realistically experienced in the physical world. Sound reproducing systems are evaluated based on the quality of the spatial display, rather than a quantification of the perceived spatial attributes resulting from the recorded signal.

As such we have no complete and systematic work from which to select the "best" attributes to model or from which to define them. We used our own experience in the domain of music production and the cited literature as motivation for selecting a set of attributes to build our model. In particular we parameterized the model with "width of source ensemble," "extent of reverberation," and "extent of immersion." Table 2 shows the attribute definitions, as given to our subjects during data collection. We describe our motivations behind parameter choices in the following paragraphs. Please note that we made no assumptions concerning the orthogonality of these three dimensions. They may be perceived concurrently, and perception of one may influence perception of the others.

Table 1. Significant spatial attributes reported in [21].*

Attributes	
Naturalness[†]	Presence[†]
Preference[||]	Low-frequency content
Ensemble width	Individual source width
Localization	Source distance
Source envelopment	Room width
Room size	Room sound level
Room envelopment	

*Most attributes are in the descriptive class.
[†]Attributes in the naturalness class.
[||]Attribute in the emotional/evaluative class.

1.3.1 Width of Source Ensemble

Width of source ensemble describes the listener's perception of how widely a group of sources is represented in the sound field, irrespective of any room characteristics. We chose this attribute because we wanted our model to capture the "wideness" of a mix, which can be manipulated by source panning, phasing, and other techniques. We were motivated by the ensemble width listed in Berg and Rumsey [21], which was defined thus for their subjects:

> 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.

Even though this attribute evaluates wideness, it cannot be related directly to apparent source width (ASW). ASW is usually tested in terms of individual sources, rather than ensembles. In contrast, our aim is to predict spatial attributes of the entire mixture, rather than its individual sources.

1.3.2 Extent of Reverberation

In extent of reverberation the listener perceives the overall level of reverberation of implied acoustic environments. Music engineers sometimes include unrealistic reverberation and room level cues for effect. Therefore it is difficult to claim strict correspondence between extent of reverberation and room level, as the latter indicates a physical space that could be actualized in the natural world. However, extent of reverberation was partially motivated by the room sound level described in [21] as:

> 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.

It has been noted in [6] and elsewhere that the reverberant sound field contributes to LEV. As discussed in Section 1.1, LEV is considered one of the primary components of spatial impression by acousticians and is a shared attribute for assessing spatial display quality. We note that the extent of reverberation might be most accurately described as an attribute of space, rather than a spatial attribute [23].

Table 2. Definitions of learning concepts as shown to subjects in study.

Spatial Attributes to Rate
The three spatial attributes that we are going to ask you to rate are the width of the source ensemble, the extent of reverberation, and the extent of immersion.
• The "width of source ensemble" of a sound is how widely spread the ensemble of sound sources appears to be. In other words, if a sound appears to radiate from a single point, then it is less wide. If it appears to radiate over a larger area, or if its source of radiation appears blurry, then it is more wide.
• The "extent of reverberation" of a sound is the overall impression of how apparent the reverberant field is. The more aware you are of the reverberation in the clip, the higher its extent.
• The "extent of immersion" is how much the sound appears to surround one's head. If you perceive the sound as primarily radiating from a single location in the front or to the side of you, then its extent of immersion is low. Rather, if the sound appears to wrap around your head, it is more immersive.

1.3.3 Extent of Immersion

Audio recordings can impart a sensation of surrounding or immersion to a listener. Music engineers may promote immersion through audio effects that cannot be reproduced by acoustic environments. While envelopment seems like an appropriately descriptive term for immersion, we avoided its use for two reasons. First, envelopment is closely associated with reverberation in acoustics, and we aimed to model the perception of reverberation separately. Second, envelopment is listed as a significant attribute in [21] with respect to the sources and the physical room. However, our concept of immersion embodies virtual environments as well as physically existing rooms. Our attribute "extent of immersion" was defined as a global relation which does not directly consider the room or the sources. We were partially motivated by presence [21], but without the requirement of a realizable acoustic environment:

> The experience of being in the same acoustical environment as the sound source, e.g., to be in the same room.

Even though the term "immersion" does not arise in Table 1, we felt that immersion was appropriately descriptive for the sensation we were considering, and would be understood by our subjects as not necessarily relating to an existing physical environment.

2 QUANTITATIVE ANNOTATION

To the best of our knowledge a quantitative model of spaciousness for audio has only been developed with respect to the quality of spatially reproducing systems. For instance, QESTRAL [24]–[27] is an artificial listener designed for evaluating the quality of reproduced audio across a broad range of program material, including television. Our model is designed for the predictive judgment of spatial quantity in musical recordings, rather than spatial quality of reproducing systems on arbitrary source material. In other words, we model how spacious a musical recording is, and our model is not motivated by estimating the spatial degradation of the source material. The model we present in this paper is trained solely on stereophonic music recordings, whereas QESTRAL is designed for larger multichannel formats.

In order to train and test a predictive model, we required human subject assessments of stereophonic music recordings along the three dimensions outlined in the last section. We performed two data collection experiments—one online and one in a laboratory. The experiments were similar in nature. They differed in that the first targeted a larger subject base, at the acknowledged cost of poorly controlled experimental conditions, and the second optimized experimental control at the cost of subject pool size. We hoped that we would find high consistency across experiments and subjects, and low correlation across dimensions in the responses. This would support the hypothesis that our chosen dimensions could be perceived consistently by subjects in different experimental conditions and with different backgrounds, and that the dimensions exhibited a well-rounded representation of perceived spaciousness. This section explains how the musical recordings were selected, segmented, and annotated in the two experiments, and presents our analysis of the data.

2.1 Materials and Methods

2.1.1 Music

The songs in our database were selected from an online music web site[1] 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 multiple genres; it should be unfamiliar, so as to avoid bias by recognition; it should represent the major parts of a song; and it should be spatially consistent. By imposing these criteria we aimed to achieve a wide breadth in the extent of spaciousness for our database.

An equal distribution of songs were selected from each of the popular genre categories on the site. These were alt/punk, classical, electronic dance, hip-hop, R&B/soul, and rock/pop. We excluded genre categories that were underrepresented on the site, such as international and jazz so as to have enough within-genre material to choose from. The genre label for each song had been selected by the artist who uploaded the song. We note that there was high variability in the interpretation of genre across the songs, and the preceding genre categories should not be taken too literally. The final list of songs was arguably representative of a broader range of categories and subcategories.

None of the songs that were picked had been commercially distributed on a large scale. Therefore they were likely to be unfamiliar to listeners. A musical segment was chosen from each song so as to fall into either a verse, a chorus, or a (nonvocal) bridge category. Twice as many bridge sections were included as verses and choruses so as to have a roughly equal number of lyrical and nonlyrical sections. We promoted spatial consistency by drawing song segments from within structural components of the music. We observed that spatial consistency was higher, for instance, if a song segment was drawn from within a verse, rather than across a verse and chorus. We avoided imposing our own biases of spaciousness upon the data set by not selecting segments based specifically upon their spatial attributes.

