UNIVERSITY OF CALIFORNIA, SAN DIEGO

Physical and Perceptual Aspects of Percussive Timbre

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Music

by

William Brent

Committee in charge:

Professor Miller Puckette, Chair
Professor David Borgo
Professor Diana Deutsch
Professor Shlomo Dubnov
Professor Shahrokh Yadegari

2010

The dissertation of William Brent is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2010

DEDICATION

For my father.

EPIGRAPH

One of the most striking paradoxes concerning timbre is that when we knew less about it, it didn't pose much of a problem.

—Philippe Manoury

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

In an ideal academic community, guidance in unfamiliar areas is always within reach, and I am grateful to several people for offering me the benefit of their expertise. First, I would like to thank Miller Puckette for enabling my current research path from the beginning, and for being so consistently approachable, patient, and helpful in getting me started. Without his guidance and the tools and resources he has made freely available, the process of learning synthesis, programming, and signal processing would have been tremendously more difficult. Within this same area, I am also grateful to Dick Moore for elucidating the finer points of DSP, Kevin Larke for sharing his time and substantial programming experience, and Jaime Oliver, with whom I began learning the C programming language. Many of the machine learning techniques implemented in my software were the result of a course on data mining by Shlomo Dubnov. Professor Dubnov also taught a course on music in MATLAB that sparked my interest in cepstral analysis—the starting point for all of my work on signal analysis. I am thankful for his contributions in these courses as well as in our personal meetings.

The perceptual experiment reported in Chapter 6 was improved under the guidance of Diana Deutsch. Her rigor, experience, and numerous examples in the music perception literature encouraged my first steps in this area. I also received the benefit of consulting with Carol Krumhansl on several occasions. To Carol, I am grateful for assistance in the area of data analysis, but also for volunteering to be my first pilot subject. Other colleagues that offered their time and support in mathematical matters include Scott DeWolf, Eric Ward, and Kristin Marshall.

Though my interactions with David Borgo and Shahrokh Yadegari were mostly centered on research concerning gesture in the performance of live computer music, this dissertation was certainly informed by our discussions in that area. Without the line of thinking that they encouraged, I would not have come to appreciate how the complex relationships between performance gestures and timbre have a direct bearing on our understanding of these musical phenomena.

My long-held interest in contemporary percussion practice is evident in the subject of this thesis. It is an extraordinary stroke of luck that the music department of UCSD hosts major figures in both computer music and percussion.

Steven Schick has actively furthered my interests in several important respects—from supporting the construction of a robotic percussion system I built in 2008, to creating performance opportunities in which I could test my developing software. A recent graduate of the percussion department, Ross Karre, has also been consistently supportive of my work, offering time, expertise, and constructive discussions on countless occasions. For recording the vast majority of percussion samples that are analyzed in this dissertation, I am thankful to percussionist Steve Solook, a graduate student and member of UCSD's resident percussion ensemble, *RedFish-BlueFish*. I also owe thanks to the remaining members of the ensemble—Justin DeHart, Bonnie Whiting Smith, Brian Archinal, and Dustin Donahue—and computer musician/percussionist Michelle Daniels, all of whom generously agreed to participate in the experiment documented in Chapter 6.

Finally, I owe thanks to my wife, Rika, and two children, Mizuki and Sen. Without their support, my accomplishments over the past five years simply would not have been possible.

Portions of Chapters 3 and 4 were previously published in the following articles:

William Brent, "A Timbre Analysis and Classification Toolkit for Pure Data", *Proceedings of the International Computer Music Conference*, 2010.

William Brent, "Cepstral Analysis Tools for Percussive Timbre Identification", *Proceedings of the 3rd Pure Data Convention, São Paulo, Brazil*, 2009.

William Brent, "Perceptually Based Pitch Scales in Cepstral Techniques for Percussive Timbre Identification", *Proceedings of the International Computer Music Conference*, 2009.

VITA

2001	B. A. in Composition, Wilfrid Laurier University, Waterloo, Ontario, Canada
2003	M. A. in Composition, Mills College, Oakland, California
2005-2010	Graduate Teaching Assistant & Research Assistant, University of California, San Diego
2010	Ph. D. in Music, University of California, San Diego

PUBLICATIONS

William Brent, "A Timbre Analysis and Classification Toolkit for Pure Data", *Proceedings of the International Computer Music Conference*, 2010.

William Brent, "Cepstral Analysis Tools for Percussive Timbre Identification", *Proceedings of the 3rd Pure Data Convention, São Paulo, Brazil*, 2009.

William Brent, "Perceptually Based Pitch Scales in Cepstral Techniques for Percussive Timbre Identification", *Proceedings of the International Computer Music Conference*, 2009.

ABSTRACT OF THE DISSERTATION

Physical and Perceptual Aspects of Percussive Timbre

by

William Brent

Doctor of Philosophy in Music

University of California, San Diego, 2010

Professor Miller Puckette, Chair

This dissertation explores relationships between perceptual dimensions of percussive timbres and measurements produced by several signal analysis algorithms. The literature of psychophysical timbre experiments since 1941 is reviewed with respect to two contrasting approaches. The earliest attempts at unraveling the interdependent aspects of timbre perception employed multiple adjective scales intended to describe various sonic features. Following developments in the technique of multidimensional scaling (MDS) in the 1960s, several researchers began to apply scaling techniques to data sets of timbre similarity judgments. At present, the majority of timbre studies are based on MDS. In spite of such advancements, the range of musical timbres has only begun to be explored from a perceptual viewpoint, and a significant gap exists in the literature for percussive instruments.

The signal analysis algorithms employed in this research are introduced in the context of timbreID—a timbre analysis software library written by the author. The library's adaptability is illustrated with respect to several musical research applications in Pure data. This flexibility is shown to be beneficial in the case of two percussive instrument classification tests, in which the effectiveness of perceptually

weighted spectral features like mel- and Bark-frequency cepstrum are evaluated alongside other standard analysis techniques from the music information retrieval literature.

In the final chapter, a perceptual experiment involving 30 diverse percussion timbres is carried out. The study confirms the importance of spectral centroid and attack duration as predictors of perceptual dimensions, and reveals two additional dimensions that may be unique to percussive timbres: "dryness" and "noisiness". A predictive model is generated using multiple linear regression, and results indicate that the noisiness dimension cannot be predicted as accurately as dimensions relating to spectral center of gravity and attack time. Thus, there is a clear need for an effective measure of perceptual noisiness for accurate description of percussive timbre.

Chapter 1

Introduction

Defining the musical attribute of timbre has proven a consistently problematic pursuit for more than a century. Part of the difficulty arises from the fact that timbre is tied to the physical source of a sound (implying complex multi-modal associations), but is also used to refer to any number of abstract qualities that are purely sonic. Motivated by possibilities in the realm of audio synthesis, it is the latter aspect that researchers have investigated in the most detail. Thus, timbre is frequently referred to as a multidimensional characteristic, and salient aspects have been identified that are at once spectral and temporal in nature. The discipline of sound design stands to benefit considerably from the discovery of additional perceptual dimensions that can be translated to reliable synthesis parameters. But the referential nature of timbre must not be ignored, even in the case of completely invented sounds. The ability to identify sound sources is fundamentally important to our survival, and knowledge that a sound is artificial does not necessarily bring this mechanism to a halt. Surely, the context of a musical performance is a special case, but it is doubtful that we could ever completely suppress the tendency to evaluate sounds around us in terms of what they may portend. With this in mind, it seems that exploring the makeup of timbre perception from a purely sonic point of view will never result in a complete picture. Thus, while the "multidimensional" label is appropriate and useful in many respects, it must be remembered that timbre might never be reliably represented numerically, even in high-dimensional spaces. The precedents set by research relating amplitude to loudness, and frequency to

pitch—connections that are not always as clear-cut as we assume they ought to be—by no means guarantee that such strong relationships can be found between what we are able to perceive and mechanically measure of sound quality.

As an alternative to "multidimensional", the term "emergent" is adopted by Stephen Handel in his writings on timbre [Han95]. The descriptor is used frequently in artificial life studies to refer to behavior that arises from the complex interaction of multitudes of simple objects. The classical artificial life example is Craig Reynolds' BOIDS algorithm that has been used to simulate flocking behavior in animation since the 1980s [Rey87]. "Emergence" is also used by Frank Sibley in an attempt to explain the relationship between aesthetic and non-aesthetic properties of art [Sib65]. When called upon to explain why we enjoy a particular piece of music, we might point to several non-aesthetic properties, such as details of its formal proportions, melodic contours, and rhythmic patterns. But such attributes only begin to explain our aesthetic experience. Different pieces with similar non-aesthetic properties could elicit radically divergent aesthetic responses. In short, it can be said that emergence describes phenomena in which the whole is greater than the sum of its parts. It also reserves the possibility that the phenomenon to which it is applied might never be fully explained or understood. In this regard, it is entirely appropriate to describe timbre as an emergent property of sound.

Although our understanding of timbre must extend beyond quantification, timbre research from the past several decades is clearly skewed toward establishing numerical representations. Those who pursue this goal are driven by the considerable benefits associated with automation, generalization, and stability. With the quantification of timbre properties, a number of possibilities open up. Vast databases of sound can be analyzed completely automatically so that composers and sound designers can easily retrieve sound materials with desired qualities. Several projects of this sort are in development, with significant levels of success [TC99][DAC07b]. An additional benefit of quantification in this area is long-term stability. Large sound databases, like the BBC sample library, are often tagged with several types of meta-data that verbally describe audio content. Though very useful in search and retrieval tasks, such data is time-consuming to produce, re-

quires translation into different languages, and must be maintained over time as the language we use to describe sound evolves.

In the realm of composition, theorists have put forward the notion of timbre hierarchies that can be used for large-scale musical organization of the type normally associated with pitch and rhythm [Ler87]. Timbre has long been used to clarify simultaneous streams of melodic constructs, and since the late 19^{th} century it has been used in orchestral settings with increasing sophistication [Bou87]. However, the idea of music utilizing timbre as a structural parameter in its own right is still quite new. This possibility is supported by experiments that have begun to establish the stability of timbre interval perception [Wes79][MC92]. In these studies, it was shown that participants were able to consistently judge both the magnitude and direction of changes along certain timbre dimensions. Such abilities may not be as reliable as those in the domains of pitch and rhythm perception, but it is conceivable that timbre sequences and inversions could be used as effective compositional devices. Lerdahl proposes the idea of timbral dissonance, consonance, and stability, and asserts that the reason timbre is not commonly used as a principal bearer of compositional form is that "unlike pitch and rhythm, it has lacked any substantial hierarchical organization." [Ler87, p. 138] In reference to percussion, Boulez confirms the need for such organization: "the percussion section, for example, shows the most visible recent transformation of the body of the orchestra ... but these instruments don't obey the hierarchy to which the others belong and so a certain number of them is necessary to create another hierarchy based essentially on timbre." [Bou87, p. 165] The quantification of timbre makes it possible to achieve timbre organization and manipulation in a controlled and repeatable fashion.

For performers—and particularly, percussionists—it is not uncommon to be faced with a piece of contemporary music that leaves instrumentation relatively unspecified. A range of instrumental freedom exists across pieces like Xenakis' *Psappha*, Ferneyhough's *Bone Alphabet*, Feldman's *King of Denmark*, and Cage's *27' 10.554"*. In *King of Denmark*, the percussionist is called upon not only to select instruments, but also to organize them into coherent sequences. It is ar-

guable that such tasks are best accomplished based solely on the intuition of the performer; however, software that is capable of automatically organizing sound sets can provide a useful starting point, and bring up interesting relationships that may not be immediately apparent. In other percussion repertoire, instruments are very strictly specified, and text-based instrument descriptions do not always suffice. For example, not all 18" cymbals sound alike. Stockhausen's music is perhaps the most notorious example of highly specified instrumentation, for which certain percussion instruments (like the Glissentrommel) are manufactured specifically. In such cases, the ability to quantify timbre characteristics can facilitate a level of consistency that would otherwise be very difficult to obtain.

It is presently possible to implement many techniques for measuring timbre characteristics in real time. In both composed and improvised performance projects, this broadens the palette of available options for controlling aspects of live electroacoustic accompaniment. Musical events can be triggered or manipulated based on a performer's subtle control over timbre as well as pitch and loudness. All of the ideas above are made possible because some aspects of timbre clearly *are* strongly related to quantifiable measurements, such as spectral envelope, spectral centroid, and attack time. With these issues in mind, it is not difficult to understand why the majority of research efforts are directed at the quantification of timbre.

This dissertation explores the relationship between perceptual dimensions of percussive timbres and measurements produced by several signal analysis algorithms. In the second chapter, the literature of psychophysical timbre experiments since 1941 is reviewed with respect to two contrasting approaches. The earliest attempts at unraveling the interdependent aspects of timbre perception employed multiple adjective scales intended to describe various sonic features. Following developments in the technique of multidimensional scaling (MDS) in the 1960s, several researchers began to apply scaling techniques to data sets of timbre similarity judgments. At present, the majority of timbre studies are based on MDS. Across both branches of research, at least two perceptual dimensions have been repeatedly discovered in independent experiments. The first relates to spectral

center of gravity, while the second is centered on attack characteristics. A third dimension relating to spectral changes over time has also surfaced, though its interpretation has not been nearly as well defined as the former dimensions. In spite of such advancements, the range of musical timbres has only begun to be explored from a perceptual viewpoint, and a significant gap exists in the literature for percussive instruments.

Chapter 3 presents several approaches to timbre quantification. Many of the techniques attempt to measure a single timbre characteristic, such as brightness, and produce a single numeric value. This process is referred to as feature extraction. Certain features are expressed as vectors in a multidimensional geometric space. In fact, spectral data itself is a feature vector with a dimensionality dependent on analysis window size. A typical number of dimensions for such information is 1024. Other feature vectors have been devised to reduce this dimensionality in order to make it more manageable. The prevailing features of this sort are the cepstrum and mel-frequency cepstrum, whose components are referred to as MFCCs. MFCCs have been extremely important in the field of automatic speech recognition, and their application to music is relatively recent [Log00].

The fourth chapter introduces a timbre analysis software library written by the author. It is illustrated in the context of several musical research applications in Pure data. The flexible analysis strategies made possible by the analysis package are shown to be beneficial in the case of two percussive instrument classification tests presented in the fifth chapter. In these tests, the effects of perceptually weighted features like mel- and Bark-frequency cepstrum are evaluated alongside other standard analysis techniques from the music information retrieval literature.

In the final chapter, a perceptual experiment involving 30 diverse percussion timbres is carried out. The study confirms the importance of spectral centroid and attack duration as predictors of perceptual dimensions, and reveals two additional dimensions that may be unique to percussive timbres: “dryness” and “noisiness”. A predictive model is generated using multiple linear regression, and results indicate that the noisiness dimension cannot be predicted as accurately as dimensions relating to spectral center of gravity and attack time. Thus, there is a clear need

for an effective measure of perceptual noisiness for percussive timbre description.

Chapter 2

Historical Overview of Timbre Studies

2.1 Experimental Design

There are two overlapping branches of psychophysical research devoted to elucidating the process of timbre perception, reflecting the dual definition of timbre noted by Handel, Hadja, and Donnadieu [Han95][Had07][Don07]. As an attribute of acoustic phenomena, "timbre" is associated with the physical source of a sound; yet it is also a catchall term that refers to all perceptual characteristics (other than pitch, loudness, and duration) of sound considered in the abstract. In connection with the former understanding of the term, the first branch of research consists of experiments based on classification tasks (as in [SF64] [Ber64], and [WG72] [CH78]), where participants' ability to recognize instrument sounds is measured with respect to various alterations of the stimuli. Berger, for instance, found that recognition accuracy of wind instrument recordings was adversely affected when using stimuli in which the attack and decay segments were removed [Ber64, p. 1890]. Wedin & Goude reported similar consequences for removing attacks, though the severity of classification impairment varied for different instruments [WG72, p. 232]. The effect of fundamental frequency on classification was explored in [HE01], where it was found that instrument recognition was severely

hampered when stimuli were separated in pitch by an octave or more. At the most basic level, studies of this sort provide a measure of our ability to accurately connect abstract sounds with their physical sources. However, the practice of systematically altering stimuli creates the potential for identifying specific aspects of perception that contribute to such categorical judgments in the first place. In this way, classification-based research could point out sonic cues that explain how we are able to link the diverse set of sounds that a violin creates with a single categorical label.

Referred to as "relational studies" [Had07, p. 251], a second branch of research aims to identify the constituent sonic parameters from which timbre perception emerges. Relational studies can be further divided according to two distinct approaches. Verbal attribute-based relational studies record judgments about a set of sounds relative to a collection of words deemed appropriate for describing timbre. This approach is exemplified by von Bismarck's thorough experiment from 1974, where groups of participants rated 35 sounds on 30 unique adjective scales, such as "dull—sharp" and "dark—bright" [vB74b]. The participants' ratings were then further analyzed in order to identify the most salient adjective scales, and acoustically measurable sonic features that correlated highly with these scales were sought. Critics of this approach will be quick to point out an obvious weakness, which is that the verbal attribute scales are prescribed by the experimenter, unjustifiably limiting the information collected from participants to be relative to particular aspects of timbre that are chosen in advance. This precludes the discovery of unanticipated relevant features of timbre.

Many consider multidimensional scaling (MDS) algorithms to be an ideal way to combat this type of bias. The second category of relational studies use different varieties of MDS algorithms to interpret timbre similarity judgments made by participants relative to several pairs of timbres played in sequence. Though it was not the first to employ MDS, John Grey's 1975 study has come to exemplify this experimental model [Gre75]. In Grey's experiment, a set of sixteen different instrument sounds playing an E-flat above middle C were arranged into all possible pairwise combinations. Participants listened to these pairs and rated their tim-

bral similarity. Results of multidimensional scaling returned a three-dimensional solution that accounted for most of the data's variance. Unlike the results of a factor analysis performed on von Bismarck's multiple adjective scales, MDS does not give any indication as to the properties of sounds distributed along any particular axis. MDS usually produces 2-4 dimensions that explain a great deal of the similarity data's variance; however, these dimensions must be interpreted after the fact by the experimenter. The three-dimensional MDS solution to Grey's data was interpreted as relating to three sonic characteristics. Along the first dimension, instruments were distributed according to spectral energy distribution. The second dimension was taken to be related to the synchrony of attack and decay times of upper partials, and the third dimension mapped sounds according to the presence of low-amplitude inharmonic upper partials during the attack.

The advantages and disadvantages of verbal attribute- and MDS-based studies are complementary. With both approaches, data reduction techniques—including factor analysis, principal component analysis, and varieties of MDS algorithms like MDSCAL, INDSCAL, and CLASCAL—are used in order to enable more intuitive interpretation of complex data. Attribute-based experiments make assumptions about the nature of timbre perception that threaten to push research results in particular directions. MDS can be a corrective to this risk, although it is important to recognize that interpretation is still required after the fact. In terms of the size of the sound set being tested (N), the MDS approach is very restrictive because it requires that participants evaluate all possible pairs in a set. If $N = 12$, there are $N(N-1) = 132$ possible pairings to listen to. With even a modestly larger set of $N = 20$, there are 380 pairs. The adjective scale approach is much less limited in this regard.

This chapter will review some of the most significant relational timbre studies that have followed these two experimental models. Many researchers have noted a clear dominance of the MDS approach [Lak00][CMSW05]. Keeping the above limitations in mind, there are nevertheless some consistencies emerging that call for further experimentation using both strategies.

2.2 Verbal Attribute Studies

The earliest studies to systematically explore connections between verbal attributes and aspects of tone quality or timbre were [Lic41], [Sol58], and [Sol59]. Motivated by the extreme imbalance of understanding between timbre and other sonic attributes like loudness and pitch, Lichte carried out a large scale investigation of the tone quality of different classes of complex tones. The spectra of the tones were designed synthetically using a tone generator capable of producing up to 16 partials. 255 participants rated pairs of tones from each of the tone classes with respect to the terms "dull", "bright", "thin", "full", "smooth", and "rough". Results indicated that there is a basis for "brightness", "roughness", and "fullness" as independent attributes of tone quality.

Following Lichte, Solomon studied 50 common verbal attributes applied to sonar recordings. His project was unique in that it used non-synthetic sound stimuli, and the vocabulary used to create attribute scales was well established semantically within the community of sonar experts that served as participants. Rather than having participants rate one element of a stimulus pair as "duller" or "brighter" than the other (as in [Lic41]), Solomon employed Osgood's concept of a semantic differential (SD) scale [OST57], where attributes are paired with their semantic opposites to form scales such as "heavy—light" and "smooth—rough". 50 participants rated the sounds individually on the collection of seven-point SD scales, and results were highly consistent across participants. Factor analysis of the data revealed seven important factors that accounted for 42% of the variance. The primary factor was interpreted as a "magnitude" dimension, with the largest loadings coming from the scales "heavy—light", "large—small", and "rumbling—whining".

2.2.1 Von Bismarck

Because it was the most comprehensive work of its time, the details of von Bismarck's study of 35 timbres must be described in more detail than given above. Von Bismarck sought to correct three problems he identified in the early

studies just mentioned. First, processes for choosing verbal attributes were not systematic. Second, the sound stimuli were not normalized for pitch, loudness, and time structure. Last, the studies did not offer many connections between the SD ratings and measurable signal characteristics.

With regard to the selection of verbal attributes, von Bismarck began with a list of 69 adjectives drawn from previous studies, and had participants rate the appropriateness of each term on a seven-point scale ranging from "very unsuitable" to "highly suitable". In von Bismarck's opinion, participants' ratings were generally consistent, and the scales with the highest mean scores were compiled to form a list of 28 SD scales. Scales for "soft—loud" and "low–high" were added to test the effectiveness of pitch and loudness normalization. The final list of 30 scales is given in Table 2.1.

Table 2.1: Thirty SD scales used in von Bismarck's 1974 experiment.

soft—loud	wide—tight
weak—strong	thick—thin
gentle—violent	clean—dirty
fine—coarse	full—empty
reserved—obtrusive	solid—hollow
low—high	colorful—colorless
soft—hard	pure—mixed
dim—brilliant	simple—complex
relaxed—tense	compact—scattered
calm—restless	interesting—boring
rounded—angular	lively—dead
dampened—ringing	pleasant—unpleasant
smooth—rough	open—closed
heavy—light	dark—bright
broad—narrow	dull—sharp

The sound set consisted of 35 synthetically produced steady-state timbres,

many of them based on vowels from the German language. They were generally divided into "noise" and "tone" categories, with tonal timbres equalized in pitch at 200 Hz. Using a chosen sound as a reference, all stimuli were equalized in loudness. Participants were able to audition the entire set for context before making any judgments on the SD scales. Two strategies were informally attempted for collecting judgments. The first presented the sound set repeatedly in various random orders, with participants rating only one SD scale per sound for each repetition. The second presented the entire sound set only once, with each sound repeating 30 times to allow participants to rate it on every SD scale. No major difference in consistency was found between the two strategies, so the official experiment was carried out using the second approach, as it was preferred by most participants. Participants were separated into two groups according to levels of musical training.

A factor analysis of the SD ratings revealed four factors that explained more than 80% of the variance. The "stumpf—scharf", or "dull—sharp" scale was consistently connected with the first factor in all cases. There was no significant difference between the ratings of musicians and non-musicians on this scale. It must be noted that in English, usage of "sharp" in a musical context is connected with pitch. "Dull—bright" may therefore be a more appropriate translation. Other relevant scales for this sound set were "compact—scattered", "full—empty", and "colorful—colorless". Contrary to expectation, the control scales for "soft—loud" and "low—high" were highly correlated with other scales, in spite of the fact that participants were instructed to ignore pitch and loudness as much as possible. Potential explanations for this unexpected result are offered, focusing on the possibility of participants misinterpreting the semantic meaning of the scales.

The multivalence of SD scales is definitely problematic. However, if scales are chosen directly by the community of musicians participating in an experiment and their judgments appear to be consistent, verbal attributes—whose meanings are constantly evolving within languages that constantly evolve themselves—can be mined for stable information. Regarding a physical correlate to the most important scale in his study, von Bismarck noted that "sharpness formation appears to be characterized by the combined effects of (1) the position of energized spec-

tral regions and (2) the magnitudes of energy in those regions" [vB74a, p. 169]. These parameters are tied to the measurement of spectral centroid. Several studies have identified either "brightness" or spectral centroid as important factors in the perception of timbre [Lic41][Gre75][PD76][IK93][Lak00][CMSW05]. Von Bismarck's subsequent article [vB74a] carefully examines the independence of sharpness (brightness) from pitch or loudness. The logic is that any primary dimension of timbre should be perceptually separable from pitch or loudness. Bismarck's study confirms this hypothesis, and further demonstrates that a quantitative doubling or halving of sharpness by spectral design produces proportional perceptual judgments [vB74a, p. 162].

2.2.2 Kendall & Carterette

Von Bismarck's study improved upon prior work considerably, but certain flaws remained. For example, the 69 SD scales used as a starting point were taken from quite varied timbre studies that investigated not only musical sounds, but speech and sonar events as well. Experiments originating from non-musical disciplines are built on assumptions that may or may not be appropriate for musical contexts. Further, participants in von Bismarck's experiment were not given the advantage of listening to his sound set before evaluating the appropriateness of the proposed SD scales. A more fundamental issue is the appropriateness of semantic differential scales themselves: certain attributes, like "brightness", have more than one possible antonym (e.g., "dullness" and "darkness"). Finally, though it is not necessarily a flaw, von Bismarck's exclusive use of static synthesized timbres situates the realm of influence of his results rather far from temporally evolving instrumental sounds.

[KC93a] and [KC93b] are verbal attribute-based studies of combined wind instrument timbres that address the issues above. Rather than looking only at previous semantic differential timbre studies, Kendall and Carterette built their list of candidate adjectives directly from a musically-oriented source: Walter Piston's *Orchestration*. All of the adjectives Piston used in reference to timbre were extracted from the book and the list was edited by the authors to eliminate re-

dundancies, as well as terms that were too strongly tied to a manner of playing or articulation (e.g., "plucked"). This resulted in a list of 61 adjectives that was to be further reduced by a formal experiment.

The adjective selection experiment utilized a small participant pool that was composed entirely of professional musicians. A subset of the full collection of woodwind timbres was chosen based on a previous experiment that enabled a perceptually-based mapping of the sounds in a three-dimensional space. The most dissimilar sounds were chosen for the subset, and participants heard them repeatedly while checking off the most appropriate adjectives from the full list of 61. Adjectives receiving two or more checks were preserved. The final list of 21 adjectives is shown in Table 2.2.

Table 2.2: Twenty-one adjectives used in Kendall & Carterette's 1993 experiment.

brilliant	brittle	crisp	edgy	full	fused	light
mellow	nasal	reedy	resonant	rich	ringing	round
smooth	soft	strong	tremulous	tense	warm	weak

After the SD scales used in an initial experiment were not successfully used by participants to distinguish between instrumental timbres [KC93a], Kendall and Carterette concluded that bipolar scales were problematic. A second experiment was carried out using verbal attribute magnitude estimates (VAME), i.e., single adjective scales. For example, rather than choosing a value along a scale from "dark—bright", participants simply rated the degree of "brightness" that a sound possessed. The follow-up experiment successfully differentiated timbres based on these VAME ratings. Thus, this approach was adopted for the more carefully selected set of musically-oriented adjectives in [KC93b].

A principal components analysis of VAME ratings produced four factors explaining 86% of the data variance. Based on the VAME scales associated with these factors, four dimensions were identified that were interpreted as related to the terms "power" (summarizing "smooth", "soft", "light", "weak", and "mellow"), "strident" (summarizing "strong", "tense", and "tremulous"), "plangent"

(summarizing "ringing", "resonant", "crisp", and "brilliant"), and "reed" (summarizing "reedy", "fused", and "warm"). Connections between these dimensions and acoustical analyses were tentatively offered, including spectral energy distribution for the "power" dimension, and spectral flux for the "strident" dimension. As a preemptive response to criticism of these relatively inconclusive results, the authors point out that decisions surrounding the interpretation of MDS-derived dimensions are often made "arbitrarily, with little basis other than intuition, and subject to biases of expectation." [KC93b, p. 495]. Though the conclusion is less satisfying, the caution Kendall and Carterette show is laudable: "The danger of reification of meaning of the dimensions of MDS configurations is real. The attribute or factor chosen to represent a dimension, such as *nasal* or *rich*, is only the most salient feature in a fuzzy set of features." [KC93b, p. 496].

2.2.3 Freed

A unique study by Daniel Freed focused on perceived mallet hardness (PMH) in percussive timbres [Fre90]. Like Kendall and Carterette, Freed's experiment involved rating actual instrument timbres—in this case metal pans, which are frequently used by contemporary percussionists as instruments. Where Kendall and Carterette specifically avoided adjectives related to the mechanical production of sound (e.g., "plucked"), Freed chose to make a single term of this type the main object of his study. According to Freed, "PMH is the timbral property that evokes an 'image' of a specific degree of mallet hardness." [Fre90, p. 311] Such an adjective is quite different in nature from those studied by von Bismarck or Kendall and Carterette. A potential criticism is that mallet hardness is only a secondary timbral descriptor; however, its timbral consequences are so frequently engaged with by percussionists that any findings in this area are clearly valuable. Generally, higher mallet hardness is known to increase the high frequency content of percussive sounds. The term "brightness" could perhaps have been used instead, but "hardness" is less susceptible to multiple interpretations, and the intuitive understanding of a physical property like mallet hardness is worth exploring separately. Making reference to [Gib66], Freed sees this as a move toward a more "ecological"

approach to perception.

As opposed to the static timbres used in [vB74b] or the complete wind instrument tones from [KC93a], acoustic analyses in Freed's experiment were limited to the attack portion of sounds—in this case, the first 325 milliseconds. Four types of analyses were chosen for investigation based on their expected correlation with physical mallet hardness: the mean and slope of the spectral level curve (tracking the area beneath the spectral envelope as it changes over time), and the mean and time-weighted average of the spectral centroid curve. Significantly, these spectral measurements were not performed on raw magnitude spectra, but spectra obtained by applying a transform designed to model the human auditory system by J. P. Stautner [Sta83]. The transform returns a spectrum indicating energy in the different critical bands. Subsequently, each band was emphasized according to the A-weighted decibel curve [Fre90, p. 314].

Nine musically trained participants rated 96 recorded sounds of four different metal pans that were each struck with 6 different types of mallets. Arranged from soft to hard, the mallet head materials used were: felt-covered rubber, felt, cloth-covered wood, rubber, wood, and metal. Participants used a nine-point rating scale ranging from "very soft" to "very hard". Results are plotted in Figure 2.1, reproduced from [Fre90].

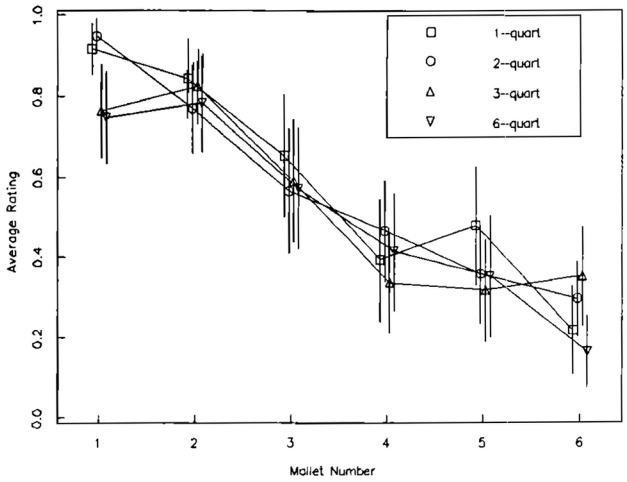


Figure 2.1: Participants' PMH ratings as a function of mallet identity.

What can be seen is that participants were able to identify mallet types based on PMH independently of pan types. The four acoustic predictors were evaluated against this data using multiple regression analysis, showing all predictors to be effective in combination. Rated individually, spectral centroid mean (in Barks) was the most effective predictor, confirming the intuitive connection between PMH and the verbal attribute "brightness".

2.3 Multidimensional Scaling

The earliest uses of multidimensional scaling in timbre similarity research are documented in [WG72] [Wes73] [Gre75], and [Plo76] following major developments in the MDS technique by Kruskal in 1964 [Kru64]. Throughout [WG72], the dimension reduction technique is alternately referred to as "multidimensional scaling", "multidimensional similarity analysis", and "principal component factor analysis", but the primary distinction is that a dimension reduction method was applied to *similarity* judgments, not multiple verbal attribute ratings. The primary goal of Wedin & Goude's study was to find correlations between what they identify as the "acoustical" and "psychological" definitions of timbre; however, it also investigates the influence of attack and decay segments of sounds, which makes its findings relevant in terms of timbre classification as well.

The stimulus set used was small, consisting of recordings of nine Western orchestral instrument timbres: flute, bassoon, violin, oboe, French horn, trumpet, trombone, clarinet, and cello. In terms of pitch and loudness, the tones shared a common fundamental of 440 Hz and were played at a "mezzo-forte" dynamic. All tones maintain 3 seconds of steady state articulation, and vibrato was allowed for violin, cello, flute, and oboe [WG72, p. 230]. Two versions of the sound set recording were prepared: one unaltered, the other with the attack and decay of each tone removed.

Four phases of experimental procedure were established and carried out on 70 participants divided into two roughly equal groups. The second group was exposed to the altered stimulus set with attacks and decays removed. In the first

experimental phase, participants listened to each tone in the set twice, then were asked to identify the instrument that produced it. Next, participants heard all possible pairs in the set, and were asked to make a similarity judgment for each on a 0 to 10 scale. A questionnaire collecting information about musical training was answered in the third phase. The final phase took place on a different day, and participants were asked to rate the similarity of the instruments used in the study, this time referred to by name only (e.g., "violin" and "cello"). The rationale for the fourth phase was to collect information about what Wedin & Goude refer to as the "cognitive structure" of each sound, which is constructed based on information that can obtained about a sound without necessarily hearing it. It was investigated in relation to the "perceptual structures" of the sounds, which consist of direct experiential information.

Data from the similarity portion of the experiment were reasonably consistent across participants, and a perceptual space was generated from the resulting matrix. For both the unaltered and transient-removed stimulus sets, three dimensional spaces were found that explained about 75% of the variance. The positions of instrument timbres in these spaces did not correspond with the "cognitive" groupings obtained from the fourth experimental phase. Violin, cello, and clarinet were clustered on the first dimension, trombone, French horn and flute were grouped on the second dimension, and trumpet, oboe, and bassoon were similar along the third dimension. Thus, the cognition-based fact that the clarinet and flute are both woodwind instruments did not predispose participants to rate the timbres as similar.

The remarkable finding from this early and limited study is that spectral energy distributions of stimuli were very clearly connected with their positions along the three dimensions produced from MDS. In Figure 2.1, it can be seen from casual inspection that the first column of spectra have consistently strong upper harmonics relative to the fundamental, the second column of spectra possess gradually decreasing upper harmonics (or a decreasing spectral slope), and spectra in the third column have a strong formant region and relatively weak fundamental strength.

