A PERCEPTION-CENTRIC FRAMEWORK FOR DIGITAL TIMBRE MANIPULATION IN MUSIC COMPOSITION

By

BRANDON SMOCK

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2014

To my mom, the hardest-working, most supportive, and most loving person I know

ACKNOWLEDGMENTS

First, I must thank my advisor Dr. Anand Rangarajan. Without him, I never would have undertaken this journey. His encouragement and support were absolutely vital to my success. I must also thank Dr. Joseph Wilson, who I think of as a second mentor. Our impromptu conversations in the lab were numerous and extremely helpful. He oversaw my non-dissertation research and went out of his way providing guidance that has greatly contributed to the scholar I am today.

I would like to thank Drs. Paul Gader and Arunava Banerjee. Encouragement I received from you as a master's student stuck with me as I moved on toward my Ph.D. I admire both of you greatly and wish I had more opportunities to work with you. I would also like to thank Dr. James Paul Sain, with whom I had numerous conversations, and who patiently and happily filled in gaps in my formal knowledge of music and theory.

To my fellow labmates past and present in the Computational Science and Intelligence lab, I want to thank you for allowing me to draw upon your vast knowledge and talents. Discussing new ideas with you was extremely helpful, and I felt that every time we collaborated, good things happened. Conversations with Taylor Glenn stand out in my mind as being particularly helpful throughout my time as a grad student. Alina Zare set a strong example to follow early on.

Thanks, of course, go to my study participants for volunteering and taking the time out of their busy schedules to help me with my research.

Finally, thank you to my friends. Your support has been invaluable, you have inspired me with your drive and abilities, and you have helped to make my time as a grad student some of the best years of my life!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

A PERCEPTION-CENTRIC FRAMEWORK FOR DIGITAL TIMBRE MANIPULATION IN MUSIC COMPOSITION

By

Brandon Smock

May 2014

Chair: Anand Rangarajan
Major: Computer Engineering

In this work, a new framework is developed for the unrestricted manipulation of timbre in musical composition. Unlike other perceptual musical attributes such as pitch or loudness, timbre is not a one-dimensional property, and is not easily manipulated in a flexible yet precisely-controlled manner. The multi-dimensional nature of timbre has been previously studied using multi-dimensional scaling (MDS) techniques. The outcome of a study involving timbre and MDS is typically referred to as a *timbre space*. Previous timbre spaces have been suitable for a high-level description of some of the underlying dimensions of timbre, but not to a satisfying degree, and have not been suitable for subsequent synthesis of timbre. In this work, four design goals are proposed for a timbre space suitable for composition, along with ways in which to satisfy them. They are: perception, synthesis, interface, and modularity. The primary focus of this work is on the first goal of perception. Understanding the mental representation of timbre at a fundamental level is vital to achieving an effective and compact system for timbre manipulation. Due to the current lack of a satisfying model of timbre perception, we devise new methods and propose new ideas to develop one. As a result, contributions of this work include an improved experimental procedure for the study of perceptual spaces, theoretical contributions to the study of timbre, a novel timbre space suitable for synthesis, a new periodic waveform analysis technique, and a detailed blueprint for the implementation of a system for timbre manipulation in composition.

CHAPTER 1
INTRODUCTION

> The musical value of the computer does not lie, of course, in its ability to duplicate exactly what a real instrument can do, but rather in yielding an extended repertory of sounds, including and going beyond the classes of sounds of actual instruments.
>
> —Risset and Mathews, 1969

Music, it could be said, is where sound meets psychology. While sound can be objectively moved, measured, analyzed, and multiplied, a person's response to music is an intensely subjective experience, an amalgamation of feelings about frequencies. Despite this subjectivity, there are commonalities to the human experience derived from possessing similar perceptive abilities. This allows us to express ourselves through music, for instance, and have a reasonable expectation that with the right background and experience, our message can be understood.

Musicians describe notes as having pitch, having loudness, having duration, and having timbre. However, these are not physical quantities belonging to the notes themselves. They are qualitative properties belonging to the notes' perception. But even if they are experienced qualitatively, there are ways in which we can measure and quantify these responses in the brain. Without much thought, an experienced listener can judge when one note is perceived to be lower or louder or longer than another. The ability to order notes by pitch, loudness, and duration speaks to the fact that though these properties have a subjective quality, they also have a quantitative aspect, describable mathematically even if our brains do not gives us access to the raw numbers.

The ability to order notes by a perceptual property, however, does not apply to timbre. This is because only one-dimensional quantities can be ordered and timbre is not a one-dimensional property. But even as a multi-dimensional attribute, timbre is ill-defined. Timbre is often said to be that about a note, which all other things being equal, allows one to distinguish between two musical instruments, such as a trombone and a clarinet. But this is a description rather than a definition. In a sense, timbre is the unconscious

aggregation of multiple simultaneous perceptions about the source characteristics of a note. But how these perceptions manifest themselves in a singular representation is not entirely clear.

A composer does not have to fully understand the psychology of timbre to use timbre. All composers use timbre because they compose for musical instruments, and musical instruments invoke timbre. But using timbre is not necessarily the same thing as utilizing timbre. The decision made by a composer that has the greatest effect on a composition's timbre is likely to be the choice of instrument itself. This is historically fairly limited because a composer has to consider how widespread an instrument is if he wants to ensure that his piece can be performed. Instruments require a great investment of time to be learned, thus limiting the number of instruments that can be widespread.

Some measure of timbral diversity can be achieved simply by switching between musical instruments in a composition. Within the same instrument, additional manipulation of timbre is also possible because instruments typically possess flexibility in their sound production. But pitch manipulation is much more accessible to composers and musicians. A typical orchestral instrument's range of pitches might cover a third of the typical range utilized in all of music, while its timbral range covers such a small fraction it can hardly be quantified.

Because timbre manipulation is not as accessible to composers, it is not specified on anywhere near the same level as pitch. Some of the primary ways in which timbre manipulation is communicated via notation are through the use of markings indicating style and articulation. Still more manipulation can be suggested by notating natural harmonics in place of typical notes. Precise control over sound production, however, is rarely notated by composers. Often, a number of choices that directly affect how the timbre of a note is perceived are left up to the performer.

A number of the technological and logistical restrictions on timbre production and specification, however, are historical in nature. In modern times, the opportunities for

a composer to exert control over timbre are much greater. Digital recording technology removes most barriers to a composer's work being heard. Therefore, composers can be more diverse in their selection of musical instruments—and thus, timbres—for which to write. As Risset and Mathews note, the computer is potentially any musical instrument. Playing it is greatly aided by programs that abstract many of the low-level details of sound production away and replace them with intuitive interfaces. Unlike the interfaces of traditional instruments, relying on the physics of resonance in a column of air or on a string, these interfaces can take any form. Thus the barrier to becoming a digital virtuoso is much lower than for traditional musical instruments.

Even with advances in digital sound synthesis and reproduction, however, timbre remains a quality of music that is not consciously directed on anywhere near the same level as pitch or loudness. Perhaps this is due to the fact that timbre itself is still not perfectly understood.

The study of timbre has been intimately related to the study of harmonics, and in this sense it can be traced at least as far back as the 17th century, when Marin Mersenne reported hearing individual harmonics and observed that a string is capable of vibrating at harmonic frequencies simultaneously [2]. In the 19th century, larger questions about auditory perception were posed and debated. In 1843, Georg Simon Ohm proposed that the ear decomposes a periodic waveform into its constituent harmonics [3]. In the same decade, August Seebeck argued that harmonics contribute to our sensation of pitch at the fundamental frequency and that the proportions of harmonics in a sound must impart timbre [3]. In 1863, Hermann von Helmholtz published his ground-breaking work On the Sensations of Tone as a Physiological Basis for the Theory of Music, applying his experimental acumen and knowledge of physics to questions of auditory and, in particular, musical perception.

Despite the foundation laid by Helmholtz, the subjective nature of timbre and its complex correspondence with acoustics remained barriers to a full theory of timbre. It

was not until 1940 that timbre began to be described explicitly as a multi-dimensional phenomenon. Modern study of timbre began in earnest in the 1970's following the maturation of multi-dimensional scaling (MDS) techniques. These techniques allowed the multi-dimensional nature of timbre to be studied directly in a spatial representation. Studies of this nature have concluded repeatedly that brightness and spectral flux are overwhelming contributors to our sense of timbral difference between notes. However, the studies themselves were not successful in illuminating timbre at a fundamental level, nor are they applicable to the subsequent generation of new sounds from the spatial representation.

One major barrier currently to timbre manipulation in a perceptual framework such as a timbre space is the lack of an explicit temporal component, so that synthesized timbres can be made of arbitrary lengths and evolve in response to changing stimuli. To this end, we propose a new timbre space and isolate instantaneous aspects of timbre from dynamic aspects of timbre. We assert that such a dichotomy exists in visual perception and argue that studying instantaneous and dynamic aspects of timbre simultaneously, as has been done in previous timbre space studies, muddles the understanding of both.

Isolated from dynamic aspects, we can explicitly investigate the multi-dimensional nature of instantaneous timbre alone. Drawing on a number of MDS techniques, we can ask if a clear orthogonal representation exists, what the meaning of its axes are, and perhaps even investigate multiple coherent orthogonal, or perhaps even non-orthogonal, representations of the same space of instantaneous timbre.

All of this is meant to serve the larger goal of achieving unrestricted timbre manipulation in music composition. Though implementation of synthesis for such a system may be possible by skirting the nature of timbre altogether, we argue that achieving precise control over an inherently perceptual phenomenon is done most effectively with perception as a primary focus. Though physics may drive musical production, perception is the only thing that matters. Though no universal representation may exist, explicitly

dealing with perception offers, in theory, the most compact and intuitive representation possible.

Additionally, the benefits of an orthogonal perceptual representation are precision and clarity. Suppose a unique perceptual representation of timbre could be identified. Methods that bypass perception create their own representation that is not guaranteed to map to the mental representation of timbre in a one-to-one manner. If elements of the representation relate to mental timbre as many-to-one, they are inefficient, including redundancy or distinctions with no bearing on perception. In the worst case, these lead to contradiction, with multiple elements of the representation trying to exert control over the same element of perception. If elements of the representation are in a one-to-many relationship with timbre, they are ambiguous and the representation is lossy. Thus a one-to-one mapping with the perception of timbre, referred to here as a perception-centric approach, is essential to a compact, consistent, non-lossy, and precise representation of timbre. This may not be possible if humans differ greatly in their timbral perceptions, but it may be possible to achieve optimally *on average*.

This work potentially represents a step forward in both our understanding of timbre and our implementation of a system for its unrestricted utilization. But it is not without its limitations. In developing the first instantaneous space of timbre, we have had to simplify some of our experiments to first demonstrate their possible utility, since the experimental procedure used in this work is relatively time-consuming. As a result, though we believe there are no inherent limitations on our experimental procedure, the data set used in this work to learn the spatial representation does not cover the entire range of instantaneous timbres. This limits us to studying only those properties that can be gleaned directly from our data set. A few properties with some potential impact on perception that could have been studied, such as the phases of harmonics, were also omitted due to their uncertainty in a small data set. With phase in particular, there exists no theory in the literature on how it might impact the perception specifically of pitched

sounds with stationary properties, which are the only sounds used in the experiment in this work. Therefore, no domain knowledge could be brought in to aid in its analysis for a small data set. Despite these limitations, we believe there is much to be learned from our work.

In Chapter 2, we review in detail the literature on timbre study, giving particular attention to work towards multi-dimensional spatial representations. In Chapter 3, we introduce our theory of timbre, which isolates instantaneous aspects from those that have a temporal component. We outline our experimental procedure for learning and interpreting a space of instantaneous timbre. In Chapter 4, we demonstrate our results, including a set of common perceptual axes and their objective and subjective interpretations. In Chapter 5, we analyze the results in more detail and place them in their proper context within the larger field of timbre study and our goal of implementing a system for unrestricted timbre manipulation. We also outline future directions for our work, detailing how to overcome the limitations artificially imposed by our experimental procedure and move even closer to achieving our primary goals.

Contributions of this work include an improved experimental procedure for the study of perceptual spaces, theoretical contributions to the study of timbre, a new timbre space suitable for synthesis, a new periodic waveform analysis technique, and a detailed blueprint for the implementation of a system for timbre manipulation in composition.

CHAPTER 2
LITERATURE REVIEW

In this work, we propose a new framework for achieving unrestricted timbre manipulation in musical composition. Achieving this goal requires knowledge in a number of areas including psychoacoustics, signal processing, and machine learning. As the set of knowledge required is rather diverse, very few comprehensive frameworks have been proposed for synthesizing arbitrary timbres from a perceptual description, particularly in musical composition. Although, a number of papers have addressed portions of such a framework in depth. We review these here.

In Section 2.1, we outline the history of formal timbre study, beginning with Seebeck and Helmholtz in the 19th century, and progressing up to analyses involving sophisticated machine learning techniques into the present day. In Section 2.2, we review multi-dimensional scaling (MDS) techniques in depth, as these have formed the basis for timbre study for more than the last 40 years. In Section 2.3, we introduce the concept of the *timbre space*, the primary result of the application of MDS techniques to formal timbre study. In Section 2.3.2, we detail some of the limitations of prior studies involving timbre space, which we attempt to improve upon in this work. In Section 2.4, we introduce ranking methods, the use of which allows us to generalize key aspects of timbre space production and expand the potential approaches used. In Section 2.5.1, we look at general models of musical sound synthesis, focusing on additive synthesis methods, which are theoretically general enough to synthesize any sound. Finally, in Section 2.5.2, we address approaches to the problem of interface, which is a key issue for any system for timbre manipulation in composition to actually be implemented.

2.1 Timbre

The first known theory on the physical basis of timbre perception was arguably proposed by August Seebeck in 1844. Among a dispute with Georg Ohm over pitch perception, Seebeck proposed that the higher harmonics of a fundamental frequency,

rather than themselves be considered distinct tones, must reinforce the perceived pitch of the fundamental tone and determine its timbre [3]. But Seebeck did little to further the idea. This is likely why it is often credited to Hermann von Helmholtz, who in his 1863 work defended Ohm's notion of pitch perception and augmented it with a highly-developed theory of how harmonics arise and contribute to timbre [4]. Through his incredible experimental acumen, it was Helmholtz, not Seebeck, who laid the foundation for our understanding of timbre. Much has been done to further explicate it since. Yet our understanding of timbre still seems to be rather incomplete.

Timbre is described by musicians using words such as bright, hollow, rich, dark, mellow, metallic, and warm. It is what gives us the ability to distinguish the same note played on a trumpet from one on a clarinet, or a piano from a violin. No matter its physical basis, timbre is fundamentally a perceptual property. The four principal perceptual properties of a note are considered to be pitch, loudness, duration, and timbre, with timbre encompassing all which is unclaimed by the first three [5]. Under the proposition by Seebeck that the relative intensities of the harmonics are what determine timbre, it is tempting to theorize that the four perceptual properties of a note have direct analogues in the four physical properties of fundamental frequency, amplitude, length, and overtone series. In 1934, Harvey Fletcher showed that the relationship between these properties is not one-to-one, however, and in fact, the perceived timbre of a note is a function of fundamental frequency, amplitude, *and* overtone series [6].

Helmholtz himself recognized that our basis for the identification of musical instruments has additional qualities not captured strictly by the relative intensities of partials, including transient behavior at the onset of a note. But he sought to focus on just those stationary properties that are primarily captured by the relative intensities of partials, what he termed *musical quality*, and in this regard he was particularly insightful. He described the relationship between several verbal timbre descriptions and their corresponding distributions of energy among the various harmonics. He considered

how some instruments produce formants, much like the vocal tract in the production of vowels, and pondered the contribution to musical quality. He even noted that the first six harmonics of a note form a major chord and indicated how the presence of a strong seventh harmonic could contribute to a dissonant musical quality. Helmholtz notably was also the first to demonstrate experimentally that phase has no noticeable effect on musical quality, although some researchers have observed phase effects for timbre in general [7].

It should be noted that Helmholtz was sure to point out that his lack of evidence that phase affected musical quality did not imply that phase has no bearing on other aspects of timbre. In fact, he mentioned that theoretical considerations led him to hypothesize that phase almost certainly has an effect on aspects of timbre other than musical quality. Plomp and Steeneken note that Helmholtz's conclusions about phase are often misunderstood [7]. Based on this and other considerations, it is reasonable to conclude that a number of researchers have overlooked the distinction Helmholtz made between *musical quality* and other aspects of timbre. This is a point to which we return later in this section.

Many connections have been made between physical properties of notes and their perceptual effects. Shepard refers to this as the *psychoacoustic* approach to the study of timbre [8]. For instance, in 1962, Fletcher et al. reported their findings that the key feature in determining the warmth of piano tones was the inharmonicity of the partials [9]. Discovering how sound can be manipulated physically to produce perceptual effects is an important aspect of timbre research. But it still does not quite explain the perceptual nature of timbre itself. Shepard refers to the study of the inherent relationships between perceived stimuli independent of their relationship to physical parameters as the *cognitive psychological* approach to timbre study.

In determining the full perceptual basis of timbre, it is insightful to draw an analogy with color perception in vision. Helmholtz notes that at least as early as 1686, it was known that all colors can be shown to be a combination of three fundamental colors.

While Helmholtz drew an analogy between color decomposition and harmonic, or frequency, decomposition, it is just as easy to draw an analogy between color and *tone color*, which is a term often used in place of timbre. That all colors can be decomposed into three fundamental colors implies that there are three orthogonal dimensions of color perception. In other words, colors cannot be ordered or organized spatially in a manner consistent with our perceptions of similarity without a space of at least three dimensions. Don Lewis and M. J. Larsen were the first, in 1940, to explicitly refer to timbre as a multi-dimensional phenomenon [10]. This is in contrast to pitch, loudness, and duration, which can all be expressed as one-dimensional (although there may be reason to consider a more general notion of pitch to be multi-dimensional [8], but with an easily identifiable one-dimensional projection that we normally refer to as pitch). Notes can be ordered by their pitch or their loudness, but not by their timbre.

In 1890, Stumpf enumerated over 20 verbal axes of timbre such as smooth versus rough, and wide versus narrow [11]. However, this enumeration does not make it clear how many orthogonal axes of timbre there are, nor how their relative strengths might be perceived in our perception of dissimilarity between different timbres. A breakthrough came with the maturation of multi-dimensional scaling (MDS) techniques in the 1960s [12–16]. These techniques allowed perceptual judgments about the dissimilarity between objects to be aggregated and fitted to a spatial representation that best explained the judgments. In other words, experiments about perceived timbre differences between pairs of sounds could be used to determine orthogonal dimensions of perceptual timbre. These experiments are covered in detail in Section 2.3. A review of MDS techniques is covered in Section 2.2.

The main problem with these experiments for fully explaining timbre perception is that they do not distinguish between fundamental and compound aspects of timbre. It has been demonstrated by numerous researchers that dynamic, or compound, attributes of sound contribute greatly to our overall perception of timbre and identification of musical

instruments. But in my opinion, Helmholtz was astute in recognizing the need for a simple term (his *musical quality*) to describe the most fundamental aspects of timbre perception. Unfortunately it appears as if numerous researchers have misrepresented Helmholtz's theory of timbre as oversimplified, as if Helmholtz intended for his theory to be a comprehensive description of timbre. Kai Siedenburg and Christoph Reuter discuss as recently as 2012 how the composer Iannis Xenakis exposed the "insufficiencies" of Helmholtz's theory with his 1956 composition *Pithopratka* [17].

On the contrary, it seems more likely that Helmholtz only intended to explicate the most fundamental aspect of timbre, what today might be referred to by some as *tone color*. Unfortunately, there is no precise dichotomy between fundamental and compound aspects of timbre in the present vocabulary, such that *tone color* only refers to this fundamental concept. To clarify this, we propose a new term that is free of any ambiguity and which captures the essence of the most fundamental aspects of tone color: *instantaneous timbre*. We mention it here briefly but expound on it more in Chapter 3.

All those aspects of timbre that are most fundamental can be described as instantaneous properties, or described as a quantity at a single point in time. This is similar to how, in color perception, those aspects of color which are most fundamental can be studied all within a single pixel. Furthermore, to study color perception at a fundamental level, it could be confounding to study the dissimilarity between objects with compound color properties, such as a shiny blue ball and a matte red block. The degree to which an object is matte is a compound property of color perception, dependent on the relationships between colors over an area. Likewise, studying notes with non-stationary properties muddles the difference between instantaneous timbre and temporal, or dynamic, timbre. Much of the methods I set out in Chapter 3 will be motivated by the desire to study instantaneous timbre, which will form the basis for a fundamental space of timbre on top of which composite timbre objects can be built.

2.2 Multi-Dimensional Scaling (MDS)

Multi-dimensional scaling (MDS) has played a major role in timbre research since 1970. Before addressing timbre spaces in more detail, it seems prudent to discuss the preceding parallel development of the MDS methods that underlie their estimation.

The origins of modern MDS can be found in a paper from Young and Householder [18] from 1938. They showed that a configuration of points in an N-dimensional space could be reconstructed given just the set of mutual distances between them. They also showed the necessary conditions for a given set of supposed mutual distances to actually correspond to a configuration of points in Euclidean space, and how to determine the minimum dimensionality of the space.

A configuration of points is defined to be invariant to rigid motion transformations. These are translation, rotation, and reflection. Thus an infinite number of sets of points correspond to the same configuration. Since only a configuration of points can be recovered given their mutual distances, the absolute coordinates of the original points are lost when represented only with their mutual distances. However, their relative coordinates are preserved, which is essentially what Young and Householder showed. A fourth kind of transformation, scaling, which multiplies all of the points by a fixed number, does not preserve the distances between the points but does preserve the ratio of the distances between the points. If only preserving the ratios is important, then a configuration of points could be defined to be invariant to four kinds of transformations.