There were a total of 50 song segments. All were stereo recordings and were in compressed MP3 format. Each song segment was 7 seconds in duration, with a 50-ms fade-in and fade-out to avoid clicks. No segments were chosen from the beginnings or endings of songs. While it would have been advantageous to use higher audio quality for our evaluations, MP3s were chosen so that we could reach a wider audience by streaming compressed audio in our online experiment. It is possible that the overall spatial quality of our data set was lower than if we had used for example PCM recordings. Since we are interested in comparatively quantifying the extent of spaciousness rather than the quality of spaciousness, we deemed this as acceptable information loss. It would have been desirable to have a larger collection of songs in our database to train and test the model. However, if we had used more song segments, it would have been at the price of collecting fewer subject ratings per song.

2.1.2 Subjects

Subjects were recruited for the online experiment by posting advertisements on nearly 20 web forums for musicians and music producers. Music-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. For the laboratory experiment 20 paid subjects were recruited by posting advertisements on several email lists targeted to music technology and music performance students. Informed consent was obtained for both experiments, which were approved by the New York University Committee on Activities Involving Human Subjects.

There were 78 participants across both studies. Online participants, of which there were 58, varied in age from approximately 18 to 65 years of age and their residences were distributed across 19 countries. They had varying degrees of experience regarding work or study in a music-related field, with some claiming up to 40 years of experience, and the participants were dispersed in the number of hours a day they spent listening to music. Subjects in the laboratory experiment were rather homogenized compared to the online experiment. They were distributed over a smaller age range (mostly 18–25 years old), they were all U.S. residents, and they were each active workers in a

[1] MP3 Music Downloads, http://www.mp3.com/.

music-related field. The laboratory subjects were asked to self-assess their level of critical-listening ability on a scale of 1 to 5. Most subjects rated themselves highly, at 4 or 5.

It is always desirable to have a large subject pool for subjective studies. However, high-quality data collection is expensive. To remain within our budget of resources, our laboratory experiment targeted a small group of subjects who were experienced listeners and would provide reliable data in a controlled environment. The online experiment was designed to collect data from a larger subject pool, at the cost of experimental control. Section 2.2 provides a quantitative analysis of the relationship between the two subject pools.

2.1.3 Environment and Monitoring

Both experiments were conducted on the same web-based interface using a series of interactive web pages. Before participants began the online experiment, they were informed that headphone use was a requirement of the study. We did not attempt to control the type of headphones or acoustic environment experienced by subjects in the online experiment. In the laboratory experiment all participants took the test in the same room (at staggered times) using the same model of high-fidelity open-back headphones, Sennheiser HD650. Subjects in both experiments were presented with a headphone calibration screen. A series of simple tones were played to facilitate volume adjustment and subjects were instructed to set their volume so that the tones were heard easily, but not too loud. For the laboratory experiment the volume was preadjusted for each subject before they went through their individual level adjustments. All subjects in the laboratory experiment except for one kept their volume unchanged from the starting level, implying that we had chosen a comfortable listening volume for the experiment.

2.1.4 Subject Training

Subjects were presented a short training session by the user interface before the experiment's official onset. On the first screen (Table 3), the term "spatial attributes" was defined. Next, participants were informed of the definitions for "source ensemble" and "reverberant field." After reading the directions, the subjects were asked to listen to a nonmusical mixture of sources (a room of applause) and focus their hearing. Table 4 shows the exact definitions and directions. Finally, definitions of the three spatial attributes were given (shown in Table 2). For each of these attributes, participants were again asked to listen to applause and concentrate their attention with respect to each attribute (Table 5). The applause track exhibited characteristics of the three spatial dimensions. Participants were not told exactly how spacious the recordings were expected to be perceived, so as to avoid biasing their judgments. After the subjects completed the training phase, a web page with a real musical example was given for practice.

2.1.5 Experiment

Subjects were asked to rate, on a bipolar 5-ordered Likert scale from "less" to "neutral" to "more," the extent of each of the dimensions for each test song. The ratings we relative to the subjects' own self-perceptions of the realizable range for each dimension. There was a GUI button activating a pop-up screen with the term definitions, in the case that a participant needed to be reminded. The order of the songs was randomized so as to eliminate any order bias across participants.

For the online experiment a web browser cookie-tracking mechanism prevented any subject with their browser cookies enabled from participating more than once. Participants were allowed easy exit out of the experiment via an exit icon in the corner of the screen. The experiment proceeded until all 50 song excerpts were assessed, or the participant exited.

In the laboratory experiment subjects were required to rate all 50 song excerpts in the data set. They had the additional benefit of an experiment investigator on hand to precisely answer questions about the terms in the experiment. The average time for completion of the laboratory experiment was approximately 30 minutes.

2.1.6 Postprocessing and Outlier Removal

The results of the combined experiments provided 2523 ratings over 50 songs and three dimensions of spaciousness. Ratings were transformed from a Likert scale to a linear numerical scale. Our work assumes that subsets of music listeners perceive spaciousness in similar ways. We removed outliers from the data set so that we could model a population of listeners whose perceptual responses approximate a normal distribution. In a normal distribution 99.73% of the data lie within three standard

Table 3. Definition shown to subjects for "spatial attributes."

Definition: "Spatial Attributes"
In this study, we ask participants to evaluate spatial attributes of recorded music. When we refer to "spatial attributes," we are referring to characteristics of the sound that leave the listener with impressions concerning the physical or virtual space that the instrumentalists are performing in.
• Impressions of space can include the size and depth of a performance hall, the materials that the hall is constructed of, the relative positions of the performers within the space, and the relative position of the listener within the space, among other things.
• They can also include impressions of space that are created through artificial means, such as panning, digital reverberation, echos, delays, or other mixing strategies.
On the next page, we will describe the specific spatial attributes that we want you to rate.

deviations of the population mean. We removed all song ratings that exceeded this threshold, as calculated for each song–dimension pair. In addition if a participant was responsible for more than one outlier rating in a given dimension, his or her ratings were removed entirely from the dimension, so as to help ensure that our data were representative of similarly perceiving subjects. The ratings for each dimension were then 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. Fig. 2 shows the sorted mean value and standard deviation for each song for the three standardized dimensions.

Table 4. Definitions shown to subjects for "source ensemble" and "reverberant field," and directions asking subjects to listen to an audio excerpt of applause.