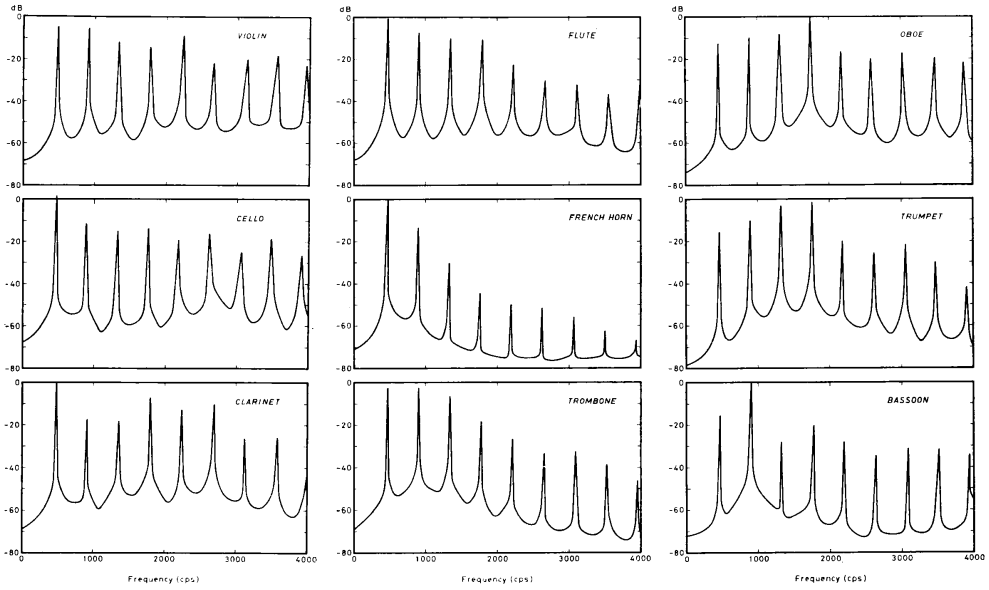


Figure 2.2: Spectra for the steady state of 9 instrument tones used in [WG72].

Wedin & Goude are modest about their conclusions, noting the "rather limited stimulus sample, and the partly imperfect control of the technical-acoustical conditions" [WG72, p. 240], but present three acoustic correlates to the perceptual space dimensions. The first dimension relates to the amount of "overtone richness", or relatively high amplitude harmonics possessed by a tone. The second dimension reflects "overtone poorness", or decreasing spectral slope. The third dimension represents tones with a low fundamental amplitude and a region of stronger upper harmonics.

Through linear regression analysis, Wedin & Goude also attempted to clarify how individual harmonics related to each axis of the perceptual space. It was found that "the highest frequency, the fundamental frequency and the middle range frequencies have the highest predictive power." [WG72, p. 238]. This information could be helpful in reducing the number of calculations required for automated timbre comparisons.

2.3.1 Grey

Though efforts toward establishing a perceptual timbre space via MDS by Wedin & Goude preceded John Grey's 1975 dissertation, *An Exploration of Musical Timbre* [Gre75], Grey's work has come to exemplify this experimental model. The single most significant reason for this was the unique nature of Grey's sound stimuli. The set contained sixteen orchestral instrument sounds, while that used in [WG72] consisted of only nine. More importantly, Grey's stimuli were time-varying and synthetically produced.

The process of producing these sounds began with conventional tape recordings. Each of the sixteen instruments played an E-flat at 311 Hz for durations ranging between a quarter- and half-second. The analog recordings were subsequently digitized and analyzed with a heterodyne filter technique. Results of this analysis yielded the detailed time-varying amplitudes of each partial for any given tone. Further refinement drastically reduced the amount of information required to resynthesize these tones using additive synthesis, making it possible to produce unprecedentedly realistic synthesized instrument tones. Thus, the complex time-evolving physical properties of the stimuli were completely determined—a feature that previous studies did not possess. The final step of proposing acoustic correlates to perceptual timbre dimensions could be based on more complete information than the steady-state spectra employed in [WG72]. In addition, controlled equalization of pitch, loudness, and duration is greatly facilitated, and previous studies are inconsistent in this regard.

Two articles following [Gre75] established some significant findings. In [Gre77], a pair of experiments are described that generate a perceptual timbre space and verify an aspect of its ability to predict similarity judgments. The first experiment collected 35 sets of similarity judgments from 20 "musically sophisticated" participants. Upon hearing each possible pairing of tones from the set of sixteen stimuli, participants rated similarity on a 30 point scale. Carroll & Chang's INDSCL MDS program processed the resulting similarity matrix, and two- three- and four-dimensional solutions were evaluated. Inconsistencies were found in the two-dimensional solution, and no clear advantage could be found

between the three- and four-dimensional solutions. Figure 2.3 reproduces Grey's three-dimensional timbre space.

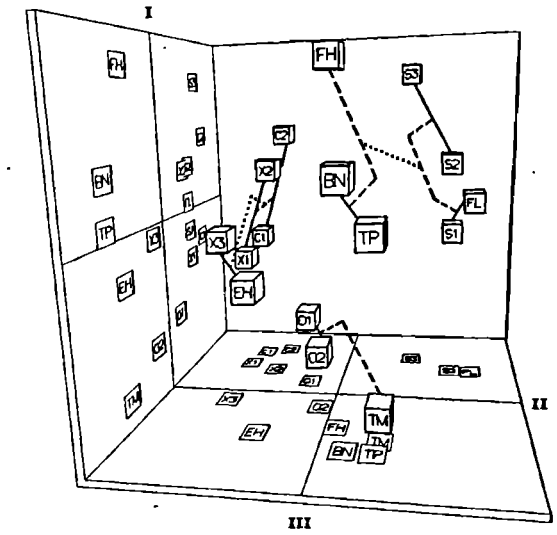


Figure 2.3: Grey's three-dimensional timbre space.

Plotted in the three-dimensional space, instruments were grouped by family to a certain degree. For instance, trumpet, trombone, and french horn are tightly clustered with respect to dimensions II and III. Grey's psychophysical interpretation of the space attributed the first dimension to spectral energy distribution (confirming previous studies regarding the importance of spectral centroid), and the remaining dimensions to temporal acoustic features. The second dimension was best explained in terms of synchronicity between upper harmonics during the rise and decay portions of the sounds. Such a pattern can also be interpreted in terms of spectral flux—the change in amplitude of all partials over time. While woodwinds had upper harmonics that entered and exited in close alignment, those of brass instruments exhibited more independent behavior. The presence of low-amplitude high-frequency inharmonic partials during the attack segment was the most relevant acoustic feature distinguishing sounds along the third axis. Clarinet and string instruments possessed these fleeting partials, while brass, bassoon, and english horn did not [Gre77, p. 1274].

The second investigation detailed in [Gre77] trained participants in an iden-

tification task over the course of several sessions. Various arrangements of the same 16 stimuli were played for participants repeatedly in each session, and classification judgments were collected. Participants' classifications became more accurate over the course of the sessions, improving from 60% to 84% [Gre77, p.1276]. Among the incorrect responses, the confusion of two sets of instruments is pointed out. A saxophone sound was confused with English horn 8% of the time, and bassoon was confused with French horn 7% of the time. Referring to the three-dimensional solution from the first experiment, it can be seen that these stimuli are in close proximity with respect to the second and third dimensions. In combination, the two experiments in [Gre77] establish and verify a perceptual timbre space that relates to both the spectral energy distribution and spectro-temporal features of the stimuli under investigation.

A separate study presented in [GG78] begins to exploit the considerable power of completely determined synthetic stimuli. Based on the perceptual timbre space generated in [Gre77], it was hypothesized that exchanging features of the distribution of spectral energy between a pair of complex tones would result in a corresponding exchange of the positions of these tones within timbre space. Thus, the effect of precisely controlled synthesis parameters was evaluated by an independent perceptual experiment—a repetition of the timbre similarity experiment that generated Grey's initial timbre space.

The first phase of this investigation was to carefully map the spectral energy distribution of one tone onto another. As each tone is described by a complex set of time-varying amplitude and frequency values, this task involves some difficult decisions. For instance, how should the spectral distribution of a sound with a great number of significant partials be grafted onto a sound with relatively few strong partials? Regarding temporal structure, which moment of the model sound should be chosen as representative of its spectral energy distribution? Grey & Gordon chose to apply the peak amplitude value of each partial of the model sound to the sound under alteration. To give an example, the strategy for mapping the features of the trumpet stimulus onto the trombone stimulus was as follows. The peak amplitudes of all harmonics in the trumpet stimulus were used to scale

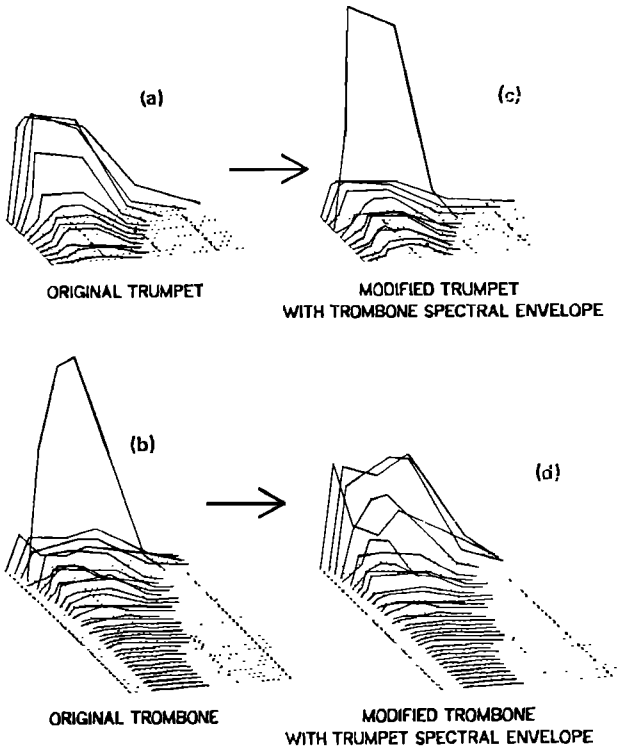


Figure 2.4: Spectral characteristics of the trumpet and trombone are exchanged.

the corresponding harmonics of the trombone stimulus so that the magnitudes of its harmonics were more like those of the trumpet, but the overall contour of each harmonic's trajectory remained similar to that of the trombone. For pairs of sounds with differing numbers of significant partials, the additional harmonics of the sound with larger spectral bandwidth were left unaltered. Thus, each sound retained its bandwidth characteristics, but important amplitude relationships between its harmonics were significantly different. Figure 2.2, reproduced from [GG78, p. 1495] illustrates the transformation.

The four pairs of stimuli that were altered in this way were trumpet—trombone, oboe—bass clarinet, bassoon—French horn, and two cello stimuli: regular and sul ponticello. Forty sets of similarity judgments were collected from 19 musically sophisticated participants, with all other experimental design parameters held constant from the previous similarity study. Again, three- and four-dimensional solutions from the MDS process returned the best fitting results. The

three-dimensional solution was used in order to facilitate comparison with the previous timbre space.

As hypothesized, the altered pairs of tones traded positions along the first axis, which was taken to represent spectral energy distribution. Because the spectral alterations also affected temporal characteristics (e.g., the synchrony of high-frequency partial trajectories), the altered tones shifted along the other two axes as well. Unaltered tones maintained their relative positions from the timbre space generated in [Gre77]. Grey & Gordon's study confirmed Grey's original perceptually based timbre relationships, and successfully predicted perceptual timbre judgments using calculated synthesis parameters.

The final section of [GG78] proposes several methods for a unitary numerical predictor of position along the first dimension of the timbre space. Amplitude information for time-evolving spectra was averaged in a number of ways, and the mean value of the time-averaged spectrum was determined to be the most effective predictive measure. Strongest results were obtained by further transforming the time-averaged spectrum by Zwicker & Scharf's loudness function [ZS65]. The authors conclude that spectral centroid "is an adequate representation of spectral energy distribution, in that it simultaneously takes into consideration the many factors which may be important: overall bandwidth, balance of levels in the lower harmonics, and the existence of strong upper formants." [GG78, p. 1498]

2.3.2 Iversen & Krumhansl

In a chapter offering an array of perspectives on the creative ramifications and state of timbre investigations, [Kru89] reports the results of a study carried out with Wessel using a newly developed MDS algorithm. Like others, she notes that the MDS approach is very appropriate for exploratory research of high-dimensional perceptual information. However, she also points out that certain aspects of timbre may be categorical, making it difficult to model along continuous dimensions. The MDS algorithm applied in her study with Wessel allowed stimuli to maintain individual dimensions that remained separate from those of the common low-dimensional solution. The degree to which a stimulus utilized its own separate

dimension was reflected by a "specificity" value. Timbres with a low specificity could be understood as fitting the common dimensions of the MDS solution quite well, while those with high specificity values required additional interpretation. Twenty-one timbres were created using FM synthesis, with the intention of modeling the sound of common orchestral instruments. Six of these stimuli were "hybrid" tones. For instance, the *trumpar* stimulus was designed to sound like a mixture of trumpet and guitar.

Krumhansl & Wessel's scaling experiment produced a space with two clearly interpretable dimensions. The first was taken to relate to rapidity of attack, and the second corresponded to brightness. A third dimension was less clear, but the authors attributed it to some aspect of the "temporal evolution of spectral components" [Kru89, p. 48]. Certain instruments, including trumpet and trombone, generated specificity values of zero, while others generated quite high values. The interpretations of high specificities given for stimuli like harpsichord and clarinet were only preliminary, but pointed to unique sonic characteristics like mechanical noise and predominance of odd harmonics. Thus, general perceptual dimensions were still generated using this MDS model, but stimuli that created problems for the fit could be identified and their additional characteristics possibly better understood.

Continuing this line of research in [IK93], Iversen & Krumhansl sought further clarification of the way in which dynamic aspects of tones contribute to perceived timbre similarity. By 1993, several studies had identified spectral centroid as "one of the main contributors to the perception of timbre" [IK93, p. 2595], and Krumhansl, Grey, and Wessel had independently found evidence that various combinations of onset characteristics constituted another important perceptual dimension. As explained above, Grey found that the synchrony of rising partials as well the presence of quiet high-frequency energy during an onset were meaningful timbre attributes, while the second dimension of Wessel's timbre space related to "the nature of the onset transient" [Wes79, p. 48]. Iversen & Krumhansl designed three experiments in order to evaluate the unique salience of onsets in timbre perception.

The stimuli were taken from the McGill University Master Samples library (MUMS), and included recordings of bassoon, cello, clarinet, English horn, flute, French horn, oboe, piano tenor saxophone, tenor trombone, trumpet (normal and muted), tuba, tubular bells, vibraphone, and violin. Like Grey, Iversen & Krumhansl's set contained sixteen orchestral instruments, but with a notable addition of two percussion instruments: tubular bells and vibraphone. Because the stimuli were recorded natural sounds rather than realistically synthesized sounds, options for varying onset and steady state characteristics were quite limited. Three sets of stimuli were prepared. The instrumental samples were unaltered in the first set, truncated to the first 80 milliseconds in the second set, and composed of only the steady state and decay segments for the third set. Stimuli in the final set were referred to as "remainders".

In each of the three experiments, participants were asked to evaluate the similarity of all possible pairings of timbres in the set. Breaking slightly from the procedure of previous MDS-based research, Iversen & Krumhansl posed the similarity evaluation to participants relative to a hypothetical task. Participants were asked to "imagine that they had a computer that allowed them to record a sound and change it in any way they wanted." [IK93, p. 2597] Their rating was then given as a function of how much they would have to change the first tone to make it sound exactly like the second. A provided rating scale ranged from "a little" to "a lot".

MDS of the similarity ratings for the first stimulus set was carried out using the same algorithm employed in [Gre77] and [GG78], which is described in [Kru64]. A two-dimensional solution was generated that generally separated stimuli with impulsive onsets from those with a more gradual attack along the horizontal axis, and arranged stimuli by brightness (i.e., spectral centroid) along the vertical axis.

Having established an initial timbre space with meaningful axes, the two remaining experiments were carried out. Dimensional scaling of similarity judgments of the onsets-only stimulus set produced a very similar timbre space, indicating that onsets alone are very relevant in timbre similarity evaluations as well as classification. Surprisingly, MDS of judgments based on the "remainders" set *also*

produced a timbre space that was very similar to that from the first experiment. Thus, a preliminary hypothesis that onsets are the most important segment of the sound in similarity tasks was rejected, and it was concluded that unique timbre attributes cannot be isolated to any particular segment of complex tones.

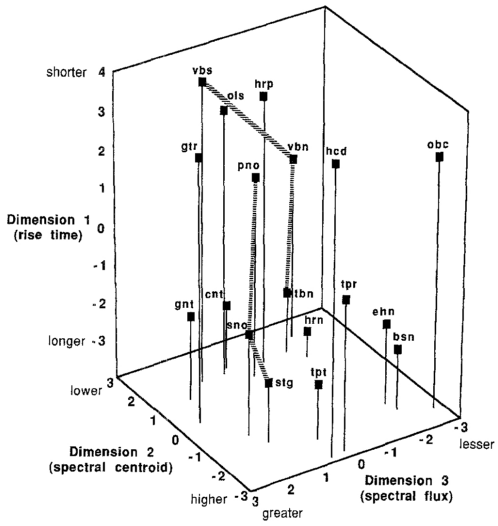
2.3.3 McAdams

Some of the most thorough and large-scale timbre studies of late have been generated by various collaborations between Stephen McAdams, Suzanne Winsberg, Jochen Krimphoff, Anne Caclin, and Stephen Lakatos, among others. Led by McAdams, a team of five researchers produced a major report [MWD+95] that advanced the MDS-based timbre experiment model in three major areas. First, synthesized stimuli were used rather than acoustic instrument recordings, including synthetic attempts at both real and invented instrument tones. Only [GG78] and [Kru89] had done this previously. Second, the CLASCAL MDS algorithm was used to measure the "specificity" of each stimulus. Stimuli with high specificity possess unique attributes that are not accounted for by the axes of the MDS solution space. Third, the large participant pool was designed according to levels of musical training. The CLASCAL algorithm was used to discover latent classes implied by the structure of the similarity data. Relationships between these classes and musical experience were sought.

The 18 stimuli were synthesized instrument tones produced by Wessel, Bristow, and Settel using FM synthesis techniques [WBS87]. An earlier study by Krumhansl [Kru89] employed the same sound set. A majority of the timbres were designed in imitation of traditional orchestral instruments, but six represented hybrid instruments, including the trumpar (trumpet/guitar), oboleste (oboe/celesta), striano (bowed string/piano), vibrone (vibraphone/trombone), obochord (oboe/harpsichord), and guitarnet (guitar/clarinet). All stimuli were equalized for loudness, duration, and pitch.

98 participants with varying degrees of musical training were recruited to create three participant categories: professional musicians, amateur musicians, and non-musicians. The number of participants in each group was 24, 46, and 28

respectively. As with previous studies, participants evaluated the similarity of all possible pairs of tones within the sound set.[1] Analysis of the similarity data identified a small set of participants with inconsistent ratings. These data were removed, leaving 88 data sets for use in the CLASCAL analysis.



Name (origins of hybrids in parentheses)	Label	Max Level (dBA)	Total Duration (ms)
French horn	hrn	72	569
Trumpet	tpt	60	520
Trombone	tbn	64	563
Harp	hrp	61	707
Trumpar (trumpet/guitar)	tpr	56	635
Oboleste (oboe/celesta)	ols	62	716
Vibraphone	vbs	59	770
Striano (bowed string/piano)	sno	61	775
Harpsichord	hcd	53	521
English horn (cor anglais)	ehn	67	507
Bassoon	bsn	65	495
Clarinet	cnt	64	496
Vibrone (vibraphone/trombone)	vbn	62	1096
Obochord (oboe/harpsichord)	obc	63	544
Guitar	gtr	57	569
Bowed string	stg	58	1071
Piano	pno	60	1008
Guitarnet (guitar/carinet)	gnt	63	557
Mean		61.5	673
Standard deviation		4.1	200

Figure 2.5: The three-dimensional timbre space produced by McAdams et al.

The two models returned from the MDS process—one in six dimensions without specificities, the other in three dimensions with specificities—were highly successful in mapping the perceptual information. The lower dimensional model was chosen because "the psychophysical interpretation of the underlying dimensions was more coherent." [MWD+95, p. 184] This timbre space is reproduced in Figure 2.5.

One of the goals of this study was to compare the perceptual space of Krumhansl's previous experiment using the same sound set.[2] The first two dimensions of McAdams et al.'s space correlated very strongly with Krumhansl's. The third dimension in [Kru89] was interpreted to be related to spectral flux. In a separate study, Krimphoff et al. [KMW94] reevaluated Krumhansl's data and identified an acoustic measure that correlated very highly with her third dimen-

[1] It is worth noting that in this case, the rating scale was presented in terms of "dissimilarity", with "very similar" on the *left*, and "very dissimilar" on the *right*.

[2] [Kru89] actually used three additional timbres that were dropped in [MWD+95]

sion. The spectral measure employed was referred to as "spectral irregularity", informally defined as the "log of the standard deviation of component amplitudes from a global spectral envelope derived from a running mean of the amplitudes of three adjacent harmonics." [MWD+95, p. 187] However, in spite of drastic measures, such as removing timbres from the set, no clear interpretation could be made for the third dimension in [MWD+95], and the spectral irregularity measure did not correlate with the position of timbres along that axis.

A listening-based analysis of timbres with high specificity values identified some of their unique features. Eleven stimuli were closely considered, and two categories of specificities were proposed. The authors concluded that some of these attributes are continuous in nature, having various degrees of intensity, while others are more discrete, depending on the simple presence or absence of an unusual feature. For instance, the harpsichord tone (possessing one of the highest specificity values) was said to have a distinct "clunk" during its offset, presumably a synthetic modeling of the instrument's mechanical noise. Counter to intuition, the hybrid instrument tones were no more likely to have high specificity than conventional instrument tones.

Results of latent class analysis were counter to McAdams et al.'s hypothesis. Out of the five latent classes uncovered by the CLASCAL analysis, most participants were members of either the first or second class. It was concluded that relationships between the musical training data and class membership were not meaningful, indicating that timbre perception is not strongly affected by experience. It is worthy of note, however, that judgments made by participants in the professional musician group were the most consistent. McAdams et al. point to the everyday aspect of timbre perception as a possible explanation of these results.

> Timbre, being composed of many of the sensory qualities that specify the identity of a sound source, may likely be used as an important auditory cue for monitoring the environment on a continual basis by listeners in their everyday lives. [MWD+95, p. 190]

More recently, a very thorough study by Caclin, McAdams, Smith, and Winsberg [CMSW05] has again confirmed spectral centroid and attack duration as important predictors of timbre. The investigation also directed considerable effort

toward better understanding of the role of spectral flux in relation to the third dimension found by Krumhansl. In three separate experiments with roughly 30 participants each, Caclin et al. gathered similarity judgments on carefully designed synthetic tones.

Stimuli for the first experiment were systematically varied in terms of spectral centroid, attack time, and spectral flux. Flux was realized as a sinusoidal variation of spectral centroid over the first 100 milliseconds of the tones. This was intended to model patterns of high frequency presence over time in natural tones. As in [Gre77], all stimuli were pitched at 311 Hz, and were equalized for loudness and duration. The spectra of all stimuli contained 20 harmonically related partials with various patterns of amplitude envelopes that decreased as a function of frequency. These patterns were used to directly control spectral centroid. In the time domain, all amplitude envelopes had attack, sustain, and decay segments.

Two- and three-dimensional timbre spaces were produced, with axes that correlated highly with spectral centroid and attack time.[3] The three-dimensional model indicated that participants used spectral flux to a very small extent in their similarity ratings, but only for tones with very high spectral flux. In Figure 2.6, strong similarity can be seen between the synthesis parameter space and the two-dimensional perceptual space.

A second experiment focused entirely on spectral flux, using three sets of stimuli: one in which spectral centroid was held constant, one in which attack time was held constant, and another in which both spectral centroid and attack time were held constant. Solutions returned from MDS analysis in all three experiments did not map stimuli as predicted by the stimulus synthesis parameters. Two crude groupings of stimuli with low and high spectral flux were formed in one case, but no patterns reflecting the systematic variation of synthesis parameters could be identified based on the similarity judgment data.

As an alternative to spectral flux, spectral irregularity was varied within a stimulus set for a third experiment. Caclin et al. define spectral irregularity as

[3] A primary experimental goal of [CMSW05] was the comparison of CLASCAL and CONSCAL MDS algorithms. In the context of this discussion, no distinction is made between timbre spaces produced by these algorithms.

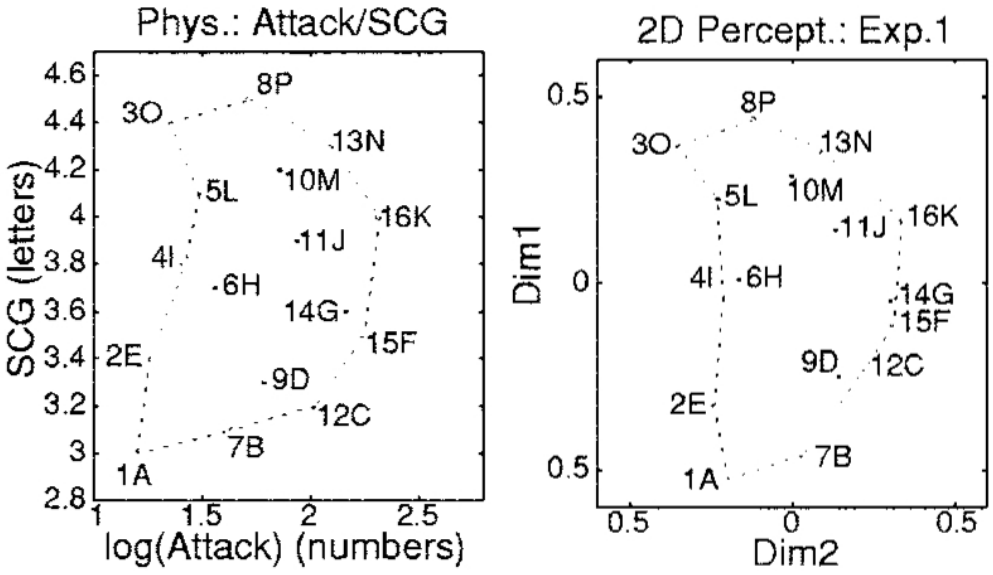


Figure 2.6: Two-dimensional synthesis parameter and perceptual spaces, from Caclin et al.

an amplitude attenuation of even harmonics relative to odd. Attenuation ranged from 0 to 8 dB. A subsequent MDS analysis produced a three-dimensional timbre space in which the third dimension correlated very highly with spectral irregularity. Though these efforts were successful, it should be noted that spectral irregularity is not a time-varying synthesis parameter like spectral flux. Hence, this experiment fails to address the original goal of accounting for Krumhansl's third perceptual dimension. Nevertheless, the work described in [CMSW05] makes an invaluable contribution to the literature, providing a much needed confirmatory study with many clear findings.

As a final point, it is interesting to note McAdams' participation in a classification-based study led by Lakatos [LMC97]. The experiment is related to Freed's perceived mallet hardness study in that both evaluate participants' ability to perceive physical characteristics of sound stimulus *sources*. In the case of [LMC97], participants were asked to select one of two graphical depictions of sound sources based on their likelihood of producing a given sound stimulus. The study by Lakatos, described in the following section, addresses this thread of research in the context of a more conventional MDS timbre similarity experiment.

2.3.4 Lakatos

Growing out of work by McAdams et al., an MDS-based study described in [Lak00] was designed in order to explore the repercussions of analyzing larger and more heterogenous sets of stimuli. In light of the consistent identification of spectral centroid and onset rise time as important correlates of timbre perception, Lakatos points out that most of the stimuli upon which these conclusions were founded were relatively similar across studies. He hypothesized that additional dimensions might be uncovered if this experimental parameter was expanded [Lak00, p. 1427]. Lakatos also intended to address the question of whether or not more recent MDS algorithms, which do not strictly assume a continuous (rather than categorical) distribution of data, offer any significant advantages. Citing a lack of data connected with the level of musical training possessed by participants, he hoped to shed some light on the effect of experience on timbre perception. Finally, by employing a program capable of mapping data according to a tree model (EXTREE), he sought patterns of categorization that related to the physical characteristics of each sound's source and manner of excitation.

34 recorded instrument tones were taken from the McGill University Master Samples collection and organized into three stimulus sets. The stimuli were chosen with the aim of representing a broad range of timbres, instrument materials, and methods of excitation. The first set, called the "harmonic" group, contained 17 tones produced by pitched instruments playing a D-sharp above middle C. The "percussive" set contained 18 tones, 7 of which were pitched. A "combined" set was made up of 20 tones—10 selected from each of the previous sets. Loudness and pitch were carefully equalized [Lak00, p. 1428].

The participant pool was composed of 18 musicians and 16 non-musicians. Participants judged all possible pairs of tones for each stimulus set in three separate hour-long sessions. The similarity rating scale ranged from "very similar" to "very different".

Similarity data from the harmonic set showed no significant differences based on musical training. A two-dimensional timbre space was generated that was quite similar to those described in [IK93]. The logarithm of attack rise time

correlated well with stimulus mappings along the horizontal dimension, separating impulsive instruments like harp, piano, and harpsichord from bowed string and wind instruments. Stimuli were spread across the second dimension in a way that correlated with the logarithm of spectral centroid. A second analysis using EXTREE to discover grouping trends indicated no clear grouping of the harmonic instruments according to instrument materials or manners of excitation.

Three dimensions were found for the percussive set, however the third dimension was uninterpretable. In spite of the drastically more heterogenous collection of timbres in the percussive set (including bamboo chimes, bongos, castanets, celesta, cuica, bowed and struck cymbals, log drum, marimba, snare drum, steel drum, tambourine, tam-tam, bowed and struck vibraphone, temple block, tubular bells, and tympani) the first two dimensions were again clearly correlated with attack rise time and spectral centroid. This suggests that the two measures are indeed fundamentally connected to our understanding of timbre.

A slight difference was found between the judgments of musicians and non-musicians. Lakatos speculated that the heavier weighting of musicians' judgments on the axes of the MDS solution was due to their increased familiarity with the unusual sound set. Clusters returned from analysis using EXTREE were clearly segregated along lines corresponding to instrument materials and excitation. Wooden and metal bars and tubes formed the first cluster, while metal plates alone formed the second. The majority of instruments in the final cluster were membranophones.

Finally, the MDS solution for data collected on the combined set (shown in Figure 2.7) was very similar to that of the harmonic set alone. Its two dimensions were well explained by attack time and spectral centroid, with logical separation of impulsive and continuous sounds on the horizontal axis. No major difference was found between judgments with respect to musical training. With the context of harmonic stimuli, the percussion stimuli were again clustered by EXTREE according to instrument material, and among harmonic timbres a group of blown aerophones emerged.

Lakatos' study engages challenges that were not addressed in previous MDS-based research, such as the use of a more diverse set of sounds that included

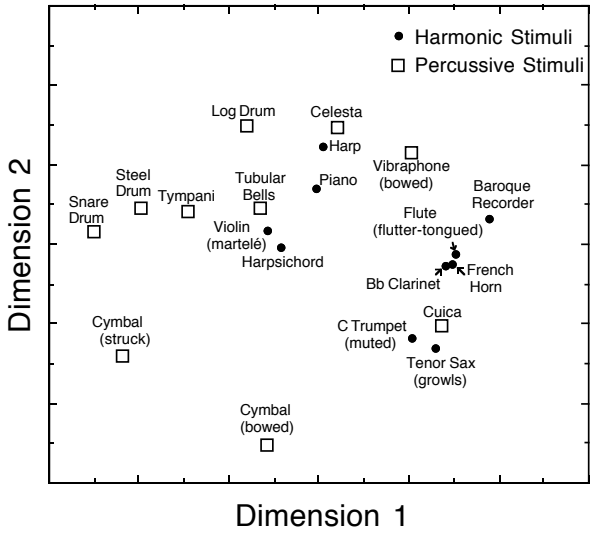


Figure 2.7: Two-dimensional timbre space for the "combined" stimulus set.

unpitched and noisy instruments. His findings were remarkable in that a new perceptual dimension did *not* emerge in response to this diversity. As discovered in previous studies (both verbal attribute- and MDS-based), the spectral center of mass was an excellent predictor of perceptual judgments. Attack duration, another well-confirmed acoustic correlate of timbre, was the only other significant predictor identified. In an effort to make sense of the useful but unexpected results, Lakatos offers the following thoughts.

> To what extent does a two-dimensional structure represent a useful descriptive model? From an intuitive perspective, it may seem unsatisfying to accept that two orthogonal dimensions capture most of the variance inherent in our rich acoustic environment. ... MDS algorithms, including CLASCAL, almost invariably generate low-dimensional solutions because they seek the most parsimonious dimensional fit to the data. [Lak00, p. 1437]

2.4 Summary

This chapter has sampled a bifurcated research tradition spanning nearly 70 years. Along the verbal attribute branch, Kendall and Carterette followed von Bismarck's well-documented efforts, and moved in a productive direction by drop-

ping the use of antonymous semantic differential scales. Their more careful choice of musically relevant adjectives serves as a reminder that adjective lists must be compiled anew for each sound set under investigation, and no assumptions can be made about the permanence of any one term over time. Their finding that spectral energy distribution and flux related significantly to participants' judgments is in line with Lichte's and von Bismarck's conclusions, and many MDS interpretations as well. Freed's study produced results that also point to the significance of spectral centroid in the prediction of a single aspect of percussive timbre—one that is conceptually tied to physical sound source characteristics. His choice to focus on perceived mallet hardness alone produced strong conclusions upon which further study can be built.

Though this review is not complete (also see [Rah66] [Jos67] [Tru71] [SMN96] [Dar05]) there are relatively few timbre experiments that explore verbal attributes. The dominance of MDS-based experiments can perhaps be explained by two understandable desires: 1) to find a general, intuitive, dimensional model of timbre, and 2) to avoid the imprecision and impermanence of language. The only dependence that MDS studies have on language is in connection with the meaning of "similar", which is a reasonably stable concept. In spite of these significant advantages, however, the attribute-based approach should not be dismissed. It is remarkable, for instance, that Lichte's 1941 study identified an attribute associated with spectral centroid as central to timbre perception. Studies of Western orchestral instrument timbres continue to confirm the importance of this spectral measure.

It appears that decades of more sophisticated studies have yielded no more than two additional correlates: attack duration and spectral irregularity. Of those studies confirming the former correlate, only Grey's work investigated it any detail beyond raw duration. His identification of asynchrony among partial rise times and the presence of weak high-frequency energy during attack deserve further investigation. McAdams et al.'s identification of spectral irregularity as relevant is relatively recent, and requires additional confirmation.

