

# Model-Based Transformation of Spatial Attributes in Digital Music

Andy M. Sarroff  
sarroff@cs.dartmouth.edu  
<http://www.cs.dartmouth.edu/~sarroff>

December, 2010

## 1 Introduction

Recorded music usually communicates one or more auditory scenes to a listener. The scenes give a listener information about the context of the performers—what kind of space they are performing in, what their positions are relative to each other, and what their positions are relative to the performing space. Often, especially in popular music, an auditory scene encompasses acoustic contexts that would never exist in the real world. At other times, auditory scenes are presented as realistically as possible.

This proposal is for a body of research that will give new control over such auditory scenes. The research sets out to provide a framework for transforming an existing musical signal and its auditory scenes into a new one, with entirely new relationships between auditory objects. The research proposes to do so using descriptive numerical vectors; a model that can generate new observable data based upon previous data; and human judgements of spaciousness in music. In doing so, it will advance the state of the art in computational perception; generative modeling; and musical signal analysis, processing, and synthesis.

A plan for research is presented below. In [section 2](#), the purpose of the research will be restated, along with key definitions. The significance of the proposed work is discussed in [section 3](#). Literature directly relevant to the research question and information on my previous work is given in [section 4](#). Approaches for solving the research problem and expected challenges are noted in [section 5](#). Finally, the main points of this proposal are summarized in [section 6](#).

## 2 Statement of Purpose

The main objective of the proposed work is to build a system that can *bidirectionally* map human perception of spatial attributes to digital musical signal. A simplified overview is given in [Figure 1](#). In the top-left vertex of the graph, a musical signal is either decomposed into, or synthesized from, a feature space. The feature space is either an input to, or an output from, a generative model of spaciousness. With feature space

as input, the model is a predictor of perceived spaciousness. When given a target value or class of spaciousness as input, a solution for an optimal feature space weighting is found, and musical audio is resynthesized from an altered feature space.

The edges in the graph represent areas of the proposed research which make one or more hypotheses:

#### **Musical Audio ↔ Feature Space**

- There exists a feature space which may adequately describe the perceived spatial properties of musical audio.
- A transformed feature space may be synthesized to produce a new musical signal with minimal audible artifact.

#### **Feature Space ↔ Generative Model**

- If the dimensions of a feature space are not orthogonal, their dependencies may be determined.
- Across all possible feature spaces for a particular musical signal and target spaciousness, there exists an optimal feature space that can be generated.

#### **Spaciousness Value or Class ↔ Generative Model**

- There exist significant subsets of humans which perceive spaciousness in a similar and consistent manner for certain subsets of music.

## **2.1 Definitions**

Definitions are given for the terminology used: “spaciousness;” “feature space;” and “generative model.”

### **2.1.1 Spaciousness**

Auditory spaciousness is described by [Blauert & Lindemann \(1986\)](#) as “the concept of type and size of an actual or simulated space.” This proposal is not concerned with *actual* spaces, as all space in a recording is *virtual*. The definition is extended to include the implied spatial dimensions of a mixture of sources and their placement in the stereophonic field. Furthermore, perceived spaciousness is not assumed to be unidimensional; spatial impression can be comprised of multiple attributes, each attribute being its own dimension. Spatial dimensions may be modeled independently or as aggregates. No assumption is made of the orthogonality of spatial dimensions to each other.

### **2.1.2 Feature Space**

A feature is a measurable property of an input signal, in this case digital audio. Many features correspond to perceived attributes of music. For instance, the Mel Frequency Cepstral Coefficient (MFCC), commonly used for music and speech analysis, is often