Components to Listen For
The two components of the music clips that we will ask you to listen to will be the source ensemble and the reverberant field.
• The "source ensemble" is the entire ensemble of instruments, including vocalists, percussion, keyboards, string instruments, and other instruments that comprise the musical recording. There may be few or many, depending on the production. Excluded from this group is any sound that cannot be associated with radiating directly from an instrument, such as echos and room reverberations.
• The "reverberant field" is the portion of the sound that does not radiate directly from the ensemble of sources. The reverberation of the room, or the sound that persists after an instrument plays a note, in addition to other sounds that are directly related to the room are included in this group. When listening for the reverberant field, you should also include effects that have been added in production such as artificial reverberation, echos, and delay.
When you press the play button below, you will hear a room of applause. Listen carefully and see if you can identify the source ensemble (the hands clapping), and the reverberant field (the sound that the room adds to the individual claps).
• You can think of each discernible hand clap as an independent source in the ensemble. In a normal musical excerpt, there would never be this many sources to listen for. However, this example should illustrate that each set of hands clapping can be thought of as a single source in an ensemble of sources.
• The reverberant field is especially apparent during the clap pauses. It is the sound of the room that persists when the sources (clapping hands) cease. Although the reverberant field is most apparent during the pauses, it may also be apparent during the clapping portions. The reverberant field in this excerpt is a natural sounding one. However, some music clips may have reverberant fields that sound more artificial.

2.2 Results

2.2.1 Laboratory and Online Experiments

It can be expected that the ratings in the online experiment would be less stable than the ratings from the laboratory, as it was not possible to control the experimental conditions for each online participant. Indeed, the average variance per song was consistently higher in the online experiment. We performed ANOVA on the average song ratings between the two experiments for each dimension. All tests passed the null hypothesis at the 99% confidence level that the distributions share the same means.

Table 5. Directions asking subjects to listen to applause while considering the attributes.

When you press "play," the same segment of applause will play.
• Try to imagine how wide the stage that the hand claps occupy is. This is the "width of the source ensemble."
• Think about the overall amount of reverberation you hear. This is the "extent of reverberation."
• If you have the sensation that the hand claps are wrapping around your head or arriving from behind you, then there is a higher "extent of immersion."

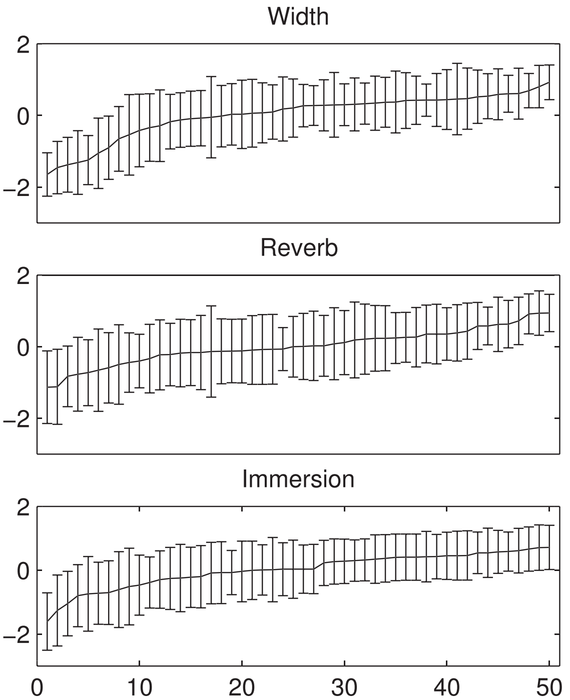

Fig. 2. Sorted means and standard deviations of standardized ratings for each song and each dimension of spaciousness.

Then we computed pairwise t-tests for each song and dimension to test the null hypothesis that the average ratings for the laboratory and online experiments share the same means for each song. Since different experimental conditions were being compared, the p values were calculated assuming unequal variance, implementing Satterthwaite's approximation for standard error [28]. Out of 50 songs and three dimensions we found that the null hypothesis could be rejected at a 99% confidence level for only two songs in the immersion dimension. This leads us to believe that the data from the online experiment come from a similar distribution as the ratings from the laboratory experiment and that they can be combined, despite the different experimental conditions.

2.2.2 Experienced and Unexperienced Listeners

We wanted to test whether ratings by more experienced listeners would be statistically different from a population with less listening experience. First we compared subjects who listen to more than 4 hours of music a day to those who do not. No statistical difference could be found. Second we tested subjects who work or study in a music-related field versus those who do not. Only the immersion dimension of one song showed a significant difference at the 99% confidence level between experienced and less experienced listeners.

2.2.3 Headphone Use

An additional concern was whether the constraint of headphones would adversely affect the reliability of ratings. Headphone listening can inhibit perceived externalization, a factor that might affect the perception of spaciousness negatively. At the same time, using headphones is necessary to minimize the effect of any unrelated environmental acoustic factors of the listening environment on the experimental framework. If headphone-inhibited externalization affects perceived spaciousness, it might be hypothesized that subjects who listen to music predominantly through headphones will be better adapted to perceive differences in spaciousness when using headphones. In our last t-test we compared (with an equal variance assumption) subjects who usually listen to music through headphones with those who usually listen to music through loudspeakers. Once again, only two instances across 50 songs and three dimensions were found to differ significantly. It is possible that our outcome might have been different had the experiment been performed on loudspeakers. However, we can assume that subjects' experience with using headphones had a negligible impact on the results.

2.2.4 Between-Song Variability

Our data set is intended to represent a diversity of perceived spaciousness along each dimension. To validate this, F statistics were calculated for the songs in each dimension. The ANOVA p values for all were nearly zero ($F_{(49,2354)}\text{width} = 29.66$, $F_{(49,2333)}\text{reverb} = 14.59$, $F_{(49,2345)}\text{immersion} = 18.74$), showing that it is highly unlikely that the average song ratings share the same means.

2.2.5 Correlation between Dimensions

Finally we wanted some sense of how closely related the ratings along each dimension were to each other. If any two dimensions are highly correlated, it is possible that a model of one may be used to predict perception of both. We averaged the subjective ratings for each song, and the Pearson's correlation coefficient R was calculated between dimensions. These coefficients are reported in Table 6. The p values associated with R indicate that all correlation coefficients were significant at the 95% confidence level. The strength of the width–reverberation correlation was low, indicating that subjects did not tend to rate songs similarly for these dimensions. There was a medium amount of correlation between reverberation and immersion, and the width–immersion correspondence was found to be highly correlated. This knowledge might be useful when considering future computational models for the perceived attributes.

3 COMPUTATIONAL MODELING

After having collected ratings for spaciousness we set out to model them computationally by finding mappings between objectively measurable audio signal characteristics and subjective human ratings. Our approach to building an objective-to-subjective mapping function is shown in Fig. 3. On the left there is an audio feature space describing components of the music recordings, along with three learning concepts. The audio features are objective signal descriptors inspired by research in MIR. The learning concepts are the three dimensions of spaciousness outlined earlier. In the middle there are necessary dimensionality reducing and parameter optimization steps. We first used a correlation-based feature selection and subset voting scheme to narrow the feature space. Then we conducted a grid search for the best parameterization of an exemplar-based machine learning algorithm. In this work we used support vector regression for machine learning, which is depicted on the right. The stages of our approach are described hereafter, followed by analysis.