Today, the term *multi-dimensional scaling* refers to any method that tries to find coordinates for points in an N-dimensional space that best preserve a set of mutual dissimilarity values between them. Points can represent any set of objects such as words, faces, colors, or foods. Dissimilarity values are usually thought of as measures of distance but they can also be measures of correlation or even just association. Much of the promise of MDS is that once points are placed into a space, the space itself can be analyzed to determine the underlying causes of the original dissimilarity values. In this way, a purely

mathematical analysis can lead to an analysis that explains what factors contribute to the dissimilarities between the points or objects in the first place.

In psychology, MDS is commonly used to estimate the dimensions of a perceptual space, such as the mental representation of human faces. There are several methods used to estimate dissimilarity values between a set of objects. In the most straightforward case, subjects are asked to numerically rate the similarity or dissimilarity between each pair of objects in a set. In this case, the subjects quantify the dissimilarity directly. However, the act of consciously quantifying what is normally a qualitative perception could lead to inconsistencies in the values provided by an individual subject.

This inconsistency can be mitigated by the use of paired comparisons. Rather than rating the objects on an unfamiliar scale that can vary over time, objects are merely compared with each other. If there are N objects, there are $N(N-1)/2$ possible comparisons. When the information is strictly which object is greater, without taking into consideration by how much, each comparison is a binary comparison. The set of comparisons can then be analyzed to determine a ranking or rating of the stimuli. A ranking is an evenly-spaced rating, so it contains no information on uncertainty in the rankings, whereas the numbers in a rating can have any relationship. Objects that are rated extremely close together have a lot of uncertainty in their order in the ranking.

Research into perceptual spaces does not rank stimuli, but rather the pairwise distances between all stimuli. Since there are $N(N-1)/2$ pairs given N objects, the number of comparisons between pairs is on the order of N^4. One method that has been used to compare and rank a set of distances is the method of triads [19]. In this method, objects are presented in sets of three and the subject is asked to determine which two objects are the most similar and which two are the most dissimilar. Among a set of three objects, there are $\binom{3}{2} = 3$ ways to pick two objects from the set, and these three pairs are ordered from most similar to least similar. Quantitative dissimilarity values are then derived from the entire set of pairwise comparisons by summing the number of times a

pair was judged more similar than another pair. However, this covers only a portion of the total number of possible comparisons between all pairs of pairs, and the proportion it covers grows smaller as the number of objects increases. Furthermore, the simple summation of favorable judgments is imprecise, and even when all of the comparisons obey transitivity, ties naturally arise between different pairs. For these reasons, we consider the use of more general ranking methods, covered in Section 2.4.

2.2.1 Euclidean or Metric MDS

Young and Householder's method to estimate the coordinates of points in a space is an exact procedure that works for distances that can be shown to correspond to an actual configuration of points in Euclidean space—sometimes called *infallible* data. Dissimilarity values derived from psychological experiments hardly fit this description. Even with distance values that are essentially Euclidean, noise or other slight errors can cause a set of distances to not correspond exactly to any configuration of points. In 1952, Torgerson published an extension to Young and Householder's method that could be used with fallible data [12]. In this case, a configuration is found that minimizes the error between the input set of distances and the distances between the points in the output configuration.

Even allowing for noisy data, however, Torgerson's method still assumes dissimilarities that are essentially Euclidean. A problem arises when the dissimilarities are instead a *function* of Euclidean distances, like those estimated from psychological experiments. One way to deal with this problem is to first convert the dissimilarities into Euclidean distances, usually by assuming the form of the function that relates the two quantities. Torgerson outlined two typical scenarios when dissimilarity values are non-Euclidean and proposed methods to try to convert them into Euclidean distances. This works as long as the function required to convert a set of similarities into a set of Euclidean distances is known or can be determined. But as Shepard [13] pointed out, the form of the function

cannot be assumed in general. Since in many situations of interest dissimilarity values are non-Euclidean, this poses a major problem.

2.2.2 Non-Metric MDS

No matter what the form of the unknown function is to convert from dissimilarity to Euclidean distance, though, the function ought to be monotonic. This means that the order, or rank, of the dissimilarity values is preserved when they are converted to Euclidean distances. Shepard, therefore, proposed to use monotonicity as the optimality criterion. In other words, the set of points that is to be found in some space of minimal dimensionality is the one that best preserves the order of the distances, rather than the distances themselves. According to Shepard it is "a surprising outcome that the two conditions of monotonicity and minimum dimensionality (which seem nonmetric or qualitative in nature) are generally sufficient to lead to a unique and quantitative solution" [13]. In fact, Shepard's method even estimates the unknown monotonic function. Because Shepard's method only relies on the ranking of the dissimilarities, it was the first method that could be used directly on any set of dissimilarity values without knowing anything about how they were derived.

Shepard's method is outlined as follows. Let us assume there are N objects and thus $N(N-1)/2$ dissimilarity values, $\{\delta_{ij}\}$, between all unique pairs of objects. First, create an initial configuration of N points in $(N-1)$-dimensional space. Shepard suggests starting with a configuration such that each point is equidistant from all other points (this is straightforward to create for any N). Rank the set of distances, $\{d_{ij}\}$, between each pair of points and compare to the ranking of the dissimilarities. Suppose the rankings are in ascending order of distance. A pair of points whose ranking is too high (are too similar) compared to their target ranking should be moved away from each other. A pair of points whose ranking is too low (are too dissimilar) should be moved toward each other. The amount by which to adjust a pair of points is proportional to the difference between their actual distance rank and their target dissimilarity rank. Thus each point is acted on by

"$N-1$ forces that are tending to pull that point towards those other points that are too distant and away from those points that are too close." After the points are adjusted, the ranking comparison is performed again and the procedure is iterated.

As Shepard points out, for any ranking of distances between N points, a configuration of points can always be found in $(N-1)$-dimensional space that conforms to these rankings. But this configuration may have little utility since the dimensionality is on the same order as the number of points. To induce the points to settle into a lower-dimensional space, Shepard performs another adjustment, a monotonic warping, during each iteration. He takes points that are close together and brings them even closer together, and pushes points that are far apart even farther apart. This is analogous to taking points that lie on a semi-circle in two dimensions and gradually stretching the semi-circle into a line, which is one-dimensional.

After convergence, even though the points may lie in a lower-dimensional space, they are still represented using $(N-1)$ coordinates. Therefore, principal components analysis (PCA) is performed to find the lower-dimensional subspace that the points have settled into. PCA identifies orthogonal axes that sequentially account for the most variance in the data. The first axis represents the direction in which the data has the most variance. The second axis represents the direction, orthogonal to the first, which has the most remaining variance, and so on. When nearly all of the variance in the data is accounted for, the rest of the axes may be discarded. This yields a lower-dimensional representation that still accounts for nearly all of the variance, or spatial extent, in the data.

It might seem as though only constraining the MDS solution to have the same ranking of dissimilarities could lead to too wide a range of potential solution configurations. However, each unique pair of points represents a dissimilarity value and each unique pair of dissimilarity values represents an inequality constraint on the solution. The number of unique pairs of a set is $O(N^2)$ so the number of inequality constraints, which is unique pairs of pairs, is $O(N^4)$. In other words, it grows rapidly as the number of points grows.

Therefore, as the number of points significantly exceeds the number of true dimensions, the seemingly weak constraint of monotonicity becomes much stronger and leads to an essentially unique solution.

Shepard's method is intuitive and effective but perhaps lacks a more formal justification. Kruskal [15] made it his goal to improve and formalize Shepard's methods. Kruskal noted that "it does not appear possible to describe [Shepard's] procedure as one which minimizes some particular measurement of nonmonotonicity." To remedy this, Kruskal proposed a measure of non-monotonicity, a sum-of-squares error he called the *stress*, given by Equation 2–1,

$$S = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}. \tag{2–1}$$

The stress measures the smallest total amount by which the set of pairwise distances between points in a configuration, $\{d_{ij}\}$, would have to be adjusted for them to match the order of the corresponding dissimilarities. Subsequently, to calculate the stress we have to calculate the nearest monotonically non-decreasing set of distance values $\{\hat{d}_{ij}\}$ to the set $\{d_{ij}\}$. This is called isotonic regression, or monotone regression, as Kruskal refers to it. A companion paper [16] details the method for performing such a regression.

Kruskal's method to find the configuration of points that minimizes the stress is iterative just like Shepard's, but how pairs of points are adjusted has a key difference. Shepard adjusts pairs of points if their distance rank is too high or low, and the amount of adjustment is proportional to the discrepancy in rank. Kruskal, instead, adjusts pairs of points if their distance value is higher or lower than their isotonically regressed distance value, and the amount of adjustment is proportional to this discrepancy. This means that a pair of points is still adjusted if it has the wrong ranking, like with Shepard, but it also means that a pair of points can be adjusted simply if another pair of points on one side of its current ranking belongs on the opposite side. In other words, pairs of points can be adjusted even if they technically have the correct ranking, because their adjustment helps

to order the other incorrectly ranked pairs of points faster. In this sense, Kruskal's method may not be radically different but may converge faster. This is one way in which a more formalized approach led Kruskal to improve upon Shepard's method.

Additionally, Kruskal does not stretch the points to induce them into a lower-dimensional space, a method used by Shepard that is reasonable but not principled. Rather, he suggests running his procedure multiple times with different dimensionalities and looking at how much the stress of the minimum-stress configuration goes down as the dimensionality increases. Whenever increasing the dimensionality results in only a marginal decrease in stress, then the proper dimensionality has been determined.

The answers produced by Shepard's and Kruskal's methods are likely to be similar but not exactly the same. Kruskal's method can be viewed more easily in terms of optimizing a precise measure of monotonicity. In fact, although no formal proof is given, Kruskal even gives a reason to believe his method may produce the maximum-likelihood configuration of points given just the rank order of their distances. It is with these kinds of considerations that Kruskal's method is seen as a clear, if modest, improvement over Shepard's.

2.2.3 Weighted Euclidean MDS

While Torgerson's method is appropriate in a number of contexts, its use in psychological applications creates an additional problem to the one addressed by Shepard and Kruskal. That is, whenever multiple subjects are asked to provide dissimilarity values, the values rarely exhibit perfect agreement. The question then becomes, how best should one combine the conflicting information from multiple subjects? Horan [20], in describing the problem and how best to approach it, used the terms *common attribute space* and *individual perceptual space*. An attribute space, which is independent of any individual subject, is a space of all the dimensions, or attributes, that contribute to dissimilarity judgment of objects in that space. An individual perceptual space is an individual subject's own version of the attribute space, where an individual may weight

some dimensions as more important than others in dissimilarity judgments, or even ignore some dimensions altogether. More detail on this model is given in Section 3.2.3. The individual perceptual space is what would result if MDS were applied to a single subject's set of dissimilarity judgments. But recovering the attribute space is the likely goal of any application of MDS in a psychological experiment.

The default approach to combining conflicting information from multiple subjects might be to average the dissimilarity values from each subject and then perform MDS as usual. Horan showed that this results in a space that distorts both the attribute space and the individuals' perceptual spaces. In 1970, Carroll and Chang [21] proposed an algorithm called INDSCAL to recover the attribute space and the weights on each dimension that lead to the individual perceptual spaces of each subject. It is interesting to note that the weights themselves can be plotted as a *subject space* to show how different subjects' dissimilarity judgments cluster.

Normally, MDS produces a configuration of points whose axes are arbitrary, due to the rotational invariance of the configuration. This hinders the task of attributing dissimilarity values to underlying causes, a procedure called *factor analysis*, because all sets of orthogonal axes must be considered. An interesting outcome of the model used by Carroll and Chang is that in theory it produces a set of principal axes. As Carroll and Chang note, "the model is by no means completely general... In particular it does not allow differential rotation as well as differential weighting of dimensions." But the model is likely a good starting point for determining the most meaningful set of axes possible. This simplifies subsequent analyses that try to determine the underlying causes of the dissimilarity values. It is interesting to note that the conflicting information provided by multiple subjects is actually complementary in this framework and provides more information for choosing the most meaningful orientation of the axes.

2.2.4 Specificities and Latent Classes

In 1993, Winsberg and De Soete [22] extended the weighted Euclidean model to incorporate specificities and latent classes, creating an algorithm called CLASCAL. Specificities were first incorporated into a weighted Euclidean model by Winsberg and Carroll in 1989, but were in other models prior [23]. A specificity is a property of a stimulus that contributes to perceived distances between stimuli but is only exhibited by a single stimulus in the set. Because it is only common to one stimulus, the presence of specificities in multiple stimuli has the potential to yield a high-dimensional, degenerate solution or warp a low-dimensional space found for dimensions shared by all stimuli. Therefore, modeling specificities explicitly can improve the model of the dimensions shared among all stimuli when the stimuli are complex and highly varied.

On the other hand, the use of specificities adds a number of additional parameters, thereby increasing the uncertainty in the estimate of each one. Therefore, one has to be careful to ensure that the use of specificities is reasonable and that there is enough data to reliably estimate the dimensions common to all stimuli. Latent classes are a simplification to the weighted Euclidean model where each subject is assumed to belong to one of a much fewer number of classes, where all subjects in a class weight the dimensions the same way. When subjects do exhibit a high degree of similarity in their distance judgments, the addition of latent classes can be seen as a useful way to reduce the total number of parameters and increase the certainty of each one's estimated value. On the other hand, if subjects exhibiting significant differences in their dissimilarity judgments are grouped together, the latent class model could possibly be a worse fit than one without the use of latent classes.

2.2.5 Isomap

MDS is useful for identifying lower-dimensional structure embedded in a higher-dimensional space. Using the pairwise distances computed between all points in the higher-dimensional space, MDS can attempt to find a lower-dimensional representation that preserves

these distances. The lower-dimensional space represents the intrinsic degrees of freedom in the data. Sometimes, though, the observed configuration of data points in the higher-dimensional space is a non-linear function of the underlying degrees of freedom. In this case, the distances between points in the high-dimensional space are not Euclidean with respect to the underlying parameters in the lower-dimensional space. However, if there are enough points, the distances in the higher-dimensional space are likely to be *locally* Euclidean, meaning they would agree with their lower-dimensional counterparts within a certain small neighborhood around each point. Often, data in this case are said to lie on a *manifold*. The shortest curve between points such that the curve stays on the manifold is referred to as the *geodesic distance* between the points. The geodesic distance between points in the higher-dimensional space corresponds to the Euclidean distance between the points in the lower-dimensional space.

In 2000, Tenenbaum [24] came up with a clever extension of Euclidean MDS to estimate the lower-dimensional embedding of points in the case of non-linear or manifold structure. Since the geodesic distances in the higher-dimensional space should correspond to the Euclidean distance in the lower-dimensional space, MDS can be used to discover the lower-dimensional structure so long as there is an estimate of the geodesic distances between all points. Tenenbaum's algorithm, called Isomap, is a scheme for estimating the geodesic distances between all points and then using Euclidean MDS to find the lower-dimensional embedding. To estimate the geodesic distances, a graph structure is created for the entire configuration of points. Each point is a node in the graph, and points are connected only to their closest neighbors in the higher-dimensional space. The edge weight connecting neighboring points is the Euclidean distance between the points, which locally should be the same in both the higher and lower-dimensional space. Then an all-pairs shortest path algorithm is computed on the graph, yielding an estimate of the geodesic distances between all points.

Since the configuration of data lying on a manifold is locally Euclidean, Isomap is often referred to as a manifold-learning algorithm, because it can discover the underlying lower-dimensional structure of such data. The outcome of Isomap clearly depends on the neighbors chosen to be connected in the corresponding graph, since only neighbors whose distances are roughly equal in the higher and lower-dimensional representations should be chosen.

2.3 Timbre Spaces

The idea behind timbre space research is to collect data on the perceived degree of dissimilarity between all pairs of a set of sounds and use MDS to fit the sounds into a space, such that the distances between sounds correspond to the reported dissimilarities. From there, further analysis can be done to uncover the individual dimensions that contributed to the perceived dissimilarities. Most methods make the assumption that the space is Euclidean and that the relevant dimensions will be orthogonal. The orientation of the space is arbitrary in basic MDS methods, which means it is indeterminate in which directions the principal axes go. Although, by assuming that the same orthogonal axes are relevant for everyone but that different people will weight different axes differently, more advanced methods can simultaneously determine the principal axes and the weights each individual places on them.

Section 2.3.1 outlines the last several decades of research into timbre spaces. This history has closely followed developments within the MDS community, with each new MDS model applied to confirm or improve upon the results achieved with the previous models. Despite the developments within MDS, the paradigm for conducting these experiments has remained remarkably unchanged since the first applications over 40 years ago. Section 2.3.2 discusses some of the primary shortcomings of conventional timbre spaces.

2.3.1 History of Timbre Spaces

Plomp [19] was the first to apply MDS to timbre, using Kruskal's algorithm on a set of dissimilarities derived from triadic comparisons. However, Plomp did not report much about his MDS findings, only mentioning that it took a three-dimensional solution to adequately explain the dissimilarities. Instead he was interested in the role of the amplitude pattern of the harmonics in determining timbre. His stimuli were the steady state portions of notes of the same pitch played on different instruments. For the same set of stimuli used to get the dissimilarities, he calculated a simple distance measure based on the difference in amplitude pattern of the harmonics. Comparing the set of perceptual dissimilarities with the calculated distances, he found a correlation value between 0.81 and 0.86. From this he concluded that the timbre of steady state tones was mostly determined by the amplitude pattern of the harmonics.

Wedin [25] was the first to report the spatial results of applying MDS to study timbre. His stimuli were notes from nine musical instruments, all playing the same pitch and approximately equal in loudness and duration. He presented stimuli in pairs and asked subjects to rate the similarity between the two stimuli on a scale of 1 to 10. Using MDS, he found a three-dimensional solution that accounted for 76% of the total variance. Each stimulus tended to fall predominantly along some dimension. Wedin explained the dimensions in terms of the different spectral envelopes of the stimuli, with each dimension having a characteristic spectral envelope.

Miller [26] was the first to apply MDS to study timbre using synthetic stimuli, and the first to do so using Carroll and Chang's INDSCAL algorithm [21]. The synthetic stimuli were designed to roughly resemble actual musical instruments (though not to the degree that they would be confused for the real thing) using a few parameters that were systematically varied. For each possible pair of stimuli, subjects were asked to quantify the dissimilarity between the two stimuli on a scale of 1 to 9, with 9 representing the most dissimilar. In the first run of the experiment, fundamental frequency was included

as a parameter that was varied to create the stimuli. Interestingly, this led to reported dissimilarities that were overwhelmingly dependent on the difference in fundamental frequency. This indicated to Miller that he needed the stimuli to all have the same fundamental frequency so that timbre-related differences could emerge. In the second run of the experiment, three timbre-related parameters were varied to create the stimuli. Unlike general MDS algorithms, INDSCAL produces a set of principal axes that are fit by assuming each subject weights the axes differently in their personal timbre space. In this case the number of harmonics included, which is related to brightness, was found to be the most salient of the three parameters, followed by envelope type, and lastly, rate of onset of the harmonics.

The study by Grey [27] is perhaps the most well-known. He analyzed notes from 16 different instruments and created synthetic reproductions where he could precisely control the underlying parameters. He equalized these for pitch, loudness, and duration. This equalization was considered to be necessary to isolate just those attributes corresponding to timbre, and would prove to be very influential in the studies that followed. Grey asked listeners to numerically rate the similarity between pairs of synthetic notes. Like Miller, he used INDSCAL to find the spatial embedding of the sounds and found a three-dimensional solution for the space of timbre.

McAdams et al. [28] conducted a timbre space study in 1995 with an extended version of the CLASCAL model. CLASCAL differs from INDSCAL by the addition of specificities and latent classes. Modeling specificities adds more parameters to the model to account for timbre factors that may be present in only a single stimulus, which in theory should prevent the spatial solution from being distorted by these factors. On the other hand, assuming latent classes for the subjects is essentially a way to remove parameters from the model to keep it from being impractical to fit. This study used 18 different synthetically-generated notes and maintained the convention of equalizing for pitch, loudness, and duration. Subjects were asked to use a 9-point rating scale to rate

the differences between different sounds. The best fit to the data found by the study was a six-dimensional solution with no specificities, but the researchers found it difficult to interpret the axes physically. The study also found a good solution that did include specificities with three primary dimensions. The first two dimensions of this solution strongly correlated with the physical properties of log-attack time and spectral centroid, respectively. The third dimension was not as easily interpretable, but found to largely depend on spectral flux. Lakatos [29] expanded the use of CLASCAL in 2000 to a mix of notes from percussive and non-percussive instruments, and found a low-dimensional space highly-correlated with spectral centroid and attack time in the first two dimensions.

One final study worth noting is that of Burgoyne and McAdams [30] who, in 2008, combined CLASCAL with Isomap to redo three of the earlier timbre space studies. Isomap is a manifold-learning technique that emphasizes local structure in determining the global configuration of a set of points. By emphasizing the local structure, which presumably can be explained with fewer dimensions, an overall reduction in complexity can be achieved in the global configuration. Burgoyne and McAdams found that Isomap collapsed the CLASCAL model into a single dimension for all three prior study stimuli, though the high-correlated physical feature was not the same in each case. The reduction in complexity is difficult to interpret in a perceptually meaningful way. Based on this result, it does not appear as if any benefit is derived from the additional use of Isomap in these kinds of timbre space studies.

2.3.2 Shortcomings of Timbre Space Research

Research using timbre spaces has yielded some insight into features that play a role in our perception and categorization of sound at the highest level. In general, perceptual spaces seem like a coherent idea with much promise for linking physical phenomena to our internal representation and perceptions. But the true explanatory and generative capacity of timbre spaces has been limited by a number of shortcomings with the prior research into them up to this point.