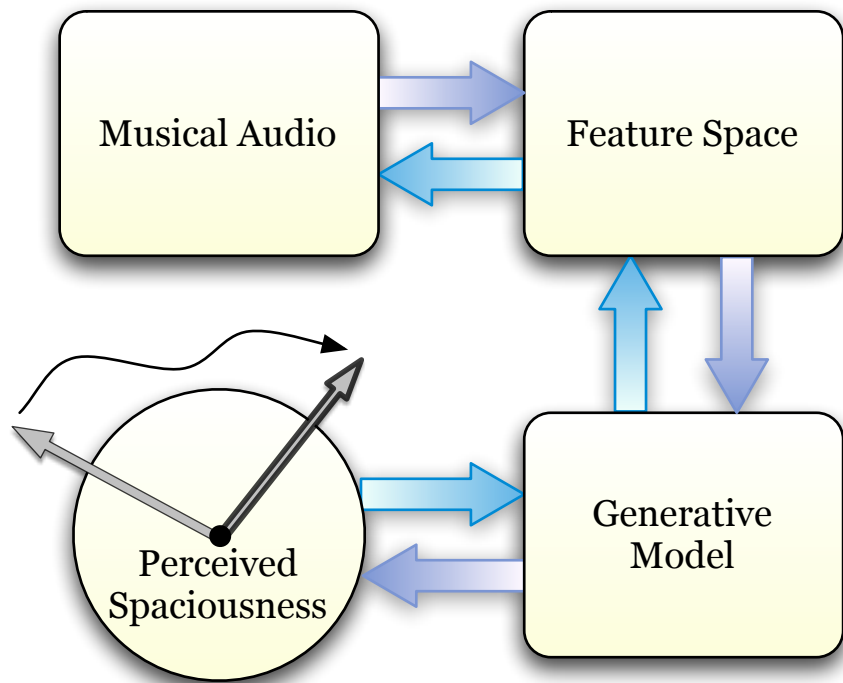


Figure 1: Overview of proposed research.

related to timbre. Feature selection can have a strong impact on the success of a model. As discussed below, there are few reported feature extractors for perceived spaciousness in musical recordings. It therefore becomes critical to build new feature extractors or select the correct combination of existing feature extractors.

### 2.1.3 Generative Model

A generative model is one which can return new observable data. This is based upon a joint probability distribution between already-observed real-world data and labels or continuous values that are associated with that data. In this case, the observable data is a feature space—the descriptive numeric vector for a musical signal. The labels or continuous values are qualitative or quantitative judgments of spaciousness associated with the musical signal. The generated output should be constrained such that it can provide an optimal re-synthesis from original musical signal to an unknown target musical signal.

## 3 Significance

Music listening activities are culturally ubiquitous, implying a motivation to model attributes of music that are perceived in common by people. This motivation is supported by several scientific journals and conferences, such as the International Society for Information Retrieval (ISMIR), the Sound and Music Computing Conference (SMC), the International Computer Music Conference (ICMC), and the IEEE Signal Processing Society (IEEE-SPS). Yet, not all perceived attributes of music are represented equally. In recent years, considerable effort has been allocated to modeling genre, emotion, tempo, timbre, and musical structure, to name a few. Noticeably absent is treatment of spaciousness. When spaciousness has been approached, it has often been in terms of reverberation or multichannel up-mixes, rather than from the perspective of computational perception or music transformation. In fact, many musical feature extraction libraries discard one channel of stereophonic input signal before processing, even though inter-channel differences are important for manipulating spatial cues.

Spaciousness is an important attribute of music production. Musicians and music engineers exploit the virtual spatial environments of their recordings to transmit important information to a listener. For example, [Västfjäll, Larsson, & Kleiner \(2002\)](#) have shown that reverberation time in recordings influences emotional interpretation of music. The skillful manipulation of spatial cues is therefore an important component of the music production process and, consequently, the effectiveness of the production.

New technologies are allowing musicians and music consumers to work with digital media in unprecedented ways. The research proposed here will enable further interaction with the recorded musical signal. Music buyers might personalize the sound of their music collection to preference. And music producers will have new tools to manipulate mixed musical signal. As people are increasingly adapting to and relying upon new methods to manage their music, this research will bring a new layer of top-level effectuality. With a successful model, the spatial attributes of a song could be modified similarly to the way an EQ knob gives music listeners primitive control over their

musical signal.