Table 6. Pearson's correlation coefficient R for averaged ratings between dimensions.

	Width–Reverberation	Width–Immersion	Reverbation–Immersion
R	0.3186	0.8745	0.5679
p*	0.0241	1.07×10^{-16}	1.70×10^{-5}

*p value of correlation—chance probability of correlation being equal to or greater than R.

3.1 Materials and Methods

3.1.1 Computing Environment

All computation was executed 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 [29].

3.1.2 Learning Concepts

The learning concepts were the same dimensions discussed earlier, namely, width of source ensemble, extent of reverberation, and extent of immersion. The averaged responses per song from Section 2 were used as the target values for the machine training and testing tasks.

3.1.3 Audio Feature Space

The entire audio data set was transformed into a multidimensional feature space. Features are descriptors of the audio signal obtained through signal filtering and analysis methods. In general one hopes to extract the most meaningful properties of an audio signal for the task at hand by reducing it to a set of descriptive numerical vectors.

Since we do not know which audio features in the literature are most relevant to perceived spaciousness, we start with a verbose set and later attempt to remove the less informative features. We batch-generated our verbose feature set using the entire feature-extracting capabilities of the MIR toolbox [30] at its current release version. Most researchers in MIR extract features from either one or the sum of channels in stereophonic recordings. We chose to use the left–right difference signal (called the side signal in MS recording) so as to preserve any spatial cues contained in the uncorrelated channel data.

By choosing to use only the side signal, we introduce some signal loss to the algorithm. It would perhaps be beneficial to use the monosummed signal in addition to the difference signal. By doing so we would be analyzing the full uncompressed signal in a transformed space, and we would also double the initial dimensionality of the feature space. Unless the left and right channels of the stereophonic mix are completely in phase or 180° out of phase, there will likely be signal redundancy between monosummed and side signals. An oversized redundant feature space will promote overfitting and feature selection bias, especially when the data set is small. We were therefore required to introduce this loss to the system.

The entire set of features, which are partitioned into categories of dynamics, rhythm, timbre, pitch, and spatial, is listed in Table 7. For most features the recording frame was decomposed and feature extraction was performed independently on each frame. Some features, such as fluctuation, were calculated on the entire audio segment. The frame-decomposed features were summarized by their means, standard deviations, and slopes, and by their estimated period frequency, amplitude, and entropy. The size of the extracted feature space was 430 dimensions by 50 songs.

The toolbox includes several well-known features that are widely used for MIR and that are part of the MPEG-7 specification, such as MFCCs, spectral centroid, and spectral flatness. These are described in detail in [31]. For the remaining features we paraphrase the feature summaries provided by the *MIRtoolbox User's Manual* [32] and provide pointers to the original source literature.

The rhythm features are largely derived from an onset detector, except for fluctuation, which is a subband-based measure of periodicity [33]. Brightness is a measure of the high-frequency energy in the signal [34], and skewness and kurtosis are the third and fourth central moments of the spectrum, respectively. Entropy attempts to quantify the amount of information in a signal by characterizing its uncertainty [35], and roughness is a measure derived from Plomp and Levelt's theory on dissonance [36]. Irregularity is the variability in amplitudes of the partials in the spectrum [37], and inharmonicity is an approximation of the energy outside of the harmonic series, as calculated from an estimated fundamental frequency. Low energy is the percentage of frames having energy below a relative threshold [38]. A chromagram places the spectral energy in pitch-class bins. Key clarity is a probability estimate associated with a frame's estimated key. Mode is an estimation of whether a frame is in a major or minor key. Finally the harmonic change detection function measures the flux around a tonal centroid [39].

We included two spatial features that were not part of the MIR toolbox. The first, wideness estimation, estimates the width of a distribution of mixed sources in a recording by calculating an azimuthal histogram of spectra using a phase cancellation strategy. The latter, reverberation estimation, attempts to measure the quantity of reverberation in a mix using the residual signal from linear

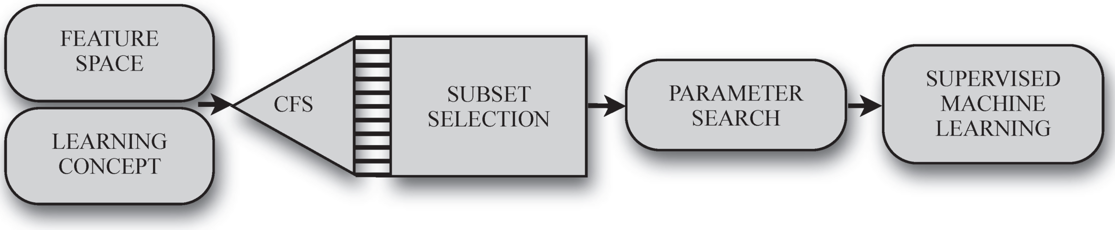

Fig. 3. Block diagram for building and optimizing mapping functions.

prediction coding. Both measurements were originally reported in [40].

The feature space was normalized to the range [0,1] and transformed into a principal-component space. The 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 attribute space. This transformation and filtering of data by principal-component analysis is an often used means of performing data cleansing on a feature space [29].

3.1.4 Feature Selection

Because the learning concepts used in this work are new, there is no literature that shows which features are the best for predicting perceived spaciousness. A large dimensional feature space can adversely affect the performance of a machine-learning algorithm and adds unnecessary complexity to the model. We therefore sought methods for reducing the feature space to a subset that was optimally robust for modeling perceived spaciousness.

For each target learning concept correlation-based feature selection (CFS) was performed with a greedy stepwise forward search heuristic. CFS chooses features that are well correlated to the learning target, yet exhibit low cross correlation with each other, and has been shown to be good for filtering out irrelevant or redundant features [41].

However, supervised attribute selection such as CFS can overfit features to their learning concept when the same data set is then used for training a model [42]. To minimize subset selection bias, a percentile-based voting scheme with 10 × 10-fold cross-validated attribute subset selection was designed. For each run the data set was randomly partitioned into ten folds. Then CFS was run ten times, leaving one fold out for each iteration.

Due to random partitioning of the data set, different feature subsets were chosen across folds and runs. We placed the features into percentile bins based upon how many times they had been selected. Eleven feature subsets with monotonically decreasing sizes were generated in this way, ranging from the full feature set to only features that were selected 100% of the time. Each of the eleven feature subsets was then used to learn a nonoptimized support vector regression algorithm on each of the learning concepts. The subset that performed best for each learning concept was voted as the final subset for further system optimization and training.