Both branches exhibit weaknesses in the area of experiments carried out on sequences of timbres. The "legato transient" between successive notes was

studied in [CH78] in a classification context, and similarity judgments of isolated timbres and melodic sequences of timbres were considered in [Ser95] in a relational context. Further work along these lines is needed. A second weakness may be a byproduct of MDS dominance: the stimulus sets for most studies are rarely composed of more than 16 timbres. This is at least partly due to a requirement of MDS algorithms that similarity ratings be obtained for all possible pairs in a stimulus set. The largest stimulus sets were those used in verbal attribute studies by von Bismarck (35) and Freed (24). Apart from MDS limitations, there is also a general restriction on the size of stimulus sets in terms of what can reliably be evaluated by participants in a single session.

Timbre homogeneity is another important issue. Stimuli used in the studies above range from actual recordings of instruments to unmistakably artificial synthetic signals. Even amongst the synthetic stimuli, most of the timbres do not stray far from the realm of Western orchestral instruments. Lakatos' combined set of 20 orchestral and percussion instruments was by far the most diverse. There are experimental justifications for this trend. For instance, in order to isolate additional perceptual dimensions, it is necessary to restrict rather than broaden the range of timbres studied. With spectral centroid, attack duration, and (tentatively) spectral irregularity established as relevant, a clear research path is the synthesis of tones that are strictly normalized along these dimensions. Such a sound set would likely be extremely homogenous, but would offer the possibility of discovering new perceptual timbre attributes.

The sounds studied above do not begin to reflect the rich palette of acoustic and synthetic timbres that contemporary composers are engaged with. Thus, in parallel with research on restricted groups of timbres, a line of study involving more diverse sound sets also needs to be established. In spite of its disadvantages, an attribute-based research approach offers greater freedom in terms of stimulus set size, which makes it an attractive option for studying heterogenous timbre sets.

Chapter 3

Objective Analysis

The experiments described in Chapter 2 indicate that our perception of timbre is too complex to be fully explained by a collection of objective measures. Stephen Handel notes that performance actions associated with the articulation of a sound may be very influential in timbre perception [Han95, p. 495]. For percussive sounds in particular, [Fre90] and [LMC97] confirm that mallet materials and the shape and nature of the resonating body itself also influence perception. Clearly, tenacious connections between action, object, and sound are formed through our musical and day-to-day experiences. These types of associations will surely vary with cultural context, and cannot be expected to remain stable. For instance, the vocabulary of performance gestures associated with timbres created by a digital musical instrument is often in constant flux. Within the context of such a performance, our cognitive map for understanding the physical origin of familiar timbres is altered. The rich network of associations that a sound carries for any particular individual will always differ subtly (and possibly drastically) from the sound's corresponding set of implications for another individual.

Such complexities make efforts toward automatically quantifying timbre seem somewhat vain. While it is important to recognize the inadequacy of quantitative descriptions of a largely qualitative sonic characteristic, several objective time and frequency domain analysis algorithms have nevertheless been applied successfully to music and speech classification tasks. The majority are based on a short-time Fourier transform, and involve various degrees of further processing.

This chapter will present two classes of algorithms from the literature that generate low and high level features. Low level features describing the frequency domain are crude measures of spectral envelope, while high level features retain a larger portion of the spectrum's original dimensionality. Cepstral techniques have proven to be the most powerful from a data reduction standpoint, and will be described in the most detail.

3.1 Low level features

The successful connection of spectral centroid with a perceptual dimension of timbre has lead to several other algorithms for summarizing spectral energy distribution along a single dimension. The desire for unidimensionality is rooted not only in the fitting of perceptual data, but also in applications where extremely large audio databases are intended to be browsed by user query, and compact information is highly prioritized. A series of spectral descriptors described in [ZR07][Tza02][TC02][PMH00] were established for the MPEG7 standard, and have been applied in projects such as IRCAM's *Studio On Line* [HBPD03]. Some of the most important sound descriptors are described below.

3.1.1 Spectral Centroid

Like any distribution of values, the magnitudes of spectral bins can be described with respect to the four central moments: mean, variance, skewness, and kurtosis [Fuj98][TC99]. The first spectral moment (mean, or centroid) is the center of mass of magnitude spectrum. It can be computed as the ratio of the sum of spectral magnitudes weighted by either bin index or frequency to the unweighted sum of spectral magnitudes. In units of frequency, centroid (C) is defined as

$$C = \frac{\sum_{k=0}^{N/2} f(k)\,|X(k)|}{\sum_{k=0}^{N/2} |X(k)|} \tag{3.1}$$

Where N is the number of points in the Fourier transform, and $f(k)$ and $|X(k)|$ are

the frequency and amplitude of the k^{th} bin respectively. Two instrument spectra and their centroids are given in Figure 3.1.

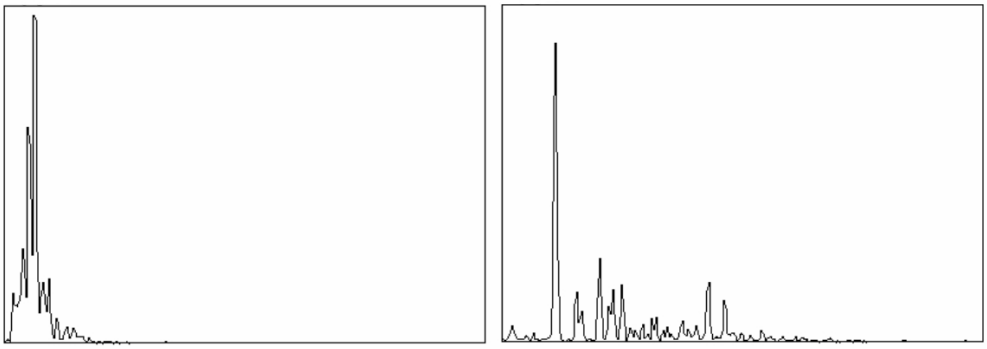


Figure 3.1: Bongo (left) and metal bowl (right) spectra, with spectral centroids of 926 Hz and 2858 Hz.

3.1.2 Spectral Spread

The second moment is spectral variance, but [Fuj98] and [ZR07] express spectral spread in terms of standard deviation. This reflects the degree to which total spectral energy is concentrated around the mean. The calculation is accomplished by centering the frequency values for each bin according to the centroid, so that the energy in bins that are distant from the centroid will be accumulated in the sum.

$$Spread = \sqrt{\frac{\sum_{k=0}^{N/2}(f(k) - C)^2 \, |X(k)|}{\sum_{k=0}^{N/2} |X(k)|}} \tag{3.2}$$

The bongo and metal bowl spectra in Figure 3.1 have spreads of 1755 Hz and 2144 Hz respectively.

3.1.3 Spectral Skewness

The third moment is skewness, which measures the symmetry of a spectrum's energy distribution. A positive skew results from distributions with a steep

slope upward on the low end and a more gradual slope downward moving toward higher frequencies. A negative skew is precisely the opposite. Mathematically, spectral skewness is defined as

$$Skew = \frac{\sum_{k=0}^{N/2}(f(k) - C)^3 \, |X(k)|}{\sigma^3 \sum_{k=0}^{N/2} |X(k)|} \tag{3.3}$$

where σ is spectral spread. The bongo spectrum in Figure 3.1 has a high positive skewness value, while the metal bowl spectrum has a lower (but still positive) value because the energy beyond the most prominent peak is more spread out, with several strong individual partials.

3.1.4 Spectral Kurtosis

The fourth moment is spectral kurtosis, which changes in connection with the sharpness of an energy distribution's peak. It is defined as

$$K = \frac{\sum_{k=0}^{N/2}(f(k) - C)^4 \, |X(k)|}{\sigma^4 \sum_{k=0}^{N/2} |X(k)|} - 3 \tag{3.4}$$

As the bongo spectrum has most of its energy in a small low-frequency band and possesses a strong central peak, its kurtosis is quite high. All of the spectral moment measures are taken relative to a central mean value, with the hope that positioning of the spectral envelope along the frequency axis will have a minimal effect on the measures. For instance, an ideal measurement of spectral spread for a band of noise would remain constant for any center frequency value. However, as frequency deviations from the centroid are raised to higher and higher powers, a frequency-dependent bias is introduced. With the highest power term used here, spectral kurtosis is the most susceptible to this bias. [Fuj98] describes the use of moment-based features going as high as the 10^{th} moment, but notes that "higher order moments tend to be noisy".

3.1.5 Spectral Brightness

Spectral Brightness is the ratio of the sum of magnitudes above a given boundary frequency *f(K)* to the sum of all magnitudes in a spectrum. Signals with a significant amount of high frequency content will have higher brightness. Typical values for *f(K)* are 1000, 1200, or 3000 Hz [LJB05][Jus00].

$$B = \frac{\sum_{k=K}^{N/2} |X(k)|}{\sum_{k=0}^{N/2} |X(k)|} \tag{3.5}$$

Referring again to Figure 3.1, the brightness value (with $f(K) = 1200$) of the bongo spectrum is 0.1, while that of the metal bowl spectrum is 0.93.

3.1.6 Spectral Rolloff

Spectral Rolloff is a second measure of high frequency content. Rather than a ratio, it is expressed as the frequency of bin K, below which a certain percentage of total spectral energy is concentrated. [TC02] specifies 85%.

$$max\{f(K) : \sum_{k=0}^{K} |X(k)| \leq 0.85 \sum_{k=0}^{N/2} |X(k)|\} \tag{3.6}$$

The rolloff frequency of the bongo spectrum in Figure 3.1 is 991 Hz, just above the spectral centroid.

3.1.7 Spectral Flatness

Spectral Flatness is the ratio of the geometric mean of magnitude spectrum to the arithmetic mean of magnitude spectrum. A very noisy spectrum without clear shape (i.e. that of white noise) has a high flatness value, while the spectrum of a single sinusoid has extremely low flatness.

$$B_f = \frac{\sqrt[N/2]{\prod_{k=0}^{N/2} |X(k)|}}{\frac{1}{N/2} \sum_{k=0}^{N/2} |X(k)|} \tag{3.7}$$

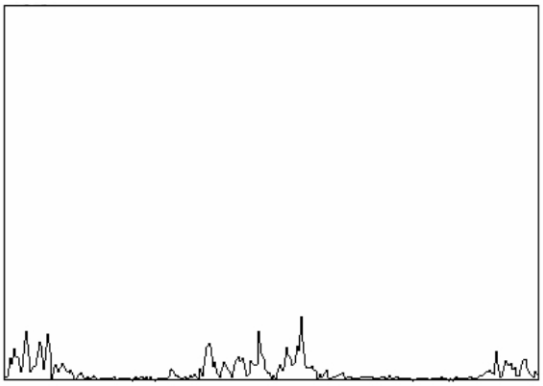


Figure 3.2: A tambourine spectrum with flatness value of 0.42.

Figure 3.2 shows the spectrum of a tambourine, several milliseconds after the initial attack. Energy is concentrated in three rough areas, where the lowest concentration is caused by the actual drum head, and the mid and high frequency concentrations emerge from the jingling of metal along the tambourine's frame. Although there are several peaks and valleys across the entire frequency range, the somewhat even energy distribution earns a relatively high flatness value of 0.42. In contrast, both of the spectra from Figure 3.1 have flatness values of 0.01 because of their strong primary peaks that eclipse remaining energy in other bins.

3.1.8 Spectral Irregularity

Spectral irregularity was explored in connection with perceptual judgments in [CMSW05], described in terms of the relative strength of even and odd harmonics. Irregularity is measured relative to the strongest peaks returned by a peak-picking algorithm [Jen99]. This measure is especially relevant for harmonic spectra, where the strongest peaks will be harmonics relative to a fundamental frequency. Irregularity can then be measured with respect to the most salient components in such a tone. Regardless of harmonicity, irregularity can also be measured relative to all spectral bins using one of two standard algorithms. Krim-

phoff's strategy is to subtract a running average of neighboring magnitudes from all but the first and last bins [KMW94].

$$I = log_{10} \sum_{k=1}^{(N/2)-1} \left| |X(k)| - \frac{|X(k-1)| + |X(k)| + |X(k+1)|}{3} \right| \tag{3.8}$$

If neighboring partials differ in magnitude by small amounts throughout the spectrum, the irregularity value will be very low. An extremely "spiky" spectrum would generate a high irregularity.

Jensen's algorithm sums squared magnitude differences with the upper neighbor of each bin, and sets this to a ratio against the sum of spectral power [Jen99, p. 94].

$$I = \frac{\sum_{k=0}^{(N/2)-1} (|X(k)| - |X(k+1)|)^2}{\sum_{k=0}^{N/2} |X(k)|^2} \tag{3.9}$$

The values returned by this formula have the useful quality of usually remaining between 0 and 1.0. Using either algorithm, the metal bowl spectrum from Figure 3.1 has a higher irregularity than the bongo spectrum. The more jagged contour of the former is visually apparent.

3.1.9 Spectral Flux

Spectral Flux is a spectro-temporal measure, given as the sum of squared difference between two successive normalized magnitude spectra. The time separation between successive frames depends on window size and overlap values. If data reduction is not a priority, the complete spectral difference between frames can be preserved to track changes in specific frequency bands. Because squaring the differences of fractional magnitude values can result in very small terms, it is also appropriate to sum the absolute value of the differences (i.e., Manhattan distance rather than Euclidean distance).

$$F = \sum_{k=0}^{N/2} \left(|X_i(k)| - |X_{i-1}(k)| \right)^2 \tag{3.10}$$

Spectral flux can be used for crude threshold-based attack detection.

3.1.10 Zero Crossing

Zero crossing is a time-domain feature, and is simply the number of times the waveform crosses zero in the window. It is computed as

$$Z = \sum_{n=2}^{N} sign[x(n)] - sign[x(n-1)] \tag{3.11}$$

where *sign* is the signum function, which returns 1 when the argument is positive, -1 when it is negative, and 0 otherwise. For pitched instruments, zero crossing rate correlates with fundamental frequency, and is thus an extremely crude indicator of pitch.

3.1.11 Log attack time

Attack time was identified as a primary perceptual dimension of timbre in studies described in Chapter 2. Limitations associated with attack detection make reliable real time capture of this measure difficult. Even in non real time, ambiguities connected with appropriate sound segmentation can cause complications. One general algorithm for this measure is the logarithm of the time elapsed between the points at which RMS amplitude is 2% and 80% of the sound's overall peak level [PMH00, p. 2].

3.1.12 Features for harmonic spectra

Additional low level features, such as the odd/even relation and the tristimulus measure have been proposed, but are most appropriate for the spectra of harmonic tones [PJ82][Jen99]. The former measure sums the energy in even and odd harmonics separately, which reflects distinctive characteristics of harmonic timbres like the clarinet. Applying such an algorithm to even and odd spectral bins would not return useful information, nor would it be appropriate to apply it to a subset of the strongest peaks, which would not be evenly spaced. The tristimulus

measure is also intended for harmonic tones, and separately measures the energy attributed to the fundamental frequency, the following three harmonics, and the remaining harmonics. Tristimulus is useful in distinguishing between the types of spectra examined in [WG72], shown in Figure 2.2, where fundamental energy and spectral slope relative to the fundamental were found to correlate with perceptual judgments. As the focus of this dissertation is constrained to percussive timbres with spectra that are generally inharmonic, these two measurements will not be explored further.

3.2 High level features

In moving from unidimensional features to multidimensional feature vectors, a range of options exists for negotiating the increased data size. Naturally, the most complete information is a matrix of time-varying magnitude spectra that covers the entire sound event. From the standpoint of creating manageable databases of sound descriptors, the interdependent parameters of time and frequency resolution must be considered in terms of memory consumption. With sampling rate $sr = 44.1$ kHz, window size $N = 1024$, and overlap $o = 2$, one second of audio requires 86 analysis frames with 512 points each, consuming 176.128 kilobytes of memory when using 32-bit floating point numbers. One option for reducing data size is to use a smaller overlap value, which results in fewer frames; however, this forfeits temporal information that is potentially critical for distinguishing one timbre from another.

The three-dimensional spectrogram of a bass drum strike in Figure 3.3 illustrates a spectro-temporal pattern that is typical of percussive sounds, where a powerful burst of high frequency energy exists for only the first hundred milliseconds of the attack, giving way to a concentration of lower energy as the main pitch of the instrument emerges. With very low overlap values, the detail in such patterns will be lost.

Strategies for reducing spectral information have similar consequences in the frequency domain. Frequency bands of a spectrum can be averaged in order

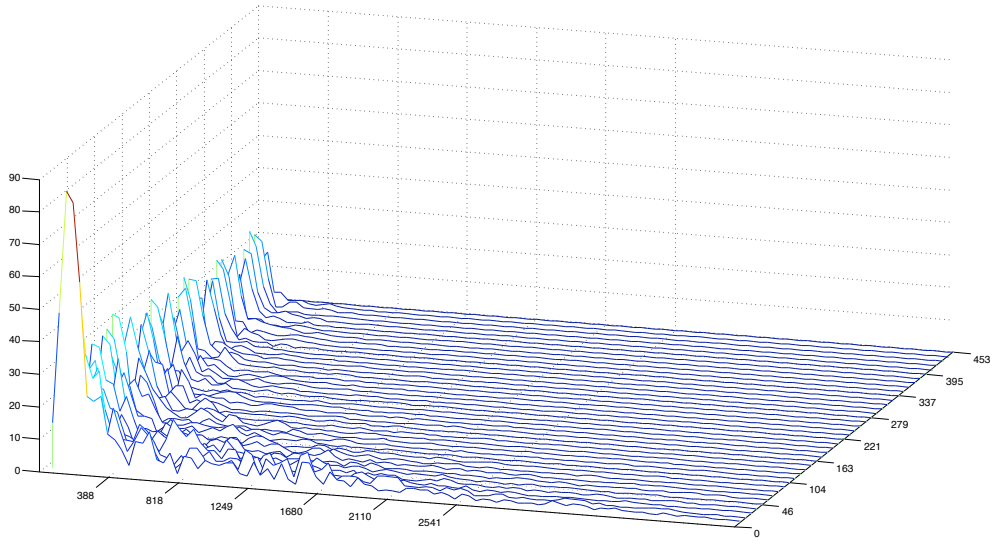


Figure 3.3: Spectrogram of a bass drum strike.

to limit the number of points for each frame, but any patterns of frequency fluctuation existing within each band will not be preserved. In this case, the choice of particular frequency bands is extremely important, as some bands are more perceptually relevant than others. For instance, the fact that half of the data in a spectrum describes frequencies above 10 kHz is clearly disproportionate in light of realistic hearing ranges, which do not typically extend to the ideal of 20 kHz. The most obvious solution to this problem is to average frequency bands that are evenly spaced according to perceptually based scales such as mels or Barks [SVN37][ZFS57]. Typical parameters for these types of techniques can reduce a 1024-point spectrum to about 30 points. Further reduction can be achieved with additional processing that measures correlation with crude spectral shapes. The most common of these techniques are based on cepstral analysis, first proposed in [BHT63], and applied to music modeling in [Log00]. This type of analysis can reduce a spectrum to only a few points that remain useful for automatic distinction between sounds.

3.2.1 Cepstral Analysis

The term "cepstrum" was originally defined in a 1963 article by Bruce P Bogert, MJR Healy, and JW Tukey, entitled "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking." [BHT63] Initial motivation for the technique came out of work on echo detection in seismological data. As the title of the article relates, its authors devised terminology that forges conceptual links between the frequency domain spectrum and the period domain cepstrum. Frequency values that typically occupy the abscissa of a spectral plot become "quefrency" values in the cepstral domain. Signal processing terminology such as filtering, phase, and analysis were translated to "liftering", "saphe", and "alanysis". For those interested in approaching the details behind this tongue-in-cheek vocabulary, the most accessible of conceptual definitions is that found in Curtis Roads' Computer Music Tutorial. According to Roads, cepstral analysis "tends to separate a strong pitched component from the rest of the spectrum." It "tends to deconvolve two convolved spectra." [Roa96, p. 518] These features explain why cepstral techniques are so heavily used in voice processing and identification. If speech is fundamentally a convolution of glottal impulses with the resonance of a speaker's oral cavity, cepstral analysis allows these components to be examined individually. By separating filtering characteristics from the effects of a particular pitched articulation, the cepstrum provides a general spectral signature of a person's voice.

In the realm of musical applications, the *Audio Oracle*, developed by Shlomo Dubnov, Gérard Assayag, and Arshia Cont, uses cepstral information to shuffle frames of an audio file by transitioning between the states of a Factor Oracle structure [DAC07a]. Data drawn from cepstral analysis is employed as a feature vector describing each 2048-sample window of the audio file. Similar vectors indicate timbrally similar frames of audio that may be suitable for creating artificial transitions. Using this information, the Factor Oracle is constructed to allow for shuffling that will generate a sequence of windows with timbral patterns that are similar to (but distinct from) the original file. For instance, if the original audio file had the following timbral sequence of events: vibraphone, oboe, vibraphone,

piano, vibraphone, piano, a reshuffled version resulting from a traversal of the Factor Oracle would have a timbrally similar sequence. In the resulting audio file, instances of vibraphone followed by piano would be more frequent than instances of vibraphone followed by oboe because there are simply more cases of the former in the original sound file. Compact cepstral information makes this process possible within reasonable computation times.

From Spectrum to Cepstrum

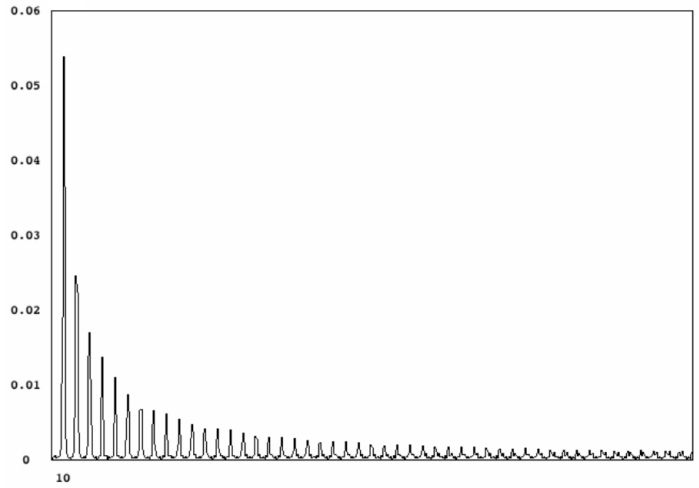


Figure 3.4: Magnitude spectrum of a 440 Hz sawtooth wave.

Compared to the results of spectral analysis, cepstral information is not as easily understood on an intuitive level. An example illustrating a transition between the frequency and quefrency domains will aid further discussion. For maximum clarity, this example will be based on a synthesized (and thus noise-free) signal with harmonic content: a 440 Hz sawtooth wave. Figure 3.4 shows its magnitude spectrum at a sampling rate of $sr = 44.1$ kHz, and window size $N = 1024$. The fundamental is thus centered around the 10^{th} bin, or $10\frac{sr}{N}$ Hz.

The real cepstrum (x_{RC}) is defined in [QT79] as

$$x_{RC}(n) = \Re[\, IFT\{ln\,|X(k)|\}\,] \tag{3.12}$$

where $X(k)$ is the frequency domain representation of a signal $x(n)$, and $\Re$ denotes the real portion of the inverse Fourier transform. Once a magnitude spectrum has been computed, the only remaining steps are to convert to a log scale and take the real portion of an inverse Fourier transform (IFT). The real cepstrum is sometimes equivalently defined as the real part of the *forward* Fourier transform (FT) of the logarithm of magnitude spectrum. At the most basic conceptual level, a cepstrum is the Fourier transform of the Fourier transform of a signal. This is why it is commonly referred to as the "spectrum of a spectrum." Rather than a time domain signal, the argument passed to the final Fourier transform in the algorithm is simply a log-frequency domain spectrum.

Quefrency Peaks

As in a spectral plot, the clearly pitched components of a signal can be easily located in the cepstrum. However, because cepstral analysis is performed on frequency domain information, the bins of a cepstral plot are connected to frequency in a different manner. Referring back to the magnitude spectrum of Figure 3.4, the peak corresponding to the fundamental frequency f of the sawtooth wave occurs at bin 10. Viewed as a signal, the evenly spaced harmonic peaks that follow can themselves be considered a consistent frequency in the spectrum. These peaks are graphically very similar to the waveform of an impulse train with decreasing amplitude. It is this "impulse train" that causes a cepstral peak. Locating a peak based on f requires a few simple operations. In the cepstral domain, bin indexes no longer refer to harmonics of the fundamental frequency of analysis; rather, they are indexes of time. A 440 Hz signal $f = 440$ is represented in time as $p = \frac{1}{f} = \frac{1}{440} = 0.0022727$ seconds, which is the period, or time required for one cycle to occur. Consequently, there should be a cepstral peak at the point along the quefrency axis corresponding to 2.2727 milliseconds. Where spectral data shows frequency, cepstral data shows period. The usefulness of the term quefrency in the place of period is debatable. At the very least it serves as a reminder that the cepstrum does not represent the conventional time domain, but something different altogether.

The quefrency bin for 440 Hz will be a time index, where each bin represents $\frac{1}{sr} = \frac{1}{44100} = 0.000022676$ seconds. To find the quefrency bin q of p seconds, p can simply be divided by the time duration of a single sample:

$$q = \frac{\frac{1}{f}}{\frac{1}{sr}} = \frac{sr}{f} = \frac{44100}{440} = 100.23 \tag{3.13}$$

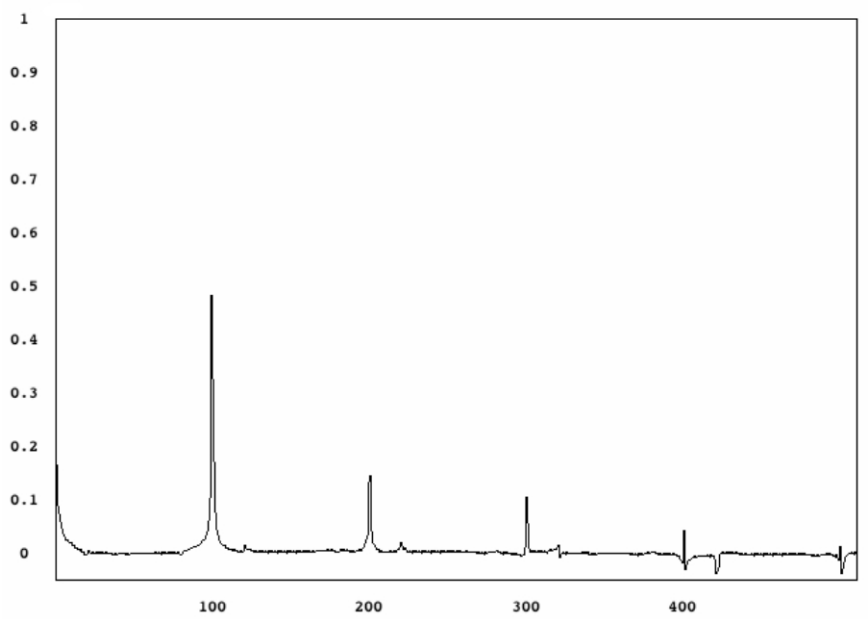


Figure 3.5: A cepstral quefrency peak resulting from a 440 Hz sawtooth wave.

In the current example, this points to bin 100, and because this is an idealized analysis, the peak is extremely well defined. If the signal's frequency content is not known in advance, it can be determined based on Equation (3.13) as well:

$$f = \frac{sr}{q} \tag{3.14}$$

Notice that the number of samples, N (which is crucial for calculating actual frequency based on spectral bin number, or vice versa) is not a part of this equation. If the sampling rate sr is fixed at 44100, cepstral bin 100 will always correspond to a frequency of roughly 440 Hz—regardless of whether N represents 1024, 2048, or 4096 samples. The ability to see the cepstral peak, however, will certainly be affected by N. Just as the Nyquist frequency is related to the highest

frequency representable by a FT, $\frac{N}{2}$ is related to the index of the highest quefrency (the longest period, or lowest frequency) that can be represented by the cepstrum. Cepstral bins beyond this point contain values that are symmetrical with the first half.

Using Equation (3.14), the frequency corresponding to the next cepstral peak at index 200 can be determined, where $f = \frac{44100}{200} = 220.5$. Peaks at indices 300 and 400 follow, which correspond to frequencies of roughly 110 and 55 Hz respectively. The original 440 Hz sawtooth wave did not possess energy at any of these frequencies. Referring once again to Figure 3.4, their presence can be explained. The "impulse train" signal that is the magnitude spectrum of a sawtooth wave is not sinusoidal, and therefore must contain partials of some sort. These cepstral harmonics cannot be assumed to relate to the original signal in any meaningful way. Therefore, when looking for the fundamental pitch of a harmonic signal, only the first cepstral peak is likely to be useful.

Combined Signal Components and Homomorphic Filtering

The concept of deconvolution comes from a class of homomorphic systems proposed by Oppenheim in 1967. According to Oppenheim, homomorphic systems "satisfy a generalization of the principle of superposition; i.e., input signals and their corresponding responses are superimposed (combined) by an operation having the same algebraic properties as addition." [OS89, p. 768] Thus, the fundamental frequency of a signal can be distinguished from its general spectral envelope, such that the convolution $f(n) * g(n)$ becomes a summation. In the case of speech, $f(n)$ is the vocal tract and $g(n)$ the glottal impulse. The move from multiplication-based convolution to summation based convolution can be expressed:

$$x(n) = f(n) * g(n)$$

$$X(k) = F(k)G(k)$$

$$ln\,|X(k)| = ln\,|F(k)| + ln\,|G(k)|$$

Then, using the definition of real cepstrum given in (3.12), after taking the real result of the IFT of each term, the final step above can be represented as

$$x_{RC}(n) = f_{RC}(n) + g_{RC}(n)$$

The cepstrum of the entire signal is the sum of the corresponding filter and impulse cepstra. This is the advantage of moving into the logarithmic domain—filter and impulse information can be combined or separated through simple addition or subtraction.

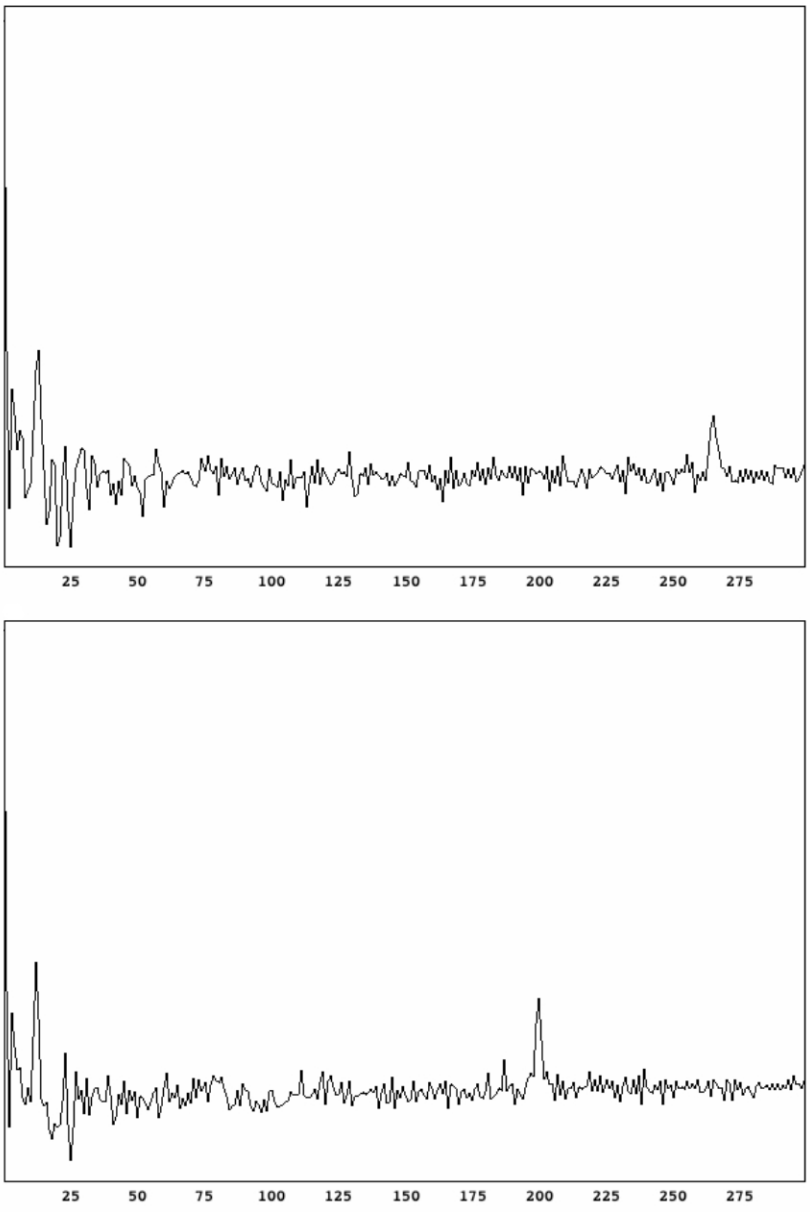


Figure 3.6: Quefrency peaks resulting from 165 Hz (top) and 220 Hz (bottom) sung vowels.

A demonstration of this concept will be most clear using vocal signals. Figure 3.6 shows cepstra based on recordings of a singer articulating the same voiced vowel at two different frequencies: 165 Hz and 220 Hz (or E_3 and A_4). The most striking difference between these cepstra is the location of quefrency peaks, which can be found at the following bin numbers: $\lfloor \frac{44100}{165} \rfloor = 267$ and $\lfloor \frac{44100}{220} \rfloor = 200$.

Overall, both have a great deal more noise than the sawtooth cepstrum from Figure 3.5, but the deconvolution is reasonable because $f_{RC}(n)$ and $g_{RC}(n)$ clearly occupy different quefrency ranges. The human voice is one of the strongest examples of the deconvolution phenomenon, and certain instruments will not give such clear results. There are no standard methods for locating an ideal upper bin limit for the region associated with $f_{RC}(n)$. Here, cepstral coefficients below the 25^{th} bin are similar in both plots, with a prominent peak around bin 15.

It can be seen that as frequency moves higher, the quefrency peak is situated *lower*. This is because cepstral analysis reveals rates of change in the log magnitude spectrum of a signal. The spectral fluctuation of $g(n)$ (the impulse component) is fast and periodic. As explained above, the quefrency peak generated by $g(n)$ is the result of evenly spaced harmonic peaks in the frequency domain. These spectral peaks (and the valleys between them) constitute periodic spectral fluctuation. Lower frequencies in the time domain cause higher periodic rates of spectral change, which is why the 165 Hz quefrency peak from Figure 3.6 is located higher along the quefrency axis than the 220 Hz peak. Higher frequency corresponds to lower period. On the other hand, the first coefficients reflect the slowest rates of spectral change, i.e., spectral envelope. The cepstrum of $f(n)$ represents the filter component, which varies very slowly in the spectral domain. It has a relatively inactive spectral signature that affects the timbre of impulses that are convolved with it (synthetically or naturally). In terms of quefrency, its variations per second are slow, and therefore information reflecting its shape will always be located at the low end of the cepstrum.