First of all, every major study since Miller in 1975 has been done using stimuli all with the same pitch. The idea that to study timbre in isolation, one must equalize the other perceptual dimensions of pitch, loudness, and duration seems to have been popularized by Grey [27]. Miller before him found that differences in fundamental frequency dominated his reported results, so he equalized for pitch [26]. One problem with this is the idea that pitch and timbre are truly independent, an assumption that Krumhansl has called into question [31]. By equalizing for pitch, one removes even the possibility that significant dimensions of timbre are dependent on fundamental frequency, which is counter to the findings of Fletcher [6]. Equalizing for pitch eliminates some of the potential for observing timbre differences related to formants. Furthermore, it has been suggested, for instance, that the clarinet possesses enough difference in timbre across its entire range of notes as to functionally be considered three separate instruments. While Grey perhaps did not intend for entire instruments to be represented with a single point in timbre space, his results do not adequately distinguish between the timbre of a clarinet, which has been demonstrated to be an incoherent concept, and the timbre of a clarinet played at a certain pitch (and possibly with other playing characteristics).

An additional problem with timbre space studies has been the use of a small number of discrete points on a scale to numerically describe the perceived difference between different sounds. Such a scale is unfamiliar to subjects prior to the study, so its use is subject to inconsistency over time. Furthermore, when Miller observed that differences in fundamental frequency overwhelmed the reported differences in timbre, it may have been due in part to the fact that he was using only a 9-point rating scale, which is essentially the same scale used in every timbre space study. A more consistent method with greater resolution would be the use of binary comparisons and a subsequent ranking algorithm, which can then be used with non-metric MDS. Wessel [32] in 1979 appears to consider the use of binary comparisons and instead chooses a rating scale like Miller because collecting an entire set of binary comparisons is tedious. No papers on timbre space appear to

consider a general ranking framework that can uncover the perceived difference amounts from only a fraction of the full set of binary comparisons. More consideration is given to the use of a ranking method in Section 2.4 and Chapter 3.

A third major shortcoming is the use of single points to represent sounds of arbitrary length. This makes learning the space more difficult and makes synthesis impractical. The sounds could be constrained to all be the same length, but then this severely limits the possible sounds that can be generated. Additionally, since specifying a sound would mean choosing a point in the space, entire sounds would have to be specified in advance. This would place a severe restriction on the flexibility of the system and its ability to synthesize sounds that evolve in real-time.

In some contexts, it may be appropriate to represent complex phenomena with single points, such as faces. Faces are usually considered to be invariant to scaling, and can all be normalized to the same size. But sound inherently is not. Timing and length are very important to how sound is perceived, which is even evidenced by the importance of attack time found in the perception of sounds in previous timbre space studies. Taken to the extreme, we could try to represent entire songs as isolated points in some very high-dimensional space. But this would make it extremely difficult to conceive of new songs. If nothing else, a static representation of notes as single points seems like it limits creativity, as the entire space of possible notes is likely to be vastly uncharted and it is unclear how to synthesize entirely new timbres without already charting the space from which they arise. A better representation of timbre space would be one where time is an explicit parameter. In this representation, instantaneous timbre properties are encoded in the space, and notes are represented as trajectories through the space. Representing notes as trajectories in a space has been proposed before but not with perceptual features, and it is the representation we adopt in this work.

A final shortcoming of timbre space research thus far is the low-dimensional solutions that have been proposed. That timbre, widely known to depend on a number of spectral

and temporal features, could be adequately described with three dimensions or less, is a very surprising result. Ethington and Punch [33] have said that three dimensions are clearly inadequate, but agree that they are a good starting place. Most studies find that spectral centroid and attack time can explain the reported differences found in the first two dimensions of timbre space. But a third dimension is consistently found to elude simple explanation, sometimes being associated with an imprecise feature such as spectral flux. Caclin et al. [34] provide a good overview of the various features that have sometimes been found to correlate with timbre space dimensions. That the third dimension is always difficult to explain ought to indicate that a three-dimensional representation is not sufficient.

Interestingly, Shepard found in one study [8] involving MDS and dissimilarity ratings from subjects that pitch requires a five-dimensional space to adequately represent perceived distances between different pitches. Only McAdams in 1995 even suggested a timbre space of at least five dimensions, though not much attention was given to its analysis. Given all of the discussion of timbre as a highly-complex perceptual attribute, it would be a very surprising result if pitch were found to be truly a more complex phenomenon than timbre.

In Chapter 3, we propose methods to address and improve upon all of these shortcomings. In the next section, we look at ranking methods, which will form the basis for one of the key ways in which we propose to improve timbre spaces. After that, we look at current models of dynamic musical sound synthesis, since we would like to be able to synthesize new notes from their timbre space representations.

2.4 Ranking Methods

Non-metric MDS requires a ranking of the pairwise distances between all points. One possible way to acquire these for a timbre space application is to ask listeners to numerically rate the dissimilarity between different sounds. But there are a few reasons why this is not preferable. One of the reasons is that the scale used to numerically rate

the differences is arbitrary and has not been learned in advance, so judgments could be subject to large fluctuations from one trial to the next.

To eliminate non-stationarity inaccuracies arising from the use of an arbitrary scale, one can instead use the method of paired comparisons, coupled with a ranking algorithm. In general, a paired comparison means asking how two items fit relative to each other on a scale, rather than asking exactly where a specific item fits on the scale. In this case we ask listeners to decide whether a pair of sounds is more similar or less similar to each other than another pair of sounds. This is an even more specific kind of paired comparison, a binary comparison.

The use of paired comparisons may still be subject to some non-stationarity. For instance, in a paired comparison where the difference between stimuli is less than the *just noticeable difference*, decisions can still be made, but have been observed to change upon repeated trials [35]. Thurstone noted that inconsistencies in paired comparison judgments between stimuli that are more similar than the just noticeable difference are indicative of one way in which the *discriminal process* is not entirely stationary, but subject to small fluctuations. One implication of this is that the entire set of paired comparisons generated by a subject for a set of stimuli cannot be guaranteed to obey transitivity, which means it could contain unintentional contradictions, particularly as the difference between stimuli approaches the just noticeable difference. Whether the contradictions are harmful or helpful, however, may be a matter of context and perspective.

From the set of paired comparisons, the goal is to determine the ranking that is in best agreement, and approaches to this generally lend themselves to one of two perspectives. The first set of approaches views the ranking as a predictor of comparison outcomes [36]. When a stimulus i is ranked higher than another stimulus j, the ranking predicts that stimulus i will be compared favorably to stimulus j. The second set of approaches hypothesizes that the stimuli each have an intrinsic value or rating that orders them, and the paired comparisons are evidence of the true ratings. Whenever a stimulus

i is compared favorably to a stimulus j, it is evidence that stimulus i has a higher rating than stimulus j. It should be noted that only a set of relative values can be recovered, although this is not a concern for typical applications.

From the perspective of the rankings as a predictor of comparison outcomes, the deviation between predicted comparison outcomes and actual comparison outcomes measures the error of the ranking. The best ranking is the one that minimizes this error. The error can be measured using strictly the number of incorrect predictions, in which case the solution is referred to as the *minimum violations ranking*, or also take into consideration the magnitude of the incorrect predictions using the differences in rank. Note that from this perspective, intransitivity among the paired comparisons is bad, as it increases the error, or uncertainty, in the ranking.

From the perspective of the objects having an intrinsic value or rating that orders them, violations of transitivity are not incorrect predictions but simply an expected result of a comparison process that can be inconsistent when stimuli are near the just noticeable difference. While violations of transitivity can add uncertainty to rankings, it could be argued that they are just as much evidence for the true stimulus ratings as non-violations, as they are evidence of extreme similarity between stimuli. Chartier et al. report that, as expected, both the Massey and Colley methods applied to a perfectly transitive set of binary comparisons result in an evenly-spaced set of ratings [37]. In the case of binary comparisons, only intransitivity can provide evidence of irregularly-spaced ratings.

Many ranking approaches come with limitations on the settings in which they will work. Two approaches that are capable of operating on an incomplete set of binary comparisons that we consider are the Colley matrix method [38] and a modified version of the minimum violations ranking method given by Ali et al. [39].

In general, methods that minimize the number of violations are formulated to operate on a complete set of binary comparisons. Thompson and Remage showed that for the case of a complete set of binary comparisons, minimizing the number of violations is justified

under a maximum likelihood framework [36]. The algorithm proposed by Ali et al., though formulated to work only on a complete set of binary comparisons, can be adapted for the case of missing binary comparisons, as we explain in Chapter 3. Still, the solution to the minimum violations ranking problem is equivalent to the problem of finding the minimum feedback arc set in a directed graph, which is known to be NP-hard [40]. This severely limits the problem sizes to which it can be applied.

The Colley matrix method is the least-squares solution to a set of linear equations. Developed to be applied to college football rankings, it is able to simultaneously determine a rating for every team that takes into account the ratings of the other teams it has played against. Colley equates his method to that of Laplace's method for estimating the location of a marker's position on a craps table. Laplace studied the problem given a uniform random sampling of positions on the table and the number of times a sample is to the left or to the right of the marker. The most likely location can be deduced using simple winning percentage, but Laplace also added a Bayesian prior of 0.5 because the best guess at the start is that the marker is in the middle of the table. In sports, if games with other teams are like samples, then with a uniform sampling of games by strength of opponent, winning percentage is the best guess at a team's strength. But Colley uses Laplace's prior and adjusts for strength of schedule, which is an attractive feature that makes the method more robust when the number of teams greatly exceeds the number of games played by each team.

One potential problem with Colley's method is the over-reliance on strength of schedule. An undefeated team with a weak strength of schedule can have its rating brought down even when this introduces intransitivity between it and higher-ranked teams it has beaten. Also, the use of a prior behaves best when teams have all played the same number of games to overcome the prior, which is true in most sports but cannot be assumed to be true in general. As a result, it may be that a hybrid approach that

starts with Colley's method and then makes iterative adjustments to reduce the number of violations of transitivity is more desirable in certain contexts.

2.5 Perceptually-Based Musical Sound Synthesis

The timbre space is the core piece of the framework proposed in this work for achieving unrestricted timbre manipulation. In this section we examine prior work made directly towards such a framework. In Section 2.5, we review dynamic models of sound synthesis aimed at representing any conceivable timbre. In Section 2.5.2, we review proposed methods for interfacing with a perceptually-based synthesis system.

A very limited amount of work has been done on comprehensive frameworks. Wessel [32] in 1979 was one of the very first to outline how musical sounds with arbitrary timbre might be synthesized in a perceptual framework. He proposed working with timbre spaces, performing psychoacoustic analysis on the dimensions to discover their underlying physical explanations, and treating each dimension as a control parameter for choosing between different timbres. However, the timbre space used in his work suffers from all of the problems we mention in Section 2.3.2, which makes it unsuitable for unrestricted timbre manipulation in composition. Wessel mentions himself that he is only considering a system capable of producing changes in timbre between different notes, not within a single note. He gives no mention of how to produce notes of arbitrary length.

Nicol [41] in 2005 proposed a framework for synthesizing sounds with arbitrary timbre from high-level descriptions of the sounds. He noted that sounds are often consciously perceived at a level far above the signal level at which versatile synthesis systems tend to operate, therefore a mapping needs to be made from high-level descriptions to low-level synthesis. He calls the foundation for his framework a timbre space, but this is a mischaracterization because his space has no basis in perception, instead being just a time-frequency representation of sound derived from signal processing (though it should be noted that a few others have used the term this way, as well). He overcomes some of the limitations of the typical timbre space paradigm by representing sounds as paths

and not points in a space. But his framework is not an adequate model of perception, merely attempting to map from a limited number of high-level perceptual descriptions to synthesis parameters. Therefore it is not suitable for the level of unrestricted timbre manipulation that we would like to achieve.

The subproblems common to all comprehensive frameworks are perception, synthesis, and interface. We give much attention to perception in Section 2.3. In the following sections, we look at approaches to synthesis and interface in the context of unrestricted timbre manipulation.

2.5.1 Models of Dynamic Musical Sound Synthesis

There are numerous methods to achieve digital sound synthesis. However, while capable of exerting some control over timbre, most of these methods are unsuitable for the unrestricted manipulation of timbre that we hope to achieve. Nicol [41] gives an overview of various synthesis methods and deems FM synthesis [42] and additive synthesis the most suitable methods for the creation of arbitrary timbre. The primary advantage of FM synthesis is the ability to specify rich timbres with a relatively small number of parameters. However, this is only an advantage from the perspective of computational requirements. Additive synthesis could easily be specified with a small number of parameters that abstract common patterns in the underlying parameters, but with more expressive control. FM synthesis in its basic form is limited in its expressive ability. A hybrid additive-FM synthesis method could be created to expand the expressive control of FM synthesis if computational constraints supersede all other considerations. However, due to their flexibility, intuitive correspondence with what we perceive, and the continued expansion of computing power available to implement them, additive synthesis methods seem like a better choice and are much more common in the literature. Therefore, we primarily focus on models of arbitrary timbre synthesis that use additive synthesis methods.

Every timbre space study performed so far has represented entire sounds as single points in a space, which is not suitable for generating arbitrary sounds, particularly of arbitrary length, without a prohibitively high number of training sounds. Furthermore, it is not clear how truly novel sounds would be generated, since the representation is not flexible enough for the underlying dimensions to be combined in new ways. A better representation with regards to synthesis would be something like a *phase space* [43].

Imagine a system whose current state is described by a set of parameters. A phase space represents all of the possible states of the system as points in a multi-dimensional space with the parameters as axes. The evolution of the system over time is represented as a trajectory in this space. Nolte [43] gives a very thorough review of the coinage of the term *phase* space and the earlier origination of the concepts in the 19th century. In the case of sounds, the instantaneous properties of a sound can be represented as points in a phase space, and their evolution over time as a trajectory. This division of instantaneous and temporal properties makes it possible to specify any sound without modeling a high number of dimensions. The dimensionality is kept low because the burden is shifted to the trajectory for capturing a significant portion of the information.

Additive synthesis fits easily within the framework of a phase space with trajectories. Early work on additive synthesis focused on methods to reduce the amount of data and processing involved to make it practical to implement in real time [44, 45]. This primarily involved linearly interpolating between a few explicit landmark parameters. This data reduction was justified by noting that hardly any perceptual difference could be observed between the fully-specified version and the piecewise linear version. Later methods have taken advantage of advances in machine learning to learn compact representations capable of capturing the range of variation for an entire musical instrument. Burred et al. [46] chose to model musical instrument sounds as a spectral envelope that evolves over the course of a note (in other words, over time). With each spectral envelope corresponding to a point in a low-dimensional space, this is a typical representation as a phase space

with trajectories. In addition to compactness and representativeness, they cite accuracy as a design criterion in their compact representation, but treat this as accuracy from a signal error perspective and do not consider *perceptual* accuracy. Therefore, they likely miss out on the data reduction achievable by only considering features that are relevant to perception.

In this work, we adopt additive synthesis and a phase space representation with trajectories, but choose to model this primarily from the perspective of perception. By making perception the primary focal point, the representation should be as compact as possible, only capturing distinctions that are relevant perceptually.

2.5.2 Interfaces for Timbre Control

In considering an interface for timbre manipulation, the primary goal is intuitive expression for composition. Real-time performance may be another goal, but usefulness in composition takes precedence. The goal is for a composer to be able to conceive of an arbitrary sound in his mind and then be able to express it effectively. This is not necessarily the same problem as that of performance, where duplication and range of expression may be more important than creativity.

An effective interface is usually one that over time is not noticed by the user, easily facilitating the translation from mental idea to physical realization. Timbre is acknowledged to be a complex phenomenon, which makes the implementation of an effective interface difficult. However, three primary types of interfaces seem to have emerged that could potentially facilitate unrestricted timbre manipulation.

The first type of interface is a decomposition of timbre into orthogonal dimensions that can be controlled directly. Every timbre space study involves learning an orthogonal representation of timbre perception. Wessel [32] in 1979 proposed that the dimensions of a timbre space be associated with acoustic parameters and that the dimensions themselves be control parameters for synthesizing notes with arbitrary timbre. Though his overall framework is not suitable for achieving unrestricted timbre manipulation, the idea of

orthogonal control implicit to his proposed interface is worth examining. That dimensions in a timbre space should be control parameters seems to imply the idea that timbre manipulation ought to be linear in the mental representation of timbre. The difficulty lies in the mapping between mental timbre space and physical synthesis properties. This mapping could be quite complicated and non-linear. But as a general rule, it seems reasonable that an interface to timbre manipulation ought to be linear in the perceptual dimensions of timbre, as Wessel proposed.

The second type of interface is a re-mapping of timbre to a space that can be navigated visually or gesturally. Musical instruments are the traditional interface between humans and musical sound synthesis. Perhaps like with Wessel, most of the actions involved with playing a musical instrument can be thought of as independent parameters controlling highly non-linear changes in synthesis parameters. But the control interface does not have to involve independent, orthogonal dimensions. Hunt and Wanderley [47] discuss a number of different mappings from gestures to synthesis parameters, though their focus is on performance, not composition. Choi et al. [48] conceived of a three-dimensional gesture control space for specifying timbre. They called it a "window" into a much higher-dimensional manifold space of sounds and proposed the use of genetic algorithms to learn the reduction to three dimensions. Whatever the associations between the window and the underlying synthesis parameters end up being, they proposed that they can be learned by a user via the process of exploration and feedback. It is an interesting contrast to specifying individual orthogonal dimensions for control, since the human brain is able to process visually up to three-dimensions simultaneously, and perhaps more gesturally.

Gestural interfaces may skip over explicitly mapping perception and go directly to synthesis, which may be a weakness in terms of expressive ability. On the other hand, it is conceivable that with the right interface, the problems of mapping from gesture to perception and perception to synthesis could be offloaded to the user to be learned through experimentation. Wessel and Wright [49] give a thorough overview of

many different types of gestural interface for synthesis. They encourage learning via experimentation and liken the act of learning to use a gestural interface well with the act of learning to use any musical instrument well, which can require a large investment of time. However, in their particular proposal they are more focused on performance than unbounded expression and limit themselves to the creation of discrete note events with their interfaces, which would not achieve our goal of unrestricted timbre manipulation.

Goudeseune [50] elaborates on methods of control for digital musical instruments, including *derivative control*, where the rate of change of a parameter is the driver rather than the instantaneous value. He also gives an extensive treatment on interpolation between high-dimensional spaces, with particular attention to interpolation properties that are desirable for music applications. The problem of mapping from a control space to a synthesis space given pairs of corresponding points in both spaces is the problem of interpolation. This interpolation is relevant for any kind of mapping, not just ones involving a visual or gestural interface.

The third type of interface is verbal, using high-level descriptions to encapsulate and simplify timbre specification that would be tedious at the physical level. This typically harnesses the vocabulary already used by musicians to categorize the notes they hear. The idea is to reverse engineer this process and determine the synthesis parameters that underlie the timbre adjectives. Ethington and Punch [33] proposed an interesting approach to this in 1994, where they learned associations between verbal and synthesis transformations. Instead of mapping absolute specifications, they learned how, starting from some sound already conceived of, its synthesis parameters could be adjusted to result in changes in the description of the sound such as "brighter" or "warmer". This avoids the problem of creating an explicit, potentially highly complex mapping between high-level description and low-level synthesis parameters.

Each type of interface does have the potential to work in a framework for unrestricted timbre manipulation, and we do not see these as mutually exclusive. In the framework

proposed in this work, we focus on creating the orthogonal axes interface to timbre manipulation and sound synthesis, as envisioned by Wessel [32], but do so with an interpolated mapping underlying the perceptual interface, as envisioned by Goudeseune [50]. We believe there is a fundamental gap that exists in the research into the perception of timbre that can be filled by this investigation. We propose that verbal descriptions or a highly expressive gestural interface be built at a layer above this interface. This hierarchical approach greatly simplifies implementation and separates an entire system for timbre manipulation into natural subproblems that should be solved independently.

CHAPTER 3
METHODOLOGY

The primary goal to be addressed by this research is the design of a framework for unrestricted timbre manipulation in musical composition. By *unrestricted timbre manipulation in composition*, we mean at a high level the ability to specify a musical note with any timbral properties. At a low level, we mean the ability to produce musical sound of an arbitrary length and alter its timbre over time as easily as we alter its loudness or pitch.

In Section 3.1, we identify four key design goals that a framework for the unrestricted manipulation of timbre should meet. In the remainder of the chapter we expand on each of the goals and our proposals for meeting them.

3.1 Design Goals

The primary design goals for a framework for unrestricted timbre manipulation are alluded to in Chapter 2 but not stated as explicitly as in this section. The four goals can be summarized as follows:

1. Perception
2. Synthesis
3. Interface
4. Modularity

The first component, *perception*, is included because timbre is an inherently perceptual property, not an acoustic or physical property. Perception refers to the mental representation of timbre. The ideal framework for timbre manipulation can facilitate an efficient translation from mental conception to actual realization in sound. This seems difficult to achieve without a firm understanding of how timbre is represented in the brain. The interpretation of timbre by a listener is also important from a compositional point of view. Therefore, facilitating effective manipulation of timbre in composition requires an understanding of is being achieved through the manipulation. This is accomplished

most effectively with a perception-centric approach, not an approach centered around the manipulation of physical properties alone. Finally, there is a compactness of representation that is only possible by considering the mental representation of timbre. This representation automatically discards any potential properties that have no bearing on perception.

The second component is *synthesis*. The notion of unrestricted timbre manipulation in composition implies the notion of unrestricted timbre *production*. An accurate representation of timbre is not useful in composition without a means to realize it. Therefore, any timbre that can be conceived of must be able to be physically realized. Consequently, an extremely flexible method of synthesis is required.

The third component is *interface*. A good representation of timbre and a means to realize it are much less useful without an effective interface. The choice of interface could depend on other goals for the application of the system. The goal of the interface in this work is primarily to facilitate composition, but performance is also taken into consideration. These two goals are likely to have subtle differences in the interfaces that best facilitate them.