3.1.5 Support Vector Regression

For each learning concept a support vector regression model was implemented with the sequential minimal optimization (SMO) algorithm developed in [43]. Support vector machines, which optimize a tradeoff between function error and function flatness, have been shown to generalize well to a number of classification and regression tasks. In support vector regression implementation an error threshold ξ is selected, below which instance errors will be invisible to the loss function. A cost constant C determines the flatness of the function and helps to prevent it from overfitting the data. The higher the value of C, the more influence errors outside of ξ have

Table 7. List of audio features and their categories.

Category	Feature			
Dynamics	Rms energy			
Rhythm	Fluctuation centroid*	Fluctuation*†	Tempo	Tempo envelope autocorrelation†
	Attack times	Attack slopes		
Timbre	Zero-cross rate	Spectral centroid	Brightness	Spectral spread
	Spectral skewness	Spectral kurtosis	Rolloff (95%)	Rolloff (85%)
	Spectral entropy	Spectral flatness	Roughness	Roughness spectrum†
	Irregularity	Irregularity spectrum†	Inharmonicity	MFCCs (13)
	Δ MFCCs (13)	ΔΔ MFCCs (13)	Low energy*	Low-energy rms
	Spectral flux			
Pitch	Chromagram†	Pitch	Key clarity	Mode
	Harmonic change detection			
Spatial	Wideness estimation‖	Reverberation estimation‖		
Summary functions‡	Mean	Standard deviation	Slope	Period frequency
	Period amplitude	Period entropy		

*Feature was not frame decomposed.

†Signal peak position and peak magnitude were calculated as features.

‖Feature was summarized by mean alone.

‡Used on frame decomposed features.

upon the function. 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 the discussion.

3.1.6 Regression Kernel

A kernel function generalizes regression models to nonlinear fits. Our support vector model employed a polynomial kernel, $K(x,y) = (< x,y > + 1)^p$, chosen as the best in an informal search. Support vector machines perform, to some extent, similarly independent of the kernel type if the kernel's parameters are well chosen [44]. In the case of a polynomial kernel the only parameter to choose is the polynomial exponent p.

3.1.7 Machine Training and Testing

An exhaustive 10 × 10-fold multiple CV grid search for the optimum values of the support vector machine cost C and its kernel exponent p was conducted. The value of ξ was set at 1×10^{-3} for the entirety of this study. The optimal parameterization was considered to be the one that yielded the lowest relative absolute error (RAE). RAE is the sum of all errors normalized by the sum of the errors of a baseline predictor. The baseline predictor we used was zero-R, which selects the mean of the target values for every instance. An RAE error of 0% would denote perfect prediction, 100% would indicate the same error as the baseline predictor, and error above 100% would indicate worse performance than the baseline predictor. The model that yielded the lowest RAE was retained and tested a final time.

3.2 Results

3.2.1 Feature Subset Percentiles and Sizes

In machine learning applications the dimensionality of the feature space is usually limited so as to minimize overfitting and model complexity. However, we had no a priori knowledge of the best feature size. We therefore imposed no constraints on the feature subspace dimensionality and used CFS and a voting scheme to choose the best feature subsets. Fig. 4 shows the results of testing feature subspaces on the nonoptimized machine. The width dimension shows a minimum with 6 features at the 50th percentile; the reverberation dimension shows 12 features at the 40th percentile, and the immersion dimension shows 14 features at the 20th percentile. Considering the small size of the data set (50 songs), the dimensionality of these feature spaces can be seen as relatively large. This may be due to the fact that most of the features were not specifically designed to extract spatial cues.

All predictors show two local minima: width at the 20th and 50th percentiles, reverberation at the 10th and 40th percentiles, and immersion at the 20th and 70th percentiles. This indicates that there might have been more than one optimal feature subset to use. The steepness of the error curves between the 0th and 10th percentiles shows that simply using the entire feature set without any feature selection would greatly inhibit the performance of the support vector algorithm. The poor model performance using the full (0 percentile) feature space illustrates model overfitting and underscores the importance of feature selection. We believe that our method for feature subset selection was successful at choosing the best features, while minimizing model complexity and avoiding selection bias.

3.2.2 Selected Features

A summary of the final feature subset for each learning concept is shown in Table 8. Features that were selected for more than one concept are shown in boldface. The width and immersion dimensions shared the most features in common. This is understandable as these dimensions also shared the highest correlation among annotations. This fact 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.

When examining the automatically chosen feature subsets we may wish to draw conclusions concerning which features are most related to spatial perception and perception of space in recorded music. However, we note that the features were selected with an unoptimized model. In addition it is difficult to speculate whether features were selected because of their perceptual relevance or their association to the source material. For instance, it is possible that the spectral flatness of the side channel influences our perception of the width of the source ensemble. However, it is also possible that the

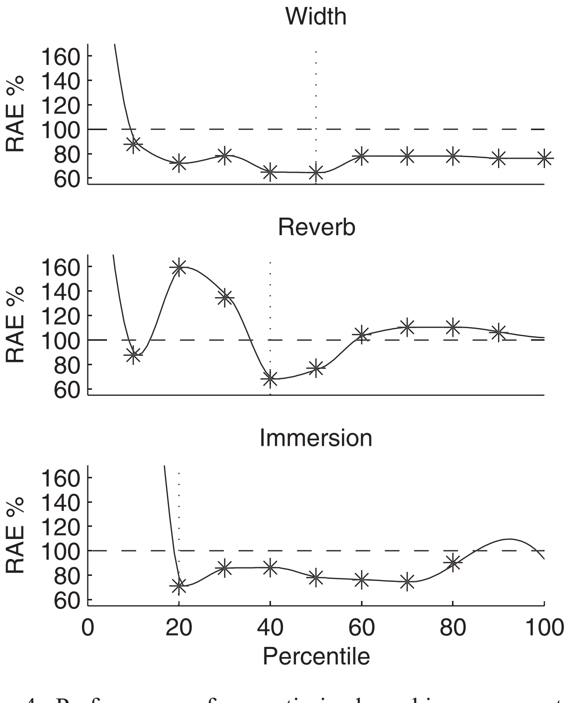

Fig. 4. Performance of nonoptimized machine on monotonically decreasing feature spaces. – – – baseline performance; ⋮ best performance.

songs in our database tended to have a correlation between spectral flatness and wideness, without signifying any meaningful perceptual relationship. Finally we note that while most features are probably not individually useful, the correct combination of features is. Therefore any discussion of the importance of a single feature to a learning concept requires significant prior investigation. Such an investigation is warranted, but was outside the scope of this work.