Figure 3.7 displays the spectra of two sawtooth waves at 440 Hz and 880 Hz. Fluctuations in the spectrum of the 440 Hz tone are much more frequent (i.e., they are more densely packed in terms of frequency) than those of the 880 Hz tone. The spacing between harmonics increases as a function of frequency. Spectrally, signals with higher fundamental frequencies will vary less rapidly. If these spectral variations are slow enough, the resulting quefrency peak will overlap with the area containing spectral envelope information in the cepstrum. This will make both $f_{RC}(n)$ and $g_{RC}(n)$ less reliable. In Figure 3.6, however, $f_{RC}(n)$ and $g_{RC}(n)$ do not

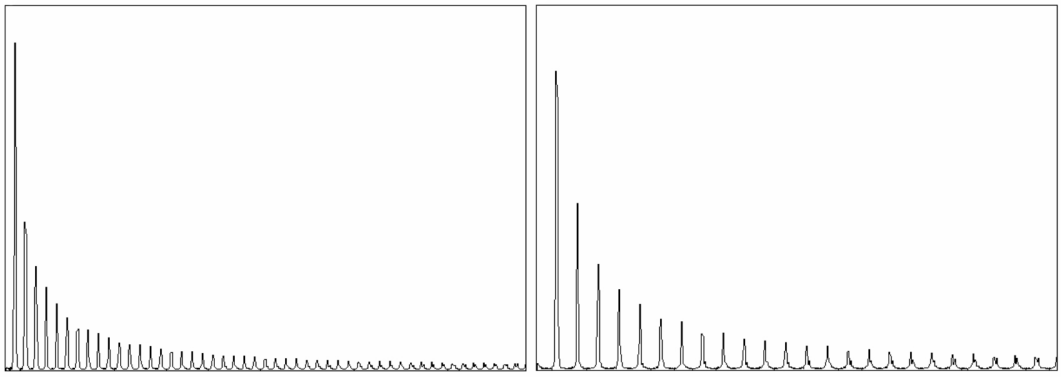


Figure 3.7: Magnitude spectra for 440 Hz (left) and 880 Hz (right) sawtooth waves.

overlap. Regardless of the frequency of the pitched component, the lowest cepstral coefficients in the two voice recordings are clearly very similar. Figure 3.8 provides a detailed view of this information.

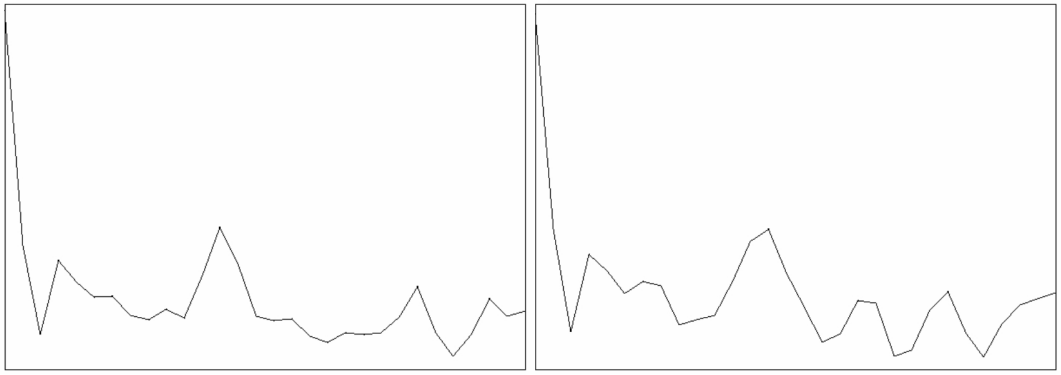


Figure 3.8: Cepstral coefficients 1 through 30 for a voiced vowel sung at 220 Hz (left) and 165 Hz (right).

Based on the first 30 coefficients alone, it is possible to verify that these two cepstra correspond to sounds with very similar timbral characteristics. By analyzing the signals of different sung vowels and compiling a database, it is also possible to discriminate between vowels based on cepstrum, independently from pitch. More specifically, an estimation of the timbral similarity of any two instances in the database can be calculated based on the Euclidean distance d, defined as

$$d = \sqrt{\sum_{n=0}^{N-1}(v_n - w_n)^2} \tag{3.15}$$

where N is the length of vectors v and w—the lower cepstral coefficients from two distinct cepstra. Based on the values of d between database instances, a similarity threshold can be chosen in order to classify new instances. The distances between an input signal's cepstrum and each instance in the database can be compared to find the closest match. With a large enough database of training examples, the events of a performance can be classified and tracked based on timbre in order to control immediate or large scale processes in electroacoustic accompaniment. In the case of inharmonic or noise-based percussion instruments, a real-time cepstral analysis tool can be used as a functional replacement for pitch tracking in score following applications, or for automatic performance transcription.

3.2.2 Mel Frequency Cepstrum

Having illustrated the process of basic cepstral analysis, its perceptually-weighted variants can now be considered. The process for computing Mel Frequency Cepstral Coefficients (MFCCs) differs from raw cepstrum computation considerably. It requires a bank of triangular overlapping bandpass filters evenly spaced on the mel scale, and the final transform is a discrete cosine transform (DCT) rather than a Fourier transform. The mel scale is based on a 1937 experiment exploring perceptual pitch relationships between tones [SVN37]. Using the experimental data of 5 participants, the authors hoped to discover a frequency unit that could be manipulated arithmetically yet remain observationally verifiable. This unit was named the mel in reference to melody. For any particular mel value, one should be able to double it, then convert both the original and doubled values to a frequency scale and confirm through experiment that the doubled mel frequency is judged to be twice as high in terms of pitch. Likewise, halving or tripling a mel value should lead to appropriately scaled perceptual results.

The experiment itself consisted of sessions in which participants were presented with two different tones produced by identical oscillators at an even loudness

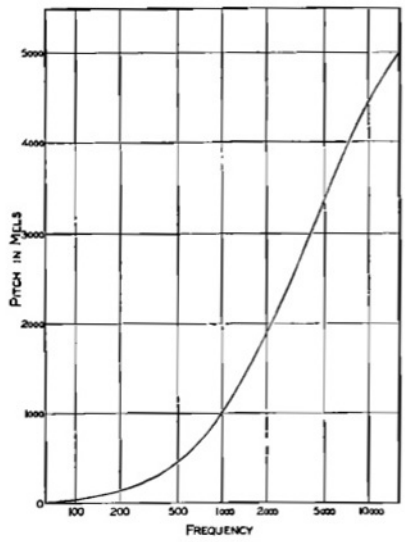


Figure 3.9: Hz plotted against mel units, from [SVN37].

level of 60 dB. The pitch of one oscillator was fixed, but the other could be varied manually. Participants were instructed to adjust the second oscillator until it was half as high in terms of pitch as the reference tone [SVN37, p. 187]. This process was carried out for reference tones at frequencies of 125, 200, 300, 400, 700, 1000, 2000, 5000, 8000, and 12000 Hz. The geometric mean of the participants' half-pitch judgements at each of these frequencies was taken and used to construct the curve shown in Figure 3.9. The arbitrarily chosen point of intersection between frequency and mels is at 1000 Hz/mels. Even mel spacing beyond this intersection translates to increasingly large spacing in Hz. Thus, warping the frequency axis of a spectrum according to an evenly spaced mel scale places more weight on lower frequency values. The general formula for calculating mels is

$$mel = 1127.01048\ ln(1 + \frac{f}{700}) \tag{3.16}$$

where f is frequency in Hz [Ber49].

Mel scaling significantly reduces the size of spectral envelope data and emphasizes lower frequency content. The extent of reduction depends on sampling rate, window size, and the mel spacing of the filterbank. The lower limit of the fil-

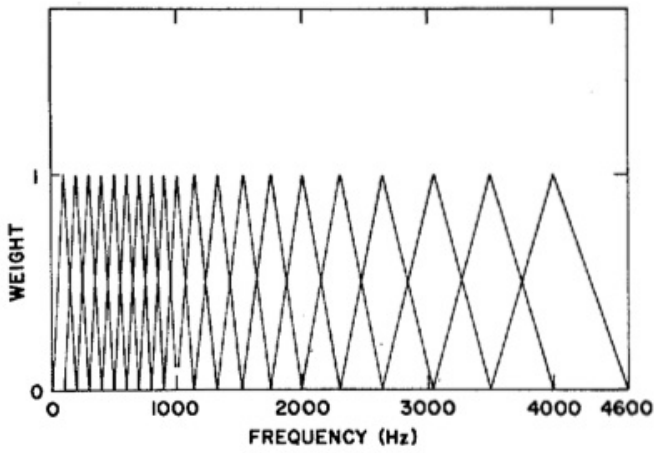


Figure 3.10: A mel-spaced triangular filterbank, from [DM80].

terbank is DC, and the upper limit should not reach the Nyquist frequency. Using Equation (3.16), the Nyquist frequency at a sampling rate of 44100 is calculated as 3923 mels. With an even spacing of 200 mels, this produces 22 mel values below Nyquist, which correspond to 20 overlapping filters (the first and last mel values are the lower and upper bounds of the first and last filters respectively). Figure 3.10 illustrates such a filterbank. Multiplying a spectrum against this filterbank compresses the first 512 bins of a 1024 point window into a smoothed 20 point estimation of the spectrum with a weighting based on the mel scale. Accordingly, Figure 3.10 clearly shows that higher frequency content is averaged over much larger bin ranges than lower frequency content. [DM80] and [RJ93] define MFCCs mathematically as

$$MFCC_i = \sum_{k=0}^{N-1} X_k \; cos[i(k + \frac{1}{2})\frac{\pi}{N}]; \;\; i = 0, \ldots, M-1 \tag{3.17}$$

where M is the desired number of cepstral coefficients, N is the number of filters, and X_k is the log power output of the k^{th} filter. [Log00] specifies log amplitude rather than power.

The DCT comprising the final step of the MFCC computation is the other fundamental difference from raw cepstrum. [Log00] proposes that the DCT ap-

proximates decorrelation obtained through Principal Component Analysis (PCA), and concludes that the use of the mel scale in music classification is "at least not harmful ... although further experimentation is needed to verify that this is the optimal scale for modeling music in the general case." [Log00, p. 8]

3.2.3 Critical Bands and the Bark Scale

In place of mels, a filterbank can be constructed using a scale based on critical bandwidths. Critical bands refer to regions of the basilar membrane that are stimulated by unique frequency ranges. An overview of multiple experiments establishing the boundary and center frequencies of critical bands is given by Zwicker, Flottorp, and Stevens in [ZFS57]. It is acknowledged that the published critical band boundaries are not fixed according to frequency, and depend upon specific stimuli. Relative bandwidths are more stable, and repeated experiments have found consistent results. In frequency, these widths remain more or less constant at 100 Hz for center frequencies up to about 500 Hz, and are proportional to higher center frequencies by a factor of 0.2. The uppermost curve in Figure 3.11 from [ZF90] shows this characteristic, while the curves beneath it provide useful points of comparison. Just-noticeable frequency difference is plotted against logarithmic frequency in the bottom curve, and the dashed curve in the middle plots the difference in frequency required to advance the point of maximum stimulation of the basilar membrane by 0.2 mm. Each of the curves can be shifted vertically to produce very close alignment with the others. Their similarity suggests that a scale based on documented critical bandwidths has physiological as well as perceptual validity. In 1960, Zwicker's letter to the editor in the Journal of the Acoustical Society of America introduced the Bark as a unit based on critical band boundaries, named after the inventor of the unit of loudness level: Barkhausen. The frequency boundaries it presents are neatly rounded versions of the values found by loudness summation experiments in [ZFS57], and are now standard reference values.

Unlike the mel scale, the Bark unit stands upon a large foundation of evidence. As Zwicker et al. put it, the "critical band has the advantage ... that it does not rest on assumptions or definitions, but is empirically determined by at least four

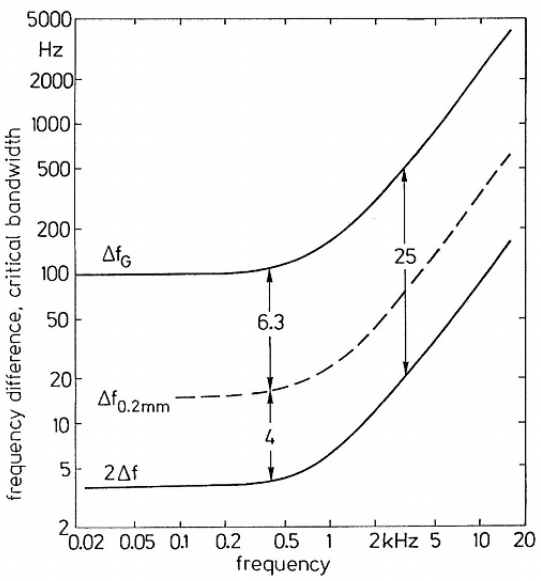


Figure 3.11: Critical bandwidths and related units vs. frequency, from [ZF90].

kinds of independent experiments." [ZFS57, p. 554] The four unique strategies for locating critical band boundaries that he refers to are threshold, masking, phase, and loudness summation. The latter is documented in most detail. Participants in the 1957 experiment were asked to match loudness between single tones and multiple-tone complexes of varying frequency width Δf. Modulating values of Δf within a frequency-dependent critical bandwidth did not affect participants' loudness judgements, but increasing Δf beyond this bandwidth resulted in increased loudness. The points in frequency at which such loudness increases occurred were correctly predicted according to proposed critical band boundaries. It was also determined that different spacings of frequencies within the complex tones—which do not affect Δf—produced unique effects. With even spacing of individual tones, loudness was higher than for spacings that bunch tones closer to either boundary of Δf. This appears to be related to the stimulus-specific nature of critical band boundaries. Also worthy of note is that the loudness summation method was not effective for locating boundaries when general loudness levels were just above the threshold of hearing. Zwicker et al. suggest that the phase based strategy is most appropriate for measurements conducted at very low loudness levels (i.e. below 20 dB SPL).

Detecting critical band boundaries through changes in phase relies on the similarity of sidebands in tones synthesized through low levels of amplitude and frequency modulation, and the fact that they significantly differ only in terms of just noticeable modulation rate, and phase—one of the FM sidebands will be 180° out of phase with its AM counterpart. Our hearing system is able to detect this difference at very low modulation rates, meaning that we are sensitive to changes in sideband phase. As modulation rates increase, however, we are unable to distinguish between the techniques (i.e., the associated phase differences are no longer noticeable), and the just detectable degree of AM and just detectable index of FM are the same. In [Zwi52], just detectable levels of modulation were measured for four participants using a collection of various carrier and modulation frequencies, and (for any given carrier frequency) critical bandwidth was taken to be twice the modulation frequency at which AM and FM became indistinguishable [ZFS57, p. 556].

In the case of masking, studied in [Zwi52], a small band of noise is placed between two tones. At very low noise sound pressure levels, the tones mask the noise. As the tones are more widely spaced in frequency, the sound pressure level at which the noise ceases to be masked remains constant until a particular tone spacing is reached, where the masking ceases at significantly lower levels [ZFS57, p. 555]. When the noise and tones are processed within separate critical bands, masking effects are decreased. The frequency spacing at which this occurs relates to the critical band.

Finally, the threshold method performed in [Gas54] tracks the way in which overall sound pressure related to the threshold of an evenly spaced tone complex varies in relation to the number of tones in the complex [ZFS57, p. 555]. Starting with a single tone and progressing with the addition of tones spaced 10 Hz apart (moving downward), threshold is repeatedly measured. A pattern is observed with respect to the the number of tones present in the complex and their appropriate individual amplitudes. For instance, with a single tone, threshold is recorded as +3 dB, while with two and four tones in the complex, each individual tone only requires 0 dB and -3 dB respectively for the complex to reach threshold as a

whole. Thus, a consistent pattern can be seen as the amplitude of individual tones decreases and the number of tones in the complex increases. But when a certain number of tones is reached, the pattern does not continue as expected [ZFS57, p. 555]. This transition point in frequency is taken to be a critical band boundary. [ZF90] describes another instance of this type of experiment, where—starting with a single tone at 960 Hz—additional tones were spaced 20 Hz apart moving upwards [ZF90, p. 134]. The same pattern was observed.

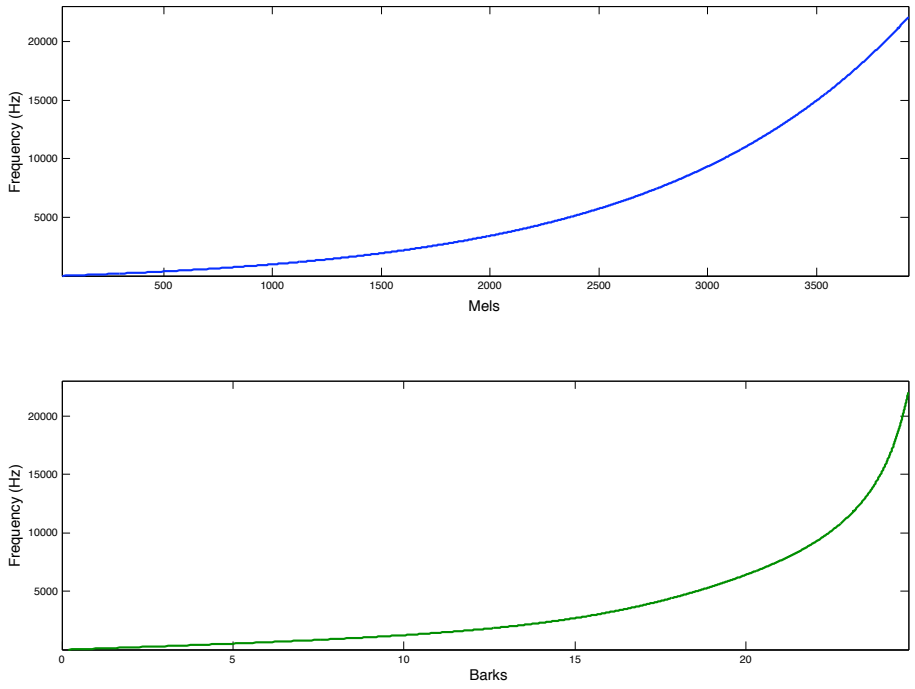


Figure 3.12: Mels (top) and Barks (bottom) plotted against linear frequency.

Despite their difference in terms of verification by independent experiments, several sources note that Barks relate very strongly to mels [Zwi60][ZF90][RMW02], the rough guide being that multiplying Barks by 100 produces a curve similar to the mel scale. The two curves are shown in Figure 3.12 plotted against linear frequency. At 3000 mels and about 19 Barks, it can be seen that the Bark scale maps to lower linear frequencies near the higher end. That is, the Bark scale ramps up to higher frequencies more suddenly than the mel scale. In the context of mel- and Bark-weighted cepstra, this means that the Bark weighting will preserve more mid

frequency detail at the expense of high frequency detail. The highest mel values determined directly by experiment correspond to 5, 8, and 12 kHz; higher values are projected based on Equation (3.16). Likewise for Barks, since there are a fixed number of critical bands that correspond to the 24 Barks, values at arbitrary subdivisions between boundaries or beyond the 24^{th} Bark must also be calculated with a general formula. Equation (3.18), taken from [Tra90], is given below, where f is frequency in Hz:

$$Bark = [26.81\frac{f}{1960 + f}] - 0.53 \tag{3.18}$$

As the two scales are quite similar, the distinction between mel- and Bark-weighted cepstrum is likely to be negligible. This hypothesis is examined in Chapter 5.

3.3 Interpreting BFCCs

The compact representation of spectral envelope offered by BFCCs requires some additional explanation in order to be understood on an intuitive level. Each coefficient reflects a correlation between a Bark-weighted spectrum and a particular basis function of the cosine transform. Applied in Equation (3.17), the cosine transform (DCT-II) is defined as

$$X_i = \sum_{k=0}^{N-1} x_k \; cos[i(k + \frac{1}{2})\frac{\pi}{N}]; \quad i = 0, \ldots, M-1 \tag{3.19}$$

where $x_{0\ldots N-1}$ is the input, and M represents the desired number of coefficients ($M \leq N$). Each Bark cepstral coefficient is simply the dot product between the input and a basis function. The basis functions for iterations 1—6 of the algorithm are given in Figure 3.13. When the input $x_{0\ldots N-1}$ is a Bark-weighted spectrum, each cepstral coefficient measures the degree of correlation between the spectrum and each basis function.

The 1^{st} basis is a vector of ones, so the value of the 1^{st} BFCC is simply the sum of all Bark spectrum magnitudes multiplied by one. If spectrum magnitudes

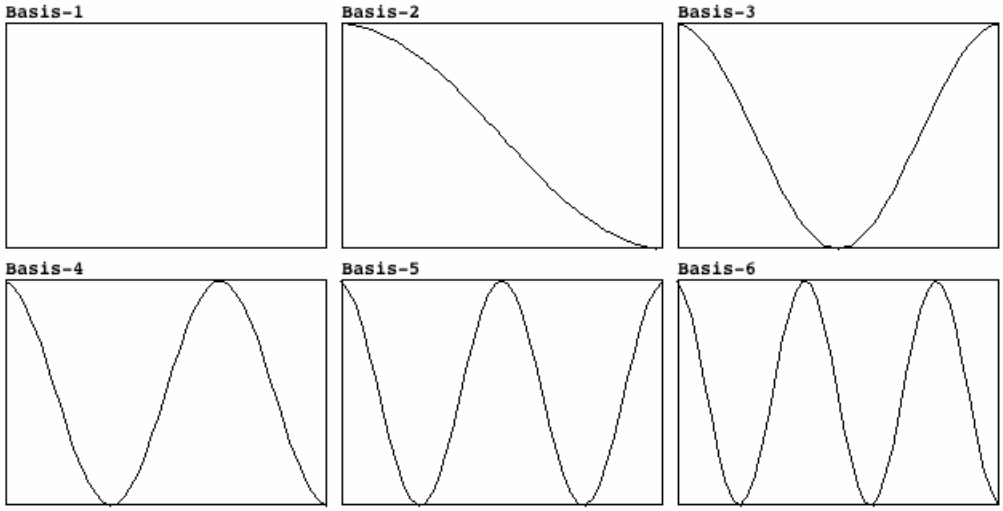


Figure 3.13: The first six cosine transform basis functions.

are not normalized before the DCT, this coefficient will reflect the amplitude of the time domain signal. Otherwise, it will always be 1.0, and of no use in classification tasks. The 2nd basis is a half cosine reaching from 1.0 to -1.0. In the unlikely case that a Bark spectral envelope perfectly resembles the contour of a half cosine, the high correlation will cause the 2nd BFCC to have a maximal value. Typical Bark spectra will not possess this exact contour, but spectra that bear a strong resemblance to the shape of the basis function will produce relatively high values for this BFCC. Musical instruments having a concentration of low- and mid-frequency energy that rolls off gently fit this general class of spectrum. As the spectral centroid of a sound moves higher, its spectrum will most likely resemble the 2nd basis function less and less. Thus, there is a strong negative correlation between the 2nd BFCC and spectral centroid. The 3rd basis dips to -1.0 at the center, so that large negative values would result from taking the dot product between this basis function and a spectrum with a symmetrical peak in the central Bark-frequency bands.

In this same fashion, the value of each additional BFCC can be conceptualized as a reflection of the degree of similarity between a Bark spectrum and the contour of a basis function. Moving higher among these functions, an increasing number of peaks and valleys occur, so that the complete set of basis functions

probes for information in all frequency regions with a resolution that depends on N. If all BFCCs are used in a distance calculation such as Equation (3.15), the resulting value of d will be identical to that produced when using Bark spectrum coefficients. This is because the cosine transform amounts to a rotation in space that preserves distances between points. If the DCT basis functions have probed as many areas of the spectrum as permitted by the resolution of the transform, there is no fundamental difference between the two forms of information. Thus, the power of the Bark-frequency cepstrum is the flexibility it offers for choosing reduced but sufficient resolution in terms of probing for crude spectral envelope shapes.

3.4 Summary

This chapter has presented two classes of algorithms for computing numerical representations of timbre. Low level features reduce high-dimensional spectral information to single numbers that reflect various aspects of energy distribution. The list of low level features given here is far from exhaustive—only those that have been implemented in the system described in the following chapter were described. While these types of measurements are extremely meager, in combination, they provide a compact description of a sound's spectral information. High level features achieve similar ends. The least processed of these features is a mel- or Bark-weighted spectrum that drops a large amount of data in frequency ranges that are less relevant perceptually. Coefficient subsets from the different types of cepstral analysis discussed measure relationships between a spectral envelope and the basis functions of Fourier and cosine transforms. Like the various individual low level features, these correlations can provide rough information about a spectrum. In a sense, the logic behind a feature vector composed of several low level features is very similar to the strategy of using a subset of BFCCs. It is not obvious that correlations with cosine transform basis functions are any more meaningful or useful than statistical measures of energy distribution. To understand the implications of these different strategies, it is necessary to examine particular sound

sets in detail.

3.5 Acknowledgements

Portions of this chapter were previously published in the following articles:

William Brent, "Cepstral Analysis Tools for Percussive Timbre Identification", *Proceedings of the 3rd Pure Data Convention, São Paulo, Brazil*, 2009.

William Brent, "Perceptually Based Pitch Scales in Cepstral Techniques for Percussive Timbre Identification", *Proceedings of the International Computer Music Conference*, 2009.

Chapter 4

timbreID

4.1 Introduction

There are relatively few general analysis toolkits for retrieving and comparing sonic timbre descriptors. MARSYAS [TC99], and MIRToolbox [LT07] are currently the most prominent. Tzanetakis & Cook's open source MARSYAS project adopts a client-server architecture, where signal analysis and pattern recognition modules perform tasks independently of a Java-based graphical user interface. The software allows users to perform a variety of tasks relative to large audio databases, including scanning through automatically segmented audio, searching for similar sounds, and classifying audio regions as either music or speech. MIRtoolbox, developed by Lartillot & Toiviainen, is an open source package of MATLAB functions and scripts. In comparison with MARSYAS, MIRtoolbox has the advantage of being developed for an environment with which researchers are already familiar, and that possesses several well-established statistical and machine learning packages for further manipulation of audio features. The authors designed MIRtoolbox functions with consideration for ease of use in graphing and batch processing tasks.

Though powerful, these packages are not immediately accessible for general creative use because they either require users to adapt to a new software interface, or assume familiarity with programming and scripting in MATLAB—an environment designed for scientific research rather than creative work. Further, they are only capable of analysis in non-real-time. At present, the most popular

programming environments used by musicians are Max/MSP, Pure Data (Pd), SuperCollider, and Csound. For some of these platforms, open source projects have been developed for the purpose of organizing sounds or querying an audio corpus based on timbral similarity. This work has enabled real-time control of sounds for live performance projects, instrument design, and broad creative exploration. CataRT and Soundspotter are the most widely recognized projects of this sort [SBVB06][CG07]. The former is available as a Max/MSP implementation, while the latter is intended for multiple platforms—including Pd. Soundspotter's Pd realization is primarily designed for real time target-driven concatenative synthesis, and all analysis takes place within a single pre-configured object. More general tools for creative work centered on timbre similarity are limited in Pd.

timbreID is a Pd external library developed by the author. It is composed of a group of objects for extracting audio features, and a classification object that manages the resulting database of information. The objects are designed to be easy to use and adaptable for a number of real-time purposes, including the generation of synthesis control streams, timbre identification, ordering of sounds by timbre, target-driven concatenative synthesis, and plotting of sounds in a user-defined timbre space that can be auditioned interactively. Just as MIRtoolbox benefits from being embedded within powerful research software, timbreID's situation within Pd creates opportunities for interfacing with many familiar real-time control and DSP objects in an environment that is applied, developed, and supported by a community of musicians. This chapter will summarize the most relevant features of the timbreID toolkit and give a sense of its flexibility by describing several different applications.

4.2 Feature Extraction Objects

In general, timbreID's feature extraction objects have four important qualities. First, each object maintains its own signal buffer based on a user-specified window size. This eliminates the need for sub-patches in Pd to set window size using the block~ object. Second, Hann windowing is automatically applied within

each object so that input signals do not need to be multiplied against a window table using the tabreceive~ object. Third, under certain circumstances, analysis timing is sample-accurate. All of timbreID's analysis objects measure the logical time between the beginning of each DSP block and incoming analysis requests, so that the desired slice of audio can be captured regardless of Pd's default 64-sample block boundaries. Thus, there is no need to set overlap values with block~ in order to define a particular time resolution.[1] Fourth, because the objects perform analysis on a per-request basis, the only computational overhead incurred during periods of analysis inactivity is that of buffering. Combined, these four qualities make signal analysis in Pd straightforward and accessible.

As a point of comparison, consider the case of capturing magnitude spectrum by chaining together the standard Pd objects rfft~, *~, +~, and sqrt~. Placed in a sub-patch with block~ at a window size of 1024 and overlap of 4, such a network will calculate a 1024 point FFT every 256 samples[2] whether or not the information is needed at any particular time. If finer time resolution is desired, block~'s overlap setting must be changed to a higher value, resulting in an even greater number of FFTs per second.

The magSpec~ object from timbreID will not perform an FFT until the user requests analysis results. Overlapping analyses can be obtained by simply connecting a metro object to magSpec~ and setting it to a rate that is smaller than the analysis window duration in milliseconds. If a request occurs at a logical time that falls between two DSP block boundaries, the slice of audio analyzed by magSpec~ will end at precisely the moment the request was made, unrestricted by Pd's default blocking resolution. For applications such as real-time classification of percussive instruments with a delay of less than 20 ms, efficiency and precision timing are critical. The feature extraction objects for timbreID were designed to meet these needs.

[1]However, only certain objects are presently capable of generating bang messages that occur between block boundaries. When clicked by a user, the graphical bang object sends its message at the start of the following DSP block, and analysis window capture will be delayed. Bang messages sent by Pd's metro or delay objects, however, *are* transmitted between block boundaries. Thus, the capacity for true sample accuracy depends upon the object initiating the analysis request.

[2]This is about 172 times a second at a sampling rate of 44.1 kHz

4.2.1 Available Features

In spite of the fact that many timbre features are very simple to calculate, most are not generated as easily as magnitude spectrum using Pd's built-in objects. For instance, based on a spectrum $|X|$, spectral flatness is calculated according to Equation (3.7) as the geometric mean of $|X|$ divided by the arithmetic mean of $|X|$. Since this operation would require writing signal blocks to graphical arrays and traversing the arrays in order to calculate the two means, it is more appropriately implemented as an external object written in C. Other features, like MFCCs, would be even more complicated to implement at the patching level.

The following objects for measuring basic features are provided with timbreID: magSpec~, specBrightness~, specCentroid~, specFlatness~, specFlux~, specIrregularity~, specKurtosis~, specRolloff~, specSkewness~, specSpread~, and zeroCrossing~. The higher level features in the set—generated by barkSpec~, cepstrum~, mfcc~, and bfcc~—are generally the most powerful for classification. The implementation details of some feature extraction objects differ slightly from the mathematical definitions provided in the previous chapter. For instance, MFCCs and BFCCs are calculated based on normalized magnitude spectrum rather than log power spectrum as specified in Equation (3.17). Although an understanding of the various analysis techniques is not required for use, a general idea of what to expect can be very helpful. To that effect, a simple demonstration and straightforward explanation of each feature is given in its accompanying help file. In order to facilitate as many types of usage as possible, non real-time versions of all feature externals are provided for analyzing samples directly from graphical arrays in Pd.

4.2.2 Open-ended analysis strategies

Independent, modular analysis objects allow for flexible analysis strategies. Each of the objects reports its results as either a single number or a list that can be further manipulated in Pd. Feature lists of any size can be packed together so that users can design a custom approach that best suits their particular sound set. Figure 4.1 demonstrates how to generate a feature list composed of MFCCs,

spectral centroid, and spectral brightness. Subsets of mel frequency cepstral coefficients (MFCCs) are frequently used for economically representing spectral envelope [LEB03], while spectral centroid and brightness provide information about the distribution of spectral energy in a signal. Each time the button in the upper right region of the patch is clicked, a multi-feature analysis snapshot composed of these features will be produced.

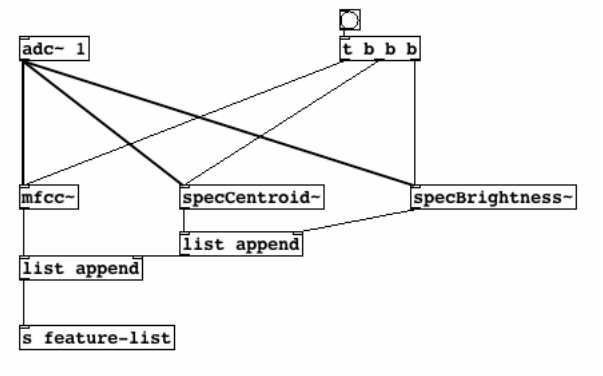


Figure 4.1: Generating a mixed feature list.

Capturing the temporal evolution of audio features requires some additional logic. In Figure 4.2, a single feature list is generated based on 5 successive analysis frames, spaced 50 milliseconds apart. The attack of a sound is reported by bonk~ [PAZ98], turning on a metro that fires once every 50 ms before turning off after almost a quarter second. Via list prepend, the initial moments of the sound's temporally-evolving MFCCs are accumulated to form a single list. By the time the fifth mel frequency cepstrum measurement is added, the complete feature list is allowed to pass through a spigot for routing to timbreID, the classification object described below in Section 4.3. Recording changes in MFCCs (or any combination of features) over time stores detailed information for the comparison of complex sounds.

These patches illustrate some key differences from the Pd implementation of libXtract, a well developed multi-platform feature extraction library described in [Bul07]. Extracting features in Pd using the libXtract~ wrapper requires subpatch blocking, Hann windowing, and an understanding of the libXtract's order of

operations. For instance, to generate MFCCs, it is necessary to generate magnitude spectrum with one object, then chain its output to a separate MFCC object. The advantage of libXtract's cascading architecture is that the spectrum calculation occurs only once, yet two or more features can be generated from the results.

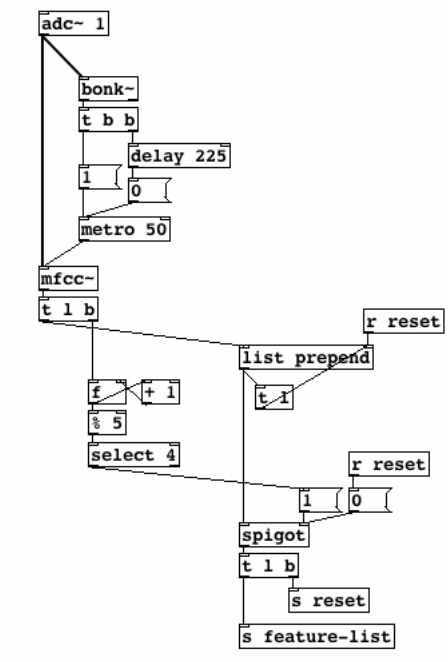


Figure 4.2: Generating a time-evolving feature list.