The final component is *modularity*. The ideal framework should have layers that accept well-defined input and produce well-defined output. By designing a framework with components that can be replaced at one layer without affecting components at another layer, this helps achieve a system where, for examples, one interface could be developed for composition and another for performance. In terms of synthesis, it is conceivable that different methods could be preferred depending on the amount of computing resources available. Even the goal of perception could benefit from a modular element, as subjective elements of perception could theoretically lead to a personalized timbre manipulation system.

In the following sections, we discuss how each of these components is addressed in the framework we propose. Out of the three layers of perception, synthesis, and interface, we

give particular attention to perception, which is the component we choose to develop for implementation.

3.2 Perception

In Chapter 2, we review the literature on the perception of timbre. A number of attempts have been made to learn the mental representation of timbre, which is referred to as a timbre space. The axes of timbre space are usually associated with acoustic or physical properties of sound. However, the representations developed so far serve as poor generative models, unsuitable as a general model of synthesis, particularly in the case of notes of arbitrary length. Neither are they satisfying as models of perception. Nearly all of the timbre spaces learned thus far have three dimensions or less, which seems odd considering how many acoustical properties are known to have an effect on timbre perception.

To develop a generative space of timbre, it seems necessary to have a representation with *time* as a parameter, so that all other parameters are explicitly associated with points in time. Attempts have been made at this kind of representation, but strictly from an acoustic perspective. What this usually ends up being is a model of additive synthesis that has a compact representation, determined using dimensionality reduction methods. This can include loss of perceptually relevant information. No model of perceptual timbre with time as an explicit parameter has been developed. The benefits of such a model include the potential for a representation that is compact as possible without loss of information and a precisely controlled generation of any conceivable timbre. Since such a model does not currently exist, its development is a primary goal of this work.

Even without considering how a system could be built for facilitating the manipulation of timbre, the comprehension of timbre as an aspect of composition seems underdeveloped. While it may be recognized that there are spectral and temporal aspects to timbre, timbre has never even been fully broken down into its constituent temporal and non-temporal aspects. Nomenclature relating to timbre is not standardized, either. Tone color, tone

quality, and texture are all sometimes used interchangeably with the word timbre. To remedy some of these problems, we propose some new nomenclature and a new hierarchical view of timbre properties.

3.2.1 Instantaneous Timbre and Dynamic Timbre

Helmholtz coined the term *musical quality* to describe a subset of timbre properties, seemingly those that could be associated with steady state sounds. Later researchers seem not to have viewed Helmholtz's work as creating a dichotomy between these and other aspects of timbre, but rather as oversimplified because it did not capture temporal aspects of timbre. However, it seems reasonable to conclude that Helmholtz did intend to create a dichotomy of timbre. Whether the idea for a dichotomy originates with Helmholtz or later theorists, we believe such a breakdown is the only way to make progress on our understanding of such a complex phenomenon.

We propose that aspects of timbre perception be broken into two distinct categories: *instantaneous timbre* and *dynamic timbre*. Instantaneous timbre captures everything about timbre that is constant, stationary, or steady state. These are properties that can be associated with a distinct point in time, much the same as instantaneous amplitude or instantaneous frequency. Tone color seems like a potential synonym for instantaneous timbre, which is something we would suggest. But this equivalence is not currently standardized. Dynamic timbre, on the other hand, captures everything about changes in timbre over time. This includes the evolution of pitch and loudness, which perhaps should not necessarily be considered entirely distinct from instantaneous timbre. Including amplitude and frequency with timbre, dynamic timbre is specified specifically as the evolution of instantaneous timbre. Thus this is a hierarchical relationship, as dynamic timbre can always be broken down into its constituent instantaneous aspects over time.

Note that such a breakdown already exists in human color vision. Pixels represent the smallest units of area over which color is perceived. There are three orthogonal dimensions of color vision that can be associated with a pixel. One possible orientation of

this space is to label the dimensions as red, green, and blue. Another possible orientation is hue, saturation, and lightness. Much like instantaneous timbre, these three orthogonal dimensions are attributes of instantaneous color and are the only dimensions that can be attributed to a single pixel. Complex properties of color such as matte-ness, or its opposite shiny-ness, that could be thought of as dimensions of color vision, cannot be attributed to single pixels and can only be described with a collection of pixels, varying either in time or space. These are elements of dynamic color, which is analogous to dynamic timbre. At an even higher level of color perception, images and objects emerge, and this analogy holds for auditory perception, as well.

Interestingly, though there are only three orthogonal dimensions of color, we have listed up to six perceptually coherent dimensions. Whether or not there are analogous structures in timbre perception is of great interest to us. It may be that there are multiple coherent orientations of instantaneous timbre space. These are the kinds of questions that cannot really be addressed without first distilling timbre down to its most fundamental properties. Therefore, we make it a key aspect of this work to develop a theory of timbre as a dichotomy between instantaneous and dynamic features.

Ethington and Punch [33] described a similar dichotomy of timbre perception but with verbal attributes. They use the term *presence* to categorize words that are associated with stationary properties, while the terms *attack* and *cutoff* are used to described words associated with non-stationary properties. While the dichotomy proposed in the present work is intended to apply to timbre at the most fundamental level, it is possible to apply it to these words, as well. Table 3-1 gives a list of words associated with timbre and their categorization as either instantaneous or dynamic. It is our hope that some of these words can be adopted to describe fundamental properties of timbre, since they are in widespread use by musicians already. But, there is no universal standard currently in use.

A set of orthogonal dimensions could serve as such a standard. Experiments performed to uncover perceptual dimensions of timbre, however, attempted to model

Table 3-1. A list of words used by musicians and how they fit into a proposed dichotomy of timbre. In this dichotomy, the most fundamental properties of timbre with no temporal component are known as instantaneous properties, while higher-order properties with a temporal component are known as dynamic.

Instantaneous	Dynamic
bright	plucked
nasal	fading
dull	vibrato
tonal	strummed
inharmonic	expanding
hollow	echoing
hissing	blown
biting	tremolo
shrill	twangy
rich	warbly

both instantaneous and dynamic aspects of timbre simultaneously. This did not lead to a clear understanding of either aspect in terms of orthogonal dimensions. Therefore, we propose to create an *instantaneous timbre space*, where each point represents a distinct combination of instantaneous timbre features only. Presumably, the axes determined by this would show some resemblance to known stationary properties of timbre, such as brightness. Dynamic timbre would then be represented as trajectories in this space. This is a classic *phase space* setup, to borrow a term from physics that has gained widespread use throughout science.

3.2.2 Timbre Space Experiment

The key aspects to be determined experimentally in this work are the number of dimensions of instantaneous timbre and their perceptual interpretations. This does not imply that a single set of dimensions is coherent for all people. We may need to develop a representation that is *most widely applicable* rather than *universal*. It may also be possible to consider multiple coherent orientations of timbre space. We wish to interpret each dimension so that we can learn to specify new timbres perceptually, connecting the spatial representation to our own perceptions, if possible.

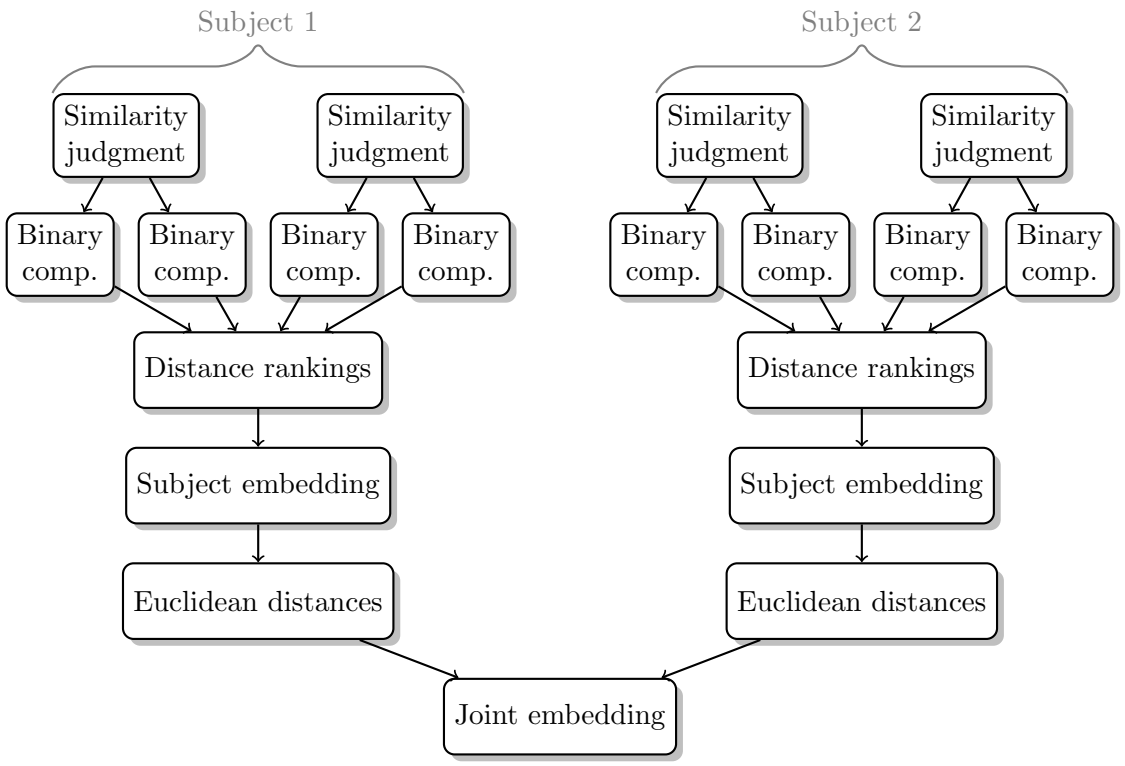

Figure 3-1. This figure shows the pipeline of derived data in the instantaneous timbre space experiment. For each group of sounds presented, the subject provides similarity judgments, which either specify which two sounds are the most similar or least similar sounds in the group. From a similarity judgment, multiple binary comparisons are derived. All of the binary comparisons are used to derive a ranking of the distances between sounds, using Colley's method. From the ranking of distances, a spatial embedding is derived using Kruskal's non-metric MDS algorithm. The Euclidean distances between points are calculated from this spatial embedding. Finally, a joint embedding is derived for all of the subjects from their individual Euclidean distances using INDSCAL.

To undertake this analysis, we propose to create an instantaneous timbre space using similar methods to those described in Section 2.3. However, there are two main differences between the proposed experiment and previous attempts, which will have a significant effect on the result. Figure 3-1 contains a diagram that outlines the experimental steps.

The first primary difference is the choice of data. We use 16 sounds, each 2 seconds in length and synthetically generated using additive synthesis with stationary parameters, as

given by Equation 3–1,

$$s(t) = \sum_{k=1}^{K} a_k \cos\left(\frac{2\pi}{F_s} f_k t\right). \tag{3–1}$$

We use 30 sinusoids for each sound, with each sinusoid k having constant amplitude a_k and frequency f_k. The fundamental frequency of the sound, f_0, could be considered a distinct parameter, although for our purposes we will assume $f_0 = f_1$. F_s represents the sampling frequency, taken to be 44,100 Hz for the sounds used in this work. Phase is not considered as a parameter in the present study.

Most of the sounds' parameters are derived from actual recordings of musical instruments, analyzed with a new periodic waveform analysis technique we describe in Section 3.3.2. These source recordings are part of the University of Iowa Musical Instrument Samples [51]. Table 3-2 gives a list and description of the 16 sounds, including what instrument each is derived from. The inclusion of the source musical instrument in the description should not be inferred to mean that a sound is representative of, or even clearly identifiable as coming from, that instrument. Rather, it is a means of suggesting what some properties of each sound might be.

The purpose of using sounds derived from real instruments is to easily generate a variety of data already known to have noticeable differences in timbre. The particular sounds used were picked to try and sample a wide range of instantaneous timbre space, but with enough similarity to avoid an entirely hollow space. Some notes of different pitch were also included, which to our knowledge has not been tried since Miller and Carterette [26] in 1975, who determined that the inclusion of pitch differences overwhelmed the perception of timbre differences. However, it seems impossible to characterize the space of instantaneous timbre without considering notes of different pitch. Two notes with their other properties held fixed could be perceived as differing by a different amount as their pitch is altered. Furthermore, it makes for a cleaner analysis to consider all instantaneous perceptual properties simultaneously and sort them out during the analysis itself, rather

Table 3-2. A description of the sounds used in the timbre space experiment. Most of the sounds were derived from analyses of actual note recordings.

No.	Instrument	Approx. pitch	Actual freq. (Hz)
1	Trumpet	A_4	442.7
2	B-flat clarinet	G_4	392.3
3	Synthetic string	E_3	164.8
4	Bass	C_3	130.4
5	Soprano saxophone	G_4	392.4
6	Piano	A_3	221.0
7	Tenor trombone	A_3	220.9
8	Piano	A_4	440.0
9	E-flat clarinet	A_3	218.1
10	Alto saxophone	A_3	219.1
11	Violin	C_4	259.4
12	Bass flute	A_3	220.8
13	Bass clarinet	A_2	109.5
14	Cello	A_3	220.1
15	Trumpet	B_3	246.5
16	Square wave	$F_3^\sharp$	185.0

than to try to create a data set containing only those isolated properties considered to be timbre, which is ill-defined to begin with.

The second primary difference is the method of feedback for the listening experiment. In typical timbre space experiments, subjects listen to a pair of sounds and numerically rate the difference between the two on a 9 or 10-point scale. However, this seems to invite inconsistency in the results, as such a scale is unfamiliar to the subjects prior to the experiment and would be difficult to apply consistently to their perceptions. Instead, we ask subjects to listen to sounds in groups of 3 or 4 and to determine which two sounds in the group are the most similar and which two are the least similar. This relative determination has the same stationarity-of-perception assumption as numerical ratings, but does not rely on a subject remembering an arbitrary and ill-defined scale. This makes it easier to stop the test and re-start it later if fatigue becomes a factor.

The purpose of choosing the most or least similar sounds in a group is to create binary comparisons between pairs of pairs of sounds. That is, we want to know how the distances between different pairs of sounds compare. Making binary comparisons is more

tedious than numerically rating the difference between sounds. But the more sounds there are in a group, the more binary comparisons are made by providing feedback about the most similar or least similar pairs of sounds. In a group of 3 sounds, there are $\binom{3}{2} = 3$ possible pairs, so choosing the most similar pair provides two binary comparisons—it indicates that the chosen pair has a smaller distance between sounds than the sounds in the other two pairs. In a group of 4 sounds, there are $\binom{4}{2} = 6$ possible pairs, so the number of binary comparisons provided by choosing the most similar pair increases to 5. Although we did not test a group of 5 sounds, the number of binary comparisons in this case would be 9. So increasing the group size could be a way to significantly reduce the amount of time required to complete the experiment, although each individual trial does become more difficult.

Given how this approach could become impractical as the number of sounds in the data set increases, it is important to consider the most efficient way to collect useful information from a subject. We developed an approach designed to extract as much information as possible as quickly as possible in the beginning, and then take more time to fill in gaps in information, if they existed, closer to the end of the experiment. Thus, the makeup of the trials in terms of the chosen groupings of sounds was not fixed at the beginning.

Some trials have answers that are very clear to the subjects, while other trials have answers that require repeated listens to arrive at. In setting up the experiment, we emphasized trials involving sounds in groups of 4 in the beginning, which are more difficult to completely assess than trials involving groups of 3. To compensate, we gave subjects the option to skip trials that were difficult to assess quickly. Subjects were instructed to not dwell on difficult assessments, though they were asked to provide information whenever possible. This served to decrease the amount of time for the experiment and increase the reliability of the information provided.

Our method is similar to the method of triadic comparisons used in early research but the use of 4 sounds in a group offers a distinct advantage. Since the goal is to rank all pairs of sounds, in theory we need binary comparisons between all pairs of pairs. This is represented with a matrix, just like the adjacency matrix for a weighted graph. Only using groups of 3 when choosing the most or least similar sounds in a group limits the binary comparisons to be between pairs that have one sound in common, which represents a subset of the entire set of possible comparisons. In fact, since a number of trials involving groups of 3 are used in this experiment, the subset of the matrix corresponding to these comparisons is readily apparent in the results shown in Figure 4-1. By adding the use of group sizes above 3, we are able to make comparisons between pairs of sounds with no sound in common, in which case we could fill out the entire comparison matrix if we wished, though this would be time-consuming is not necessary.

From the binary comparisons, the goal is to determine a ranking of all the pairwise distances between sounds. Then non-metric MDS, which is ranking-based, is used to determine the N-dimensional embedding of all the sounds. In the method of triadic comparisons, the ranking is determined by summing the number of times a pair of sounds is judged to be either more similar than another pair of sounds. This is equivalent to winning percentage in sports. But winning percentage only provides an accurate ranking if the entire binary comparison matrix is filled out completely. In the case of triadic comparisons, this is already impossible because the group size is limited to 3. However, even with a group size of 4, we would prefer to not have to fill out the entire binary comparison matrix, as this would take a very long time. Therefore, we want to view this as a general ranking problem given a potentially sparse matrix of binary comparisons. We choose to solve this using the Colley matrix method [38], for reasons we discuss in Section 2.4.

3.2.3 Determining and Interpreting Perceptual Axes

Given the instantaneous timbre spaces produced by the MDS methods, the goal is to interpret or explain the dimensions of the spaces perceptually. For the individual subject embeddings, this is complicated by the fact that the spaces have no fixed orientation. Therefore, it is not clear what, if any, are the principal axes of the spaces. Nor is it clear what perceptual criteria ought to be used to determine the principal axes.

Carroll and Chang [21] proposed an algorithm to fit a model that presupposes the existence of a set of principal perceptual axes common to all subjects. In this model, individuals differ in their personal perceptual spaces by placing different weights on the common set of axes. The distance between two points j and k in the personal space of subject i is given by Equation 3–2,

$$d_{jk}^{(i)} = \sqrt{\sum_{t=1}^{r} w_{it}(x_{jt} - x_{kt})^2}, \tag{3–2}$$

where r is the number of dimensions of the space and w_{it} is the weight placed on dimension t by subject i. The coordinates of points $\{y_j^{(i)}\}$ in the individual space are related to the coordinates of points $\{x_j\}$ in the common space by Equation 3–3,

$$y_{jt}^{(i)} = w_{it}^{1/2} x_{jt}. \tag{3–3}$$

Carroll and Chang claim that axes found using this weighted Euclidean model for MDS should be considered to be psychologically important. They do provide some evidence to support their assertion, but the idea that each person uses the same set of orthogonal axes with different weights on them to characterize their own perceptual space is an assumption that appears to have gone unchallenged and untested in the field of timbre research for the last 40 years. The weighted Euclidean model is certainly a convenient method to combine MDS results from multiple subjects. It may even be more reliable than the spatial embeddings determined for subjects individually because it uses more information and thus can potentially eliminate noise and errors. But the model

should not serve as a *de facto* psychological theory. Therefore, we approach the model as a tool that may be able to determine a set of principal axes, and use complementary methods to validate our observations.

Supposing that we have determined a perceptually meaningful axis in instantaneous timbre space, we want to use every method at our disposal to interpret its meaning. An interpretation involves either a perceptual description in words or an analytical explanation in terms of the acoustic feature that varies along the axis. We describe several methods that could be used to obtain an interpretation, either perceptually or acoustically.

3.2.4 Analytical Methods of Perceptual Axis Explanation

We describe two analytical methods of perceptual axis explanation. The first method is to collect the values, $\{p_j\}$ for each of the n sounds projected onto an axis and the values of some acoustic feature, $\{q_j\}$, of the sounds, and compute the correlation r_{pq} between the two sets, given by Equation 3–4,

$$r_{pq} = \frac{n\sum p_j q_j - \sum p_j \sum q_j}{\sqrt{n\sum p_j^2 - (\sum p_j)^2}\,\sqrt{n\sum q_j^2 - (\sum q_j)^2}}. \tag{3–4}$$

This is a relative measure of how well an acoustic feature explains an axis.

We propose several acoustic or physical features to test as possible interpretations of the axes produced by INDSCAL. The proposed acoustic features are ones thought to be relevant perceptually. However, it is not clear *a priori* the precise manner in which some of these features ought to be computed from the sounds. Therefore, tests involving the acoustic features could be tests of the features themselves, and not just the axes that they are being used to interpret.

In this work, we test seven acoustic features which can be easily determined or computed from the sounds' additive synthesis representation, given by Equation 3–1. The features are listed in Table 3-3. The values of each of the features, labeled $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_7$, for each of the 16 sounds is listed in Table 3-4.

Table 3-3. The seven acoustic features and their formulas, based on the additive synthesis representation in Equation 3–1.

Acoustic feature	Label	Formula
Fundamental frequency	$\vec{p}_1$	f_1
Log fundamental frequency (semitone offset)	$\vec{p}_2$	$log_2(f_1/440) * 12$
Spectral centroid	$\vec{p}_3$	$\sum_{k=1}^{K} a_k f_k$
Odd harmonics proportion	$\vec{p}_4$	$\sum_{k \text{ is odd}}^{K_{k=1}} a_k$
Loudness	$\vec{p}_5$	$\sum_{k=1}^{K} a_k \mathcal{L}(f_k)$
Loudness-weighted spectral centroid	$\vec{p}_6$	$\sum_{k=1}^{K} a_k f_k \mathcal{L}(f_k)$
Inharmonicity standard deviation	$\vec{p}_7$	$\sqrt{\frac{1}{K}\sum_{i=1}^{K}(\frac{f_i}{i \cdot f_1} - \frac{1}{K}\sum_{i=1}^{K}\frac{f_i}{i \cdot f_1})^2}$

Table 3-4. The values of the seven acoustic features for each of the 16 sounds.