3.2.3 Parameter Search

The error surfaces for the parameterization of each machine are given in Fig. 5. These show the RAE for each value in the grid search for optimum 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 improved results significantly. It is worth noting that the flattest error surface, that for extent of reverberation, is also the one that had the best R value (as discussed in Section 3.2.4), indicating robustness against parameter choices. In addition it was the dimension that exhibited the deepest local minimum in the feature subset search (Fig. 4). This may indicate that, of the three learning concepts, the feature space used for reverberation was the best matched.

3.2.4 Model Performance

The results of testing the optimized models are listed in Table 9. The RAE value was lowest for the width and immersion dimensions, at 62.63% and 64.36%, respectively. The reverberation dimension performed slightly worse, with 67.20% RAE. However, by all other measurements of accuracy, the predictor for extent of reverberation performed best.

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 predicted and actual values. The coefficient of determination R^2 was highest for reverberation, indicating that the function accounted for 62% of the variance in the test set.

By examining RAE for all models we see that the worst predictor was at least 32% more effective than the baseline predictor. Therefore we believe that the accuracy of the models suggests high predictive capability for each of the perceived spatial dimensions. These results are encouraging and suggest that objective measurements of recorded music may be mapped successfully to perceived spaciousness, and perhaps to dimensions of music perception that have not been examined in this work.

Table 8. Selected features after running on a nonoptimized machine.*

Concept	Feature*	Summary Function[†]	Feature	Summary Function[†]	Feature	Summary Function[†]
Width (50 %)[‖]	**Tempo envelope autocorrelation PM**	**PF**	**Spectral flatness**	**PA**	**Wideness estimation**	**M**
	Reverberation estimation	**M**	**Δ MFCC 5**	**S**	**ΔΔ MFCC 11**	**M**
Reverberation (40 %)	MFCC 3	M	MFCC 3	PE	MFCC 3	S
	ΔΔ MFCC 13	PA	Key clarity	S	Chromagram PM	PF
	Harmonic change detection	PA	Spectral flux	PA	Pitch	PA
	Δ MFCC 10	S	Δ MFCC 10	PF	Δ MFCC 13	S
Immersion (20 %)	MFCC 6	PE	Spectral centroid	PE	**Tempo envelope autocorrelation PM**	**PF**
	Spectral flatness	**PA**	Spectral kurtosis	SD	**Wideness estimation**	**M**
	Reverberation estimation	**M**	Mode	PE	Pitch	PF
	Δ MFCC 7	S	**Δ MFCC 5**	**S**	Δ MFCC 11	S
	Δ MFCC 11	M	**ΔΔMFCC 11**	**M**		

*Features in **boldface** were picked by algorithm for more than one learning concept.

[†] PM—peak magnitude; M—mean; SD—standard deviation; S—slope; PF—period frequency; PA—period amplitude; PE—period entropy.

[‖] Feature subset's percentile group.

4 DISCUSSION

Our findings are significant to engineers and music listeners—modeling such an important attribute of music gives us a new perspective from which to analyze and process digital musical signals. The work we present leads us to encouraging conclusions. However, there are several limitations to the study, which we will discuss in the following paragraphs, as well as suggestions for future work.

The choices for the three spatial attributes were motivated by work in related fields, as well as our own intuitive experience with recorded musical signals. Because the attributes are nonstandard and have not been examined in formal experiments, it is difficult to know whether they are semantically and perceptually valid. Informally our subject feedback showed us that the spatial definitions were understood by our participants. However, further attribute elicitation and validation should be done for the discovery of significant perceived attributes of spaciousness as they relate to produced musical signals.

The mid/side decomposition of stereophonic audio is a lossless transformation of the left and right channels into an alternate space. It is well known that both the mid and side components of stereophonic sound carry spatial cues. Two of the primary challenges to successful machine learning are avoiding feature selection bias and overfitting. These are more difficult to prevent when using a small data set or a large feature space. Because our data set was relatively small, we needed to minimize the size and redundancy of the feature space. We excluded the mid signal from our model, thereby introducing informational loss to the system. It is possible that the model's performance might have been different if we had used the mid signal instead. We encourage further work that will examine the comparative impact of the mid channel on the model's performance or that concurrently will investigate both signal components on a much larger data set.

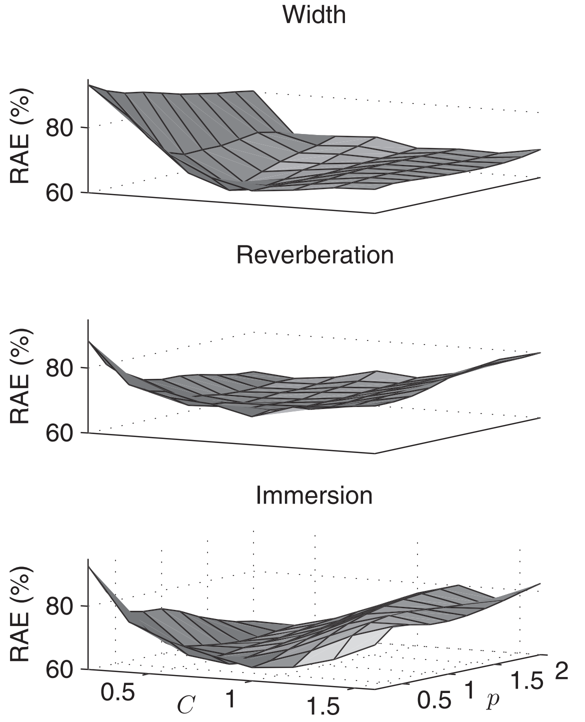

Fig. 5. Relative absolute error surface for machine parameter grid search of kernel exponent p and machine cost C.

Our feature space included only two spatial features. In recent years there have been increased efforts to extract spatial information from audio. These include features for the assessment of spatial quality for multichannel reproduction [26], [45], envelopment resulting from surround sound recordings [46], and spatial analysis of binaural recordings [47]. Many of the metrics in the literature convolve the signal with a head-related transfer function at different angles of rotation, and employ various forms of the interaural cross-correlation function or proportional energy and temporal measurements. Because our experiment required the use of headphones and disallowed free-field listening, we were not sure how robust such metrics would be for the model. However, given the limited number of spatial features in our system and our strategy of using an initial verbose feature set, it may have been advantageous to include other features. We welcome such further efforts with more spatial features in general or smaller, more targeted feature sets.

Our resources were limited, that is, we were not able to employ a large number of experienced listeners for the laboratory experiment. This constraint also limited the number of songs we could use in our database. With more resources we would have conducted a larger scale laboratory experiment, included more songs in the database, and used a higher quality audio format. It is difficult to speak about the scalability of our model to larger populations without a broader experiment. We hope that others in the research field will examine similar MIR-inspired methods of modeling on larger and more diverse populations.