While timbreID objects are wasteful in this sense (each object redundantly calculates its own spectrum), they are more efficient with respect to downtime. As mentioned above, features are not generated constantly, only when needed. Further, from a user's perspective, timbreID objects require less knowledge about analysis techniques, and strip away layers of patching associated with blocking and windowing.

4.2.3 Details of Analysis Algorithms

Certain analysis techniques can be implemented with subtle differences. For instance, spectral flux is frequently defined as the sum of squared difference between successive frames of magnitude spectra [TC02]. When the change in any

particular bin has a magnitude less than 1, squaring the difference can result in very small values. In that case, taking the absolute value might be a better strategy than squaring for keeping all values positive in the final summation. Additionally, as an alternative to the compact sum of differences, the complete list of bin-by-bin fluctuations could be very useful for tracking change in any particular frequency band. Accordingly, functions in the specFlux~ external report the sum of either the absolute value of bin flux or squared bin flux, as well as the complete list of raw flux values.

Like specFlux~, each feature extraction object has its own unique parameters that are explained in accompanying help files. For instance, all high level feature objects offer a spectrum normalization option that may not be appropriate in some analysis scenarios. In order to have maximum control over these types of details, all feature extraction and classification functions were written by the author, and timbreID has no non-standard library dependencies.

4.3 The Classification object

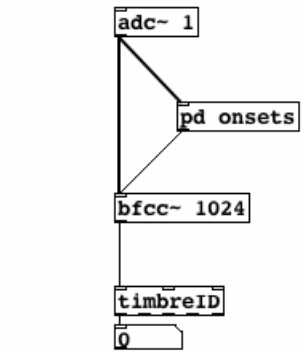


Figure 4.3: timbreID in a training configuration.

Features generated with the objects described in Section 4.2 can be used directly as control information in real-time performance. In order to extend functionality, however, a multi-purpose classification external is provided as well. This object, timbreID, functions as a storage and routing mechanism that can cluster

and order the features it stores in memory, and classify new features relative to its database. Apart from the examples package described in the following section, an in-depth help patch accompanies timbreID, demonstrating how to provide it with training features and classify new sounds based on training. Figure 4.3 depicts the most basic network required for this task.

Training features go to the first inlet, and features intended for classification go to the second inlet. Suppose the patch in Figure 4.3 is to be used for percussive instrument classification. In order to train the system, each instrument should be struck a few times at different dynamic levels. For each strike, an onset detector like bonk~ will send a bang message to bfcc~—the Bark-frequency cepstral analysis object. Once a training database has been accumulated in this manner, bfcc~'s output can be routed to timbreID's second inlet, so that any new instrument onsets will generate a nearest match report from timbreID's first outlet. A match result is given as the index of the nearest matching instance as assigned during training. For each match, the second outlet reports the distance between the input feature and its nearest match, and the third outlet produces a confidence measure based on the ratio of the first and second best match distances.

For many sound sets, timbreID's clustering function will automatically group features by instrument. A desired number of clusters corresponding to the number of instruments must be given with the "cluster" message, and an agglomerative hierarchical clustering algorithm will group instances according to current similarity metric settings. Afterward, timbreID will report the associated cluster index of the nearest match in response to classification requests.

Once training is complete, the resulting feature database can be saved to a file for future use. There are four file formats available: timbreID's binary .timid format, a text format for users who wish to inspect the database, ARFF format for use in WEKA[3], and .mat format for use in either MATLAB or GNU octave.

[3]WEKA is a popular open source machine learning package described in [HDW94]

4.3.1 timbreID settings

Nearest match searches are performed with a k-nearest neighbor strategy, where K can be chosen by the user. Several other settings related to the matching process can also be specified. Three different similarity metrics are available: Euclidean, Manhattan (taxicab), and Correlation. For feature databases composed of mixed features, feature attribute normalization can be activated so that features with large ranges do not inappropriately weight the distance calculation. This is accomplished by scanning the database for the extreme values of each attribute, then scaling according to the following formula:

$$N_j = \frac{F_j - min_j}{max_j - min_j} \tag{4.1}$$

where F_j is the j$^{\text{th}}$ attribute of a feature F, and max_j and min_j are the maximum and minimum values for that attribute column across the entire database. Specific weights can be dynamically assigned to any attribute in the feature list in order to explore the effects of particular proportions of features during timbre classification or sound set ordering. Alternatively, the feature attributes used in nearest match calculations can be restricted to a specific range or subset. Or, the attribute columns of the feature database can be ordered by variance, so that match calculations will be based on the attributes with the highest variance.

4.4 Applications

Further aspects of timbreID's functionality are best illustrated in context. This section describes six of the example patches that accompany the timbreID package.

4.4.1 Plotting Cepstrograms

Using the non-real-time magSpec and bfcc objects, spectrogram and cepstrogram plots can be created directly in Pd. The cepstrogram example patch

offers several options for analyzing audio loaded to the graphical sample array, such as window size, overlap, and spectrum normalization. When a plot request is initiated, each iteration of an until loop generates BFCC data that is written to a table, then displayed as a column of colored pixels using Pd's native struct and scalar objects. Just as spectrograms graph changing spectral information in a visually convenient way, cepstrograms can make cepstro-temporal patterns apparent that would be otherwise difficult to identify.

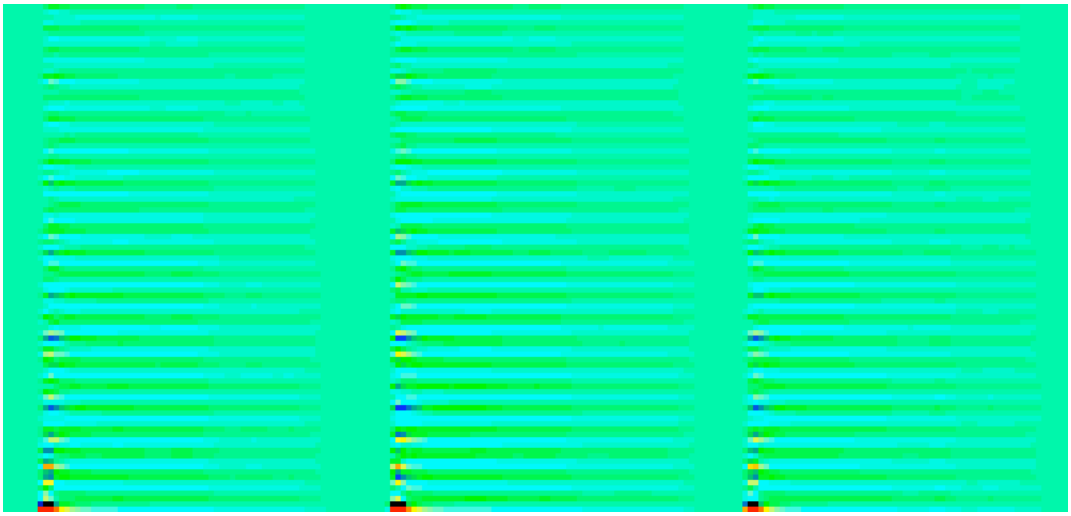


Figure 4.4: Cepstrogram of three glockenspiel tones.

Figure 4.4 shows three strikes of the same glockenspiel bar at roughly the same loudness. Spectrum normalization was disabled to eliminate changes due to noise. Pixels along the vertical dimension represent BFCCs 0—46 from bottom to top, with the color scheme indicating values of each coefficient. The lower and upper value limits are represented by black and red respectively. From left to right, the horizontal dimension represents time. In addition to illustrating consistent patterns over time as each note decays, this cepstrogram also reveals that the first few analysis frames of each tone have very consistent coefficient values from one instance to the next. It can be seen that patterns in the first 15 coefficients do not vary widely between the three tones, and will likely be sufficiently unique for accurate classification. In Figure 4.5, two quiet strikes of the same nipple gong are plotted. Here, temporal patterns near the attack emerge more slowly,

and are again unique and consistent. Thus, as a research tool, the cepstrogram example patch offers a convenient method for exploring the nature of sounds in the cepstral domain. Patterns identified using this system or the accompanying spectrogram patch can be exploited for more accurate and informed classification in performance-oriented patches.

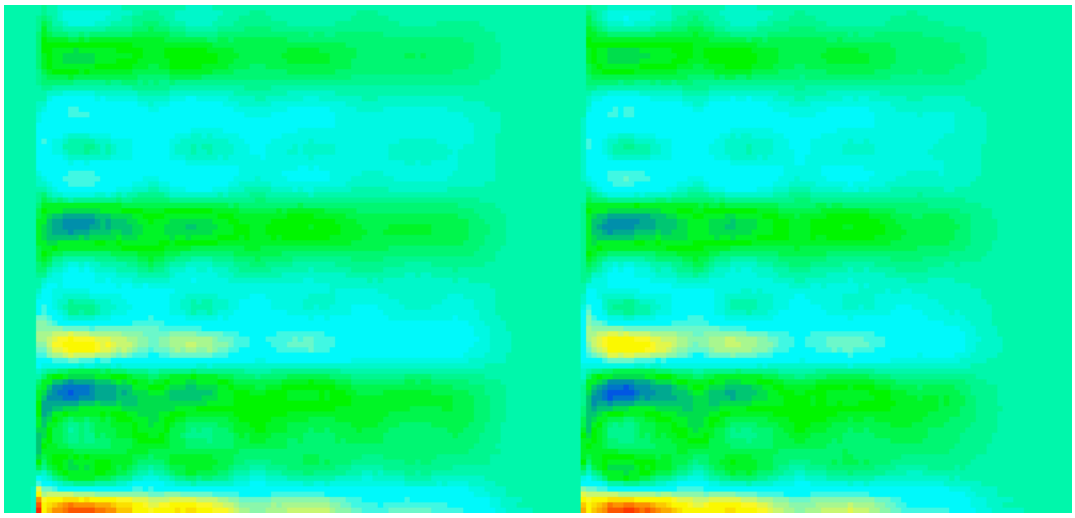


Figure 4.5: Cepstrogram of two nipple gong tones.

4.4.2 Percussive Instrument Recognition

Percussive instrument classification is already possible using bonk~ [PAZ98], but timbreID objects allow for customized analysis strategies and offer some additional options that can lead to increased accuracy. The instrument recognition example illustrates how to use classification information to create mappings between live sounds and a set of samples. Apart from onset detection and sample playback subpatches, Figure 4.6 shows the simple patch in its entirety. The training process briefly described in Section 4.3 applies here as well. After opening the spigot for routing bfcc~'s analysis information to timbreID's training inlet, 5—10 training instances should be provided for each instrument that will be played live. If three instruments are used, a "cluster 3" message can be sent to timbreID so that similar instances in the database are grouped under a common cluster index.

With only three instruments, it is very likely that the clustering algorithm will produce accurate groupings; however, results of the process can be verified using the "cluster_list" message, which prints the members of each cluster to Pd's post window.

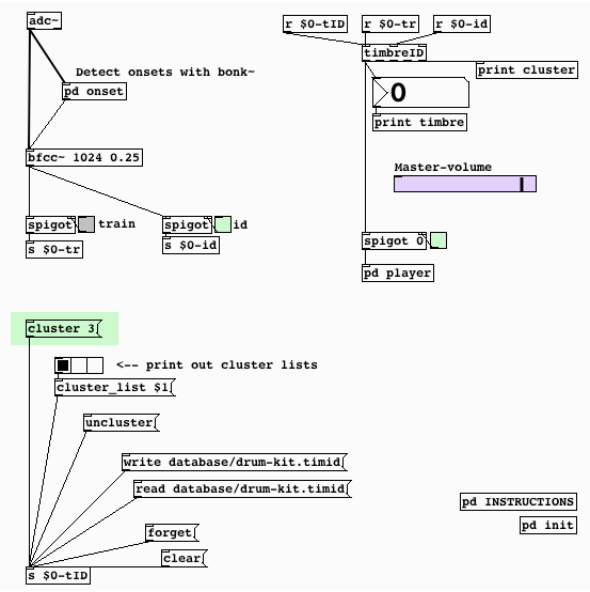


Figure 4.6: An instrument recognition and sample mapping patch.

Once training and clustering are complete, the identification spigot can be opened so that any incoming sounds will be classified by the system, producing a cluster index number at timbreID's first outlet. The sample playback subpatch then uses these indexes to locate the onsets of different samples loaded to an array. Input/output mapping is immediate and discrete—striking each instrument will trigger playback of a consistently corresponding sample. Synthesis control that relies on continuous data requires constant rather than attack-based analysis. This strategy is described in the following example.

4.4.3 Vowel Recognition

As described in Chapter 3, under the right circumstances cepstral analysis can achieve a rough deconvolution of two convolved signals. In the case of a sung voiced vowel, glottal impulses at a certain frequency are convolved with a filter

corresponding to the shape of the vocalist's oral cavity. Depending on fundamental frequency, the cepstrum of such a signal will produce two distinctly identifiable regions: a compact representation of the filter component at the low end, and higher up, a peak associated with the pitch of the note being sung. Mel or Bark frequency cepstral techniques do not produce the filter and source deconvolution because their skewed and smoothed spectra fail to preserve the evenly spaced harmonics that generate quefrency peaks. Furthermore, perceptually weighted cepstral techniques de-emphasize high frequency content that may be useful in classification. In the case of cepstral analysis, the lower coefficients should hold their shape reasonably steady in spite of pitch changes, making it possible to identify vowels no matter which pitch the vocalist happens to be singing. As pitch moves higher, the cepstral peak actually moves *lower*, as the so-called "quefrency" axis corresponds to period—the inverse of frequency. If the pitch is very high, it will overlap with the region representing the filter component, and destroy the potential for recognizing vowels regardless of pitch.[4]

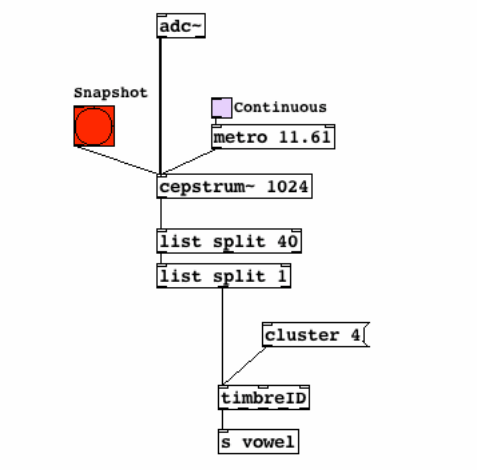


Figure 4.7: Sending training snapshots and continuous overlapping cepstral analyses to timbreID.

Having acknowledged these limitations, a useful pitch-independent vowel recognition system can nevertheless be arranged using timbreID objects very easily.

[4]These qualities of cepstral analysis can be observed by sending cepstrum~'s output list to an array and graphing the analysis continuously in real-time.

In an informal evaluation, it was discovered that regular cepstrum was indeed more effective for vowel classification than BFCCs or MFCCs. The vocal spectra that were inspected contained weak high frequency energy that Bark cepstral analysis fails to take into account. Figure 4.7 shows a simplified excerpt of an example patch where cepstral coefficients 2 through 40 are sent to timbreID's training inlet every time the red snapshot button is clicked. Although identical results could be achieved without splitting off a specific portion of the cepstrum[5], pre-processing the feature with two instances of Pd's list splitting object keeps timbreID's feature database more compact. The choice of cepstral coefficient range 2 through 40 is somewhat arbitrary, but it is very easy to experiment with different ranges by changing the arguments of the two list split objects.

In order to train the system on 3 vowels, about 5 snapshots must be captured during training examples of each sung vowel. For distinguishing background noise, 5 additional snapshots should be taken while the vocalist is silent. Next, the "cluster" message is sent with an argument of 4, which automatically groups similar analyses so that the first vowel is represented by cluster 0, the second vowel by cluster 1, and so on. The cluster associated with background noise will end up as cluster 3. It is not necessary to ensure that the same number of training instances are provided for each vowel. If 7 training examples are given for the first vowel and only 5 for the others, the clustering algorithm should still group the analyses correctly. Clustering results can be verified by sending the "cluster_list" message, which returns a list of any particular cluster's members via timbreID's fourth outlet.

To switch from training to classification, cepstrum~'s pre-processed output must be connected to timbreID's second inlet. The actual example patch contains a few routing objects to avoid this type of manual re-patching, but they are omitted here for clarity. Activating the metro in Figure 4.7 enables continuous overlapping analysis. If finer time resolution is desired for even faster response, the metro's rate can be set to a shorter duration. Here, the rate is set to half the duration of the analysis window size in milliseconds, which corresponds to an overlap of

[5]The alternative would be to pass the entire cepstrum, but set timbreID's active attribute range to use only the 2^{nd} through 40^{th} coefficients in similarity calculations.

2. As each analysis is passed from cepstrum~ to timbreID, a nearest match is identified and its associated cluster index is sent out of timbreID's first outlet. The example patch animates vowel classifications as they occur. Extending this configuration of objects beyond the example, it is possible to use such classification data as control information so that an improvising vocalist can effect changes in computer-generated processes or synthesis with the shape of his or her mouth.

4.4.4 Target-based Concatenative Synthesis

Some new challenges arise in the case of comparing a constant stream of input features against a large database in real-time. The feature database used for vowel recognition only requires about 20 instances. To obtain interesting results from target-based concatenative synthesis, the database must be much larger, with thousands rather than dozens of instances. In addition to the systems mentioned in the introduction, this type of synthesis can be achieved using *Guidage* [DAC07b], and is practiced live by the artist sCrAmBlEd?HaCkZ! using his own software design. The technique is to analyze short, overlapping frames of an input signal, find the most similar sounding audio frame in a pre-analyzed corpus of unrelated audio, and output a stream of the best-matching frames at the same rate and overlap as the input.

The example included with timbreID provides an audio corpus consisting of five minutes of bowed string instrument samples. As an audio signal comes in, an attempt at reconstructing the signal using grains from the bowed string corpus is output in real time. In the example case of speech input, timbre relationships with the synthesized output are quite meaningful, and it approaches actual intelligibility. A variation on this process is to draw grains from an audio corpus that is entirely separate from that which is analyzed and compared with real time input. In both cases, timbreID compares input features with a database and returns the index location of the nearest matching grain. For the alternate technique, this index is simply used to read into a different audio corpus, so that the input and output timbres are not related directly. In this case, the strength is that the behavior of the system is very consistent, allowing an improvising musician to learn how to

produce any desired audio output.[6]

In these types of applications, timbreID's third inlet can be used in order to search large feature databases. Classification requests sent to the third inlet are restricted by a few additional parameters. The search for a nearest match is carried out on a specified subset of the database by setting the "search_center" and "neighborhood" parameters. Suppose timbreID is storing a training database with 5000 instances representing unique grains in the audio corpus. With search_center and neighborhood set to 1500 and 2000 respectively, timbreID will compare features sent to its third inlet with grains 500 through 2500—i.e., 1000 grains above and below search_center.

The concatenative synthesis example provides options for different grain sizes and analysis rates, but with default settings, the process of computing a BFCC feature for the input signal, comparing it with 2500 instances in the feature database, and playing back the best-matching grain occurs at a rate of 43 times per second. Using a 2.91 GHz Intel Core 2 Duo machine running Fedora 11 with 4 GB of RAM, the processor load is about 17%. By lowering the neighborhood setting, this load can be lowered. However, reducing processor load is not the only reason that restricted searches are useful. A performer may also wish to control which region of the audio corpus from which to synthesize.

A third parameter, "reorient" causes search_center to be continually updated to the current best match during active synthesis. Starting with the search_center and neighborhood parameter values above, if the first match request returns grain 2200, search_center is set to that value, and the new database range for the following search will be between grains 1200 and 3200. With matches occurring 43 times per second, the search range adapts very quickly to changes in the input signal, finding an optimal region of sequential grains from which to draw.

In an effort to combat sharp discontinuities from one grain to the next, the "max_matches" parameter can be altered, so that timbreID considers previous matches as well as the current nearest match before reporting the most appropriate grain. With max_matches set to 10, if the previous match was grain 2200 and the

[6]Audio examples demonstrating the results of these processes can be accessed at http://www.williambrent.com.

current nearest match to the input feature is 1800, timbreID will look at the 10 next-nearest matches to the input feature. If one of them is closer to the previous match than the current match is to the input feature, it will be output from timbreID as the most appropriate grain even though it is actually not the most proximate grain in feature space. The previous match itself is not eligible as a current match in order to avoid repeated playback of a single grain.

For identification applications like the vowel classification example, bfcc~'s spectrum normalization is best kept active so that notes sung at different dynamic levels will not throw off the identification process. In concatenative synthesis, however, making classification vulnerable to dynamic variation may result in a more naturally evolving output signal, so normalization is disabled in the example patch.

4.4.5 Timbre ordering

The timbre ordering examples use two different approaches to sound segmentation: the first reads in pre-determined onset/offset times for each of 51 percussion instrument attacks, and the second automatically divides loaded samples into grains that are 4096 samples in length by default. Onset/offset labels for the first example were generated manually in Audacity, exported to a text file, then imported to a table in Pd. The percussive sound set included with this example is small, and is intended to provide a clear demonstration of timbreID's ordering capabilities. Figure 4.8 shows a region of the patch that includes the table where ordering information is stored and 5 sliders that control feature weighting.

Ordering is always performed relative to a user-specified starting point. With 51 instruments, when an instrument index between 0 and 50 is supplied along with the "order" message, timbreID will output the ordering list at its fourth outlet for graphing. Using the 5 feature weight sliders, it is possible to boost or cut back the influence of any particular feature in the ordering process. The features implemented in this patch are temporally evolving spectral centroid, spectral flatness, zero crossing rate, loudness, and BFCCs.

After hearing the results of a particular ordering, the levels of the feature

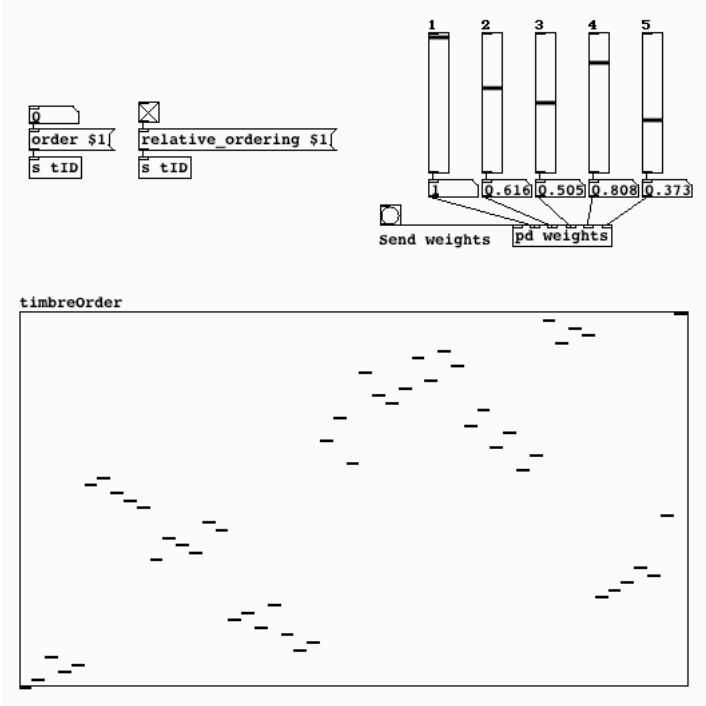


Figure 4.8: Fifty-one percussion sounds ordered based on a user-specified weighting of 5 features.

weight sliders can be changed in order to produce a new ordering and gain an understanding of the effects of various features in the process. An ordering is shown in the graph of Figure 4.8, where the y axis represents instrument indices 0 through 50, and the x axis indicates each instrument's position in the ordering. It begins at instrument 0 with a drum and progresses through other drum strikes followed by snares, a sequence of cymbal strikes, and a sequence of wooden instruments. Ordering the set by starting with a wooden instrument will produce a different result that retains similarly grouped sequences. An expanded version of this patch could be useful as a compositional aid for exploring relationships between sounds in a much larger set, offering paths through the sounds that are smooth with respect to different sonic characteristics.

Two types of ordering are available: "raw" and "relative". The graph in Figure 4.8 was produced with relative ordering, which starts with the user-specified instrument, finds the nearest match in the set, then finds the nearest match to that match (without replacement), and so on. The point of reference is always shifting.

Raw ordering begins with the given instrument, then finds the closest match, the second closest match, the third closest match, and so on. Orderings of this type start with a sequence of very similar sounds that slowly degrades into randomness, and usually finish with a sequence of similar sounds—those that are all roughly equal in distance from the initial sound, and hence, roughly similar to each other.

The second ordering example loads and segments arbitrary sound files. Loading a speech sample generates sequences of similar phonemes with a surprisingly continuous pitch contour.

4.4.6 Mapping sounds in timbre space

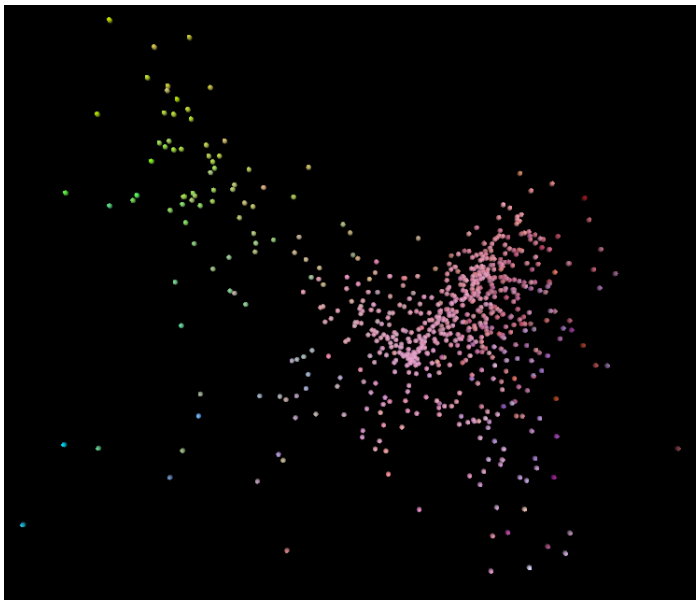


Figure 4.9: Speech grains mapped with respect to the 2nd and 3rd BFCC.

Another way to understand how the components of a sound set relate to one another is to plot them in a user-defined timbre space. CataRT is the most recognized and well developed system for this task; timbreID makes it possible within Pd using GEM for two- and three-dimensional plotting. In the provided example, axes of the space can be assigned to a number of different spectral features, zero crossing rate, amplitude, frequency, or any of 47 Bark-frequency cepstral coefficients by clicking the labeled radio buttons. By editing the analysis sub-patch,

additional features can be included. Using automatic granular segmentation, an audio file can be loaded and analyzed at a specified grain size. Figure 4.9 shows speech grains plotted in a space where values of the second and third BFCCs are mapped to the horizontal and vertical dimensions respectively. RGB color can be mapped to any available features as well. Here, it is set to the values of BFCCs 4—6.

Mousing over a point in the space plays back its appropriate grain, enabling exploration aimed at identifying regions of timbral similarity. The upper left region of Figure 4.9 contains a grouping of "sh" sounds, while the central lower region contains a cluster of "k" and "ch" grains. Other phonemes can be located as well. In order to explore dense regions of the plot, keyboard navigation can be activated to zoom with respect to either axis (or both simultaneously), and move up, down, left, or right in the space.

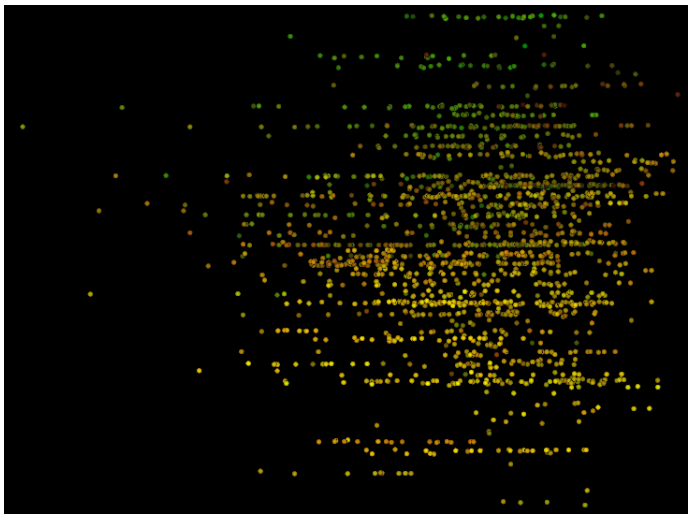


Figure 4.10: String grains mapped with respect to amplitude and fundamental frequency.

Figure 4.10 shows a plot of string sample grains mapped according to RMS amplitude and fundamental frequency. Because the frequencies in this particular sound file—a recording of a Bach violin partita—fall into discrete pitch classes, its grains are visibly stratified along the vertical dimension. Crescendos and decrescendos are easily performed by moving the mouse along a single pitch stratum. In conjunction with sophisticated instrument controllers such as those described

in [OJ08] and [Oli10], the performative tendencies of this mainly research-oriented patch can be developed to create full-fledged digital musical instruments that navigate timbre space.

Figure 4.11 shows 60 percussion instruments clustered into 12 groups, with each group shown in a distinct color. In this case, segmentation was accomplished using labels generated manually in Audacity. The plot illustrates that there are five instances of each instrument class (cluster), and that they are fairly separated in the chosen spectral centroid/BFCC 2 space. Sounds can also be displayed and rotated in three dimensions with appropriate depth to reveal separations between clusters that are proximate or even overlapping from a two-dimensional perspective. This aids in assessing whether a group of sounds can be reliably classified using any particular combination of features. If an instrument's instances cluster well and are relatively distant from other instrument clusters, it is an indication that real-time classification will be reliable. The green circle in the upper right corner of the plot is controlled by the position of the mouse. When points come into contact with this circle, their corresponding samples or grains are played back. The radius can be resized to achieve a desired level of control over sound playback.

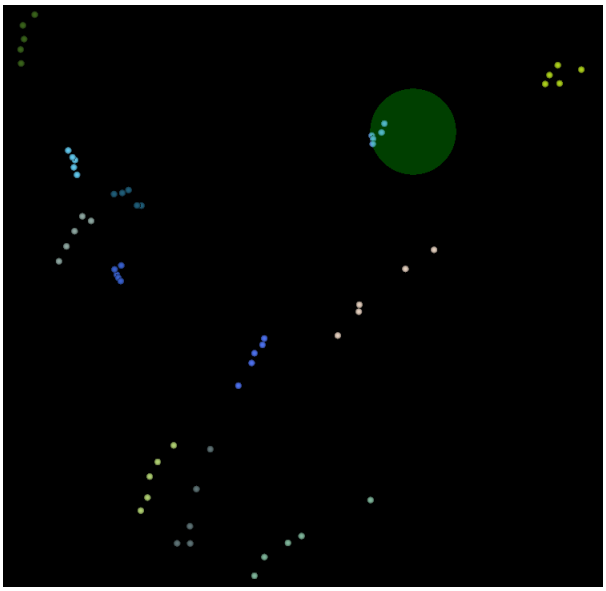


Figure 4.11: Sixty percussion samples colored by cluster.

Plotting grains from a frequency chirp can be instructive for understanding

relationships between various features. The negative correlation between spectral centroid and the 2nd BFCC was noted in Section 3.3. In this example, it is possible to see the correlation directly. As spectral centroid increases along the abscissa, values of the 2nd BFCC drop in a smooth curve. In the Bark-frequency domain, as the sinusoidal peak sweeps to higher frequencies, less and less of its area overlaps with a Bark-spectral envelope resembling the first cosine transform basis function. Plotting centroid against rolloff and zero crossing rate shows completely linear relationships. These features are essentially identical when applied to such a simplified spectral envelope.

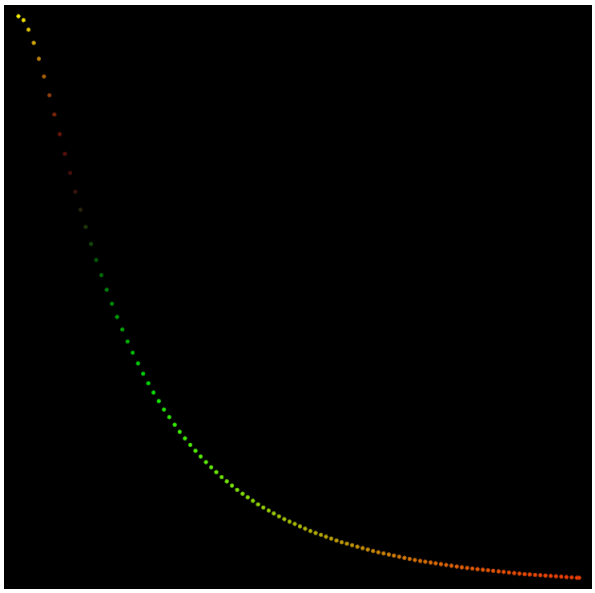


Figure 4.12: Grains from a 20-20,000 Hz frequency chirp plotted with respect to spectral centroid and the 2nd BFCC.

Mapping is achieved by recovering features from timbreID's database with the "feature_list" message, which is sent with a database index indicating which instance to report. The feature list for the specified instance is then sent out of timbreID's fifth outlet, and used to determine the instance's position in feature space.

4.5 Conclusion

This chapter has introduced some important features of the timbreID analysis/classification toolkit for Pd, and demonstrated its adaptability to six unique tasks. The example patches are simple in some respects and are intended to be starting points that can be expanded upon by the user. Future development will be focused on adding new features to the set of feature extraction objects, implementing a kD-tree for fast searching of large databases in order to make concatenative synthesis more efficient, and developing strategies for processing multiple-frame features of different lengths in order to compare sounds of various durations.

4.6 Acknowledgements

Portions of this chapter were previously published in the following article:

William Brent, "A Timbre Analysis and Classification Toolkit for Pure Data", *Proceedings of the International Computer Music Conference*, 2010.