Sound	$\vec{p}_1$	$\vec{p}_2$	$\vec{p}_3$	$\vec{p}_4$	$\vec{p}_5$	$\vec{p}_6$	$\vec{p}_7$
1	442.7	0.10	1884.6	0.53	1.50	2991.5	0.00004
2	392.3	-1.98	1308.5	0.89	1.29	1777.2	0.00063
3	164.8	-17.00	1629.9	0.79	1.03	1830.1	0.04834
4	130.4	-21.05	687.0	0.73	0.90	1054.2	0.00049
5	392.4	-1.98	1382.7	0.39	1.44	2053.3	0.00010
6	221.0	-11.92	1346.1	0.51	1.30	1914.4	0.00640
7	220.9	-11.93	1350.0	0.52	1.39	2166.0	0.00014
8	440.0	0.00	881.5	0.77	1.28	1253.6	0.00675
9	218.1	-12.15	1122.6	0.91	1.22	1791.8	0.00052
10	219.1	-12.06	1153.5	0.49	1.26	1842.1	0.00019
11	259.4	-9.14	1472.6	0.60	1.40	2375.1	0.00119
12	220.8	-11.93	604.0	0.83	0.97	767.2	0.00125
13	109.5	-24.08	1329.8	0.73	1.19	2011.1	0.00020
14	220.1	-11.99	494.4	0.59	0.90	579.9	0.00143
15	246.5	-10.02	1592.0	0.53	1.47	2541.7	0.00008
16	185.0	-15.00	952.3	0.91	1.19	1559.6	0.00052

To calculate loudness, we define a function called the loudness proportion, $\mathcal{L}(x)$, shown in Figure 3-2. This function is derived from ISO 226:2003[1], which defines the required decibel level for a sine wave at each frequency to be heard at a pre-defined loudness. We invert one of these equal-loudness curves to yield a function that weights each frequency by its perceived loudness. Since ISO 226:2003 is only defined for discrete

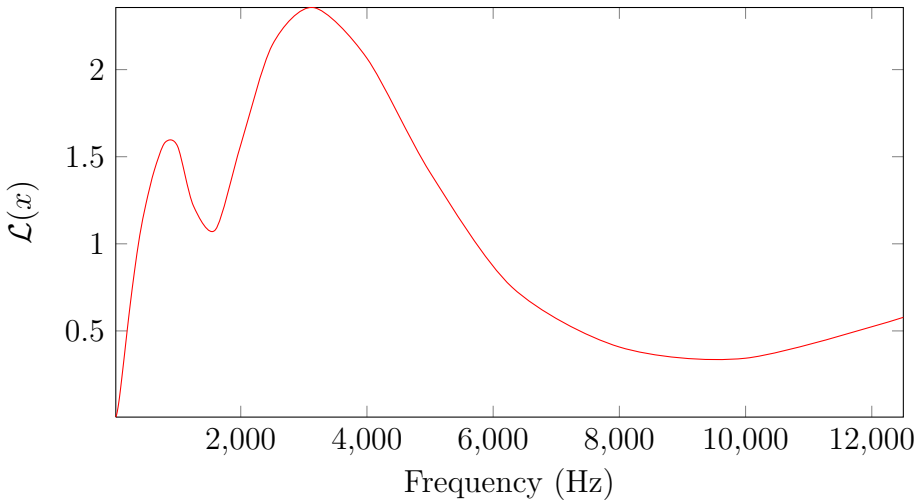

Figure 3-2. $\mathcal{L}(x)$, the loudness proportion, which is a weighting of frequencies by their perceived loudness, derived from ISO 226:2003[1], the equal-loudness-level contours. The curve is normalized to have a mean value of 1.

frequency values, we interpolate at intermediate frequency values using a piecewise cubic spline. We normalize the curve to have a mean value of 1.

These features are based on properties that are known to be relevant to perception. However, the precise mathematical form most relevant to perception is unknown. The functions here are based on the best information available. Therefore, the second analytical method of axis explanation is to learn acoustic features of the sounds relevant to perception directly from their projections in instantaneous timbre space. There are a number of machine learning methods capable of doing this. Unfortunately, this is not practical in our case because the number of sounds in our study is too small compared to the number of possible acoustic parameters. Thus, this would lead to overfitting.

3.2.5 Exploratory Methods of Perceptual Axis Explanation

One way to explain a perceptual axis is to have sounds be displayed on an axis at their projected values and listened to by a test subject to try to determine what perceptual features the subject can associate with the variation observed along the axis. In some ways this is a better explanation of an axis because it remains purely in the domain of perception. However, it may be a difficult task in the case that the axis corresponds to

a perceptually coherent dimension that does not already have a verbal description that the subject is aware of.

Another exploratory method is to apply knowledge of the domain to manually search for patterns in the data that fit the variation along a perceptual axis. In our case with only 16 sounds, knowledge of the domain is crucial to pare down the total number of acoustic parameters that can potentially vary. Assuming appropriate domain knowledge is applied, a pattern found manually in this manner can be objectively verified using the correlation method described in Section 3.2.4. In a sense, a manually-determined pattern and a pre-specified acoustic feature are not much different, and this method could be thought of as analytical because it has an analytical verification step.

3.2.6 Finding Axes Via Least-Squares Projection

The procedure of interpreting axes using objective features of the sounds can be inverted to look for the axes that best correlate with acoustic features. This is most useful when the acoustic features can be demonstrated to be relevant to perception. The procedure works as follows.

An acoustic property that has a linear relationship with a perceptual axis can be considered to be a projection of the sounds onto that axis. Therefore, if any acoustic property has a linear relationship with the perceptual space, there should be an axis along which the sounds can be projected to recover the values of the acoustic property. The best possible alignment of an acoustic property, treated as an axis, and a perceptual space can be determined by assuming the property is a one-dimensional projection and using least-squares regression to find a set of weights that best reproduce this projection.

Let A be the matrix of points representing sounds in instantaneous timbre space. Let $\vec{p}$ be a set of values for some acoustic property of the sounds and let $\vec{b}$ be the mean-centered version of the same values. Then we want to find the projection $A\hat{x} = \hat{b}$ such that $||\hat{b} - \vec{b}||$ is minimized. This is equivalent to minimizing $||A\hat{x} - \vec{b}||$, which is minimized using the pseudo-inverse of A, and the value of $\hat{x}$ which minimizes $||A\hat{x} - \vec{b}||$ is

given by Equation 3–5,

$$\hat{x} = \left(A^T A\right)^{-1} A^T \vec{b}. \tag{3–5}$$

So $\hat{x}$ defines the axis that best matches the acoustic feature $\vec{b}$, $\hat{b}$ is the projection of the sounds onto that axis, and $||\hat{b} - \vec{b}||$ is the error between the feature and it's best-matching projection. It should be noted that while this is a linear method for matching $\vec{p}$ with an axis in the space, $\vec{p}$ can be any non-linear combination of the acoustic properties of the sounds.

Care does need to be taken when examining the correlation values found using this method. As the number of dimensions approaches the number of points in the dimension, random values assigned to the points could be found to correlate highly with the space. This is illustrated in Figure 3-3, which shows the correlation between a random set of N values and the best-matching projection in a random D-dimensional space of N points, as N and D are varied. In our case, with 16 points and 5 or 6 dimensions, the expected random correlation is around 0.5. This represents a baseline that any correlation must be above using this method to be considered at all meaningful.

We use this procedure to look for principal axes of instantaneous timbre space that may not be discovered by INDSCAL. We also use this procedure to verify that the axes of the joint embedding determined by INDSCAL have a correspondence with axes in the individual embeddings of each subject, since correspondence between the two is central to the justification of the principal axes found by INDSCAL being meaningful.

3.3 Synthesis

The most important consideration for synthesis in a framework for unrestricted timbre manipulation is flexibility. A synthesis method must be able to recreate a wide range of timbres and do it in a way such that one timbre can be seamlessly morphed into another. In Chapter 2, additive synthesis is identified as the preferred method.

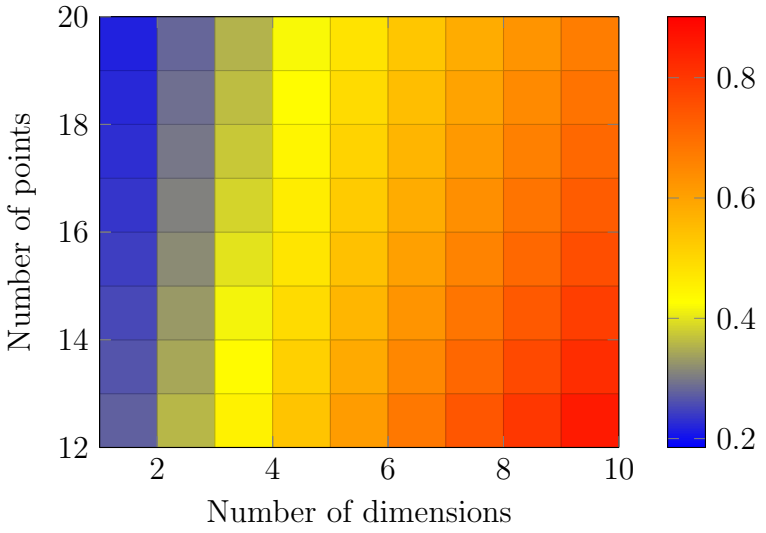

Figure 3-3. Illustration of the expected correlation between a randomly-generated set of values and the best-matching projection (or axis) found in a space with randomly-generated points, as the number of points and dimensions is varied. For the case of 12 points in 10 dimensions, the expected correlation is 0.9. This means that in such a configuration, correlations must be above 0.9 to be more meaningful than random.

The inverse of synthesis is analysis. A synthesis method is best utilized when there is a method to decompose a signal into, or fit a signal to, a set of synthesis parameters.

3.3.1 Additive Synthesis Representation

In its most general form, additive synthesis represents a signal as simply the sum of a set of sine waves of different amplitudes, phases, and frequencies. As the same can be said of the Fourier transform, additive synthesis in its most general form is clearly capable of representing any possible signal (and timbre). However, one problem with simply using the Fourier domain is there is no explicit representation of time or the evolution of parameters over time. Each sinusoid has a constant amplitude and frequency. Since the only way to impact how a sound evolves over time in this representation is to alter the amplitudes and phases of the sine waves, any sound to be generated must be specified in its entirety prior to synthesis.

To facilitate the synthesis of sound that can evolve in real-time, the sinusoids must have time-dependent, rather than constant, amplitudes and frequencies. In its naive form,

this version of additive synthesis suffers from a huge increase in storage requirements. We can imagine rather than t values to represent a signal of length t, we have on the order of t^2 values to represent t sinusoids and their evolution over time. But luckily for a system for which this increase is prohibitive, the situation is not as dire. Allowing sinusoids to have time-dependent amplitudes decreases the number of sinusoids needed to represent a signal. In fact, technically only one well-crafted sinusoid is needed, but this sinusoid would have a rapidly changing amplitude to match the values of the signal it is supposed to represent. As a result, this single sinusoid would not be informative of the underlying timbre, so we would lose the ability to morph between timbres in an intuitive way.

The solution, then, is to have a relatively small number of *slowly-evolving* sinusoids. This definition is not terribly precise but it means that the number of sinusoids should be much less than the length of the signal, and the time-dependent amplitudes and frequencies of the sinusoids should be smooth, without any discontinuities or otherwise unpredictable jumps. The form of this model is given by Equation 3–6,

$$s(t) = \sum_{k=1}^{K} a_k(t) \cos\left(\frac{2\pi}{F_s} \sum_{u=1}^{t} f_k(u) + \phi_k\right), \tag{3–6}$$

where t is the current time step, K is the number of sinusoids, F_s is the discrete sampling rate, $a_k(t)$ is the amplitude of sinusoid k at time t, $f_k(t)$ is the frequency of sinusoid k at time t, and ϕ_k is the phase of sinusoid k.

One constraint on the model that was mentioned is that $K \ll T$, where T is the total number of time steps. Another constraint that we adopt is given by Equation 3–7,

$$f_k(t) \approx k \cdot f_0(t), \tag{3–7}$$

where $f_0(t)$ is the fundamental frequency at time t. In using this model, we wish to restrict ourselves to mostly tonal notes, corresponding to approximately periodic waveforms. This can be thought of as a quasi-harmonic model of additive synthesis.

We intend for this model to include some slight inharmonicity as exhibited by certain strings, so that the error between $f_k(t)$ and $k \cdot f_0(t)$ can increase as k increases.

It would not be difficult to extend this model to include noise and more severe inharmonicity for synthesis purposes, although it would be more difficult to analyze sounds to fit to such a model. A number of methods have been proposed for fitting to harmonic models of additive synthesis, but these methods become less applicable as the amount of inharmonicity increases. In Section 3.3.2, we present a new method suited to analyze quasi-periodic waveforms that actually becomes a better estimate of the time-varying parameters in the presence of certain types of inharmonicity.

3.3.2 Masked Inverse Fast Fourier Transform

In this section, we present a new method called the Masked Inverse Fast Fourier Transform (MIFFT) for analyzing quasi-periodic waveforms and fitting them to our model for additive synthesis. As presented, the method assumes that the fundamental frequency is already known or can be estimated using a number of methods that have been proposed elsewhere. The fact that the fundamental frequencies of the sounds to be analyzed in this work are known in advance means that we do not need to incorporate a step to estimate the fundamental frequency, so we do not focus on this portion of the method.

The idea behind the MIFFT is somewhat philosophical in nature. In signal analysis, time and frequency are considered to be complementary yet opposing domains in which the certainty of a signal's value in one domain results in uncertainty in the other domain. Methods such as wavelets or the Short-Time Fourier Transform (STFT) try to strike a balance between the two. But neither method incorporates adaptation to the inherent time and frequency separation that may already exist in a signal.

In both the time domain and frequency domain, the value of a signal at index i is highly correlated with the values at $i-1$ and $i+1$. It may also be the case that somewhat high correlations are also seen at some multiple of an offset, so that the value at index i is correlated with the value at $i+kT$ and $i-kT$, where T is a fixed offset, and $k \in 1, \ldots, K$,

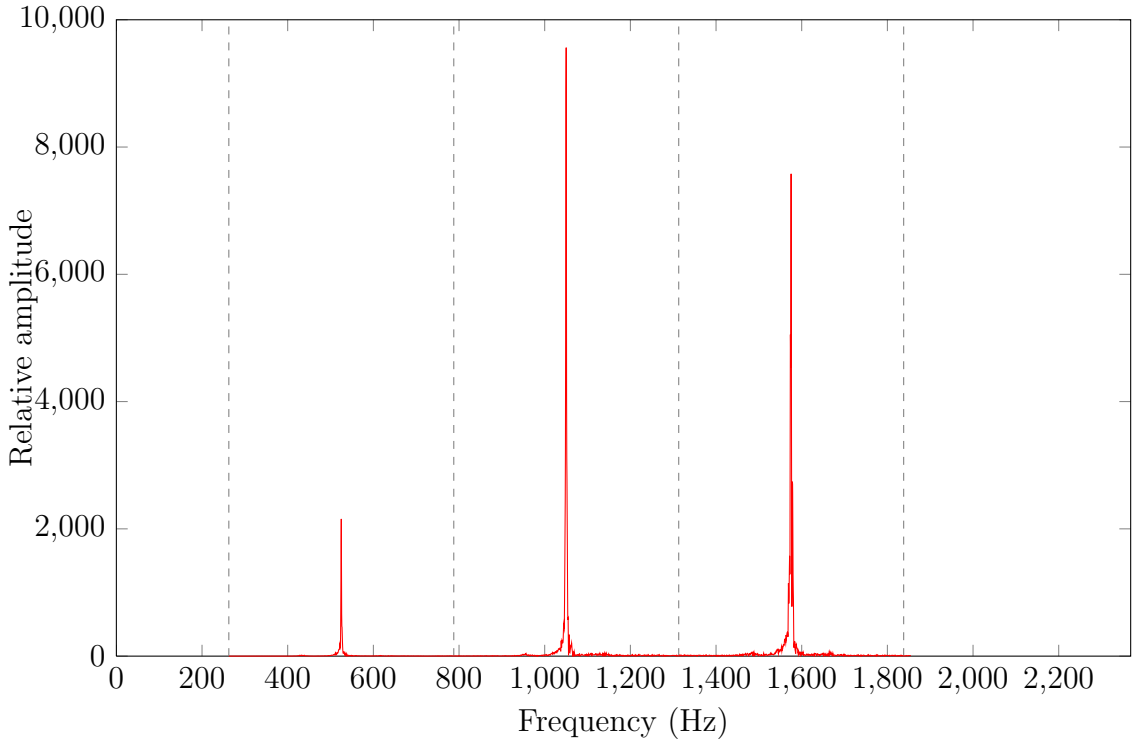

Figure 3-4. An illustration of disjoint support among the first three harmonics of a C_5 trumpet note, using a portion of the Discrete Fourier Transform of the signal corresponding to the sound. Each large peak in energy corresponds to a harmonic. The dashed lines represent the boundaries of each harmonic's support. Note that the energy for each harmonic is concentrated well within the boundaries of support, demonstrating that each harmonic has essentially disjoint support.

where K is an integer. Portions of a signal that are uncorrelated are said to have *disjoint support*. An illustration of disjoint support among harmonics for a trumpet note is shown in Figure 3-4. We use the term disjoint support rather than a term such as *independent* because we want to emphasize the fact that signals are represented as combinations of basis functions. It is in the basis functions that portions of a signal are said to have disjoint support, meaning the value of one basis function for that signal is uncorrelated with the value of another.

Combining the analyses of two portions of a signal that are correlated helps to increase the certainty of our estimate of both portions. However, combining the analyses of two portions of a signal with disjoint support provides no additional help to the analysis

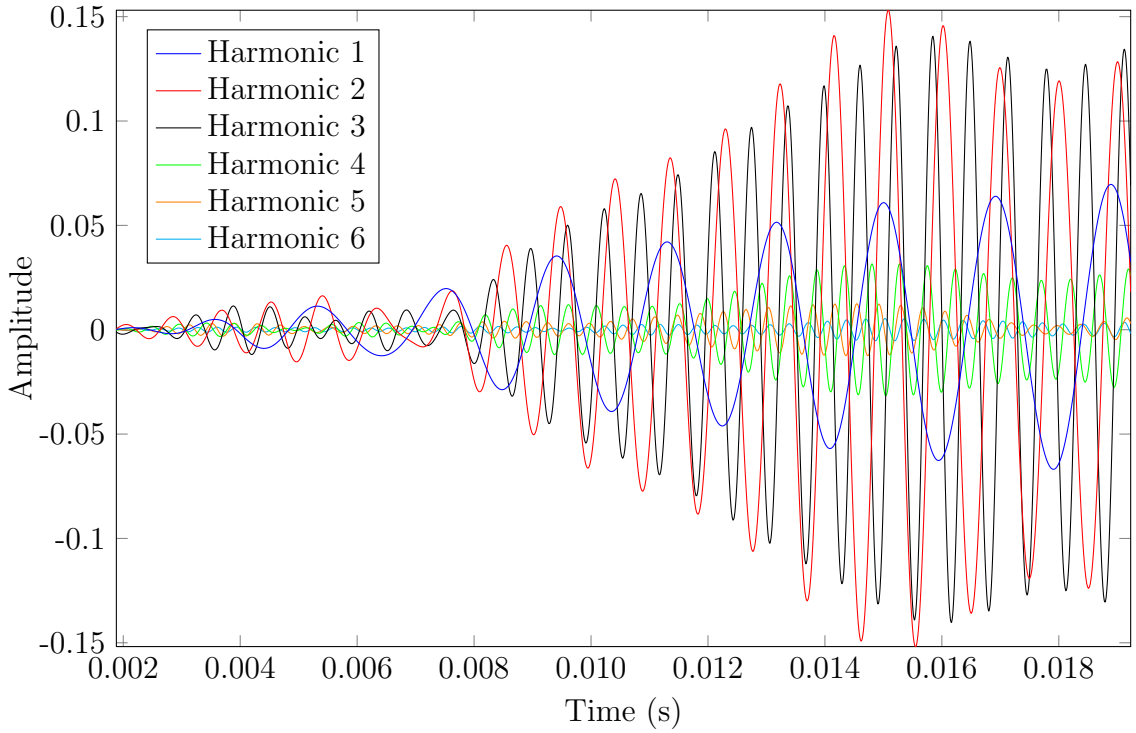

Figure 3-5. A portion of the time-domain signals for the first six harmonics during the onset of a C_5 trumpet note, resolved using the Masked Inverse Fast Fourier Transform (MIFFT) method.

of either, and maybe even weakens the analysis of both. Therefore, the analyses of portions of a signal with disjoint support ought to proceed independently.

We use this notion of disjoint support and independent analysis to motivate a method for quasi-periodic waveform analysis. In a quasi-periodic waveform, the different harmonics that compose the waveform have approximately disjoint support in the frequency domain, an example of which is shown in Figure 3-4. Therefore, we can apply a simple mask to the coefficients corresponding to each harmonic and separate them for independent analysis. Because we wish to analyze their time-dependent properties, after we separate them in the frequency domain we can then transform them back to the time-domain for further analysis, as is shown in Figure 3-5.

One interesting consequence of this idea is that the signals corresponding to strings with inharmonicity where $f_k(t) > k \cdot f_0(t)$ have even less overlapping support in the

frequency domain, making the assumption of disjoint support even more valid. This is a clear advantage of our method over short-time analysis methods where the analysis is based on a fixed window in the time domain. Methods using a fixed-window are justified from the perspective of disjoint support when frequencies are time-invariant and perfectly harmonic, but this is less valid as inharmonicity and time-dependent frequency variation increases.

In the MIFFT method, each harmonic has a corresponding mask indicating which frequencies constitute its support. Each reconstructed harmonic, $h_k(t)$, is given by Equation 3–8,

$$h_k(t) = \frac{1}{N} \sum_{n=0}^{N-1} m_n^{(k)} X_n \cdot e^{i2\pi nt/N}, \tag{3–8}$$

where the $\{X_n\}$ are the Fourier coefficients of the signal in the frequency domain and $m_n^{(k)}$ is the value of the mask for the kth harmonic for the nth frequency. For each mask, we have,

$$m_n^{(k)} = \begin{cases} 1 & \text{if frequency } n \text{ is in harmonic } k\text{'s support} \\ 0 & \text{if frequency } n \text{ is not in harmonic } k\text{'s support} \end{cases}, \tag{3–9}$$

where

$$\sum_{k=1}^{K} m_n^{(k)} \leq 1, \tag{3–10}$$

so that any Fourier coefficient can belong to the support of at most one harmonic.