Despite the limitations, our model showed good predictive accuracy and encourages further development of this and related frameworks. For instance, this paper underscores an interest in the semantic relationship between individual features and perceived spaciousness. We used support vector regression with a polynomial kernel to model the learning concepts. Because of the nonlinear nature of kernel methods, we cannot interpret directly the meaningfulness of individual features to the model. Yet other methods of machine learning (such as

Table 9. Model performance.*

	Width	Reverberation	Immersion
RAE(%)	62.63	67.20	64.36
Correlation coefficient R	0.73	0.79	0.76
Coefficient of determination R^2	0.53	0.62	0.58

*All results are averaged from multiple CV.

decision trees) allow the examination of individual features as they relate to the outcome of the model. We look forward to future work with other machine-learning algorithms, as well as semantic analyses of the most successful feature sets.

By parameterizing an important perceived attribute of music and mapping it to measurable quantities of digital audio, a meaningful way of accessing music is provided. We believe that more research on computational models might yield methods for processing signals for top-down control of perceived spaciousness. We can use the parametric EQ knob as a metaphor for such signal modeling and processing. With additional research in this area, the spatial perception of music might be controlled directly by turning one or several spatial perception knobs. Finally this paper presents a framework for modeling a perceived attribute of music that is directly tied to music production. We would like to explore generalizations of this framework to other production-related percepts. These might include concepts such as vocal clarity, rhythmic salience, or listening fatigue.

5 CONCLUSION

We have presented a system for computationally modeling three dimensions of spaciousness in recorded music. First we discussed the concept of spaciousness in the context of previous work in other music-related fields. We outlined a parameterization of spaciousness by three dimensions—the width of the source ensemble, extent of reverberation, and extent of immersion. By conducting two human subject experiments a newly annotated set of stereophonic music recordings was generated along the three dimensions of spaciousness. Lastly a support vector regression function and nonlinear kernel were used to map the music annotations to a set of signal descriptors. Automatic feature selection was used in conjunction with exemplar-based support vector regression to build an algorithmic model of spaciousness. The model was evaluated by multiple CV and found to predict spaciousness more than 32% above the baseline predictor. We therefore conclude that perceived spaciousness of musical recordings can be modeled and predicted effectively along an arbitrary numerical continuum.

6 ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under grant IIS-0844654 and by a Challenge Fund grant and a Dean's grant for student research from NYU's Steinhardt School.

REFERENCES

[1] A. Marshall, "A Note on the Importance of Room Cross-Section in Concert Halls," *J. Sound Vib.*, vol. 5, pp. 100–112 (1967).

[2] W. Keet, "The Influence of Early Lateral Reflections on the Spatial Impression," in *Actes du 6e Congrès International d'Acoustique* (Tokyo, Japan, 1968).

[3] M. Morimoto and Z. Maekawa, "Auditory Spaciousness and Envelopment," in *Proc. 13th Int. Cong. on Acoustics*, vol. 2 (1989), pp. 215–218.

[4] M. Morimoto, H. Fujimori, and Z. Maekawa, "Discrimination between Auditory Source Width and Envelopment," *J. Acoust. Soc. Jpn.*, vol. 46, pp. 449–457 (1990).

[5] A. H. Marshall and M. Barron, "Spatial Responsiveness in Concert Halls and the Origins of Spatial Impression," *Appl. Acoust.*, vol. 62, no. 2, pp. 91–108 (2001).

[6] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among Interaural Cross-Correlation Coefficient ($IACC_E$), Lateral Fraction (LF_E), and Apparent Source Width (ASW) in Concert Halls," *J. Acoust. Soc. Am.*, vol. 104, pp. 255–265 (1998).

[7] M. Barron, "The Subjective Effects of First Reflections in Concert Halls—The Need for Lateral Reflections," *J. Sound Vib.*, vol. 15, pp. 475–494 (1971).

[8] J. S. Bradley and G. A. Soulodre, "The Influence of Late Arriving Energy on Spatial Impression," *J. Acoust. Soc. Am.*, vol. 97, pp. 2263–2271 (1995).

[9] H. Haas, "The Influence of a Single Echo on the Audibility of Speech," *J. Audio Eng. Soc.*, vol. 20, pp. 146–159 (1972 Mar.).

[10] D. de Vries, E. M. Hulsebos, and J. Baan, "Spatial Fluctuations in Measures for Spaciousness," *J. Acoust. Soc. Am.*, vol. 110, pp. 947–954 (2001).

[11] M. Morimoto and K. Iida, "Appropriate Frequency Bandwidth in Measuring Interaural Cross-Correlation as a Physical Measure of Auditory Source Width," *Acoust. Sci. Technol.*, vol. 26, pp. 179–184 (2005).

[12] J. S. Bradley and G. A. Soulodre, "Objective Measures of Listener Envelopment," *J. Acoust. Soc. Am.*, vol. 98, pp. 2590–2597 (1995).

[13] H. Furuya, K. Fujimoto, C. Y. Ji, and N. Higa, "Arrival Direction of Late Sound and Listener Envelopment," *Appl. Acoust.*, vol. 62, pp. 125–136 (2001).

[14] P. Evjen, J. S. Bradley, and S. G. Norcross, "The Effect of Late Reflections from above and behind on Listener Envelopment," *Appl. Acoust.*, vol. 62, pp. 137–153 (2001).

[15] N. Ford, F. Rumsey, and B. de Bruyn, "Graphical Elicitation Techniques for Subjective Assessment of the Spatial Attributes of Loudspeaker Reproduction—A Pilot Investigation," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol 49, p. 543 (2001 June), convention paper 5388.

[16] R. Mason, N. Ford, F. Rumsey, and B. De Bruyn, "Verbal and Nonverbal Elicitation Techniques in the Subjective Assessment of Spatial Sound Reproduction," *J. Audio Eng. Soc.* (*Engineering Reports*), vol. 49, pp. 366–384 (2001 May).

[17] J. Berg and F. Rumsey, "Identification of Perceived Spatial Attributes of Recordings by Repertory

Grid Technique and Other Methods," presented at the 106th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 47, p. 525 (1999 June), preprint 4924.

[18] J. Berg and F. Rumsey, "Correlation between Emotive, Descriptive and Naturalness Attributes in Subjective Data Relating to Spatial Sound Reproduction," presented at the 109th Convention, of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 4, p. 1106 (2000 Nov.), preprint 5206.

[19] J. Berg and F. Rumsey, "Verification and Correlation of Attributes Used for Describing the Spatial Quality of Reproduced Sound," in *Proc. AES 19th Int. Conf. on Surround Sound* (Schloss Elmau, Germany, 2001 June 21–24).

[20] J. Berg and F. Rumsey, "In Search of the Spatial Dimensions of Reproduced Sound: Verbal Protocol Analysis and Cluster Analysis of Scaled Verbal Descriptors," presented at the 108th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 48, pp. 359, 360 (2000 Apr.), preprint 5139.