Chapter 5

Classification Performance Evaluation

5.1 Examining Percussive Timbres

This chapter presents the results of a performance evaluation exploring the effectiveness of timbreID's feature extraction objects in a classification task. Analysis was performed on percussive sounds that are investigated from a perceptual point of view in the following chapter, where correlations between the two data sets are considered. Percussive timbres were chosen for several reasons. To begin, research investigating these sounds is limited, in spite of the importance of timbre as an organizational parameter in contemporary percussion composition. In terms of sonic characteristics, a great number of percussive sounds have a brief or even absent steady state, making the attack portion an obvious common target for obtaining relevant acoustic analyses within a diverse sound set. The cepstrograms from Chapter 4 (Figures 4.4 - 4.5) illustrate the stability of cepstral characteristics during the attack segment across several strikes of the same instrument. Further, looking forward to the perceptual study in the following chapter, the attack segment duration of percussive instruments is more varied than might be assumed, providing opportunities for investigating the role of attack time—an aspect of timbre that has been established as perceptually relevant in classification

tasks [CLA+63][IK93]. With regard to spectrum, very few percussion instruments exhibit anything like a true harmonic structure. Even instruments that have an unambiguous pitch, like the vibraphone, are quite difficult to analyze reliably with standard pitch tracking algorithms that depend upon the presence of harmonically related partials. Certain established acoustic measurements that have been employed as timbre descriptors, like spectral irregularity and the tristimulus measure, are most appropriate for harmonic tones like the violin, clarinet, or flute. Further study of percussive timbres may turn up similarly appropriate measures for inharmonic spectra, or illustrate how harmonically-oriented features may be put to use effectively when considering different types of spectra. Regarding pitch, with the exception of keyboard instruments like the xylophone, marimba, and vibraphone (which, granted, are foundations of the percussion family), most percussion instruments have very limited means for producing a range of pitches. From one perspective, keyboard instruments themselves can be viewed as collections of individual instruments that are capable of producing only a single pitch. Most membranophones and idiophones are also tuned to only one pitch. Thus, classification difficulties associated with interactions between pitch and timbre, such as those noted in [HE01][KI92] must be accepted as inevitable, making artificial schemes for normalizing pitch unnecessary. In fact, the presence of pitch in percussive tones can be understood as part of the timbre itself. Where perceptual studies of Western orchestral instruments require a control for pitch (a strategy that imposes boundaries upon the musical relevance of subsequent findings), the study of percussive instruments presents a naturally occurring experimental control. Finally, the palette of timbres that fall under the category of percussion is enormously varied, and desired ranges of timbre similarity can therefore be easily constructed by the experimenter.

5.2 Method

Several samples of percussion instruments were recorded for training and testing the system. In an effort to compile a range of timbres typical of con-

temporary percussion, instruments were chosen from the orchestration of Varèse's *Ionisation*, spanning the three standard instrument categories of skins, metals, and woods. The performance tests exploring these timbres range in difficulty, from the ideal condition of diverse, isolated timbres with optimally placed analysis windows, to more difficult scenarios with a more similar timbre set, additional background noise, and suboptimal analysis timing scenarios. Beginning with highly favorable circumstances allows the most accurate results possible, so that gradual degradation of these circumstances will make reasons for the inevitably reduced accuracy in real-time applications more clearly understood.

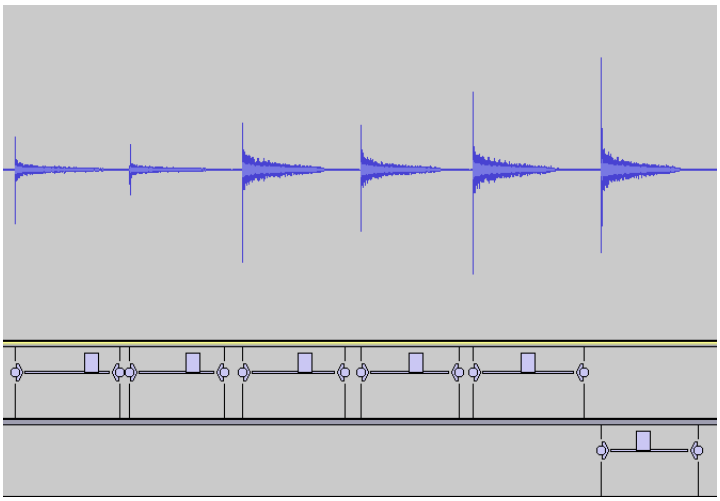


Figure 5.1: Training and testing instances of tam tam strikes.

All recordings were taken in a soundproof studio using high quality GRACE m802 preamplifiers and a Neumann KM140 cardioid microphone. The sounds were initially digitized at 96 kHz, 24 bits per sample, and were subsequently downsampled to 44.1 kHz, 16 bits per sample using *SoX*.[1] Each instrument was struck repeatedly at several different dynamic levels in order to produce a sufficient number of training and testing instances. The percussionist used suitable mallets for each instrument. After downsampling, the audio was manually labelled in Audacity to mark the precise onset of each instrument strike. As these recordings were very high quality with a low noise floor, instrument onsets were clearly visible as

[1] *SoX* is an open source audio utility available at http://sox.sourceforge.net.

deviations from zero in the waveform. Two sets of label files were created: one defining onset and offset points for 5 unique training instances of each instrument, and a second defining onset and offset points of a single test instance for each instrument. The test instances in the second label file were not selected according to any consistent criteria. In most cases, the strike following the fifth training instance was used. A typical range of dynamics for training and testing instances is shown in Figure 5.1, where 5 tam tam strikes are labeled to be used for training, and a sixth strike is labeled for testing.

5.2.1 Instruments

Two different sound sets were used, referred to below as "diverse" and "similar". The instrument list for the diverse set of timbres is shown in Table 5.1. Because mallet type and point of contact cause a very perceptible difference in timbre for certain instruments, sounds that vary in these respects are treated as distinct timbres even when produced by a single instrument. There is a great deal of diversity in the set, but some pairs of instruments are clearly more similar than others; for instance, the orchestral crash cymbals and suspended cymbal, the snare drum and military drum (with snares), and the two tam tams. Such similarities were included in order to introduce the possibility of misclassifications in an otherwise timbrally scattered sound set.

The similar set of timbres was generated using six different striking methods on five different tam tams and gongs (three of which are listed in Table 5.1). Each instrument was struck on the edge, the band between the edge and center, and the center using both a drumstick and a standard felt orchestral mallet. These timbres were judged to be perceptually distinguishable by the author in consultation with a percussionist, with the highest similarity existing between timbres generated by the three different strike points on a single instrument using one mallet type. Spectral plots for example strikes of instruments in both the diverse and similar sets are provided in Appendix A.

Table 5.1: Thirty percussion instruments used for performance evaluation.

Clave (low)	Orchestral crash cymbal
Clave (high)	Military drum
Cowbell	Tambourine
Snare drum	Triangle
Temple block	Castanet
Temple block	Glockenspiel (middle G)
Tom (12.5”)	Medium nipple gong (felt, center)
Bass Drum (37”)	Large tam tam (drumstick, middle band)
Bongo (low)	Guiro
Bongo (high)	Lion’s Roar
Brake Drum	Sleigh Bells
Sus. cymbal (edge)	Large tam tam (felt, edge)
Sus. cymbal (bell)	Medium tam tam (drumstick, middle band)
Large nipple gong	Medium nipple gong (drumstick, edge)
Maraca	Wooden Plank (poplar, 0.75” x 2.5” x 18.75”)

5.2.2 Analysis Strategies

A major objective of this study is to identify the most effective analysis strategy for classification purposes. Analysis methods were chosen taking into account restrictions that are likely to exist in actual performance contexts. Ideally, multiple short-time overlapping analyses can be taken over the entire course of each sound in order to extract as much time-varying information as possible. In real-time applications, however, it is useful to obtain analysis and classification results immediately upon the onset of a new instrument strike. One approach is simply to analyze a single frame of audio containing the onset of the new sound. With a short enough analysis window, results can be obtained within 20 milliseconds or less after an attack, enabling synthetic accompaniment to adapt to a musician's actions roughly synchronously. A more advanced and slightly more time-consuming approach is to perform a series of overlapping analyses following the onset. While this does not come close to describing the entire sound, it does capture the complex spectro-temporal structure of the attack segment, which can significantly improve classification accuracy. The cost is not only increased latency, but increased data size as well. This is significant in terms of storage and additional computation during distance calculation. In response to these drawbacks, a third approach is to summarize multiple frame analysis information with the mean and standard deviation of each feature (or each individual coefficient in a multiple-component feature vector like Bark-frequency cepstrum) over time. All of these strategies bypass the difficult question of how to directly compare the timbres of sounds with different durations. In light of the constraints of real-time considerations and the temporal characteristics of percussive timbres in general, the range of options for analysis has been limited to the attack segment only.

A variable that impacts all of these strategies is analysis window placement. Because this evaluation makes use of manually produced label points, it is possible to place the analysis window so that it begins precisely at the onset of each sound. In real-time performance applications, it is impossible to pinpoint attacks so accurately. Thus, it is important to produce training instances based on analysis windows that are displaced from actual onsets, and to perform multiple

classifications of each test sound at various displacements as well. To investigate reliability in the face of this real-time complication, repeated classification tests were run with the analysis window beginning between 0 and 896 samples (20.3 ms) displaced from the actual onset, moving in steps of 128 samples (2.9 ms). This variable will be referred to as onset displacement (OD). Accuracy results presented in the figures below are based on the average accuracy across all eight OD settings.

A second major objective is to identify the most consistently useful feature or combination of features for classification, with consideration for keeping training database information compact. The 15 features that were evaluated are listed in Table 5.2. Several features require consideration of additional parameters, which are specified here. The boundary frequency for spectral brightness was set at 1200 Hz. The concentration used for spectral rolloff was 85%. Because spectral flux requires two analysis frames, it was measured starting 256 samples (5.8 ms) after the common analysis point for all other features. Manhattan distance was used to compute the spectral difference between frames. Spectral irregularity was calculated using Jensen's algorithm, defined in Equation (3.9). Magnitude spectra for all spectral and cepstral features were normalized before further processing in order to strip the effects of amplitude variation. Filterbanks for mel- and Bark frequency cepstra were constructed using 81 mel and ½ Bark spacing, respectively, producing 47-point vectors for both MFCCs and BFCCs. These filterbanks were applied to magnitude spectra rather than log power spectra as specified in [DM80]. Conventionally, energy is weighted and summed in each band of the filterbank [YJO+00]. Here, in order to reduce high-frequency bias, weighted energy was *averaged* in each band according to filter width. The timbreID object for computing BFCCs provides this option. The question of whether to sum or average is one that hinges upon how strictly the Bark-frequency model is intended to be realized. The human ear does not average energy within critical bands. Thus, the averaging approach is less true to the concept of Bark weighting. However, the crude triangular filter shapes used in a typical Bark weighting algorithm are no less artificial. From this perspective, the degree of adherence to perceptual models can be justifiably sacrificed if classification performance can be improved. Here,

the averaging of weighted magnitude spectrum produced slightly better results. A common window size of 1024 samples was chosen for all feature calculations as a reasonable compromise between time and frequency resolution. The table-reading versions of timbreID feature extraction objects (described in Section 4.2.1) were used in order to process all tests in non-real-time.

Table 5.2: Fifteen features used for performance evaluation.

Zero Crossing Rate	Spectral Flux	Magnitude Spectrum
Spectral Rolloff	Spectral Centroid	Bark Spectrum
Spectral Brightness	Spectral Spread	Cepstrum
Spectral Flatness	Spectral Skewness	Mel-Frequency Cepstrum
Spectral Irregularity	Spectral Kurtosis	Bark-Frequency Cepstrum

Numerical features, such as spectral brightness, will be referred to as *low level*, in contrast with *high level* feature vectors like Bark-frequency cepstral coefficients. When evaluating combinations of low level features, the database was normalized as described in Section 4.3.1 in order to equalize the range of all features before distance calculation. In the case of multiple-frame analysis, components of high level feature vectors were repacked according to coefficient number using a patched network of list manipulation objects in Pd (i.e., information was organized into time-varying coefficient tracks). Capturing the temporal evolution of individual coefficients enables comparison between coefficient subsets of different multiple-frame high level features. Because all cepstral features are examined with respect to data reduction and their vector lengths range from 46 to 511 coefficients, specific ranges of interest were selected. As described in Section 3.3, use of the entire set of BFCCs will produce distance values that are identical to those generated when using an unprocessed Bark-weighted spectrum. Therefore, at maximum, only the first 15 coefficients of the cepstral features were used in comparisons between all features. The first coefficient was always excluded as it is essentially only an indicator of amplitude. It is specifically the accuracy of 15-point cepstrum, BFCC, and MFCC subsets that is evaluated here. The full range of available bins were

used for magnitude and Bark-weighted spectrum in order to illustrate differences in accuracy between complete and data-reduced spectral measurements. For magnitude spectrum, bins corresponding to DC and Nyquist were removed, leaving 511 total coefficients.

5.3 Results

5.3.1 30 Diverse Timbres

Single-frame onset analysis

Average scores across all onset displacement (OD) settings for single-frame analysis are shown in Figure 5.2. Scores are expressed as percentages, where 100% indicates perfect classification of all 30 test instances across each of the 8 OD settings (240 total classifications). As expected (with respect to dimensionality), a drastic difference in performance can be seen in comparing the high and low level features. Of the latter, spectral rolloff (ROL), zero crossing rate (ZCR), spectral centroid (CEN), and spectral brightness (BRI) produced the highest scores at 42.92%, 32.5%, 31.67%, and 27.92% respectively. As spectral centroid has been documented elsewhere as a useful predictor of perceptually based timbre spaces, its relatively higher performance here is not surprising. On the other hand, the roughly equal accuracy generated using zero crossing rate is unexpected given the extreme simplicity of the measure. Used in combination, the complete set of 9 equally weighted low level features (CLL) attains an accuracy of 91.67%, approaching that of the high level features.

The high level features, on average, do not produce more than two errors in this test, so it is impossible to identify one as more effective than any other. Although magnitude spectrum achieves 100% accuracy, it also requires 511 points. The remaining high level features have far fewer components but perform nearly equally well. A more detailed view of performance with respect to data reduction reveals some significant differences. A second round of single-frame tests was run in which the subset of coefficients used for distance calculations was varied

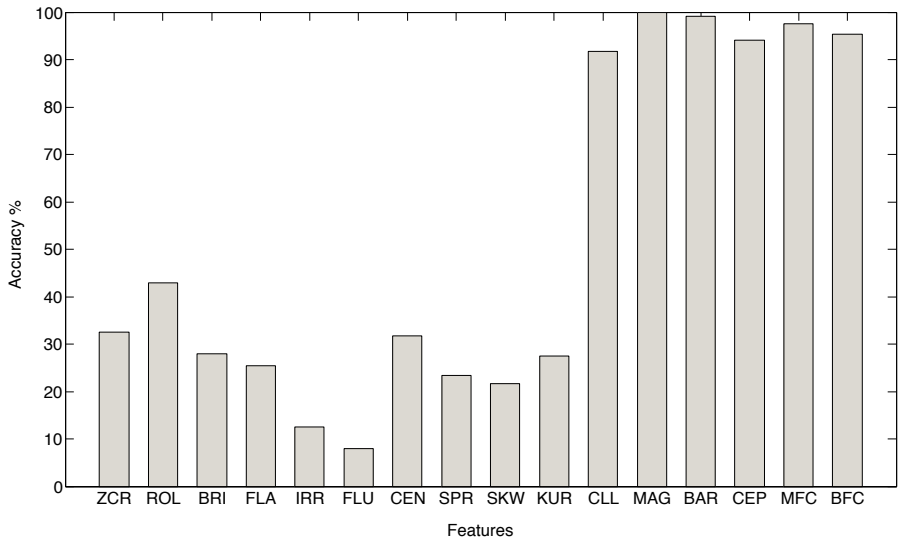


Figure 5.2: Scores for individual low level features, combined low level features, and high level features.

systematically for all high level features. The number of coefficients used ranged from one to the full set available for each feature, incrementing by one coefficient per test. Figure 5.3 plots accuracy against coefficient range (CR) used for all high level features. Mel- and Bark-frequency cepstral coefficients show a significant advantage over other features for the smallest range of coefficients. With only two coefficients, classification using BFCCs reaches 80% accuracy while magnitude spectrum—with no inherent data reduction capabilities—is only able to produce 35%. Scores using MFCCs more or less follow the same trajectory as BFCCs, while cepstrum generates scores in between. Clearly, all of the high level features involving a transformation beyond straight magnitude spectrum exhibit superior accuracy using a small subset of coefficients. MFCCs and BFCCs stand out as particularly effective, reaching 95% accuracy with 6 coefficients. Cepstrum outperforms both Bark and magnitude spectrum by over 10% when using only 6 coefficients. Though not shown in Figure 5.3, classification using magnitude spectrum reaches a stable plateau of 96% accuracy with at least 55 coefficients. At this sampling rate and window size, the 55^{th} bin corresponds to 2326 Hz. Thus, for this particular analysis strategy and sound set, spectral information above 2326 Hz does not appear to contribute significantly to classification accuracy.

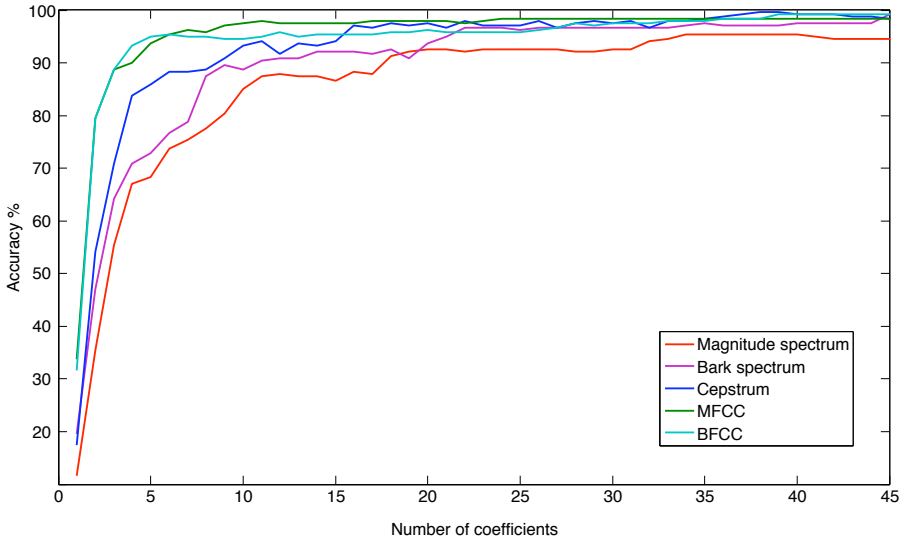


Figure 5.3: Accuracy vs. coefficients for all high level features.

Figure 5.4 shows accuracy stability for BFCCs with respect to OD and CR. The 90%+ scores using a CR of 1—15 do not deteriorate relative to OD, indicating that the strategy of providing training examples that are displaced from the actual onset is effective.

Multiple-frame analysis

Though it is much more data intensive, a short burst of analyses using any or all of the 15 low level features is easily managed in real time using a patching network similar to that shown in Figure 4.2. The performance improvement can be clearly seen in Figure 5.5, where the most significant change is among the low level features. Five analysis frames were used, with a hop size of 128 samples (2.9 ms). Once again, the relatively high classification accuracy achieved using a simple multiple-frame zero crossing measurement (58.3%) is very surprising. Compared to the single-frame case, accuracy for combined low level features improved from 91.67% to 97.5%, bringing it within the range of accuracy produced by high level features. The multiple-frame strategy using Bark spectrum, MFCCs, and BFCCs, and unprocessed cepstrum produced scores similar to those given in the single frame case, where no high level feature averaged more than 2 errors. Again, such a

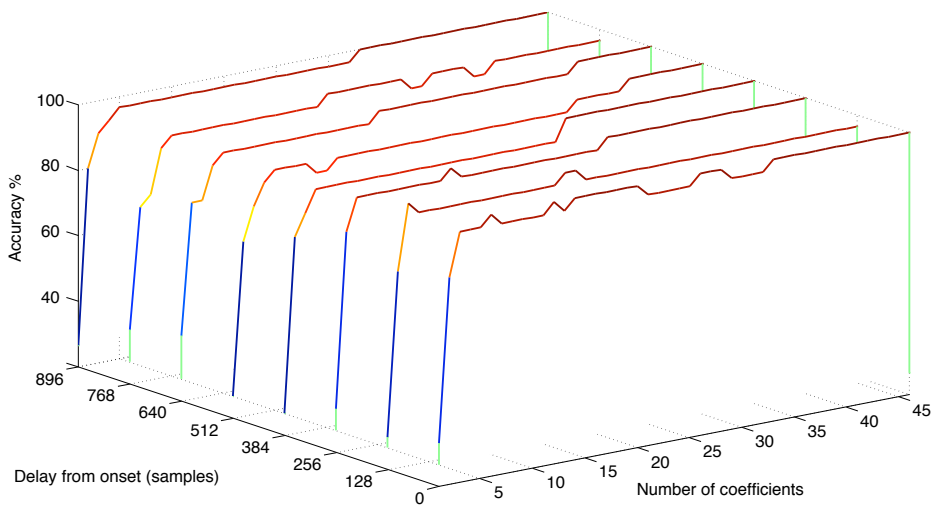


Figure 5.4: CR vs. OD vs. accuracy for BFCCs.

low error rate does not enable the identification of any individual high level feature or analysis technique as superior.

As in the previous section, scores are plotted against the number of coefficients used in order to make some finer discriminations between high level features in a multiple-frame analysis context. In this case, coefficients are not single numbers, but 5-component vectors with unique meanings for each feature. The vector for a single magnitude spectrum bin traces changes in energy within its narrow frequency band (43 Hz wide at this sampling rate and window size). Magnitude spectrum is a raw measure without any design for data reduction; however, tracking individual bins in this manner gives rise to the possibility of using selected bin tracks as a compact timbre descriptor. This type of data reduction has the advantage of requiring only the removal of information, and no further processing. Bark spectrum is already a reduced form of spectrum, with an additional characteristic of emphasizing more perceptually relevant frequency ranges. Although it has far fewer coefficients than magnitude spectrum, the resulting tracks refer to meaningful ranges more consistently across the low and high ranges of the feature. For multiple-frame BFCCs, each the 47 bin tracks describe energy distribution changes in reference to a basis function of the cosine transform. For instance, the multiple-frame vector for the second BFCC (recall that the first BFCC is never

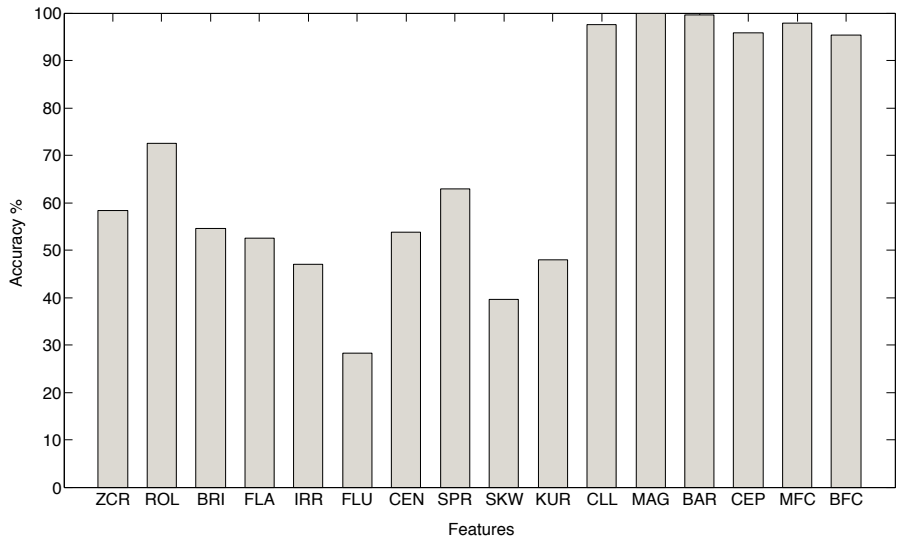


Figure 5.5: Scores for individual low level features, combined low level features (CLL), and high level features using multiple frame analysis.

used), which is calculated relative to a half-cosine basis function, would trace the relative presence or absence of a broad band of low frequency spectral energy over time. Thus, as more coefficients are added, more time-varying data about these types of gross spectral energy distributions are included for classification.

Figure 5.6 shows results obtained using 5 analysis frames, plotted as a function of the number of coefficient vectors used. The data reduction capabilities of all cepstral features are apparent. Using only 6 coefficients, the three cepstral features produce 95%+ accuracy in comparison with the 83.75% earned by magnitude spectrum. Note that magnitude and Bark spectrum perform almost equally well, and require roughly 20 coefficients to achieve scores of 95%. Considering that the 20^{th} bin corresponds to frequencies of only 818 Hz and 1268 Hz for magnitude and Bark spectrum respectively, it is significant that the temporal evolution of energy in the bins below is unique enough to generate such high classification scores for this sound set. The selection of these first 20 bins is completely arbitrary, a by-product of the fact that the lowest range of cepstral coefficients are being investigated here. It may be possible that a smaller subset of different coefficients can generate equal accuracy.

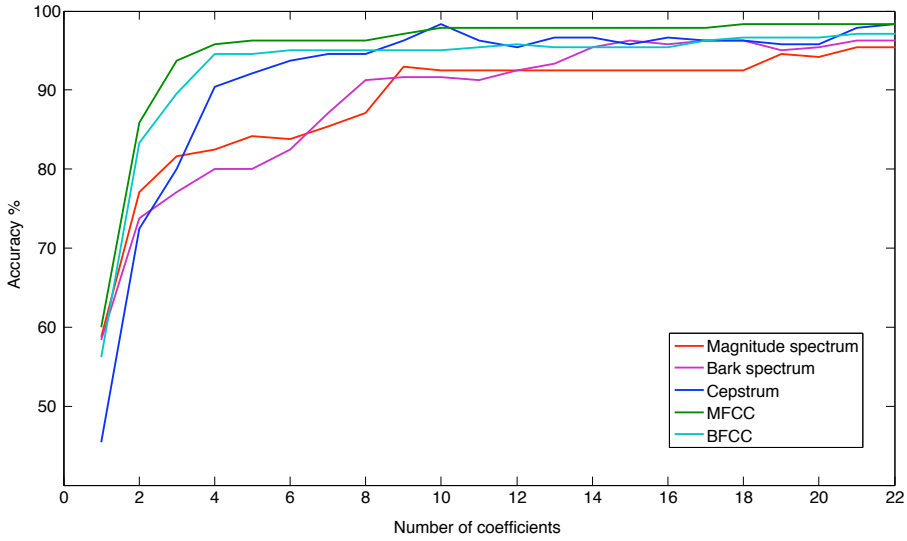


Figure 5.6: Accuracy vs. coefficients for all high level features using multiple-frame analysis.

Summarized multiple-frame analysis

The final strategy explored in this test is simply a statistical summarization of the time-evolving measurements from the previous section. The information generated by any feature over 5 frames can be expressed compactly by taking the mean and standard deviation of the data, reducing the feature length from 5 to 2 points. Scores for summarized individual and combined low level features are lower than those based on the complete 5-frame features, but significantly higher than scores generated using a single analysis frame in most cases. This can be seen in Figure 5.7, which shows scores for all three analysis strategies at once.

Figure 5.8 shows a trend similar to that found using single frame and complete multiple frame strategies. Using fewer than 10 coefficients, all cepstral measures outperform both magnitude and Bark spectrum, with an advantage shown by the perceptually weighted cepstra. The information presented thus far indicates that the optimal analysis strategy and feature for this diverse set of timbres is a 5-frame analysis using the first 15 Bark- or mel-frequency cepstral coefficients. This translates to a feature vector with $5 \times 15 = 75$ components. If further data reduction is needed, the mean and standard deviation of each coefficient across 5 frames creates a vector with $2 \times 15 = 30$ components that performs roughly equally

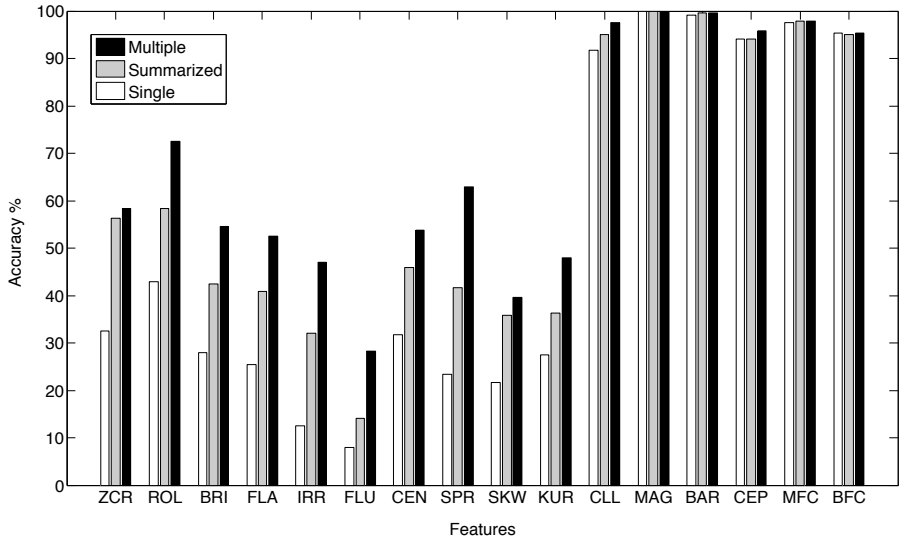


Figure 5.7: Scores for individual low level features, combined low level features (CLL), and high level features using single (white), summarized multiple-frame (grey), and complete multiple-frame (black) analysis.

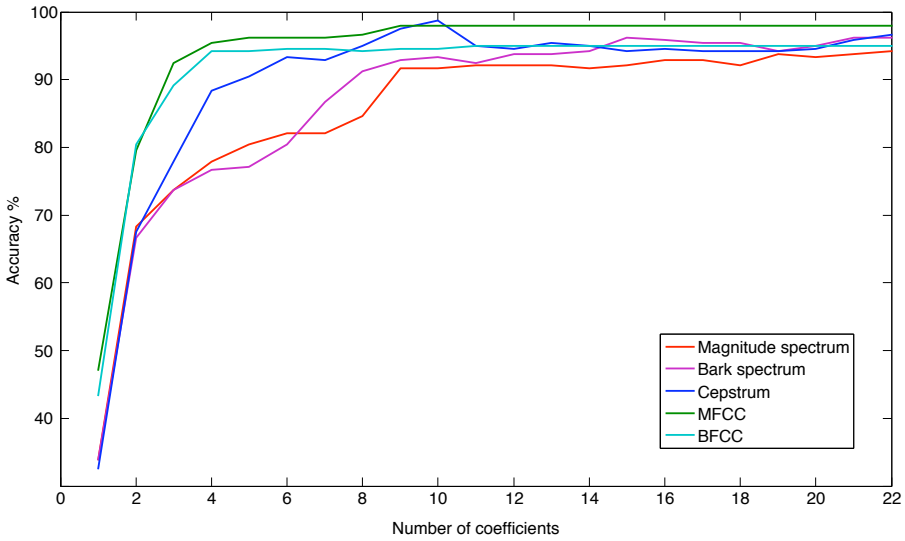


Figure 5.8: Accuracy vs. number of coefficients for summarized high level features.

well. In cases where increased latency is acceptable and 10 analysis frames can be used, data summarization will result in greater reduction of data size.

The Bark-weighted spectral plots of six instrument attacks from the diverse set are shown in Figure 5.9, providing a sample of the contrasting energy distributions existing among these timbres. As can be seen from the instrument list in

Table 5.1, the timbres contrast from a perceptual point of view as well (e.g., very few participants would have any difficulty distinguishing a snare drum from a triangle). Results from a more difficult test using the similar timbre set are presented in the following section.

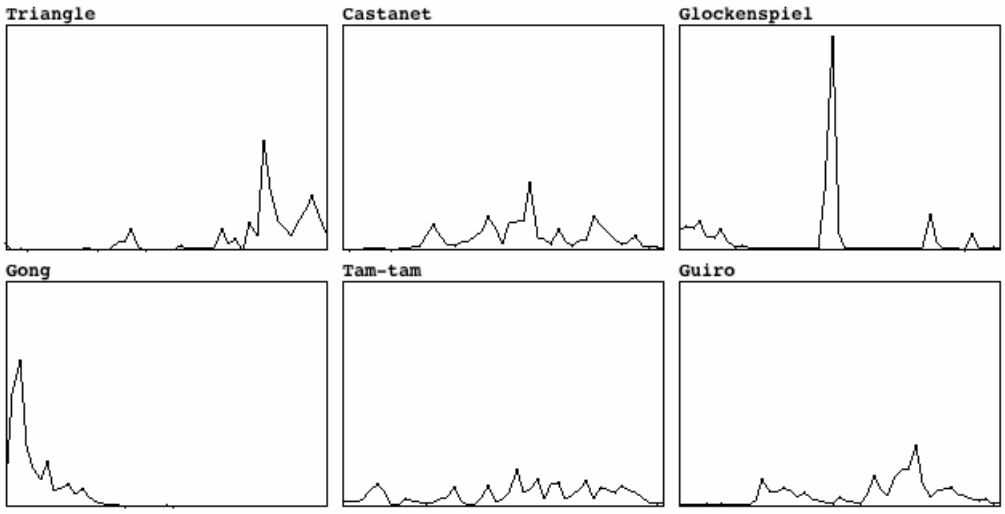


Figure 5.9: Bark cepstra for six timbres from the diverse set.

5.3.2 30 Similar Timbres

As described in Section 5.2.1, the 30 timbres in the "similar" set were generated using a collection of five inharmonic metals: 3 tam tams and 2 gongs. Each instrument was struck using a drumstick and felt orchestral mallet in three locations—edge, between the edge and center (referred to below as "middle"), and center. Figure 5.10 shows Bark-weighted spectra for the attacks of six of these timbres, generated by strikes on the edge, middle, and center of one instrument. The upper and lower rows show spectra from strikes using a drumstick and orchestral felt mallet respectively. In comparison with Figure 5.9, is it clear that classification by spectral energy distribution will be more problematic with this sound set. Sounds produced with the felt mallet possess spectral distributions with most of the energy concentrated in the lower frequency range, while drumstick strikes generate more broadband spectra. Thus, there are two general classes of spectra

with 15 instances each. Spectra for the complete set of similar timbres is provided in Appendix A. The increased difficulty of this test will provide further insight regarding the reliability of robust high level features evaluated in the previous section. In spite of the very similar Bark-spectral envelopes of these instruments, with a concentration of high energy due to the drumstick attack, it is easy to distinguish between them perceptually.