While this proposal emphasizes its usefulness to the musical domain, an effective model could be applied to other, non-musical, domains. Communications systems and multi-modal human-computer interfaces are increasingly exploiting spatial auditory displays. Model-based transformation of spatial attributes would be a powerful resource in these domains. Additionally, generative models for perceived spaciousness could be used by robots, autonomous agents, and other sensing systems. Finally, this research will advance the state of the art for computational perception, generative modeling, and feature synthesis.

## 4 Background

Background references and discussions are provided below for some of the main areas of the proposed research. In addition, I include my experience in related work.

### 4.1 Spaciousness

Research on the dimensionality of perceived spaciousness, along with qualitative and quantitative methods of evaluation, have been conducted in two principal fields. In 1967, [Marshall](#) determined that “spatial responsiveness,” is a desirable property of concert halls. Since then, research on the physical acoustics of concert spaces has investigated how listeners perceive and how to evaluate such environments. Concert halls (and other listening environments) have been primarily parameterized into dimensions of Apparent Source Width (ASW) ([Keet, 1968](#)) and Listener Envelopment (LEV) ([Morimoto & Maekawa, 1989](#); [Morimoto, Fujimori, & Maekawa, 1990](#)).

Audio quality evaluation has also prompted research, especially for multi-channel sound reproduction systems. Because the quality of these systems hinge on the believability and enjoyability of the displays, there must be an empirical system for qualitative evaluation of such systems. Experiments with various attribute elicitation techniques are reported in [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](#)). Papers about methods of sound capture for optimal spaciousness are also easily found, for instance, in [Theile \(2001\)](#).

Despite these treatments, research on the perceived spaciousness of an *audio recording*, sans reproducing system or physical listening environment, has been limited. The literature widely acknowledges the importance of perceived spaciousness, yet few approaches in terms of digital musical signal analysis are found. (One exception might be new research on Spatial Audio Coding for sound reproduction systems.) When they are, the topics are usually relegated to narrower points, such as re-panning a mixture of sources ([Avendano, 2003](#)). As such, there are also few reports on feature extractors for spaciousness.

## 4.2 Generative Modeling

Generative modeling has been used in several music-related topics. Many of these involve generating new music and sound, such as [Birchfield, Mattar, & Sundaram \(2005\)](#); [Birchfield \(2003\)](#); [Paiement, Grandvalet, Bengio, & Eck \(2007\)](#), rather than the transformation of existing music. Others discuss using methods such as Hidden Markov Models (HMMs) for generating content based upon temporal dependencies in music, for example [Paiement, Grandvalet, & Bengio \(2009\)](#). Generative modeling appears to be growing in popularity as a means for dealing with auditory and visually based signals, however approaches rarely emphasize transforming the input signal into a new one. [Visell \(2004\)](#), suggests such applications.

## 4.3 Feature Synthesis

Matt Hoffman and others at the Princeton Sound Laboratory have built a modular, open source framework, FeatSynth ([Hoffman & Cook, 2007](#)). The system minimizes a distance metric between a target feature vector and the feature vector of a set of synthesized parameters by searching through a parameter space ([Hoffman & Cook, 2006](#)). The usefulness of such a framework for “resynthesis with modification” is suggested throughout the related publications. In an unrelated work, [Le Groux & Verschure \(2008\)](#) synthesize new music from the principal components of a feature vector. Despite these examples, feature synthesized music appears to be in its early stages.

## 4.4 Author’s Previous Work

Research on objective measurements for spaciousness were published in ([Sarroff & Bello, 2008](#)). In that work, two objective measurements, one for wideness of source distribution and another for the extent of reverberation of a musical signal, were proposed and experimentally validated. A medium-scale human-subject study was initiated in 2008 to collect data about perception of spaciousness along three dimensions. The studies were performed on-line and in-laboratory and confirmed consistency of response for the dimensions being evaluated. Together, the human subject studies have collected nearly 3,000 responses.<sup>1</sup> Subsequent to collecting and analyzing human response data, an SVM regressor was trained and evaluated. Those results are reported in [Sarroff & Bello \(2009\)](#). The comprehensive research and results of these stages was reported in the my Master’s Thesis ([Sarroff, 2009](#)).