[21] J. Berg and F. Rumsey, "Systematic Evaluation of Perceived Spatial Quality," in *Proc. AES 24th Int. Conf. on Multichannel Audio: The New Reality* (Banff, Alta., Canada, 2003 June 26–28).

[22] J. Berg and F. J. Rumsey, "Validity of Selected Spatial Attributes in the Evaluation of 5-Channel Microphone Techniques," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 50, p. 522 (2002 June), convention paper 1593.

[23] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, pp. 651–666 (2002 Sept.).

[24] F. Rumsey, S. Zielinski, P. Jackson, M. Dewhirst, R. Conetta, S. George, S. Bech, and D. Meares, "QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction Using an Artificial Listener," presented at the 125th Convention of the Audio Engineering Society, (*Abstracts*) www.aes.org/events/125/125thWrapUp.pdf, (2008 Oct.), convention paper 7595.

[25] R. Conetta, F. Rumsey, S. Zielinski, P. Jackson, M. Dewhirst, S. Bech, D. Meares, and S. George, "QESTRAL (Part 2): Calibrating the QESTRAL Model Using Listening Test Data," presented at the 125th Convention of the Audio Engineering Society, (*Abstracts*) www.aes.org/events/125/125thWrapUp.pdf, (2008 Oct.), convention paper 7596.

[26] P. Jackson, M. Dewhirst, R. Conetta, S. Zielinski, F. Rumsey, D. Meares, S. Bech, and S. George, "QESTRAL (Part 3): System and Metrics for Spatial Quality Prediction," presented at the 125th Convention of the Audio Engineering Society, (*Abstracts*) www.aes.org/events/125/125thWrapUp.pdf, (2008 Oct.), convention paper 7597.

[27] M. Dewhirst, R. Conetta, F. Rumsey, P. Jackson, S. Zielinski, S. George, S. Bech, and D. Meares, "QESTRAL (Part 4): Test Signals, Combining Metrics, and the Prediction of Overall Spatial Quality," presented at the 125th Convention of the Audio Engineering Society, (*Abstracts*) www.aes.org/events/125/125thWrapUp.pdf, (2008 Oct.), convention paper 7598.

[28] F. E. Satterthwaite, "An Approximate Distribution of Estimates of Variance Components," *Biometr. Bull.*, vol. 2, no. 6, pp. 110–114 (1946).

[29] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. (Morgan Kaufmann, San Francisco, CA, 2005).

[30] O. Lartillot, P. Toiviainen, and T. Eerola, "A Matlab Toolbox for Music Information Retrieval," in *Data Analysis, Machine Learning and Applications*, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. (Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin Heidelberg, 2008), pp. 261–268.

[31] H. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond* (Wiley Online Library, 2005).

[32] O. Lartillot, MIRtoolbox 1.3.3 software, Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyäskylä, Finland (2011 June), *MIRtoolbox User's Manual*, https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/MIRtoolboxUsersGuide1.3.3.

[33] H. Fastl, "Fluctuation Strength and Temporal Masking Patterns of Amplitude-Modulated Broadband Noise," *Hear. Res.*, vol. 8, pp. 59–69 (1982).

[34] P. Juslin, "Cue Utilization in Communication of Emotion in Music Performance: Relating Performance to Perception," *J. Experim. Psychol.: Human Perception and Perform.*, vol. 26, pp. 1797–1813 (2000).

[35] C. E. Shannon, "A Mathematical Theory of Communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, pp. 3–55 (2001 Jan.).

[36] R. Plomp and W. J. M. Levelt, "Tonal Consonance and Critical Bandwidth," *J. Acoust. Soc. Am.*, vol. 38, pp. 548–560 (1965).

[37] K. Jensen, "Timbre Models of Musical Sounds," Ph.D. thesis, Datalogisk Institut, Copenhagen University, Copenhagen, Denmark (1999).

[38] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 293–302 (2002 July).

[39] C. Harte, M. Sandler, and M. Gasser, "Detecting Harmonic Change in Musical Audio," in *Proc. 1st ACM Workshop on Audio and Music Computing Multimedia* (New York, NY, 2006), pp. 21–26.

[40] A. Sarroff and J. Bello, "Measurements of Spaciousness for Stereophonic Music," presented at the 125th Convention of the Audio Engineering Society, (*Abstracts*) www.aes.org/events/125/125thWrapUp.pdf, (2008 Oct.), convention paper 7539.

[41] M. Hall, "Correlation-Based Feature Selection for Machine Learning," Ph.D. thesis, Dept. of Computer

Science, University of Waikato, Hamilton, New Zealand (1999 April).

[42] A. J. Miller, *Subset Selection in Regression* (Chapman & Hall/CRC, Boca Raton, FL, 2002).

[43] A. J. Smola and B. Schölkopf, "A Tutorial on Support Vector Regression," *Statist. and Comput.*, vol. 14, pp. 199–222 (2004).

[44] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, Cambridge, MA, 2001).

[45] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "Objective Measures of Listener Envelopment in Multichannel Surround Systems," *J. Audio Eng. Soc.*, vol. 51, pp. 826–840 (2003 Sept.).

[46] S. George, S. Zielinski, and F. Rumsey, "Feature Extraction for the Prediction of Multichannel Spatial Audio Fidelity," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1994–2005 (2006 Nov.).

[47] B. Supper, "An Onset-Guided Spatial Analyser for Binaural Audio," Ph.D. thesis, Dept. of Music and Sound Recording, University of Surrey, Guildford, UK (2005).

THE AUTHORS

A. M. Sarroff

J. P. Bello

Andy M. Sarroff was born in New York, NY, in 1978. He received a Bachelor of Arts in Music degree from Wesleyan University in 2000 and subsequently began a career as a music engineer. He later joined the Music Audio Research Laboratory at New York University and received a Master of Music degree in 2009. He is currently a Ph.D. student in the Computer Science Department at Dartmouth College, Hanover, NH.

His research interests include music informatics; music signal analysis and processing; and machine learning, listening, and perception.

●

Juan P. Bello received a Ph.D. degree in electronic engineering from Queen Mary University London, UK, where he was also a postdoctoral researcher and technical manager of the Centre for Digital Music. Since 2006 he has been an assistant professor of music technology at New York University, New York, and a founding member of its Music and Audio Research Laboratory (MARL).

Dr. Bello teaches and researches computer-based analysis of audio signals and its applications to music information retrieval, digital audio effects, and interactive music systems. He is a member of the IEEE and the International Society for Music Information Retrieval (ISMIR), and a regular reviewer and contributor to digital signal processing and computer music journals and conferences. His work has been supported by scholarships and grants from Venezuela, the UK, the EU, and the US, including, more recently, a CAREER award from the National Science Foundation. He is also a researcher and member of the Scientific and Medical Advisory Board of Sourcetone, a music and health startup.