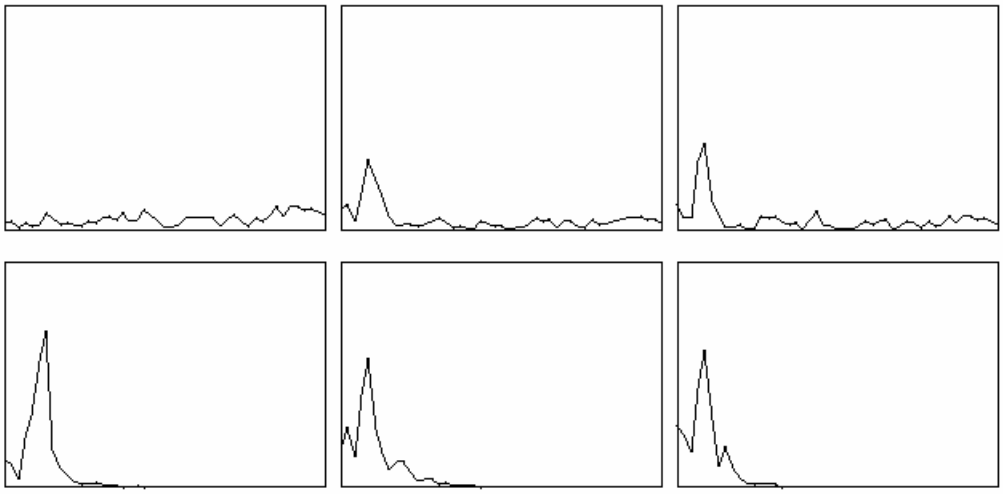


Figure 5.10: Bark cepstra for six timbres from the similar set.

Single-frame Analysis

Single frame analysis was carried out as before, with results averaged across all OD settings shown in Figure 5.11. Results are lower overall, and the most significant proportional change is the performance of unprocessed cepstrum relative to the other high level features. Using the diverse sound set, the largest difference in accuracy between these measures was 1.2%, or 0.36 incorrect classifications. Here, the smallest gap between cepstrum and any other high level feature is 31.7%—roughly ten misclassifications. An almost identical drop in accuracy occurred for the combined low level features. The accuracy gap between magnitude spectrum and the other cepstral measures also widened significantly, amounting to differences of 15.4% and 7.9% for mel and Bark cepstrum respectively. Scores for Bark spectrum were only 2.1% lower than magnitude spectrum. In the con-

text of more similar spectra, the effectiveness of data-reduced features appears to break down. The combined low level features and cepstral measures all provide a compact amount of information about rough spectral envelope contours. When the contours become similar, the finer level of detail provided by Bark spectrum, or ideally, magnitude spectrum is needed for the highest levels of accuracy. Nevertheless, considering the relatively limited information provided by only a single frame of 15 mel or Bark cepstral coefficients, their 80%+ accuracy is quite good.

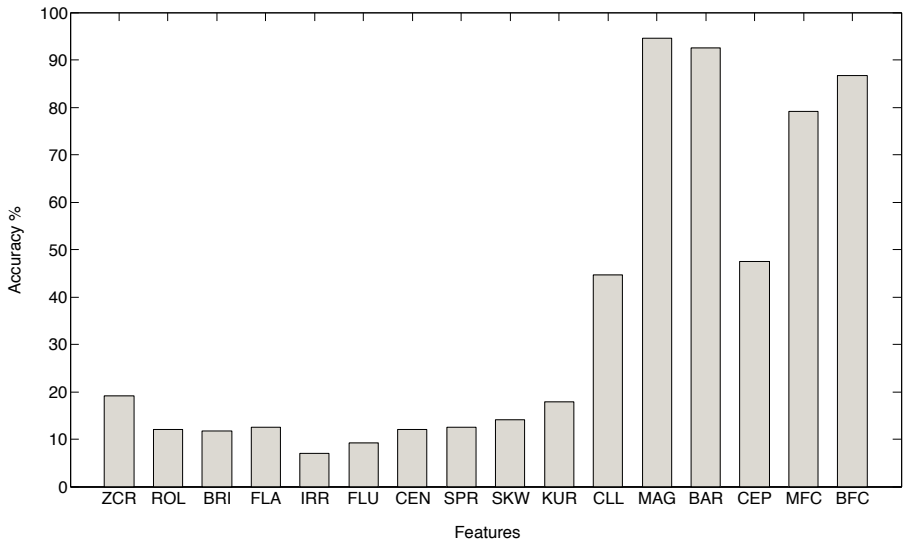


Figure 5.11: Scores for individual low level features, combined low level features (CLL), and high level features using single frame analysis.

Multiple-frame Analysis

Figure 5.12 shows the improved scores achieved using multiple frames. Among the low level features, zero crossing rate remained one of the most robust, and the vector of 10 combined features (CLL) achieved a score 5.4% higher than that of the 15 raw cepstral coefficients. While scores for magnitude spectrum remained the same, the gaps between it and mel and Bark cepstrum were roughly halved from the single frame case, to differences of 5.8% and 2.9% respectively.

A major break from previous patterns can be seen in Figure 5.13. Very small sets of time-varying Bark-weighted and regular magnitude spectrum coefficients produced scores well above 90%. Using 5 coefficients, MFCCs and BFCCs

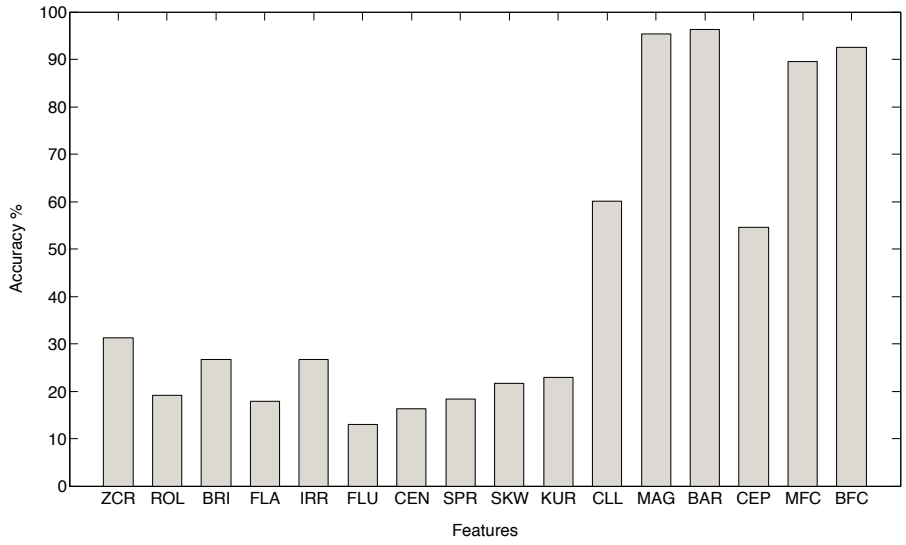


Figure 5.12: Scores for individual low level features, combined low level features (CLL), and high level features using multiple-frame analysis.

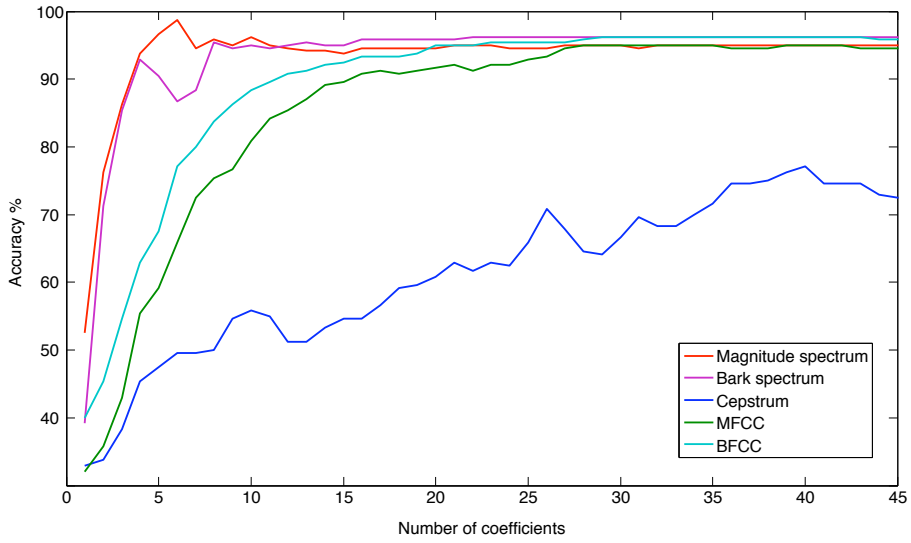


Figure 5.13: Scores vs. CR for high level features using multiple-frame analysis.

scored 37.5% and 29.17% *lower* than magnitude spectrum respectively. The success of the spectral features in such a limited frequency range can be attributed to unique spectro-temporal patterns in the lowest modes of the instruments being tested. While BFCCs provide a good summary of the entire frequency range, the spectral features capture the most relevant information for these instruments at a

far greater resolution. Though Bark-weighted spectrum is a data reduced form of magnitude spectrum, the weighting preserves low frequency detail, which explains its effectiveness here.

Figures 5.14—5.16 show spectrograms of the lowest 15 magnitude spectrum bins across 8 frames for the same instrument being struck in three locations.[2] All three timbres possess a peak near 86 Hz, and more prominent peaks near either 431 Hz or 301 Hz. The center strike exhibits the most stable peak over time, corresponding to the sustained pitch of the instrument. In contrast, the primary peaks of the middle and edge strikes reach their maxima and decay during the attack segment. With only 5 coefficients, only tracks for the lowest mode near 86 Hz are used in classification. Figure 5.16 shows that for the center strike this area holds its shape during the 8 frames, while the edge and middle strikes shown in Figures 5.14 and 5.15 dip significantly in this area over time.

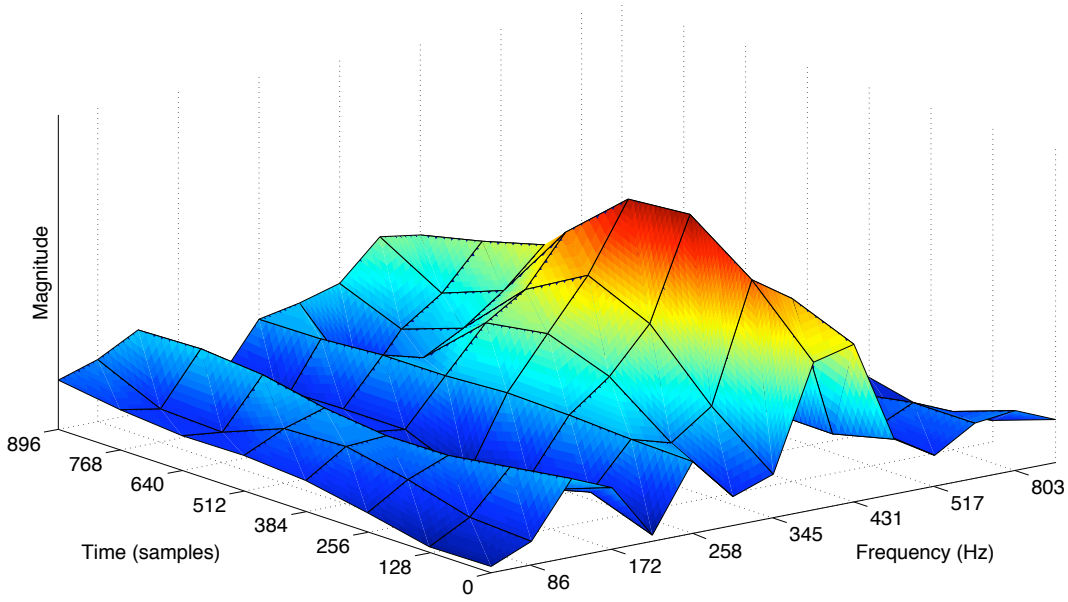


Figure 5.14: Low frequency spectrogram of a tam tam drumstick strike (edge).

Classification mistakes resulting from these features are instructive as well. Using 5 coefficients, multiple frame magnitude spectrum scored an average of 96.67%, with a total of seven errors among tests at all OD settings, and a max-

[2]However, the multiple frame features used for classification are only 5 frames in length.

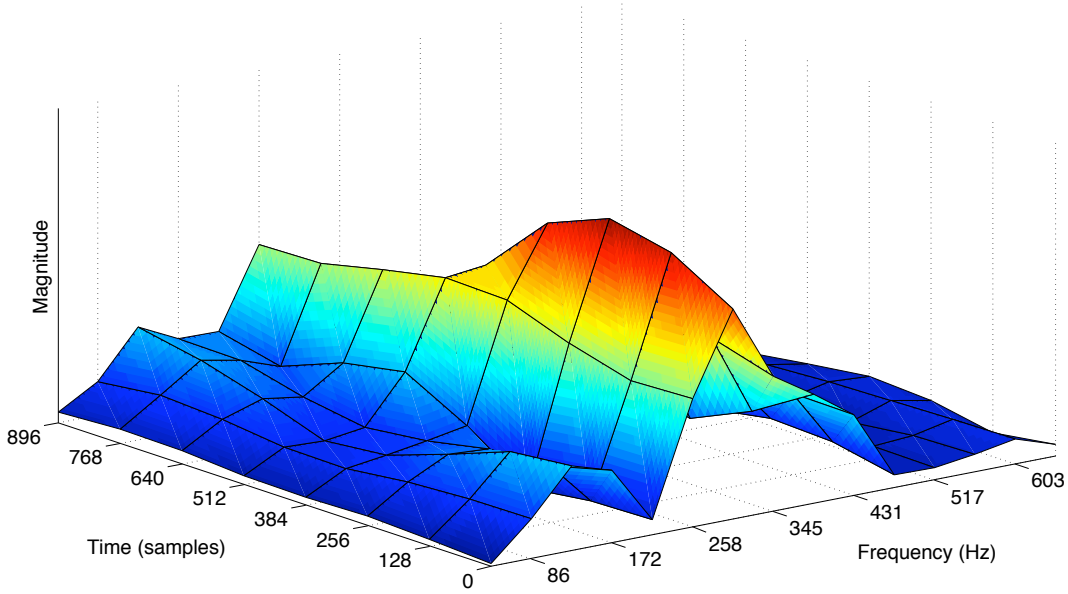


Figure 5.15: Low frequency spectrogram of a tam tam drumstick strike (middle).

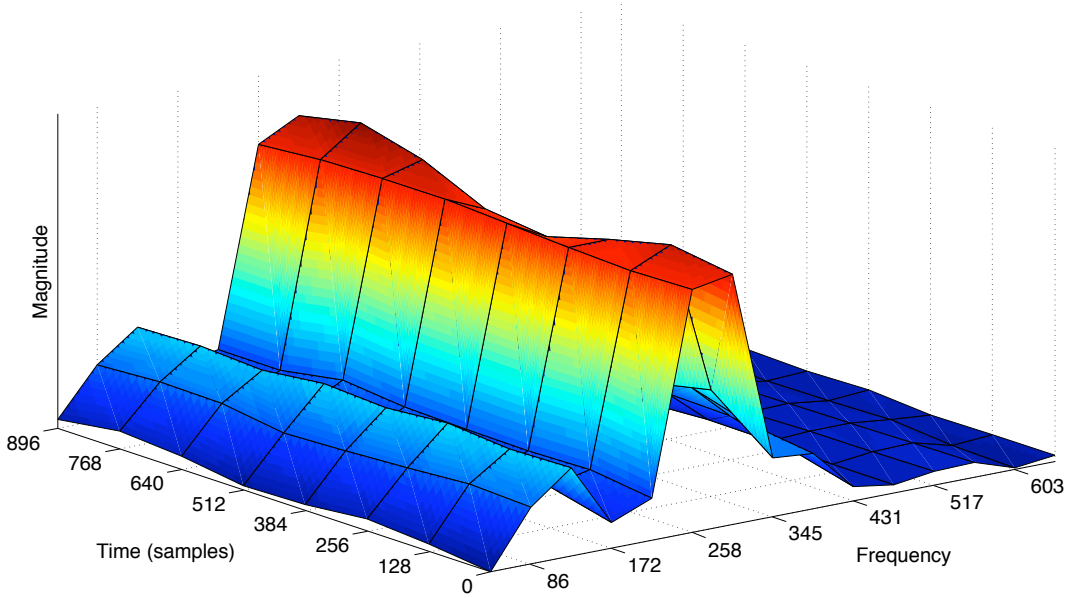


Figure 5.16: Low frequency spectrogram of a tam tam drumstick strike (center).

imum of two errors in each test. Four out of the seven errors were confusions between the edge and center felt mallet strikes of the second tam tam. Of the remaining errors, two were confusions between felt mallet strikes of the center of

the second tam tam and the middle of the third tam tam, and one confused edge and middle strikes of the first gong using a drumstick. Thus, five out of the seven errors failed to identify the *location* of the strike, but not the instrument or mallet type. In no cases was a drumstick attack mistaken for a felt mallet attack, as the corresponding spectra have very different energy distributions, with the felt mallet producing mainly low- and mid-frequency energy.

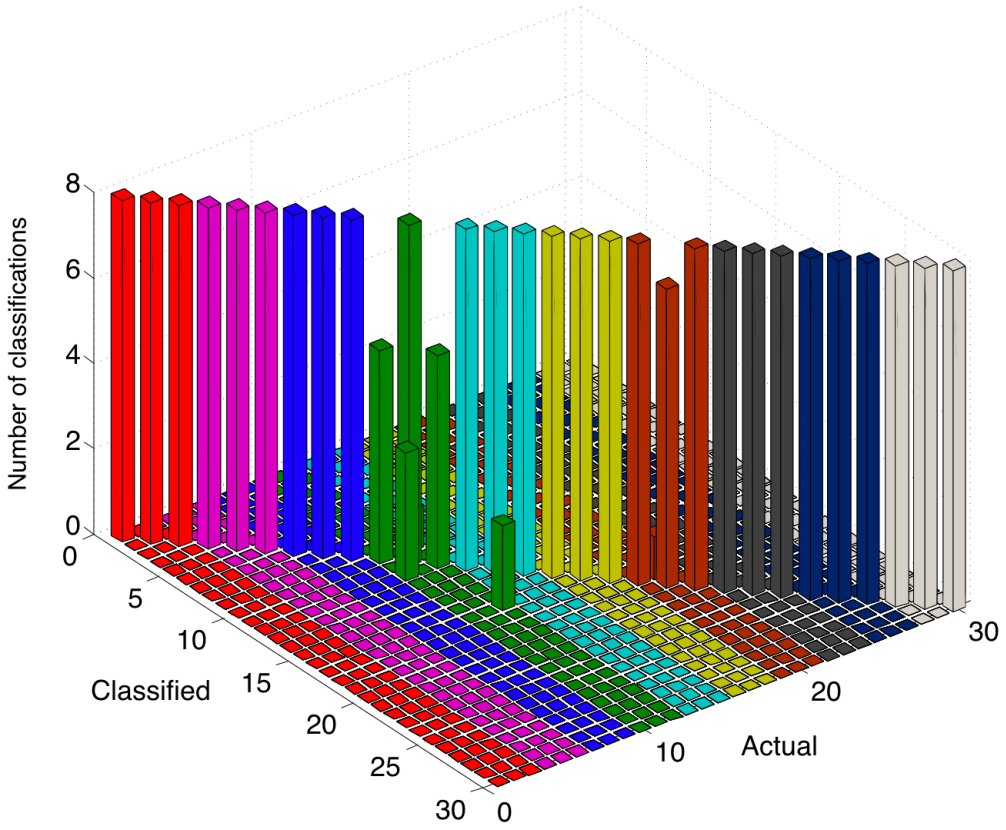


Figure 5.17: Confusion matrix for classifications using 5 multiple-frame magnitude spectrum coefficients.

A three-dimensional confusion matrix for classifications using magnitude spectrum is presented in Figure 5.17. The uniform diagonal is another reflection of the 96.67% accuracy rate, where deviations (such as those in the green band corresponding to timbres 10—12) illustrate the nature of misclassifications. Most misclassifications remain near the diagonal because neighboring timbres share the same instrument and mallet type. Unique instrument/mallet combinations are

colored differently, where each group of three represents edge, middle, and center strikes using the same instrument and mallet. Thus, the two instrument-based errors are easily identified visually in the green region as the bar most displaced from the diagonal. The gong edge/middle confusion is on the opposite side of the main diagonal in the brown region (timbres 19–21). At a glance, it is clear that the majority of misclassifications are location-based.

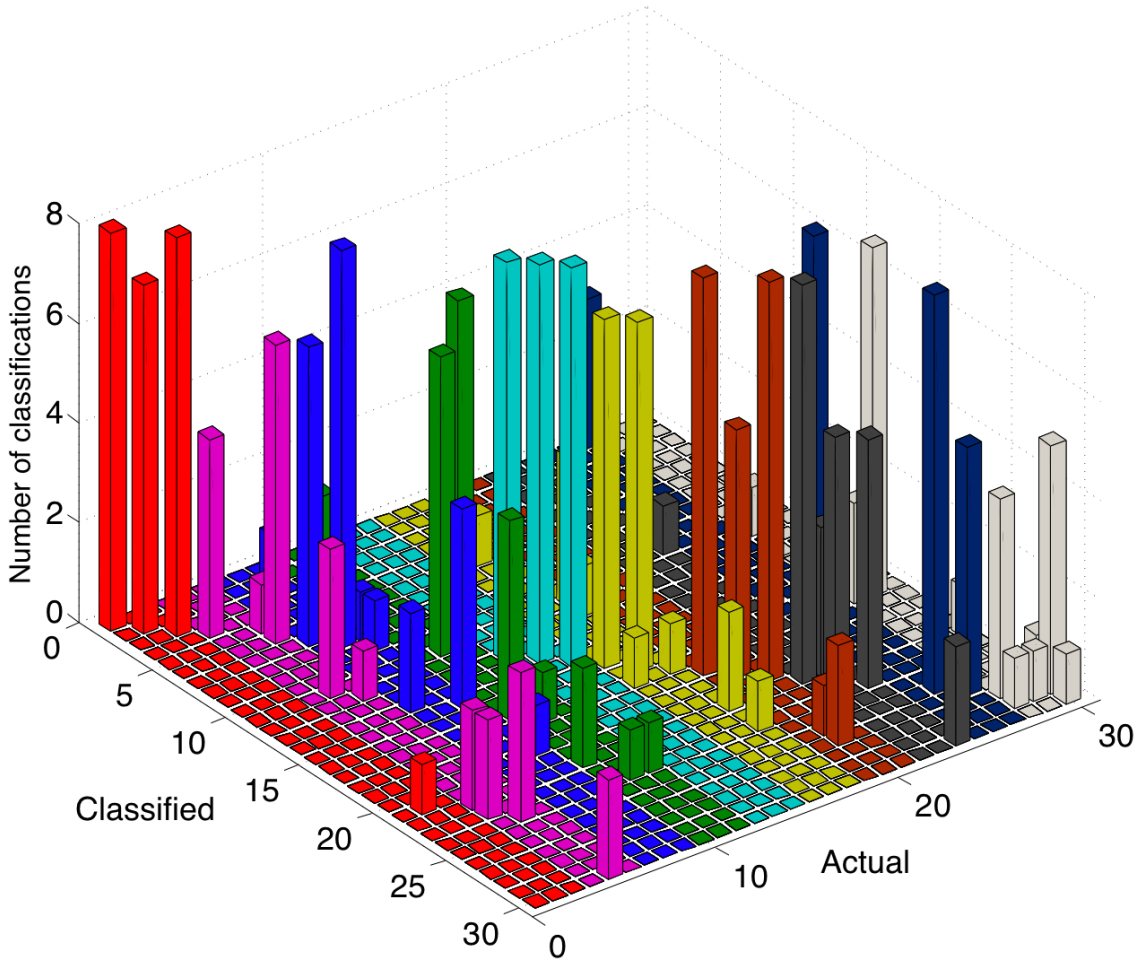


Figure 5.18: Confusion matrix for classifications using 5 multiple-frame Bark cepstrum coefficients.

Misclassifications resulting from the use of 5 BFCCs are not only more numerous—as illustrated previously in Figure 5.13—they are also more likely to make errors of a less subtle nature for particular instrument/mallet combinations. In Figure 5.18, the three rightmost timbres (grey) refer to the second gong struck

with a felt mallet. Deviations from the diagonal are clustered nearby with only one outlier from an incorrect instrument altogether. However, displacements in other areas are far more scattered, illustrating the inferiority of this feature with respect to the nature of misclassification.

5.3.3 Signal distortion

By adding various levels of white noise, signal interference from other instruments can be simulated in a controlled manner. Results from this test give an indication of how strongly classification accuracy will be affected when microphone leakage occurs. White noise is an imperfect simulation—actual instrument resonances will not have such even spectral characteristics. However, this approach has the advantage of being straightforward and uniform. Figures 5.19 and 5.20 display the results of adding white noise at -36 dB and -42 dB to the diverse and similar sound sets using the optimal multiple-frame analysis approach. At -36 dB, this distortion completely masks some of the quieter instrument samples. In the case of the similar timbre set, the testing conditions are therefore extremely unfavorable.

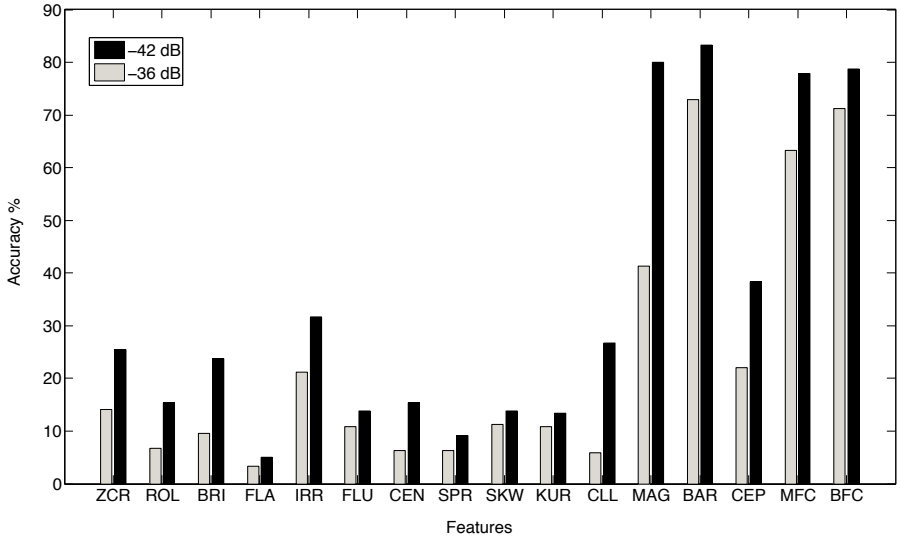


Figure 5.19: Accuracy for all features when adding white noise at -36 dB (grey) and -42 dB (black) to the diverse timbre set.

For the diverse timbre set, scores are lower overall, with average accuracy of some low level features indistinguishable from random chance (3.33%). For in-

stance, the accuracy of spectral flatness, which produced scores over 50% in tests on the unaltered diverse set, does not outperform chance at either noise level. The spectral flatness of white noise on its own is very near the maximum value of 1.0. Therefore, this feature is especially susceptible to the type of interference tested here. Of the most effective low level features identified previously (zero crossing rate, spectral rolloff, spectral brightness, and spectral centroid), zero crossing rate and brightness fare better than others. The relatively high score of spectral irregularity may be due to its higher invulnerability to doses of broadband noise, as frequency peaks rising above neighboring spectral noise should still generate relatively unique values.

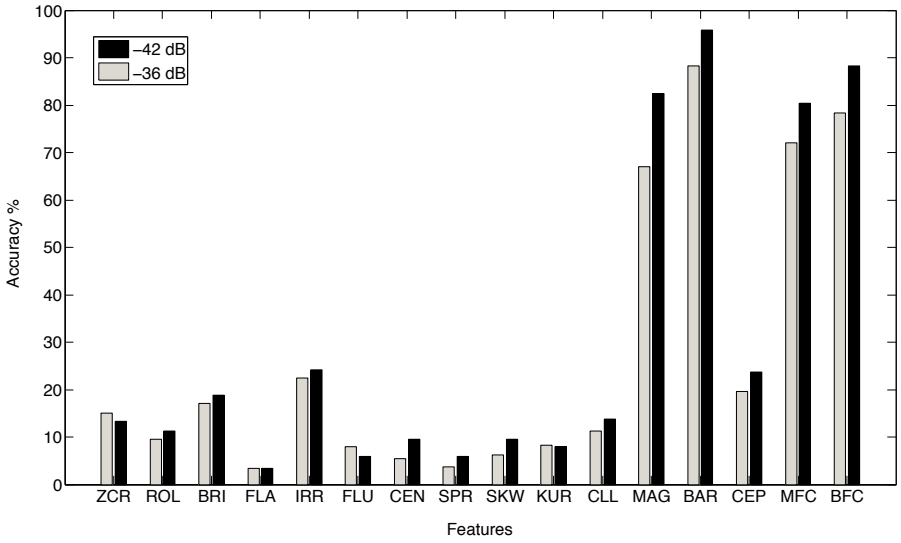


Figure 5.20: Accuracy for all features when adding white noise at -36 dB (grey) and -42 dB (black) to the similar timbre set.

The 6 dB difference in noise level causes a roughly 2:1 difference in accuracy when using magnitude spectrum and cepstrum. Proportional differences between the scores of all high level features are noticeably different than before, with Bark spectrum, MFCCs, and BFCCs significantly outscoring magnitude spectrum at the higher noise level. This difference can be attributed to the fact that magnitude spectrum preserves high frequency noise, while for Bark- and mel-weighted features, this spectral region is drastically attenuated. Likewise, raw cepstrum is also susceptible to high frequency noise, and its performance is further diminished

from the already poor scores given in Figure 5.12. Cepstrum scores amount to about half those of magnitude spectrum at both noise levels. Computed according to Equation (3.12), cepstrum is the only high level feature making use of a log magnitude spectrum. Large doses of noise across the entire frequency range are amplified in the logarithmic frequency domain, resulting in a spectral envelope representation that differs greatly from those stored during the training process. The presence of noise in linear magnitude spectra is less prominent.

Identical distortion applied to the similar timbre set produces predictable results. Spectral flatness remains completely ineffective as a classifier at both noise levels. As with the noise-treated diverse set, zero crossing rate, spectral brightness, and spectral irregularity are the most effective low level features. In comparison with other high level features, raw cepstrum is further crippled, and is outperformed by an individual low level feature (spectral irregularity) for the first time. Mel and Bark cepstrum hold roughly the same ∼80% score range shown in the case of the diverse set at -42 dB, with scores reduced by about 10% at the higher noise level. As in Figure 5.19, scores for magnitude spectrum are eclipsed by the perceptually weighted features, though less drastically in the case of noise at -36 dB. Scores generated by perceptually weighted features are thus superior due to their de-emphasis of high frequency noise. It can be concluded that these features are the most robust in the face of signal distortion.

5.4 Conclusions

These tests have confirmed some important characteristics of audio features and analysis techniques in the context of the timbreID analysis package. First, the general inferiority of individual low level features is apparent in the case of each analysis strategy and both timbre sets. However, using 5 overlapping analysis frames, some individual features produced accuracy in the range of 60% and 70% for the diverse timbre set. In this case, spectral rolloff generated the highest average score of 72.5%, or 21.75 correct classifications out of 30. In an equally weighted combination, the ten low level features achieved 97.5% under the same circum-

stances. Using the summarized multiple-frame approach, results were consistently between the bounds of single and multiple frame scores.

High level features performed very reliably in tests on the diverse timbre set. Distinctions between these measures were made relative to data reduction capabilities. In general, Bark spectrum stayed in very nearly the same score range as magnitude spectrum, and is much more compact (46 vs. 511 points). The three cepstral techniques exhibited superior performance using fewer than 10 coefficients for all analysis strategies. Thus, cepstral techniques were optimal in the classification test of diverse timbres. Mel- and Bark-weighted cepstral features were consistently more effective than raw cepstrum in all cases. In all testing scenarios, mel and Bark cepstrum performed essentially identically. In a break from the conventional calculation of these cepstral techniques (which is based on the sum of weighted log power spectrum bins in each band of the filterbank [DM80]), the average of weighted magnitude spectrum bins was used. Based on the study presented here, multiple-frame mel or Bark cepstra calculated in this manner are the most appropriate features for real-time classification of a diverse collection of percussive timbres.

A more difficult test performed using similar metallic timbres produced some contrary results. Scores generated using 15 MFCCs or BFCCs were in the same range as those for magnitude spectrum. However, for the 5 metal instruments investigated, temporal patterns in low frequency modes as captured by 5–10 magnitude spectrum bins were more effective in classification than the general spectral summarization offered by similarly sized subsets of MFCCs or BFCCs. These results show that additional investigation of spectral characteristics among similar sound sets can lead to carefully composed feature vectors that outperform the lowest BFCCs. Under circumstances of signal distortion, it was shown that the de-emphasis of high frequency content produced by perceptual scale weighting offered an advantage over full magnitude spectrum. Averaging rather than summing weighted energy in filter bands further de-emphasized high frequency content. The modular design of timbreID feature extraction objects greatly facilitates the process of composing custom features as needed for any of these conditions. Under cir-

cumstances where higher spectral resolution can capture unique spectro-temporal patterns, custom timbre descriptors may also reduce processor load. In the case of the similar timbre set, the steps of multiplying against a Bark-weighted filterbank and performing a DCT were shown to be unnecessary.

In summary, with a timbrally diverse sound set, a small subset of BFCCs can produce very high scores, and requires no additional research. For more similar sound sets, BFCCs produce useful scores, but can be outperformed using a subset of magnitude spectrum bins if the sound set can be considered in detail before classification.

Chapter 6

A Perceptual Timbre Space for Percussive Sounds

The studies reviewed in Chapter 2 established at least two consistent perceptual dimensions of timbre: brightness and attack time. Although there was a great deal of variation in the particularities, a third dimension relating to spectral changes over time was also identified in several cases. With the exception of [Fre90] and [Lak00], the majority of studies explored pitch-based orchestral instruments in the categories of strings, winds, and brass. This chapter presents a study investigating 30 percussive timbres selected from the orchestration of Varèse's *Ionisation*. In the previous chapter, these timbres were referred to as the "diverse" set. The full list of instruments are given in Table 5.1. A primary objective of the experiment described here is to discover dimensions of timbre that are specific to percussion instruments, and to confirm established dimensions that may be common to all types of instruments. A second objective is to identify correlations between perceptual dimensions and the physical measurements employed in Chapter 5, with the intent of creating predictive models of subjective judgments.

6.1 Method

In light of the experimental history outlined in Chapter 2, it was decided to record participant judgments on verbal attribute scales. The advantages offered

by multidimensional scaling (MDS) were considered to be outweighed by its main disadvantage: a severe restriction on the size of the stimulus set. In this case, the investigation of 30 sounds would have required $30 \times 29 \times 0.5 = 435$ trials. A unique feature of this study is the relatively large stimulus set, which allows for further investigation of context effects with respect to common perceptual dimensions. This was a stated goal of Lakatos' experiment (Section 2.3.4), which studied 20 diverse sounds. Most other MDS studies consider roughly half the number of timbres used here. A second motivation for choosing a verbal attribute strategy is that the chosen adjectives can provide insight regarding particular factors that contribute to the primary perceptual dimensions discovered. As noted by Lakatos, MDS algorithms are designed to produce a parsimonious model of data, which can be "unsatisfying" in the context of a phenomenon as rich as timbre perception. [Lak00, p. 1437] Because the underlying timbre characteristics that govern general similarity judgments are not known when interpreting the dimensions of a space generated by MDS, in most cases only a single attribute is proposed for each axis. The use of verbal attribute scales offers the possibility of understanding the underlying aspects of primary dimensions in more than one respect.