# 5 Approaches and Expected Challenges

Using the vertices and edges in [Figure 1](#) as a guide, this section comments on possible methodological approaches and poses some expected challenges of the proposed research.

---

<sup>1</sup>The online site is still active and can be accessed at [http://study.smusic.nyu.edu/~andy/NYU\\_Spatial\\_Perception\\_Study/](http://study.smusic.nyu.edu/~andy/NYU_Spatial_Perception_Study/).

### 5.1 □ Musical Audio

The research will necessitate a sizable collection of musical audio for analysis and evaluation. The collection should be diverse enough to represent varying genres, recording techniques, and attributes of spaciousness. In previous work, I collected music from web-based distribution sources. However, the collection was not consistently-high audio quality. Perceived spaciousness can be affected by quality of audio or reproducing system.

### 5.2 □ Feature Space

A set of features known to be descriptive of spaciousness does not exist. Feature selection can be a tricky process, especially when avoiding feature selection bias in machine learning tasks. New methods for feature extraction of spatial attributes may be explored. Alternatively, ideas may be derived from new research in spatial audio coding, or new features might be designed with evolutionary methods, such as genetic algorithms. Perhaps “deep learning” methods may be explored for building new feature representations.

### 5.3 □ Generative Model

Deriving a successful model for the joint probability distribution of observable data and labels will be a research challenge. The generative model should not output just *any* observable data, but data that meets some criterion for smooth transformation from the original signal. As noted above, there are recent examples of generative modeling, largely for producing new musical content. These examples may provide a starting point for building the correct generative model for transformation of spatial attributes.

### 5.4 □ Perceived Spaciousness

It is not yet known exactly which dimensions of spaciousness should be modeled, and if they should be modeled independently or as aggregates to a higher order of perceived spaciousness. Additionally, a database of “ground truth” labels or values needs to be collected. A fundamental challenge to machine learning tasks is ensuring that there is enough data for training, testing, and validation. Most likely, some large scale system for the collection of labels or values will be initiated. There is a wealth of literature on large-scale collection of data from human subjects.

### 5.5 Musical Audio ↔ Feature Space

Once an appropriate feature set has been selected, feature extraction is largely straightforward. However, representation of the feature space may not be. For instance, the space may be summarized by any number of functions. Many analysis methods for musical signal discard temporal dependencies and throw all features into a “bag of features.” This is not necessarily the best approach. If the feature space is large, it may be necessary to reduce its dimensionality.

As noted above, there now exists at least one open source framework for feature-based synthesis. As the framework has not yet been used for the explicit purposes posed here, it is unknown whether it will be suitable for the proposed model or if extension and/or redesign will be necessary.

### **5.6 Feature Space ↔ Generative Model**

Given that a feature space is not too large for the sample size, is well selected, and is representative of the attributes that are to be modeled, its input to and output from a model can be straightforward. If the feature space is transformed before input, transformation coefficients must be kept so that the model's output can be appropriately modified.

### **5.7 Generative Model ↔ Perceived Spaciousness**

Whether spaciousness is best represented as class labels or continuous values is another question, perhaps only to be answered after enough human subject data is analyzed.

## **6 Summary**

This paper proposes a plan for research in model-based transformation of spatial attributes in digital music. The purpose of the research is to build a bidirectional system for mapping human perception of spaciousness to digital recordings of music. By building such a system, humans will have increased control over digital musical content, including ability to synthesize perceptually relevant musical material with new spatial attributes. This research, while presented as a solution to increasing expectations of music makers and music consumers, is also highly relevant to other domains. Spatial attributes of the physical environment inform the way that humans localize objects, communicate with others, and interpret context and surroundings. The virtualization of real environments in communication systems, auditory displays, and human computer interaction systems means that any perceptually relevant control of spatial attributes can be highly useful. Additionally, this research will promote the state of the art for computational perception technology and sound sensing systems, including robots and contextually-aware mobile technology.

I wish to express that this research proposal is not static in concept or direction. Rather, any conversation regarding modifications or alternative research trajectories are encouraged and warmly regarded.