3.4 Interface

Interface is an important consideration for unrestricted timbre manipulation. In this work we prioritize facilitating composition over performance, which is a potentially important distinction. Both require a large degree of expressiveness, but other considerations differ between the two. Real-time performance requires a fluidity and compactness that is not demanded by composition. On the other hand, composition requires an intuitive

interface for efficiently translating from ideas to implementation. Once the ideas are implemented by composition, performance needs to imitate them in real-time, but not re-create them from scratch.

3.4.1 Hierarchical or Layered Approach

The different requirements for different goals are one reason why we propose a modular and hierarchical approach to interface. As we discuss in Section 2.5.2, multiple interfaces make sense for timbre manipulation. A verbal interface is useful because it corresponds to the language musicians currently use to specify timbre. A gestural interface is advantageous because it gives a person access to a number of manipulable timbre dimensions simultaneously, like a musical instrument but with a much broader palette. An interface based on orthogonal axes of perception is desirable because it would be the most direct and compact representation, with control parameters that are guaranteed to interact independently.

In a hierarchical approach, these multiple interfaces can be built on top of one common interface, extending the fundamental interface with high-level specifications. The common interface, then, is the only one that needs to deal directly with the mental representation or synthesis. This modularity makes it easier to substitute different high-level interfaces for different purposes.

Our proposed hierarchy is outlined in Figure 3-6. The single arrows in this diagram represent a one-to-one correspondence between objects in one layer and objects in another layer. The multiple arrows represent single objects at a higher layer mapping to a sequence of objects at a lower layer, building the evolution of parameters at a high level on top of their instantaneous specification at a lower level. For the directness, compactness, and independence of parameters, we choose the orthogonal dimensions of timbre space as the common interface with perception and synthesis. In addition to higher-level interfaces being built on top of it, this interface can be used directly. This can be useful, especially when one wants to avoid the ambiguity or contradiction that can come from a combination

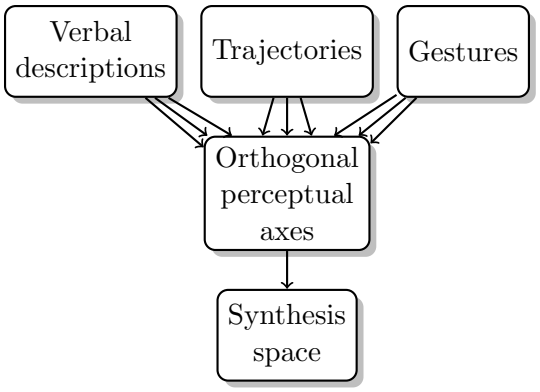

Figure 3-6. The proposed hierarchy for the entire system, serving as an interface from high-level timbre specification to low-level synthesis. Any number of high-level objects can be used to interface with the instantaneous timbre space. Each object at the highest-level, such as a gesture or verbal description, maps to a number of points in the instantaneous timbre space and specifies their order. This abstraction greatly simplifies the specification for the user. Points in instantaneous timbre space then map directly to synthesis parameters, preferably in a one-to-one manner so that the evolution of a sound could be specified in real-time if necessary.

of verbal attributes or gestural parameters. Trajectories are specifications of coordinates in the orthogonal axes over a period of time, which is just a way to encapsulate an entire interaction with the orthogonal dimensions as control parameters.

3.4.2 Orthogonal Axes as Control Parameters

The orthogonal axes are the primary interface on which we focus in this work. The experimental procedure for developing the axes is given in Sections 3.2.2 and 3.2.3. An individual's timbre space has an arbitrary orientation, so a key question in developing this interface is whether or not a certain set of orthogonal axes can be considered in some sense more principal than another, or even whether multiple sets of principal axes exist.

A coordinate in instantaneous timbre space directly corresponds to a set of instantaneous synthesis parameters. Therefore, a path (or trajectory) in this space specifies the evolution of instantaneous timbre over time, which in turn specifies a sound. In theory, this is the most perceptually compact representation possible, since only dimensions relevant to perception are included, while no dimension relevant to perception is omitted.

The representation used in our work does omit some dimensions of perception that are not demonstrated by the 16 sounds used in the experiment. These dimensions include noise and silence. These two dimensions could easily be incorporated in a future version of the work with a larger data set. However, some dimensions of instantaneous timbre may only emerge within the context of dynamic stimuli, in which case they would not be discovered by our method. We plan to investigate this possibility in the future.

Using the orthogonal axes corresponds to the simultaneous manipulation of several linear parameters. This can be difficult to do in real-time without an interface at a higher-level which can incorporate multiple types of gestures. However, if real-time interaction is not a concern, like perhaps for composition, the parameters can be programmed. The joint specification of all of the programmed parameter changes can be thought of as a path through instantaneous timbre space. On the other hand, a higher-level interface could ease the interaction even when real-time manipulation of timbre is not a concern.

CHAPTER 4
RESULTS

The primary goal of this work is to develop a framework for utilizing timbre as freely as loudness or pitch in musical composition. The focus experimentally in this work is developing the perceptual component of the framework and increasing our understanding of timbre as a complex perceptual phenomenon. In this chapter, we report the primary experimental results and the data derived from them, and analyze them in detail.

4.1 Timbre Space Experiment

The procedure used for this experiment is described in Section 3.2.2. In all, six subjects were asked to listen to a set of 16 sounds in small groups of 3 or 4 and identify the two most similar and two least similar sounds in each group. Providing information on the two most similar sounds in a group of 4 sounds says that among the $\binom{4}{2} = 6$ unique pairs, the two sounds in the most similar pair are nearer to each other than the two sounds in any of the other 5 possible pairs. This provides comparison information between this pair and the other 5 pairs. Taken together, all of the trials of this sort fill out a comparison matrix that is P-by-P, where $P = \binom{16}{2} = 120$. The comparison matrix contains $\binom{120}{2} = 7120$ unique entries.

The P-by-P comparison matrix, M, for subject 2 is shown in Figure 4-1. A white square can be thought of as a victory and indicates that the pair of sounds in the row is more similar than the pair of sounds in the column. A black square can be thought of as a loss and indicates that the pair of sounds in the row is less similar than the pair of sounds in the column. A gray square represents the lack of any comparison information. The matrix is anti-symmetric so that $M(i, j) = -M(j, i)$.

The comparison matrix for each subject is somewhat sparse, but there is enough information to yield a ranking of the pairs from most similar to least similar. Colley's matrix method is used to do the ranking. The same comparison matrix ordered from most

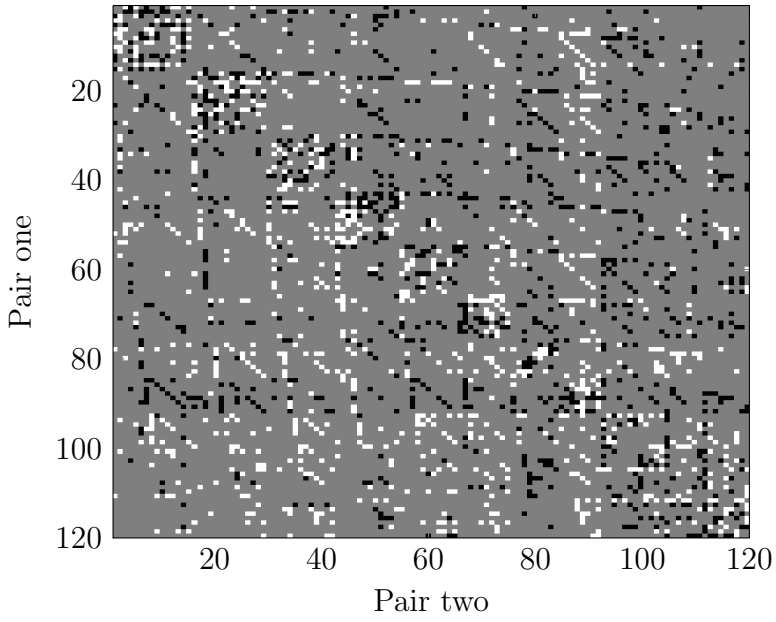

Figure 4-1. The P-by-P comparison matrix for subject M. Each row and each column represents a pair of sounds. Each square indicates the outcome of the comparison between the row pair and the column pair. A white square indicates that the sounds in the row pair are judged by the subject to be more similar than the sounds in the column pair. A black square indicates that the sounds in the row pair are judged by the subject to be less similar than the sounds in the column pair. A gray square indicates that there is no comparison between the two pairs.

similar to least similar is given by Figure 4-2. The pairs at the top of the ranking have the most evidence in favor of their being the most similar pairs.

Kruskal's non-metric MDS algorithm is used to compute the best spatial embedding in D dimensions given the ranking of distances between pairs of points. The error of the embedding given the rankings is referred to by Kruskal as the stress. Figure 4-3 shows the stress of each embedding for each subject for a range of dimensionalities.

We choose to compute 5-dimensional spaces. One of the reasons for this is that the reduction in stress after 5 dimensions starts to become so insignificant that beyond that overfitting to the noise in the data is likely to start occurring. Another reason is that with only 16 sounds, using more than 5 dimensions would start to impair our ability to

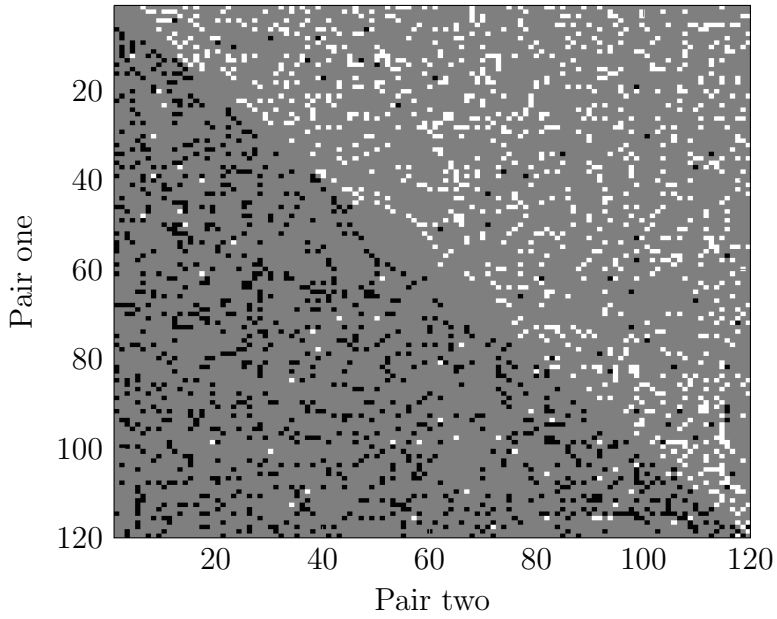

Figure 4-2. The same comparison matrix as in Figure 4-1 except with the pairs ordered from most similar to least similar by Colley's matrix method. Black squares above the main diagonal and their corresponding white squares below the main diagonal represent violations of transitivity, since they indicate that a pair that is ranked more similar than another pair is judged by the subject to be less similar.

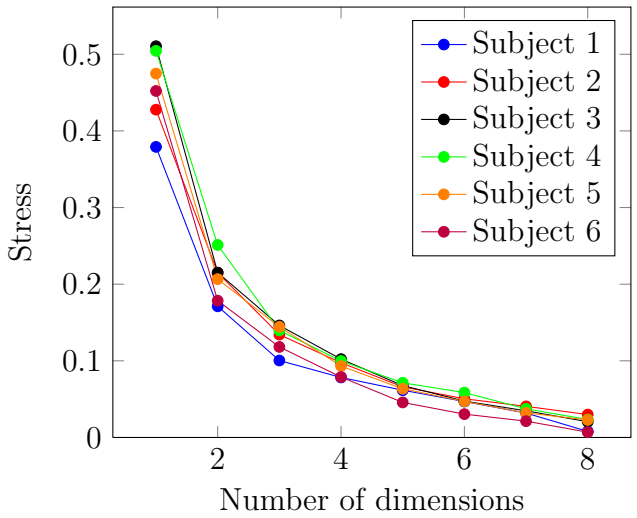

Figure 4-3. Plot of the stress calculated for the configuration found by Kruskal's non-metric MDS for each subject for each number of dimensions. The reduction in stress appears to stop being significant after a dimensionality of either 5 or 6.

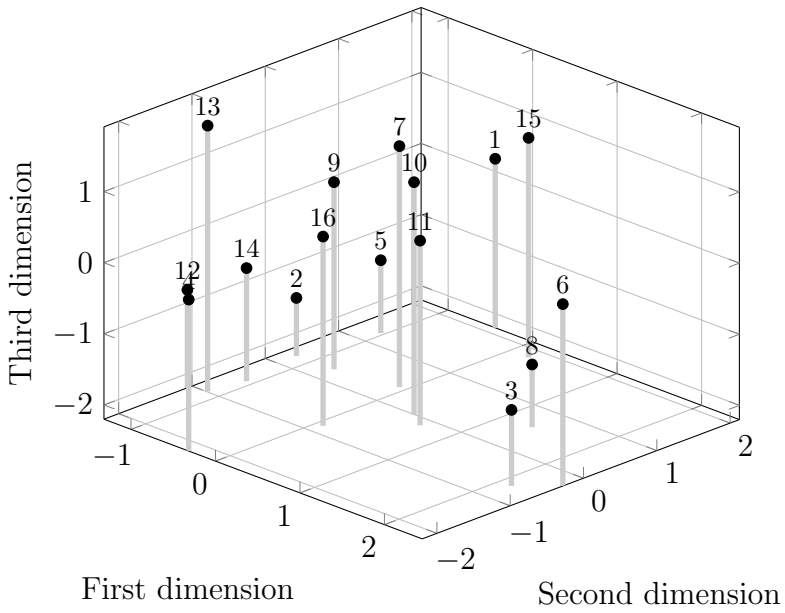

Figure 4-4. First three dimensions of the individual instantaneous timbre space for subject 2.

distinguish structure in the space from random variation, an issue that is discussed in Section 3.2.6.

Figure 4-4 shows the first three dimensions of the 5-dimensional solution found for subject 2. The subject whose spatial embedding has the lowest stress in 5 dimensions is subject 6. Figure 4-5 shows the first three dimensions of subject 6's spatial embedding.

4.2 Joint Instantaneous Timbre Space

INDSCAL is used to combine spatial information from multiple subjects into one common space. Even though we choose 5 dimensions for the individual embeddings, we are free to choose a different number of dimensions for the common space. Having more than 5 common dimensions would make sense if subjects differ in which dimensions they perceive most strongly, since no subject would include all of the common dimensions in their individual embedding. Having less than 5 common dimensions could be useful because, with the energy more concentrated, it may cause INDSCAL to find clearer perceptual dimensions than when the energy is spread out over a larger number of

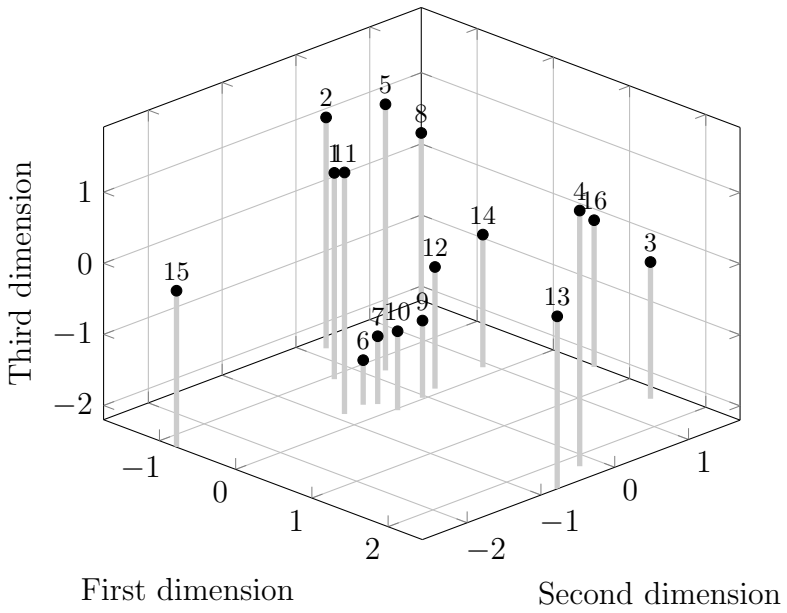

Figure 4-5. First three dimensions of the individual instantaneous timbre space for subject 6.

dimensions. To investigate these possibilities, we examine INDSCAL solutions for 4, 5, and 6 dimensions.

Figure 4-7 shows the first three dimensions of the 4-dimensional joint embedding computed by INDSCAL. Figure 4-8 shows the first three dimensions of the 5-dimensional INDSCAL embedding, and Figure 4-6 shows the first three dimensions of the 6-dimensional INDSCAL embedding.

INDSCAL also produces a set of weights that each subject applies (actually they apply the square roots of the weights) to the common dimensions to yield their individual spaces. Table 4-1 gives the square roots of the weights computed by INDSCAL for six subjects and four dimensions of instantaneous timbre. The values are normalized to sum to 1 so that they are easier to compare. Table 4-2 gives the square roots of the weights for the case of five dimensions and Table 4-3 gives the square roots of the weights for six dimensions. In all three cases, the input to INDSCAL is the Euclidean distances computed in the 5-dimensional individual embeddings.

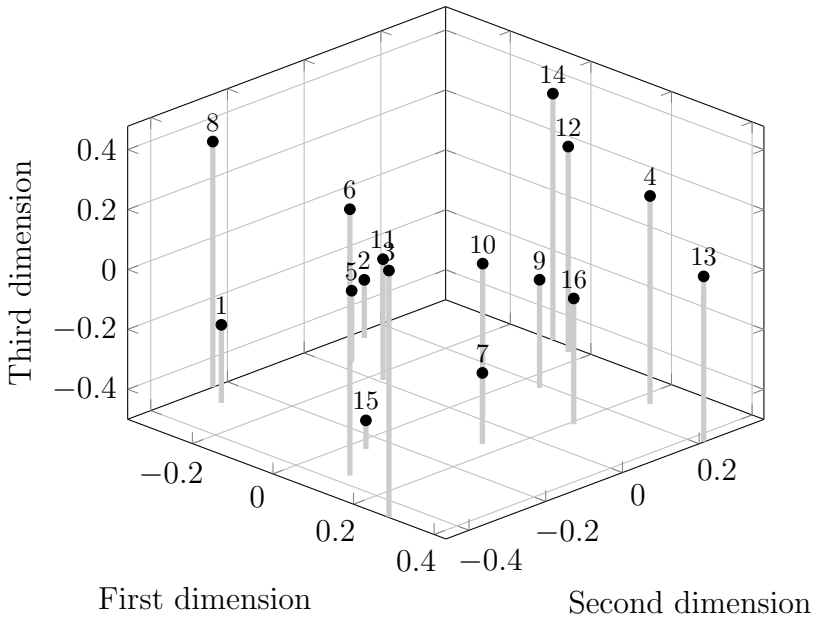

Figure 4-6. First three dimensions of the 4-dimensional INDSCAL embedding of joint instantaneous timbre space.

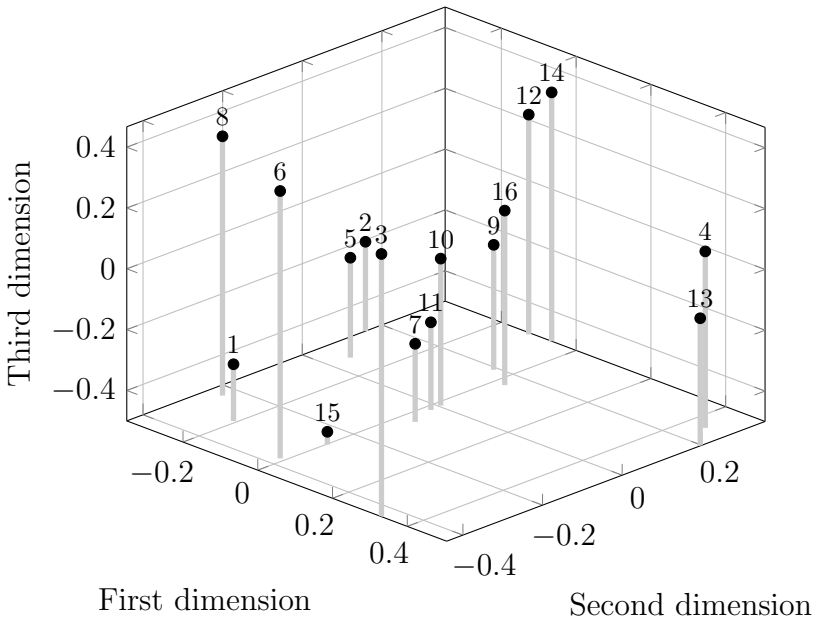

Figure 4-7. First three dimensions of the 5-dimensional INDSCAL embedding of joint instantaneous timbre space.