## References

- Avendano, C. (2003, Oct.). Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications. In *Applications of signal processing to audio and acoustics, 2003 IEEE workshop on*. (p. 55-58).  
4.1
- 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.  
4.1
- 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.  
4.1
- 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.  
4.1
- Berg, J., & Rumsey, F. (2003). Systematic evaluation of perceived spatial quality. In *Proceedings of AES 24th international conference on multichannel audio*, Banff, Alberta, Canada.  
4.1
- Birchfield, D. (2003). Generative model for the creation of musical emotion, meaning, and form. In *ETP '03: Proceedings of the 2003 ACM SIGMM workshop on experiential telepresence* (pp. 99–104). New York, NY, USA: ACM.  
4.2
- Birchfield, D., Mattar, N., & Sundaram, H. (2005). Design of a generative model for soundscape creation. In *International computer music conference, barcelona, spain*.  
4.2
- Blauert, J., & Lindemann, W. (1986, Aug). Auditory spaciousness—some further psychoacoustic analyses. *Journal of the Acoustical Society of America*, 80(2), 533–542.  
2.1.1

- 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.  
4.1
- 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)  
4.1
- 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)  
4.1
- 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.  
4.1
- 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.  
4.1
- 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.  
4.1
- Hoffman, M., & Cook, P. (2006). Feature-based synthesis: mapping acoustic and perceptual features onto synthesis parameters. In *Proceedings of the International Computer Music Conference (ICMC'06)*.  
4.3
- Hoffman, M., & Cook, P. (2007). The FeatSynth framework for feature-based synthesis: design and applications. In *Proceedings of the International Computer Music Conference, Copenhagen, Denmark*.  
4.3
- Keet, W. (1968). The influence of early lateral reflections on the spatial impression. In *Reports of the sixth international congress on acoustics*, Tokyo.  
4.1

- Le Groux, S., & Verschure, P. (2008). Perceptsynth: Mapping Perceptual Musical Features to Sound Synthesis Parameters. *IEEE ICASSP 2008*, 125–128.  
4.3
- 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.  
4.1
- 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).  
4.1
- Morimoto, M., Fujimori, H., & Maekawa, Z. (1990). Discrimination between auditory source width and envelopment. *J Acoust Soc Jpn*, 46, 449–457. (in Japanese)  
4.1
- Morimoto, M., & Maekawa, Z. (1989). Auditory spaciousness and envelopment. In *Proceedings of 13th ICA*.  
4.1
- Païement, J., Grandvalet, Y., & Bengio, S. (2009). Predictive models for music. *Connection Science*, 21(2), 253–272.  
4.2
- Païement, J., Grandvalet, Y., Bengio, S., & Eck, D. (2007). A generative model for rhythms. In *Nips'2007 music, brain & cognition workshop*.  
4.2
- Rumsey, F. (1998). Subjective assessment of the spatial attributes of reproduced sound. In *AES 15th international conference: Audio, acoustics and small space*, Copenhagen, Denmark.  
4.1
- 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.  
4.1
- Saroff, A. M. (2009). *Spaciousness in recorded music: Human perception, objective measurement, and machine prediction*. (Master's Thesis, New York University)  
4.4

- Sarroff, A. M., & Bello, J. P. (2008, October 2–5). Measurements of spaciousness for stereophonic music. In *Proceedings of the aes 125th convention*. San Francisco, CA.  
4.4
- Sarroff, A. M., & Bello, J. P. (2009, July). Predicting the perceived spaciousness of stereophonic music recordings. In *Proceedings of the 6th sound and music computing conference (smc-09)*. Porto, Portugal.  
4.4
- Theile, G. (2001). Natural 5.1 music recording based on psychoacoustic principles. In *Proceedings of the 19th international conference of the audio engineering society, schloss elmau, germany*.  
4.1
- 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.  
3
- Visell, Y. (2004). Spontaneous organisation, pattern models, and music. *Organised Sound*, 9(2), 165.  
4.2
- Zacharov, N., & Koivuniemi, K. (2001, July 29–August 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)  
4.1