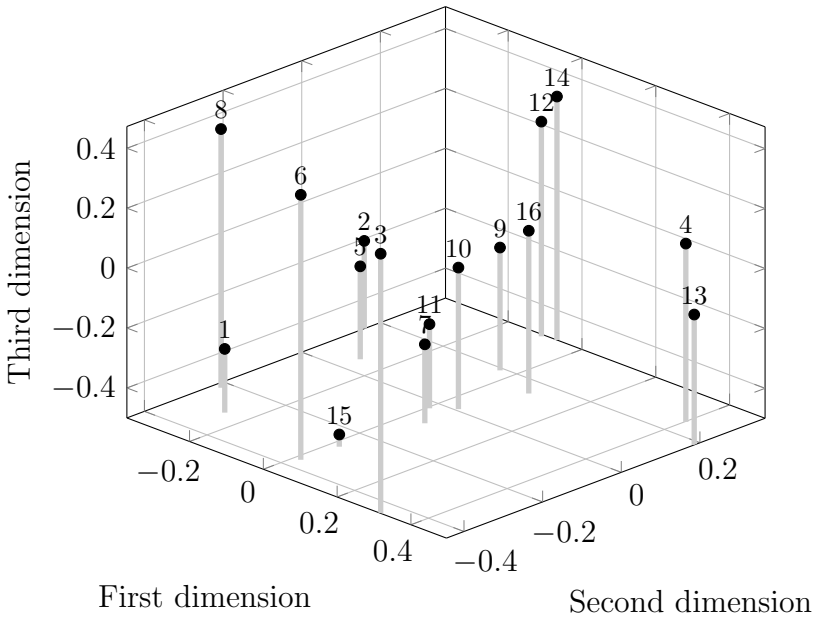

Figure 4-8. First three dimensions of the 6-dimensional INDSCAL embedding of joint instantaneous timbre space.

Table 4-1. The square roots of the weights computed by INDSCAL for six subjects and four dimensions of instantaneous timbre, normalized to sum to 1.

Subject	$\sqrt{w_1}$	$\sqrt{w_2}$	$\sqrt{w_3}$	$\sqrt{w_4}$	Total
1	0.056	0.046	0.047	0.033	0.180
2	0.044	0.053	0.046	0.030	0.173
3	0.050	0.041	0.049	0.015	0.154
4	0.047	0.050	0.042	0.021	0.160
5	0.057	0.040	0.012	0.046	0.155
6	0.057	0.036	0.042	0.043	0.178
Total	0.310	0.266	0.236	0.187	1.000

Table 4-2. The square roots of the weights computed by INDSCAL for six subjects and five dimensions of instantaneous timbre, normalized to sum to 1.

Subject	$\sqrt{w_1}$	$\sqrt{w_2}$	$\sqrt{w_3}$	$\sqrt{w_4}$	$\sqrt{w_5}$	Total
1	0.048	0.041	0.041	0.028	0.017	0.175
2	0.040	0.047	0.040	0.026	0.014	0.169
3	0.043	0.036	0.043	0.012	0.011	0.146
4	0.039	0.045	0.037	0.018	0.044	0.183
5	0.052	0.036	0.009	0.041	0.020	0.158
6	0.051	0.032	0.036	0.039	0.012	0.170
Total	0.274	0.236	0.208	0.164	0.118	1.000

Table 4-3. The square roots of the weights computed by INDSCAL for six subjects and six dimensions of instantaneous timbre, normalized to sum to 1.

Subject	$\sqrt{w_1}$	$\sqrt{w_2}$	$\sqrt{w_3}$	$\sqrt{w_4}$	$\sqrt{w_5}$	$\sqrt{w_6}$	Total
1	0.043	0.036	0.036	0.023	0.025	0.013	0.175
2	0.036	0.042	0.035	0.027	0.013	0.013	0.168
3	0.035	0.032	0.041	0.011	0.024	0.008	0.152
4	0.031	0.041	0.034	0.008	0.025	0.039	0.179
5	0.051	0.032	0.005	0.034	0.014	0.017	0.153
6	0.045	0.028	0.030	0.037	0.030	0.006	0.174
Total	0.240	0.211	0.182	0.140	0.131	0.096	1.000

Each subject's personal weighting of the joint embedding should approximately re-produce their individual embedding. Thus for each subject there are two personal spaces we can examine, the original individual embedding produced by Kruskal's algorithm, and the weighted version of the joint embedding produced by INDSCAL. The individual embeddings are difficult to compare directly because they have an arbitrary orientation and cannot be linearly fit to each other using least-squares since not every linear transform preserves the configuration of points in the space (shearing does not). Therefore, the weighted embeddings, which are an oriented approximation of the individual embeddings, are useful as a way to compare each subject's personal instantaneous timbre space.

Figure 4-9 shows the first two dimensions of every subject's 6-dimensional weighted embedding. As can be seen from this figure and the weights in Table 4-2, the first two dimensions of joint instantaneous timbre space are exhibited very similarly by all six subjects. Table 4-10 shows the third and fourth dimensions of the same weighted embeddings. In these dimensions, differences are more readily apparent. While subjects 1 and 2 remain very similar, subjects 3 and 5 exhibit significant differences in their spaces.

It cannot be assumed that the weighted embeddings are accurate reproductions of the individual embeddings for each subject. The weighted embeddings all represent a non-uniform but linear scaling of the dimensions of the joint embedding and have the same axes. We can test to what extent these axes exist in the individual embeddings using

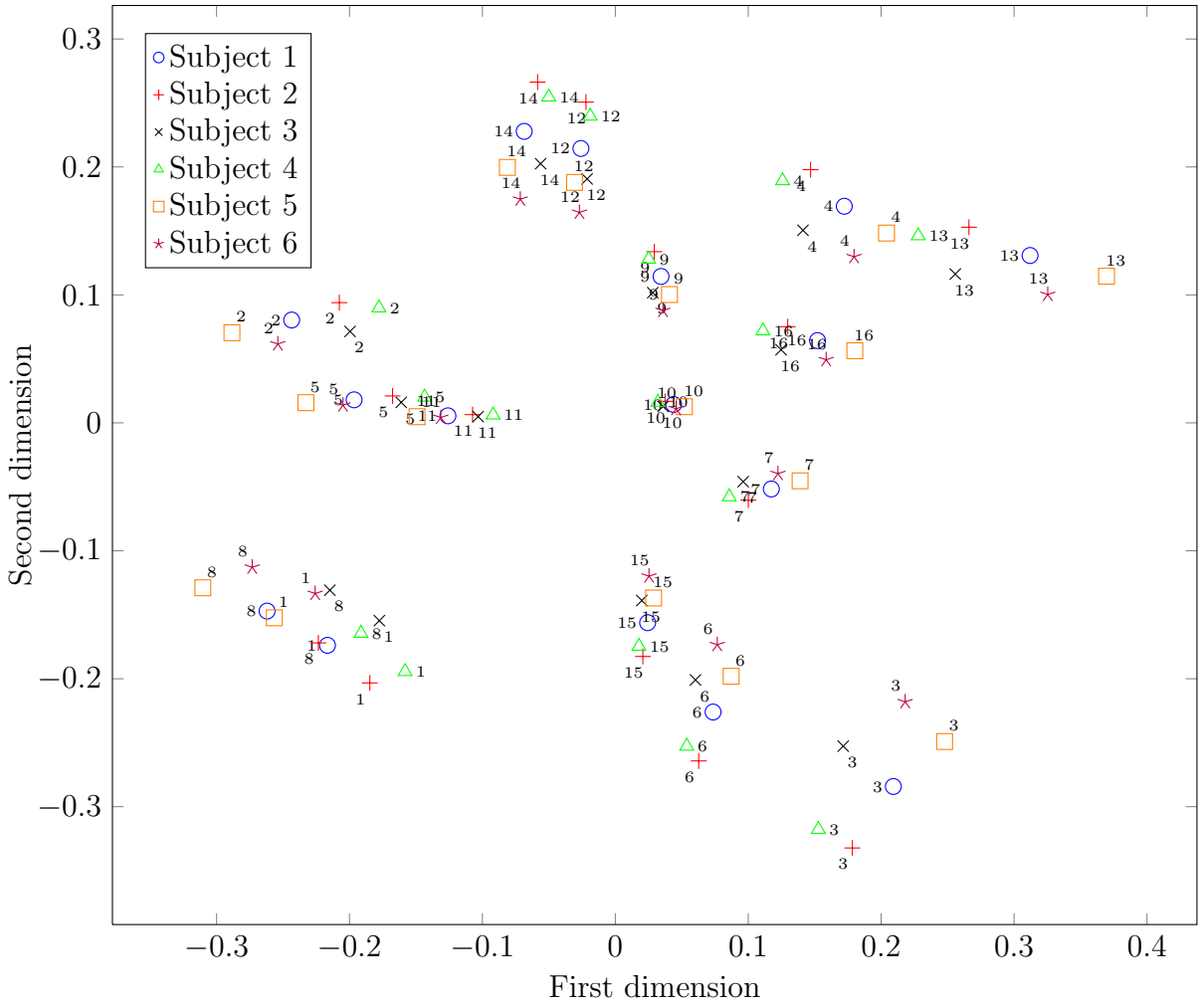

Figure 4-9. Visual comparison of the first two dimensions of each subject's weighted embedding as produced by INDSCAL. Subjects do not exhibit much difference in these dimensions

least squares. Table 4-4 shows the correlation between the principal axes determined by INDSCAL and the closest-matching axis in each of the individual spaces.

4.2.1 The Perceptual Dimensions of Timbre

One of the central aims of this work is to explain perceptual dimensions of instantaneous timbre. Since INDSCAL claims to discover the axes that are the most meaningful perceptually, these axes are a good place to start. As is explained in Section 3.2.3, there are a number of ways in which we can attempt to explain these axes. Table 4-5 shows the correlations between each axis and seven different acoustic properties of the sounds. The

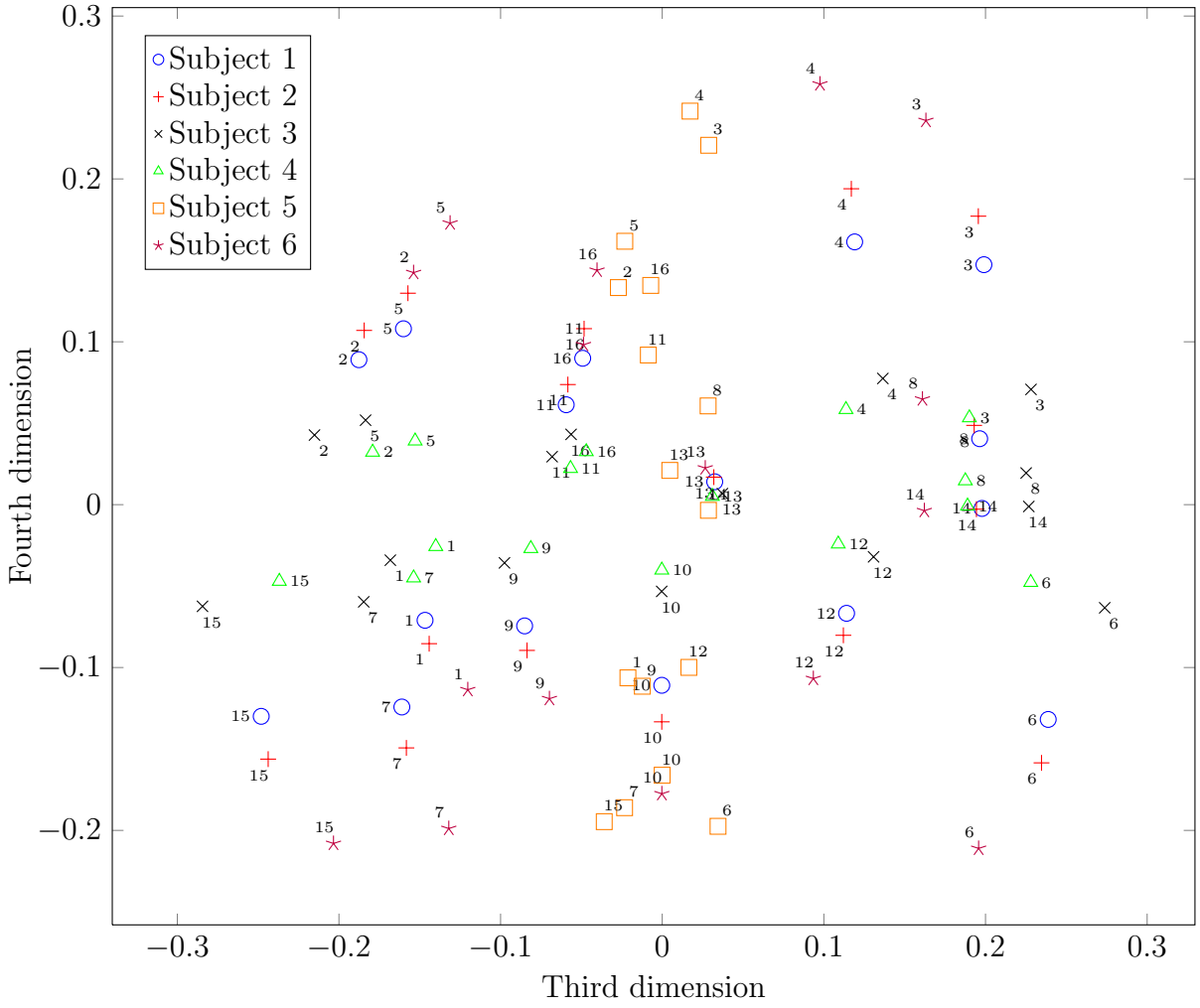

Figure 4-10. Visual comparison of the third and fourth dimensions of each subject's weighted embedding as produced by INDSCAL. Subjects start to exhibit noticeable differences in these dimensions.

Table 4-4. The correlation between the principal axes $\vec{d}_1, \vec{d}_2, \ldots, \vec{d}_6$ determined by INDSCAL and the individual embeddings of the six subjects.

Subject	$\vec{d}_1$	$\vec{d}_2$	$\vec{d}_3$	$\vec{d}_4$	$\vec{d}_5$	$\vec{d}_6$
1	0.96	0.96	0.88	0.72	0.84	0.39
2	0.89	0.96	0.90	0.78	0.54	0.43
3	0.87	0.88	0.93	0.40	0.87	0.30
4	0.86	0.98	0.93	0.34	0.89	1.00
5	0.98	0.89	0.34	0.90	0.61	0.60
6	0.98	0.85	0.82	0.95	0.99	0.28

Table 4-5. The correlations between the principal axes $\vec{d}_1, \vec{d}_2, \ldots, \vec{d}_6$ determined by INDSCAL and seven acoustic properties $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_7$ of the sounds.

	$\vec{p}_1$	$\vec{p}_2$	$\vec{p}_3$	$\vec{p}_4$	$\vec{p}_5$	$\vec{p}_6$	$\vec{p}_7$
$\vec{d}_1$	0.93	0.95	0.06	0.17	0.40	0.09	0.28
$\vec{d}_2$	0.30	0.30	0.72	0.32	0.53	0.60	0.54
$\vec{d}_3$	0.32	0.35	0.51	0.18	0.69	0.65	0.44
$\vec{d}_4$	0.01	0.10	0.15	0.38	0.36	0.28	0.36
$\vec{d}_5$	0.31	0.38	0.30	0.33	0.05	0.35	0.01
$\vec{d}_6$	0.03	0.08	0.04	0.30	0.15	0.02	0.03

acoustic properties are $\vec{p}_1$ = fundamental frequency, $\vec{p}_2$ = log fundamental frequency, $\vec{p}_3$ = spectral centroid, $\vec{p}_4$ = odd harmonics proportion, $\vec{p}_5$ = loudness, $\vec{p}_6$ = loudness-weighted spectral centroid, and $\vec{p}_7$ = inharmonicity factor standard deviation. These properties are described in more detail in Table 3-3.

From the correlation analysis, the first axis is clearly related to pitch, with the correlation between $\vec{d}_1$ and $\vec{p}_2$ being 0.95. The other axes are not as clear from the correlation analysis. The analysis suggests a significant relationship between the second axis and spectral centroid, indicating that $\vec{d}_2$ may be linked to brightness. The third axis has a potentially significant relationship with loudness. Interestingly, the fourth and sixth axes show no significant correlation with any of these features.

Another method of investigating an axis is to simultaneously observe the sounds projected onto the axis and listen to the sounds, then attempt to find a perceptual pattern that explains the variation along the axis. For the second axis, brightness does seem to have a strong correlation with the axis perceptually. Interestingly, the inharmonic sounds, which are not necessarily the brightest, are at the far end of the axis with the brightest sounds. This explains in part why inharmonicity has a fairly strong correlation with this axis and why spectral centroid may have had a weaker correlation than it should have. The third axis, too, seems perceptually to have a strong correlation with brightness, except this time the inharmonic sounds are placed at the end of the spectrum with the least bright sounds.

Table 4-6. The new correlations between the principal axes $\vec{d}_1, \vec{d}_2, \ldots, \vec{d}_6$ determined by INDSCAL and seven acoustic properties $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_7$ of the sounds after removing sounds 3, 6, and 8 from the correlation calculations.

	$\vec{p}_1$	$\vec{p}_2$	$\vec{p}_3$	$\vec{p}_4$	$\vec{p}_5$	$\vec{p}_6$	$\vec{p}_7$
$\vec{d}_1$	0.91	0.94	0.30	0.25	0.37	0.22	0.19
$\vec{d}_2$	0.51	0.51	0.89	0.54	0.90	0.91	0.67
$\vec{d}_3$	0.58	0.58	0.83	0.23	0.90	0.81	0.64
$\vec{d}_4$	0.00	0.10	0.26	0.26	0.29	0.29	0.21
$\vec{d}_5$	0.24	0.33	0.29	0.38	0.09	0.32	0.15
$\vec{d}_6$	0.04	0.15	0.08	0.27	0.15	0.05	0.19

As a result of these observations, the most inharmonic sounds, sounds 3, 6, and 8 (as can be seen in Table 3-4), were removed and the correlations were re-calculated. Table 4-6 presents the results. The second and third dimensions now show a very strong correlation with spectral centroid, loudness, and loudness-weighted spectral centroid. This indicates that these two dimensions both ought to be interpreted as brightness but with outliers.

This same method of investigation with the fourth axis also yielded an observable pattern. In the data set, there are a relatively large number of A notes. The fourth axis groups all of the A notes, along with one B note, on one side of the axis, and places the rest of the non-A notes on the other side of the axis. Given the prevalence of A notes in the data, it is not surprising that A versus non-A emerged as a defining distinction. But it also speaks to a perceptual phenomenon, which is that notes an octave apart are often confused for one another. Thus, there is an inherent perceptual similarity between notes an integer number of octaves apart, and this shows up in the fourth axis.

The fifth INDSCAL dimension required a number of listens because its variation was not as apparent as the variation of pitch and brightness observed in the first four axes. But a fairly clear perceptual axis did eventually emerge. On the lower (though this is an arbitrary orientation) side of the axis, sounds could be described as smooth or upbeat. On the higher side of the axis, sounds could be described as biting and negative. These traits are not known to be correlated with any particular acoustic properties, therefore an analytical verification cannot be performed to determine a correlation coefficient.

Table 4-7. A summary of the interpretations of the dimensions of instantaneous timbre found by INDSCAL.

Dimension	Interpretation	Primary method used
1	Absolute pitch	Analytical
2	Brightness with outliers	Perceptual and analytical
3	Brightness with outliers	Perceptual and analytical
4	Relative pitch	Perceptual and analytical
5	Smoothness, positivity, bite	Perceptual
6	Unknown	

Table 4-8. The correlation between seven acoustic properties $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_7$ of the sounds and their best-matching axes in the joint embedding.

$\vec{p}_1$	$\vec{p}_2$	$\vec{p}_3$	$\vec{p}_4$	$\vec{p}_5$	$\vec{p}_6$	$\vec{p}_7$
0.98	1.00	0.92	0.75	0.93	0.94	0.84

The sixth INDSCAL dimension was found to be very difficult to explain perceptually and no conclusion can be drawn. As can be seen from Table 4-4, however, this dimension seems to primarily exist in the individual instantaneous timbre space of only one subject, with whom there is nearly perfect correlation. Therefore it is not surprising that this dimension is difficult to explain. It would be extremely interesting if this one subject were able to provide a perceptual explanation, but this test was not able to be done.

A summary of the interpretations for the dimensions of instantaneous timbre found by INDSCAL is given by Table 4-7.

Carroll and Chang argue that the orientation found by INDSCAL is the only orientation the joint embedding can have, since according to the model there is no linear connection between any other axes in the joint embedding and axes in the individual embeddings. However, it is still interesting to test if the acoustic properties correlate with axes in the joint embedding that are not aligned with the principal axes found by INDSCAL. Table 4-8 lists the correlations found between the acoustic properties of the sounds and their best-matching axes in the joint embedding. We also test the individual spaces to see if there are axes in these spaces that correlate with the acoustic properties thought to be relevant to perception. These correlations are listed in Table 4-9.

Table 4-9. The correlation between seven acoustic properties $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_7$ of the sounds and their best-matching axes in the individual embeddings of each subject.

Subject	$\vec{p}_1$	$\vec{p}_2$	$\vec{p}_3$	$\vec{p}_4$	$\vec{p}_5$	$\vec{p}_6$	$\vec{p}_7$
1	0.94	0.97	0.82	0.53	0.88	0.87	0.76
2	0.96	0.95	0.92	0.67	0.95	0.95	0.87
3	0.91	0.93	0.87	0.68	0.91	0.87	0.69
4	0.87	0.91	0.92	0.50	0.85	0.89	0.71
5	0.95	0.98	0.86	0.47	0.76	0.76	0.69
6	0.97	0.98	0.83	0.65	0.90	0.89	0.66

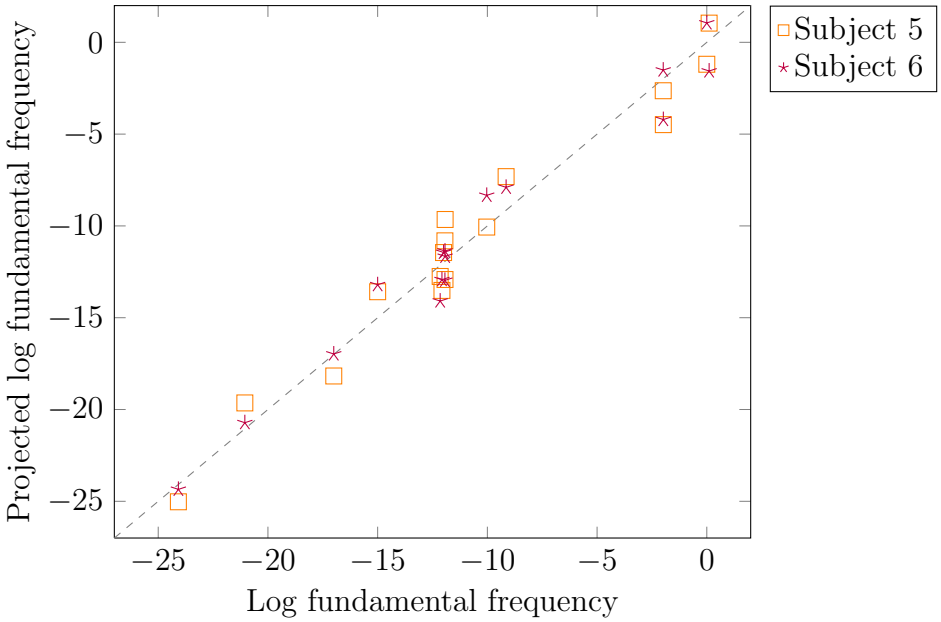

Figure 4-11. The x-axis values are the log fundamental frequencies of the 16 sounds. The y-axis values are the best-matching projections in the individual embeddings of instantaneous timbre space for subjects 5 and 6. The correlations between the sets are listed in column 3 of Table 4-9.

Figure 4-11 shows the log fundamental frequency of each sound plotted against the best-matching projection of the sounds found in the individual spaces of subjects 5 and 6. Out of all of the subjects, these two had the highest correlation between their spaces and log fundamental frequency.

The inharmonicity feature does not appear to have a particularly strong correlation with the instantaneous timbre spaces, but due to the relative lack of inharmonic sounds in the data set, this does not necessarily mean the feature calculated in this work is not an appropriate measure of perceptual inharmonicity. An observation about this feature

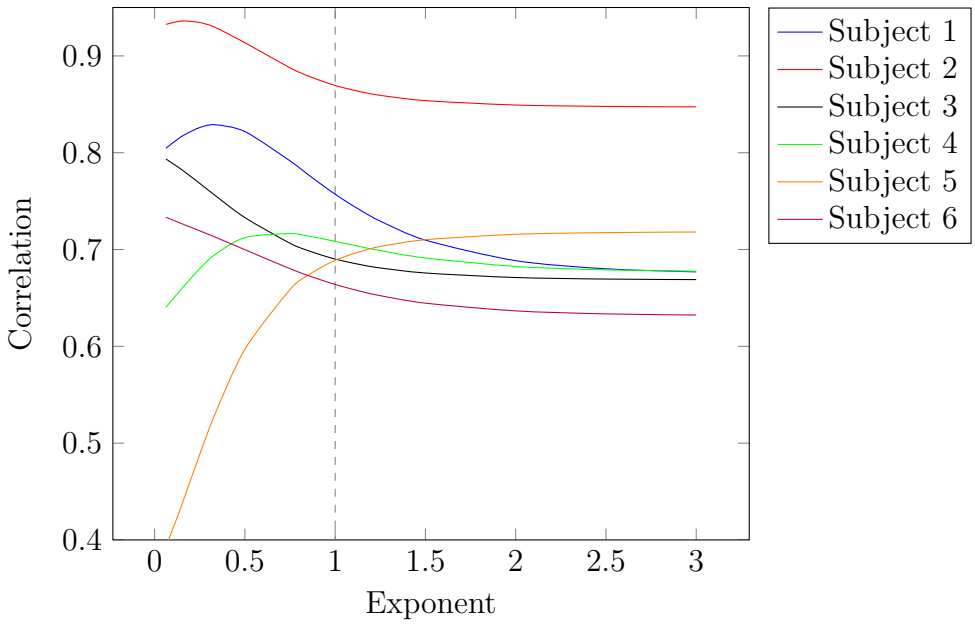

Figure 4-12. The inharmonicity is calculated as the proportionally-weighted standard deviation of $f_i/(i * f_0)$ for all of the harmonics. This graph shows how the correlation of this statistic with subjects' individual instantaneous timbre spaces changes as the value is raised to different exponents. The correlation peaks at different exponents for each subject.

is shown in Figure 4-12. The inharmonicity is calculated as the standard deviation of $f_i/(i * f_0)$ for each harmonic, where f_i is the frequency of harmonic i and f_0 is the fundamental frequency of the note. The standard deviation is not the only possible choice of summary statistic. Another is the mean or the variance. Since the variance is the square of the standard deviation, a number of powers of the standard deviation were examined to see if they gave any improvement in correlation over the standard deviation for each of the subjects' instantaneous timbre spaces.

Each subject has a different peak exponent value. The exponents each represent non-linear warpings of the inharmonicity feature when viewed as a projection in timbre space. The fact that each subject has a fairly high peak correlation suggests that the particular feature being used for inharmonicity may be ranking the inharmonicities correctly but not matching perception in an optimal way. Thus other ways of calculating inharmonicity may need to be investigated.

CHAPTER 5
CONCLUSIONS

In this work we have proposed a comprehensive, perception-centric framework for the digital manipulation of timbre in composition. This has been an inherently inter-disciplinary approach, drawing on knowledge from domains such as computer science, music, signal processing, and psychology. In addition to our contribution to the development of a system for unrestricted timbre manipulation, we have made a number of potential contributions to signal processing, experimental procedure, and musicology.

Among our primary contributions is the development of a new representation for timbre, which is the first to simultaneously represent perceptual dimensions of timbre and be suitable for synthesizing sounds from this representation. Previous attempts at a representation of timbre using psychological experiments have not succeeded in identifying fundamental dimensions of timbre and have not been suitable for synthesis. On the other hand, representations of time-varying acoustic properties that are suitable for synthesis have chiefly been developed as models of existing musical instruments and used the minimization of signal reconstruction error as their primary development objective. We believe this approach is fundamentally incapable of developing a correspondence between time-varying properties and perceptual dimensions. Thus our approach overcomes the limitations imposed by previous attempts.

Like some of the representations based on time-varying acoustic parameters, we have proposed a phase space representation, in which all attributes at a single moment in time are treated as points in the space, and paths in the space represent the evolution of these attributes over time. Except in our case, these attributes are perceptual in nature. In this work, the notion of instantaneous timbre, while hinted at in prior work, has been made explicitly distinct from dynamic timbre. This creates a new hierarchy of timbre that has implications for both a system to achieve its manipulation and for its study in general.

To learn a space of instantaneous timbre, we have proposed a study that applied machine learning techniques to data collected from human subjects. As has been mentioned, this general type of psychological study has been undertaken before. But this is the first time such a study has been attempted exclusively with instantaneous timbre. We have also improved upon earlier methods in a number of ways. We expanded the range of variation in the data to include notes of different pitch. Procedurally, we have collected responses from subjects as relative comparisons rather than absolute judgments, which we argue eliminates a potentially large source of inconsistency in the observations. At the same time, we have improved upon the prior method of triadic comparisons by expanding the group size in a trial to more than three at a time. This allows the entire matrix of binary comparisons to be accessible, rather than a subset of the matrix corresponding to pairs with one sound in common. We have compensated for the increased time it takes to collect responses and improved upon the method of triadic comparisons by framing the problem of ordering the pairs by their similarity as a general ranking problem, given an incomplete binary comparison matrix. This is a more principled approach than simple winning percentage, as used in triadic comparisons, and leverages advances in ranking theory.

In general, we have been able to draw a number of conclusions from our experiments. Investigation into the nature of instantaneous timbre indicates that there are at least 5 coherent dimensions readily apparent just from a collection of 16 sounds. Prior studies insisted that pitch must be held fixed to study timbre. However, we propose that the opposite may be true, and we have shown that timbre differences can emerge just as easily when pitch is included as a source of variation. The primary dimension of perception indicated in our study is based on absolute pitch. But brightness also emerged clearly, as well as a new dimension of timbre not mentioned in prior literature, that is associated with words such as smoothness, positivity, and bite. Relative pitch emerged as a dimension, as well, which is not surprising given that the perceptual quality of harmony

is based on intervals. A study indicating the existence of multiple dimensions of pitch has been done prior, but never has a dimension of relative pitch been observed in the context of timbre.

While we uncovered new dimensions of timbre that have not been observed in prior studies, there are additional dimensions of timbre that we may have missed. Some of these could be due to a small data set that lacked enough variation in other dimensions for them to emerge clearly. We also consider the possibility that there are instantaneous dimensions of timbre that can only be observed with dynamic stimuli. This possibility is something we plan to investigate in the future.

Based on our observations, we believe we have provided further indication that INDSCAL is a useful method for determining coherent axes of perception. However, we have done so by explicitly examining the role of subjectivity and observing the extent to which these axes actually exist in the individual spaces of the subjects. It is of great interest to us whether the INDSCAL space can be deemed more reliable than the individual spaces, by leveraging information from multiple sources, or whether the INDSCAL space is a muddled version of the individual spaces. The clear utility of the INDSCAL result seems to indicate the former, but we cannot determine this conclusively from the present study, and future work may be required to do so.

There are a number of other future directions suggested by our work that we would like to investigate. Our overarching goal of a system for timbre manipulation required us to first propose new concepts and methods related to the study of timbre itself. Both basic research into the study of timbre and the development of a system that uses this perceptual knowledge have many more questions to be explored. To facilitate research into a number of these, we believe that first and foremost a larger study similar to the one proposed in this work but incorporating more sounds ought to be undertaken. Increased diversity in the sounds would allow more perceptual dimensions to emerge and allow us to be more certain of the form of these dimensions.

For instance, inharmonicity was clearly present in 3 out of the 16 sounds used in the study. This feature had a marked influence on the ability of INDSCAL to resolve a dimension of brightness. This indicates to us that inharmonicity is a strong perceptual feature, but that not enough variation in inharmonicity was included for it to emerge as its own dimension. Also, analytical methods of interpreting the axes could be greatly improved with the use of more sounds. We were able to explain some axes by correlating them with acoustic features thought to be relevant to perception. But we were not able to automatically learn the features that best explain an axis due to the lower number of sounds compared to the number of acoustic features. Discovering the acoustic basis for a perceptual axis would make it easier to synthesize new sounds that correspond to a perceptual feature. But future work in this area is only possible with more data.

Perceptual interpretations of instantaneous timbre dimensions would be useful, as well. We have used some exploratory methods, bolstered by knowledge of the domain, to interpret axes perceptually. But with more data, more systematic investigation could be done. INDSCAL and related methods of weighted combination are currently the best methods for learning orientations of timbre space where principal axes emerge. But it is not clear if the axes learned by INDSCAL are the only set of principal axes possible. In color vision, at least two completely different coherent sets of axes exist. Additional perceptual experiments could help to illuminate this.

For instance, a one-dimensional MDS experiment could be set up where specific verbal descriptions of the sounds are used as the criteria for comparison, rather than an all-encompassing notion of similarity. This would create perceptual axes that could then be analytically correlated with timbre space embeddings to see if they are present and how they are oriented. Conversely, an axis could be tested for its perceptual coherence by presenting sounds auditory, presenting their projections visually, and asking subjects to identify a pattern that explains the variation observed in the projections onto the axis. Identification of the pattern could be done by identifying words to describe it, as has been

done in this work, but it could also be done implicitly to simply test the coherence of an axis before attempting to fully interpret it. This would entail leaving out some of the sounds from the projection and asking subjects to place these sounds onto the axis based on the pattern they have identified. If the axis corresponds to a perceptually coherent dimension, a pattern should emerge, and the accuracy in identifying the locations of the test sounds on the axis should measure its coherence. This is another example of how more analytical analysis could be done in the context of perception.

We believe we have made a persuasive case that the hierarchy of instantaneous and dynamic timbre proposed in this work is necessary and consistent with perception. But much of the form that dynamic timbre takes within our model still needs to be investigated. Dynamic timbre is represented in our model as paths through instantaneous timbre space. But an enumeration of dynamic timbre properties and their representations as paths still needs to be done. We have emphasized developing the instantaneous spatial representation in this work because it is a necessary pre-requisite to studying the dynamic representation. But as soon as a satisfying instantaneous spatial representation is achieved, more attention can be given to dynamic considerations.

The hierarchy between instantaneous and dynamic timbre has implications for the future development of the framework for timbre manipulation in general. Developing this framework was identified as one of our primary motivations in undertaking this research. We have laid out a detailed blueprint to achieve this goal, and taken a large step toward implementing the centerpiece of the framework. In the future, we would like to implement the entire system. This entails developing a low-level mapping from instantaneous timbre space points to synthesis parameters and a high-level mapping from paths in instantaneous timbre space to simple interfaces and abstractions. Development of the mapping to synthesis parameters seems like the next logical step. To do this well, more points in the instantaneous timbre space need to be associated with synthesis parameters. Currently, we have associations between 16 points only, for the 16 sounds in our experiment. Thus

we need to expand the number of sounds in our experiment not only to improve our instantaneous timbre space representation, but also to improve our ability to map from this space to alternate representations. It may be that as more points are added to the space and we are more certain of the axes, it will become no longer necessary to place the points in the space via the entire experimental procedure outlined in this work. Instead, there may be quicker ways to add points to a pre-existing space. Once enough points are in place, a number of machine learning methods are at our disposal to map between instantaneous timbre space and synthesis space.

The final piece to achieving a system for digital timbre manipulation in composition is then the interface. We have identified several types of interfaces we think could be useful, and believe all of them could be implemented. Future work in this area is needed to map from high-level abstractions and timbre objects to the orthogonal perceptual representation. Tests need to be done to assess the effectiveness of each interface for different purposes.

In all, we believe this system is no longer so far-fetched, and look forward to continue work in this area to achieve it, and to increase our understanding of timbre as a perceptual phenomenon.

REFERENCES

[1] ISO, "226:2003 Normal equal-loudness-level contours", International Organization for Standardization (2003).

[2] C. Plack, A. Oxenham, R. Fay, and A. Popper, *Pitch: Neural Coding and Perception*, volume 24 (Springer New York) (2005).

[3] R. Turner, "The Ohm-Seebeck dispute, Hermann von Helmholtz, and the origins of physiological acoustics", The British Journal for the History of Science **10**, 1–24 (1977).

[4] H. Von Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music* (Longmans, Green, and Company) (1912).

[5] S. McAdams and A. Bregman, "Hearing musical streams", Computer Music Journal 26–60 (1979).

[6] H. Fletcher, "Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure", The Journal of the Acoustical Society of America **6**, 59–69 (1934).

[7] R. Plomp and H. Steeneken, "Effect of phase on the timbre of complex tones", The Journal of the Acoustical Society of America **46**, 409–421 (1969).

[8] R. N. Shepard, "Geometrical approximations to the structure of musical pitch.", Psychological Review **89**, 305 (1982).

[9] H. Fletcher, E. D. Blackham, and R. Stratton, "Quality of piano tones", The Journal of the Acoustical Society of America **34**, 749–761 (1962).

[10] D. Lewis and M. Larsen, "The measurement of masked auditory thresholds.", Journal of Experimental Psychology **27**, 601 (1940).

[11] S. E. Olive, "The preservation of timbre: Microphones, loudspeakers, sound sources and acoustical spaces", in *Audio Engineering Society Conference: 8th International Conference: The Sound of Audio* (Audio Engineering Society) (1990).

[12] W. Torgerson, "Multidimensional scaling: I. Theory and method", Psychometrika **17**, 401–419 (1952).

[13] R. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function. I.", Psychometrika **27**, 125–140 (1962).

[14] R. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function. II.", Psychometrika **27**, 219–246 (1962).

[15] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", Psychometrika **29**, 1–27 (1964).

[16] J. Kruskal, "Nonmetric multidimensional scaling: A numerical method", Psychometrika **29**, 115–129 (1964).

[17] K. Siedenburg and C. Reuter, "Beyond Helmholtz: 150 years of timbral paradigms", in *12th International Conference on Music Perception and Cognition* (2012).

[18] G. Young and A. Householder, "Discussion of a set of points in terms of their mutual distances", Psychometrika **3**, 19–22 (1938).

[19] R. Plomp, "Timbre as a multidimensional attribute of complex tones", Frequency Analysis and Periodicity Detection in Hearing 397–414 (1970).

[20] C. Horan, "Multidimensional scaling: Combining observations when individuals have different perceptual structures", Psychometrika **34**, 139–165 (1969).

[21] J. Carroll and J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition", Psychometrika **35**, 283–319 (1970).

[22] S. Winsberg and G. De Soete, "A latent class approach to fitting the weighted Euclidean model, CLASCAL", Psychometrika **58**, 315–330 (1993).

[23] S. Winsberg and J. D. Carroll, "A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model", Psychometrika **54**, 217–229 (1989).

[24] J. Tenenbaum, V. De Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction", Science **290**, 2319–2323 (2000).

[25] L. Wedin and G. Goude, "Dimension analysis of the perception of instrumental timbre", Scandinavian Journal of Psychology **13**, 228–240 (1972).

[26] J. Miller and E. Carterette, "Perceptual space for musical structures", The Journal of the Acoustical Society of America **58**, 711 (1975).

[27] J. Grey, "Multidimensional perceptual scaling of musical timbres", The Journal of the Acoustical Society of America **61**, 1270–1277 (1977).

[28] S. McAdams, S. Winsberg, S. Donnadieu, G. Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes", Psychological Research **58**, 177–192 (1995).

[29] S. Lakatos, "A common perceptual space for harmonic and percussive timbres", Attention, Perception, & Psychophysics **62**, 1426–1439 (2000).

[30] J. Burgoyne and S. McAdams, "A meta-analysis of timbre perception using nonlinear extensions to CLASCAL", Computer Music Modeling and Retrieval. Sense of Sounds 181–202 (2008).

[31] C. L. Krumhansl, "Why is musical timbre so hard to understand", Structure and Perception of Electroacoustic Sound and Music **9**, 43–53 (1989).

[32] D. Wessel, "Timbre space as a musical control structure", Computer Music Journal 45–52 (1979).

[33] R. Ethington and B. Punch, "SeaWave: A system for musical timbre description", Computer Music Journal 30–39 (1994).

[34] A. Caclin, S. McAdams, B. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones", The Journal of the Acoustical Society of America **118**, 471 (2005).

[35] L. L. Thurstone, "A law of comparative judgment.", Psychological Review **34**, 273 (1927).

[36] W. Thompson and R. Remage, "Rankings from paired comparisons", The Annals of Mathematical Statistics 739–747 (1964).

[37] T. P. Chartier, E. Kreutzer, A. N. Langville, and K. E. Pedings, "Sensitivity and stability of ranking vectors", SIAM Journal on Scientific Computing **33**, 1077–1102 (2011).

[38] W. N. Colley, "Colley's bias free college football ranking method: The Colley matrix explained", Princeton University (2002).

[39] I. Ali, W. D. Cook, and M. Kress, "On the minimum violations ranking of a tournament", Management Science **32**, 660–672 (1986).

[40] N. Alon, "Ranking tournaments", SIAM Journal on Discrete Mathematics **20**, 137–142 (2006).

[41] C. A. Nicol, "Development and exploration of a timbre space representation of audio", Ph.D. thesis, University of Glasgow (2005).

[42] J. M. Chowning, "The synthesis of complex audio spectra by means of frequency modulation", Computer Music Journal 46–54 (1977).

[43] D. D. Nolte, "The tangled tale of phase space", Physics Today **63**, 32 (2010).

[44] L. H. Sasaki and K. C. Smith, "A simple data reduction scheme for additive synthesis", Computer Music Journal **4**, 22–24 (1980).

[45] K. W. Schindler, "Dynamic timbre control for real-time digital synthesis", Computer Music Journal 28–42 (1984).

[46] J. Burred, A. Robel, and T. Sikora, "Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds", Audio, Speech, and Language Processing, IEEE Transactions on **18**, 663–674 (2010).

[47] A. Hunt and M. M. Wanderley, "Mapping performer parameters to synthesis engines", Organised Sound **7**, 97–108 (2002).

[48] I. Choi, R. Bargar, and C. Goudeseune, "A manifold interface for a high dimensional control space", in *Proceedings of the 1995 International Computer Music Conference*, 385–392 (1995).

[49] D. Wessel and M. Wright, "Problems and prospects for intimate musical control of computers", Computer Music Journal **26**, 11–22 (2002).

[50] C. Goudeseune, "Interpolated mappings for musical instruments", Organised Sound **7**, 85–96 (2002).

[51] L. Fritts, "University of Iowa electronic music studios", (2011), URL http://theremin.music.uiowa.edu/MIS.html.

BIOGRAPHICAL SKETCH

Brandon Smock was born in 1983 in West Palm Beach, Florida. In 2006, he graduated *cum laude* from the University of Florida with a Bachelor of Science degree in computer science. In 2008, he earned his Master of Science in computer engineering from the University of Florida. In 2014, he earned his Doctor of Philosophy in computer engineering from the University of Florida. His doctoral dissertation, *A Perception-Centric Framework for Digital Timbre Manipulation in Music Composition*, was done under the supervision of Dr. Anand Rangarajan.

During his time as a doctoral student, he worked as a research assistant in machine learning in the Computational Science and Intelligence laboratory under the supervision of Drs. Paul Gader and Joseph Wilson. He twice attended the IEEE Geoscience and Remote Sensing Symposium, contributing proceedings papers entitled *Reciprocal Pointer Chains for Identifying Layer Boundaries in Ground-Penetrating Radar Data* and *Optimal Fusion of Alarm Sets from Multiple Detectors Using Dynamic Programming*. He also served as a teaching assistant for the CISE department's course on discrete mathematics.

His post-graduation plans include pursuing a career as a research scientist or starting his own company. He intends to accept a position as a post-doc in the Computational Science and Intelligence lab at the University of Florida to continue his research in the area of optimal detector fusion. In the future, he envisions continuing his research in psychoacoustics and digital synthesis of timbre, and expanding his research in areas related to optimal automated decision-making.