The disadvantages of a verbal attribute approach have already been discussed in Chapter 2. In short, the chosen adjectives—no matter how carefully they are selected—may influence the aspects of timbre to which participants attend. There is also no assurance that participants will use the given scales similarly and consistently. Regarding the former point, a body of MDS-based timbre research has sufficiently established recurring perceptual dimensions on a nonverbal basis. Thus, it may be constructive at this point to risk influencing participant judgments with verbal suggestions in order to gain further understanding of the meaning behind common dimensions like "brightness". This is something that was recognized by Plomp, one of the earliest advocates of MDS in timbre studies:

> When at some time these [MDS] experiments have given a clear picture of the multidimensionality of timbre perception in its dependence on the physical parameters, it would be of great interest to investigate the relationship between the dimensions found and the verbal categories by which the timbre differences can be described. [Plo70, p. 414]

Regarding language consistency, the participants in this study were all percussionists sharing a common performance practice. Therefore, the relative stability of terminology should be quite high. Five of the six participants rehearse and perform together on a regular basis, where nuances of instrument articulation are frequently discussed in relation to timbre. While language will always introduce a degree of ambiguity, the scenario here can be considered optimal.

Adjectives for the rating scales were chosen from the collection given in [KC93b]. As described in Section 2.3.2, Kendall & Carterette's original list of 61 adjectives was drawn from a musically relevant text. This list was edited by the present author to remove terms that reference the physical qualities of instruments (e.g., "metallic" and "wooden"). Several adjectives from [vB74b] were included, and adjectives deemed appropriate for unpitched percussion timbres (e.g., "noisy") were also appended. The complete list of 100 adjectives is given in Appendix B. One of the participants in the present study examined the collection and identified 15 adjectives as the most relevant descriptors of percussive timbre. These adjectives are given in Table 6.1.

Table 6.1: Fifteen adjectives used for the VAME rating scales.

Sharp	Rough	Deep
Dry	Pure	Rich
Bright	Round	Noisy
Shrill	Warm	Dull
Dead	Brilliant	Thin

Note that some adjective pairs are potential synonyms and antonyms. Following the logic for attribute scale design given in [KC93a], antonymous scales were avoided and 15 independent 7-point VAME rating scales were employed. This avoids the ambiguity of scales with many potential antonyms, or no clear antonym at all. Participants were instructed to interpret the left and right limits of each scale as "not at all" and "extremely" respectively (e.g., "not at all bright" and "extremely bright"). Under this approach, antonymic relationships should be apparent

as strong negative correlations between judgments on different scales. Synonyms will likewise be identified as positive correlations.

6.1.1 Participants

Participants were 6 graduate students from the UCSD music department. Five participants were percussion performance majors and members of the resident contemporary percussion ensemble. One participant was a percussionist studying within the computer music area of the department. All participants were native English language speakers, and none were paid for participation in the experiment.

6.1.2 Apparatus

Stimuli were presented via two interactive Pd patches. The first patch was designed merely to familiarize participants with the sound set, and did not record judgments. The patch generated unique random orderings of the complete sound set, and participants could play back individual sounds by pressing the space bar. After each sound was played back, the next sound in the random sequence was cued up automatically. Once all 30 sounds had been played back, the patch automatically reorganized the set in a different random order so that participants could continue to listen to the stimuli informally if desired.

The second patch, pictured in Figure 6.1, allowed participants to rate each sound on 15 verbal attribute scales. As in the previous patch, the playback order was randomized and unique for all participants. Sounds were played back via the space bar (or the GUI "PLAY" button), and in this case could be repeated as many times as desired by the participant. It was not possible to repeat a sound until active playback had completed, which forced participants to listen to each sound in its entirety. A progress bar at the bottom of the window provided feedback regarding the remaining duration of each sound. A volume adjustment slider was provided as well.

After participants made ratings on all 15 scales and clicked the "RATE" button, the next sound in the sequence was automatically cued up, and it was not

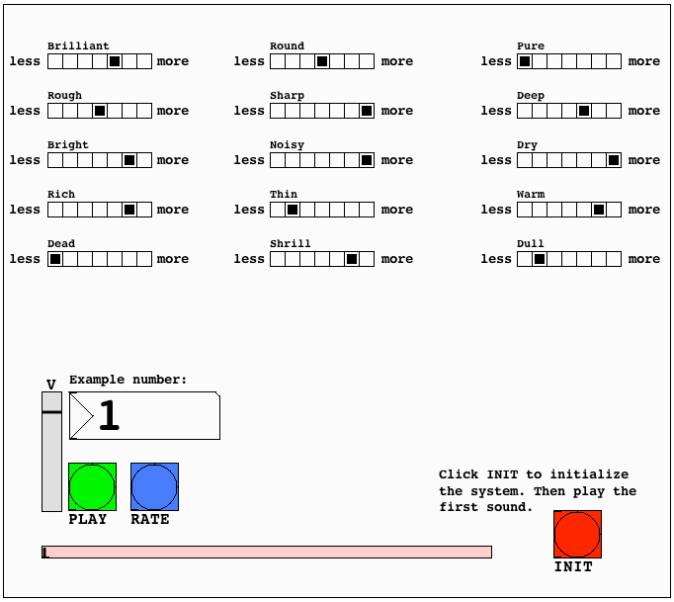


Figure 6.1: User interface for auditioning stimuli and recording judgments.

possible to edit previous ratings. In order to avoid ordering effects, the positions of the 15 scales were randomized on each trial. The initial states of the 7-point scales were also randomized on each trial. Labels to the left and right of the scales were provided as reminders of the meanings for minimum and maximum values. This was added after a pilot stage revealed occasional confusion. The left—right/less—more orientation was apparently more intuitive for some scales than for others.[1] For each trial, the patch recorded the following information to a separate text file: the stimulus number, the rating values on all 15 scales, the number of times the stimulus was played back, and the trial duration.

6.1.3 Stimulus Materials

As noted previously, this collection of timbres possesses characteristics that are of interest in the context of a perceptual study; however these features also present some procedural difficulties. First, because the majority of instruments are unpitched, there is no clear strategy for normalizing pitch in the set. Even

[1]For instance, making a rating on the right side of the scale for a very "bright" sound seemed natural for some pilot participants, but going far to the right for an extremely "deep" sound was less intuitive.

among the pitched instruments—including the claves, bass drum, tom, glockenspiel, and wooden plank—not all are capable of producing a range of pitches. In [Lak00], pitches for some stimuli were corrected using pitch-shifting software. Here, unaltered stimuli are prioritized, and the degree of pitch-shifting necessary for normalization would have caused severe changes in timbre. Interactions between pitch and timbre are thus unavoidable in the case of pitched instruments, and not problematic for the remaining unpitched stimuli.

With respect to duration, again there is no ideal method for normalization. While resonating instruments like tam tams and triangles can be dampened in order to control duration, other instruments like the castanet, claves, and snare drum offer no means for extending their extremely short durations. Further, because duration and loudness are interrelated when dealing with timbres possessing varied amplitude envelopes, loudness normalization was also problematic. Most of the sounds in this set have attack segments of 1 ms or less, but several are much longer. For instance, the loudest point in a lion's roar tone may be half a second after the initial onset. Typical articulations of maraca, guiro, and sleigh bell may also have amplitude peaks that are displaced by 100 ms or more from the onset. One possibility is to truncate the durations of all instruments to that of the shortest instrument, making loudness normalization much more manageable. While this is also worthy of investigation, in the present study a choice was made to avoid such artificial modifications and investigate stimuli with durations that are typical in actual musical practice. Thus, loudness could not be normalized as precisely as in other cases. All instruments were recorded at several different dynamic levels, and mezzo-forte instances were selected for the stimulus set. Sounds ranged in duration from 202 ms (castanet) to 7862 ms (tam tam), with a mean duration of 2274 ms.

6.2 Procedure

Participants were told that they would be rating 30 timbres on 15 attribute scales. In the first stage of the experiment, participants interacted with the intro-

ductory patch to hear all of the sounds in the set. They were required to listen to the entire set at least once, and most participants listened to the set two times. A list of the 15 adjectives from Table 6.1 was shown in the bottom portion of the patch, and participants were instructed to imagine making ratings using these descriptors as they listened casually.

In the second stage, ratings were made on all scales at every trial, following [vB74b]. The positions of the scales shifted randomly on the screen, and participants were instructed to proceed from top to bottom, left to right. Thus, the order of scale ratings was different for each trial. The interface for each scale was an instance of Pd's graphical radio button object with 7 elements. The desired point on the scale could be chosen by clicking with the mouse. Participants were told to listen to each sound as many times as necessary. It was recommended that they listen to the sound at least once for every adjective scale. No time limit was imposed, but at least one 10-minute break was required halfway through the test. Additional breaks were allowed if needed. Both stages of the experiment were carried out using high quality AKG 702 reference headphones in a soundproof, non-reverberant room. In total, the experiment lasted approximately one hour.

6.3 Results

6.3.1 Consistency of Ratings

One method for discerning the similarity of ratings between participants is to look at the standard deviations of judgments for each instrument. The average standard deviation for randomly generated ratings on a 7-point scale is about 1.93. Because participants are not likely to have a uniform approach to using the scales, it is relevant that standard deviations of participant ratings were far below the chance value for many scales. In very rare cases, ratings did not deviate at all. This occurred in extreme examples, such as ratings of "brightness" for triangle and orchestral crash cymbals, "roughness" for guiro, and "shrillness" for bass drum. Regarding guiro, the extreme consistency indicates that in certain cases, the nature of a particular adjective may lead to ratings that are more descriptive of the

physical properties of the instrument than its purely sonic characteristics. Percussionists are aware that the guiro's grooved surface is literally rough to the touch; of course, this has nothing to do with the psychophysical sensation of roughness documented by Terhardt [Ter70].

Table 6.2 shows rating scale adjectives ordered according to mean standard deviation across all instruments. The most consistently used scales were relatives of "bright", "sharp", and "noisy". Although sharpness is discussed in [vB74b] as synonymous with brightness, use of the term among these participants is in reference to attack quality. A sound with a short attack is said to be sharper than one with a gradual attack.

Table 6.2: Adjectives ordered by mean standard deviation.

	Mean std. dev.
Deep	0.98
Sharp	0.99
Bright	0.99
Pure	1.03
Noisy	1.11
Thin	1.12
Dead	1.20
Round	1.20
Warm	1.22
Rich	1.23
Dry	1.25
Rough	1.30
Brilliant	1.32
Shrill	1.32
Dull	1.43

A different perspective on rating consistency is given by examining correlations between ratings for all possible pairings of participants. With 6 participants,

Table 6.3: Adjectives ordered by mean inter-participant correlation.

	Mean correlation.
Deep	0.70
Pure	0.65
Noisy	0.64
Sharp	0.64
Dry	0.62
Bright	0.60
Rich	0.58
Shrill	0.55
Thin	0.55
Brilliant	0.53
Dead	0.52
Round	0.50
Rough	0.44
Warm	0.34
Dull	0.23

15 unique pairings are possible, and mean values can be determined for each adjective. The mean correlation coefficient is given for each adjective in Table 6.3. The top 6 adjectives in both Tables 6.2 and 6.3 include "bright", "deep", "noisy", "pure", and "sharp", indicating that these rating scales were used most consistently. In both cases, "dull" is the least consistently used adjective.

6.3.2 Adjective correlations

Correlations between VAME scale ratings point to adjectives with similar or opposite meanings in the context of timbre. Tables 6.4 and 6.5 respectively show roughly synonymous and antonymous adjective pairs with corresponding correlation coefficients. Correlations were generated from the mean judgments of all 6 participants, and only pairs with correlations greater than 0.7 are shown. All p-values are below 1.0e-05.

Table 6.4: "Synonymous" adjective pairs.

	Correlation
Noisy—Rough	0.91
Round—Warm	0.91
Brilliant—Shrill	0.91
Dead—Dry	0.91
Bright—Brilliant	0.90
Bright—Shrill	0.89
Dead—Dull	0.85
Deep—Warm	0.84
Deep—Round	0.84
Bright—Sharp	0.79
Dry—Dull	0.74
Sharp—Thin	0.73

The strong relationship between "bright" and "sharp" seems to contradict the point made earlier—that among these participants sharpness refers to attack

Table 6.5: "Antonymous" adjective pairs.

	Correlation
Pure—Rough	-0.91
Pure—Noisy	-0.89
Bright—Deep	-0.86
Round—Sharp	-0.84
Round—Bright	-0.82
Bright—Warm	-0.80
Deep—Sharp	-0.79
Brilliant—Dull	-0.76
Sharp—Warm	-0.76
Round—Thin	-0.75
Shrill—Warm	-0.75
Deep—Thin	-0.75
Dead—Rich	-0.74
Dry—Rich	-0.73
Deep—Shrill	-0.72
Round—Shrill	-0.72

quality, not high frequency content. This is likely the result of a forced correlation due to the fact that most timbres in this set with shorter attacks happen to have high spectral centroid and brightness values as well. However, the maraca, guiro, and sleigh bells are exceptions that should provide opportunities to understand how the two dimensions differ. Thus, while this correlation is strong, it should not be assumed that brightness and sharpness are completely dependent. Furthermore, it must be stressed that in general, even though adjectives with strong correlations will be referred to as "synonyms" for convenience, ratings on each scale contain unique information. Though the timbre set here is relatively diverse, it is not sufficiently diverse to make truly general claims about the interchangeability of different adjectives applied to timbre.

6.3.3 Physical Correlates of Perceptual Judgments

In Chapter 5, it was shown that 15 BFCCs (i.e., BFCCs 2—16) were more effective than combined low level features in a classification test. Because the 2nd BFCC has a strong negative correlation with spectral centroid, it also relates to adjective scales that are generally linked to brightness. The 3rd BFCC (which measures Bark spectral correlation with a full cosine and is therefore a good indicator of strong mid-range frequency content) correlates fairly well with several adjective scales as well. Higher BFCCs produce very few correlations with magnitudes above 0.4. In the context of this experiment, the assortment of low level features specified in Table 5.2 were more useful than the systematic spectral summarizations offered by BFCCs.

As before, all analysis was performed on the attack segment using a window size of $N = 1024$. Undoubtedly, this approach is flawed in some regards, as participants made ratings based on hearing entire instrument articulations. However, such inconsistencies are unavoidable when investigating a heterogenous set of stimuli that possess various durations. The attack segment was chosen as an obvious target of analysis, as several studies have noted the importance of attack in classification experiments [IK93]. Log attack time (ATT) was added to the ten low level features employed previously, measured manually as the time from onset to

peak amplitude. The remaining low level analyses were performed on peak regions rather than actual onsets for sounds with longer attack times (including maraca and lion's roar), in order to capture the most characteristic segment of all sounds.[2] High correlations with this set of features are more numerous than those for the lowest BFCCs, with coefficient magnitudes greater than or equal to 0.5 found between each feature and at least one rating scale. Table 6.6 shows correlations with magnitudes greater than 0.5 and p-values below 0.01. Spectral flux and log attack time yield the weakest relationships, while some measures of high-frequency content produce coefficient magnitudes above 0.8. Spectral moment measures (CEN, SPR, SKW, & KUR) also fare well, with several magnitudes above 0.7.

It is apparent that some adjective scales possess very few strong relationships to physical measures. These are "dead", "dry", and "rich", although richness appears to have some relationship to spectral fourth moment (KUR). Patterns expected to follow from synonym/antonym relationships pointed out in the previous section can be seen in several cases. Approximate synonyms like "noisy" and "rough" have similar correlation patterns across measures of spectral flatness (FLA), irregularity (IRR), and flux (FLU), while the contrary scale, "pure", shows a precisely opposite pattern. Spectra of "pure" tones, like the glockenspiel, generally have only a few strong isolated frequency peaks during the attack, which generate very high irregularity values. A "noisy" tone will generate lower irregularity values because, without strong isolated peaks in the spectrum, neighboring bins are more likely to share values in a similar range. "Round" and "warm" have similar patterns across spectral rolloff (ROL), brightness (BRI), and all spectral moments, and as expected, "bright", "brilliant" and "shrill" correlate in an opposite direction from "round" and "warm" in all cases. Spectral skewness, which is positive when the spectral energy distribution tail on the high frequency side slopes toward zero more gradually, appears to be a potential predictor of warmth and roundness judgments that are linked to a predominance of low frequencies. A negative skewness indicates the opposite spectral distribution, such as that resulting from high-frequency bursts associated with striking a tam tam with a drum-

[2]For instance, a weak pre-attack exists for maraca articulations when pellets inside the hull gather prior to the main accentuation of the sound. Here, the perceptual onset was analyzed.

Table 6.6: Judgment/low-level feature correlations (*mag.* ≥ 0.5, $p < 0.01$).

	ZC	ROL	BRI	FLA	IRR	FLU	CEN	SPR	SKW	KUR	ATT
Shrill	0.56	0.65	0.66	-	-	-	0.61	0.64	-	-	-
Rough	-	-	-	0.58	-0.66	0.51	-	-	-	-	0.52
Pure	-	-	-	-0.51	0.69	-0.53	-	-	-	-	-0.53
Dull	-0.56	-0.55	-	-	-	-	-0.53	-	-	-	-
Dead	-	-	-	-	-	-	-	-	-	-	-
Warm	-	-0.58	-0.84	-	-	-	-0.57	-0.62	0.70	0.64	-
Sharp	-	0.53	0.71	-	-	-	0.50	0.65	-0.75	-0.70	-
Thin	-	-	0.57	-	-	-	-	-	-0.59	-0.57	-
Noisy	-	0.53	-	0.62	-0.65	0.52	-	0.50	-	-	-
Round	-0.57	-0.68	-0.85	-0.57	-	-	-0.68	-0.70	0.77	0.68	-
Brilliant	0.61	0.66	0.68	-	-	-	0.64	0.63	-	-	-
Dry	-	-	-	-	-	-	-	-	-	-	-
Bright	0.63	0.72	0.79	0.54	-	-	0.70	0.74	-0.61	-0.50	-
Deep	-	-0.50	-0.73	-	-	-	-0.52	0.53	0.60	0.55	-
Rich	-	-	-	-	-	-	-	-	-	0.54	-

stick. Thus, negative skewness values are likely to result in positive correlations with brightness judgments and the brightness measure. However, skewness is measured relative to spectral centroid while the brightness measure is based on a fixed boundary frequency. Certain signals may produce positively skewed spectra that are nevertheless concentrated beyond the brightness boundary. Such possibilities point toward the advantage of multifaceted spectral distribution analysis methods, even though some measures may seem redundant at first glance.

6.3.4 Principal Components Analysis

With potentially unique but highly intercorrelated rating scales, principal component analysis (PCA) offers a means of identifying the most fundamental dimensions upon which participants made their ratings. In this case, the 15 adjective dimensions can be summarized by only three principal component dimensions that explain 90.45% of the variance existing in the original mean judgment data. The loadings of each adjective scale on the three dimensions after varimax rotation are given in Table 6.7. The leftmost column shows that for the first principal component (PC1), information is drawn most heavily from brightness-related scales, including "shrill", "sharp", "round", "brilliant", "bright", and "deep". Scales load negatively and positively as expected. For instance, "deep" and "bright" load oppositely. Of the six brightness-related scales just mentioned, most weigh on the first PC dimension fairly exclusively (i.e., their weight is disproportionately concentrated on this dimension). The second PC axis is slightly less conclusive. Its most significant loadings come from the "dry" and "dead" scales, which load primarily on this dimension alone. Thus, PC2 can be referred to as the "dryness" dimension. Scales for "rich", "dull", and "brilliant" have a lesser—but still significant—influence on this axis. While "dull" loads exclusively on PC2, its potential antonym from Table 6.5, "brilliant", is also a significant contributor to PC1, the brightness dimension. Both the "brilliant" and "rich" scales load negatively on PC2. Finally, the third PC axis is determined mostly by noise-related scales, including "noisy", "pure", and "rough". As expected "rough" and "noisy" load negatively, while "pure" has a positive loading. All three scales load nearly ex-

clusively on PC3. In summary, the three principal component dimensions point to three primary perceptual dimensions for these percussive timbres: brightness, dryness, and purity.

Table 6.7: Adjective loadings on the first three principal component dimensions after varimax rotation. Loading magnitudes of 0.3 or greater are highlighted in blue.

	PC1	PC2	PC3
Shrill	0.35	-0.23	-0.05
Rough	0.01	0.02	-0.52
Pure	0.03	0.00	0.56
Dull	-0.08	0.30	0.03
Dead	0.03	0.48	-0.04
Warm	-0.28	0.00	0.02
Sharp	0.37	0.08	0.02
Thin	0.28	0.20	0.04
Noisy	0.06	-0.01	-0.60
Round	-0.35	-0.05	0.04
Brilliant	0.31	-0.30	0.03
Dry	0.11	0.61	-0.09
Bright	0.37	-0.13	0.00
Deep	-0.45	-0.03	-0.05
Rich	-0.14	-0.30	-0.20

At this point it is worthwhile to consider the connections between consistent scale use and the dimensions revealed by PCA. Recall that the five most consistently used scales were “bright” “deep”, “noisy”, “pure”, and “sharp”. The first and third perceptual dimensions were identified based on ratings from all of these scales, which supports the validity of PCA results. That is, the scales that were used most consistently by participants are also those that contribute to determining positions on the PC axes. And, although rating scales for “dry” and “dead” (the main rating scales for the second perceptual dimension) were less consistent

according to the orderings given in Tables 6.2 and 6.3, they did earn high placement in comparison with the remaining scales. For instance, the "dry" scale was the fifth element when ordering according to average inter-participant correlation.

Table 6.8: Correlations between adjectives and physical measures after decorrelation from PC1 ($mag. > 0.5$, $p < 0.01$).

	ZCR	ROL	FLA	IRR	FLU	SPR	ATT
Rough	-	-	0.62	-0.66	0.51	-	0.57
Pure	-	-	-0.58	0.69	-0.56	-	-0.54
Dull	-	-0.50	-	-	-	-	-
Dead	-0.54	-0.51	-	-	-	-	-
Sharp	-	-	-	-	-	-	-0.60
Noisy	-	0.53	0.61	-0.67	-	-	-
Deep	-	-	-	-	-	0.54	-

Sequential decorrelation of the first three rotated principal components from all judgments and measurements further elucidates relationships between the most salient adjective scales and physical measures, and reveals some remaining structure in the data that may amount to a fourth perceptual dimension. Expunging the brightness PC removes the majority of strong correlations between brightness-related adjective scales and measurements. Judgment/measurement correlations taken after PC1 decorrelation with a magnitude above 0.5 and p-value below 0.01 are shown in Table 6.8. No correlations meeting these criteria were found between any measures and the "bright", "brilliant", "round", "shrill", or "warm" scales. The "deep" scale, however, remains connected with spectral spread. Thus, there may be aspects of "deep" ratings that are independent of brightness and spectral centroid. The "sharp" scale also remains, but its only strong correlation is with log attack time—relationships with spectral rolloff, brightness, and centroid are erased. The largest correlation magnitudes in Table 6.8 are linked to noisiness, with the "rough", "pure", and "noisy" scales retaining correlation coefficients similar to those shown in Table 6.6, prior to PC1 decorrelation.

Table 6.9: Correlations between adjectives and physical measures after decorrelation from PC1 and PC3 ($mag. > 0.5$, $p < 0.01$).

	ZCR	KUR	ATT
Dead	-0.51	-	-
Sharp	-	-	-0.62
Rich	-	0.55	-

Because the correlations in Table 6.8 are associated most clearly with purity/noisiness, the third PC is decorrelated next. Having removed both the brightness and purity dimensions, we expect to find relationships involving any of the scales loading on PC2, i.e., "dry", "dead", "rich", "brilliant" or "dull". Indeed, both the "dead" and "rich" scales turn up correlation magnitudes above 0.5 when compared with measurements for zero crossing rate and spectral kurtosis respectively. Table 6.9 shows the results. Note that the negative correlation between the "sharp" scale and log attack time remains from the previous table. When the final (2^{nd}) PC is decorrelated from this data set, the only remaining correlation with a magnitude greater than 0.5 and significance of $p < 0.01$ is found between log attack time and the "sharp" scale ($mag. = -0.65$, $p = 0.001$). While sharpness and brightness judgments were shown previously to be correlated, the strong relationship that remains after correcting for brightness indicates that sharpness is indeed used to refer to quickness of attack. In all, these manipulations show patterns that point to one very strong perceptual dimension for brightness, a strong dimension for purity/noisiness, a dimension related to dryness, and a remaining dimension for sharpness of attack that is completely independent of the three PC dimensions. Based on these findings, future experiments with a different set of diverse timbres could be composed with the intent of avoiding overlap between dimensions and forced correlations (such as that found here between brightness and sharpness).

6.3.5 A Predictive Model

Having identified groupings of adjective scales that determine stimuli positions along the three perceptual dimensions, the next task is to exploit the correlations with physical measures shown above for predicting these key judgments. Table 6.6 indicates that there is no shortage of measures relating to brightness-oriented judgments. Apart from a link between richness and spectral fourth moment, direct correlates of deadness, dryness, and richness were not identified. The strongest relationships to purity/noisiness judgments are with spectral flatness, irregularity, and log attack time. It is clear from the magnitude of correlations that brightness judgments will be more easily predicted than those relating to the remaining perceptual dimensions.

Table 6.10: Mean squared errors for predicting 15 adjective scales using 6-fold cross validation.

	MSE
Shrill	1.97
Rough	2.54
Pure	2.54
Dull	0.89
Dead	0.90
Warm	0.45
Sharp	1.03
Thin	1.74
Noisy	3.50
Round	0.54
Brilliant	1.23
Dry	2.31
Bright	0.76
Deep	1.56
Rich	1.62

Multiple linear regression was performed to create multiple predictive models using a stratified 6-fold cross validation strategy. Stratification was carried out by ordering all stimuli by spectral centroid, dividing this list into 5 equal groups, and populating validation sets with one stimulus from each group. Spectral centroid was chosen as the basis for stratification because of its relationship with the primary perceptual dimension of brightness. The mean squared error across all folds for predicting each adjective scale is given in Table 6.10.

As expected, the error for predicting judgments on the "bright" scale is quite low, as are those for antonymous scales like "warm" and "round". Other rough synonyms and antonyms for brightness produced higher error rates. The heaviest loadings on the second PCA dimension were from "dry", "dead", and "rich". Table 6.6 shows these scales to have the fewest number of significant correlations with physical measures. However, the weighted combinations of all features produced by linear regression exhibit some predictive power for the "dead" scale. Sharpness predictions were also strong, but predictions relating to purity/noisiness were quite poor.

6.4 Conclusions

In the current and previous chapters, predictions related to percussive timbre characteristics were explored from two different perspectives. In the case of automated classification of a diverse set of timbres, it was shown that high accuracy rates could be achieved with multiple frame analysis and a small subset of Bark-frequency cepstral coefficients, or alternatively, using a combined feature composed of nine low level features describing spectral energy distribution and one feature that measures waveform zero crossings in the time domain. Under more difficult classification conditions, such as the use of a very similar set of timbres and the introduction of various levels of white noise, the BFCC subset was conclusively more robust than the combined low level features. It was also shown that if unique features of members of a sound set can be identified in advance, elementary measures like magnitude spectrum sub-band analysis can perform more reliably

than BFCCs. However, in automated classification applications, the systematic approach offered by BFCCs appears to be the most appropriate general method.

In contrast, the vague spectral summarizations of individual BFCCs proved less useful in predicting expert judgments of timbre relative to 15 VAME scales. The application of standard statistical measures of data distributions to spectral envelopes, and the additional measures presented in Chapter 3 yielded relevant information for these purposes. A specific focus on percussive timbres has uncovered two perceptual dimensions that have not emerged in the majority of previous research investigating pitched orchestral instruments: dryness and noisiness. While some significant correlations were found between the noisiness dimension and physical measures like spectral flatness and irregularity, these measures were not adequate for producing accurate predictions of noisiness judgments. For more robust predictions, a reliable measure of perceptual noisiness must be devised. Dimensions of timbre that have been identified in previous studies, such as brightness and attack quality, were found here as well. By adopting language that is actively used by a group of percussionists, it was shown that these dimensions relate to several adjectives, including "bright", "brilliant", "shrill", "deep", "round", and "sharp". Regarding the dryness dimension, there was an extreme disparity of stimulus durations in the set used for testing. Participants may have been predisposed to describing instruments with very brief durations like the castanet as "dry" in comparison with resonant instruments like the tam tam. If aspects of dryness beyond mere duration exist, future investigation of this possible dimension will require strict normalization of duration.

Furthermore, this study has only explored relationships between perceptual judgments and the first 23 milliseconds of instrument attacks. This was accepted as the least flawed method for analyzing and comparing sounds of greatly varied duration. It will be useful to uncover any differences in findings between this study and an experiment in which unnatural alteration of stimuli was deemed acceptable. As noted earlier, stimuli could be truncated to a common length to allow for very precise normalization of duration and loudness. Development of a method for normalization that does not impose such artificial changes would be extremely

beneficial, increasing the rigor of a study such as this while preserving its aim of investigating percussive sounds in a natural state. Perhaps more importantly, approaches for exploring timbre relationships in something closer to actual musical context must be advanced.

Finally, the question of whether certain dimensions may be better understood in a categorical rather than continuous sense must be explored further. The commonly uncovered dimension of brightness appears to be well served by continuous measures like spectral centroid. However, it seems unlikely that attack quality, which has turned up in several MDS studies, can be wholly explained by the simple duration of the attack segment. Log attack time is one significant measure, but as Grey's careful observations regarding synchrony of partial rise patterns illustrate, phenomena surrounding instrument attacks carry a great deal of complexity. It is too early to say that such patterns can be adequately expressed along continuous dimensions. In the case of percussive tones, with many possible instrument and mallet combinations, it is quite possible that *categories* of attack quality exist. Noisiness may be similarly problematic. Subjective judgments about sounds that are perceptually differentiable with respect to these characteristics certainly depend on more than what can be measured physically. Many percussion instruments carry unpredictable cultural associations. For instance, it seems a daunting task to map every thought that the brake drum, military drum, sleigh bells, and lion's roar might evoke in a listener. Thus, while the study of timbre has shown that a great deal can be understood through general signal processing algorithms, it is clear that our endlessly shifting relationships to sounds will require ongoing research that grapples with considerable ambiguity.

Appendix A

Spectra of Timbre Sets

Figures A.1 and A.2 show magnitude spectra for the 30 stimuli from the diverse percussive timbre set discussed in Chapters 5 and 6. All analyses were performed on the attack portion of the sounds using a 1024 point window. For each plot, the abscissa shows the entire frequency range, and magnitude is represented on the ordinate. Figures A.3 and A.4 show spectra for all stimuli from the similar timbre set, discussed in Chapter 5. Row pairs show drumstick and felt mallet strikes of the same instrument alternately. Columns correspond to instrument strike location, which is either at the edge, in between the edge and center (labeled "middle"), or center.

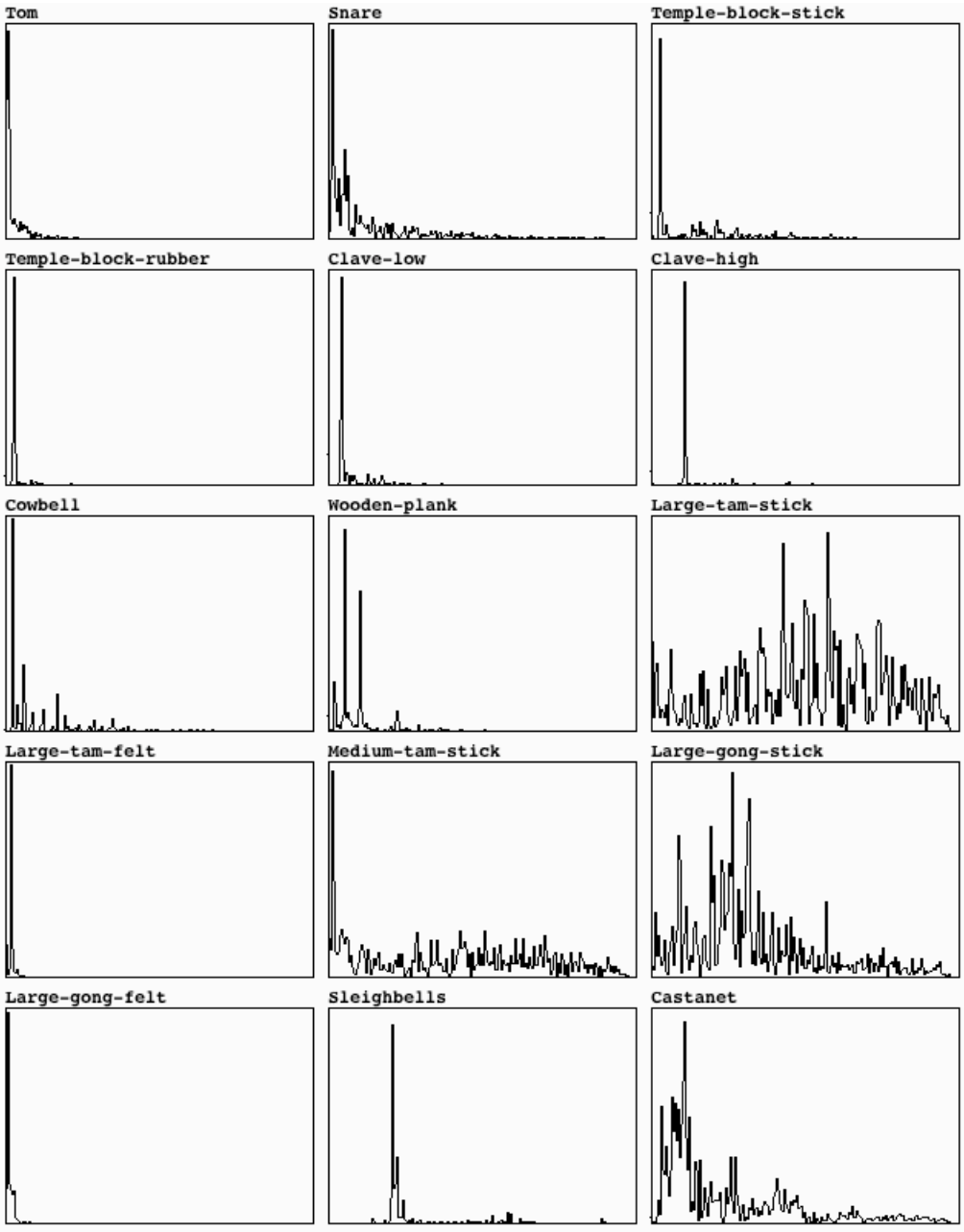


Figure A.1: Magnitude spectra for timbres 1—15 in the diverse set.

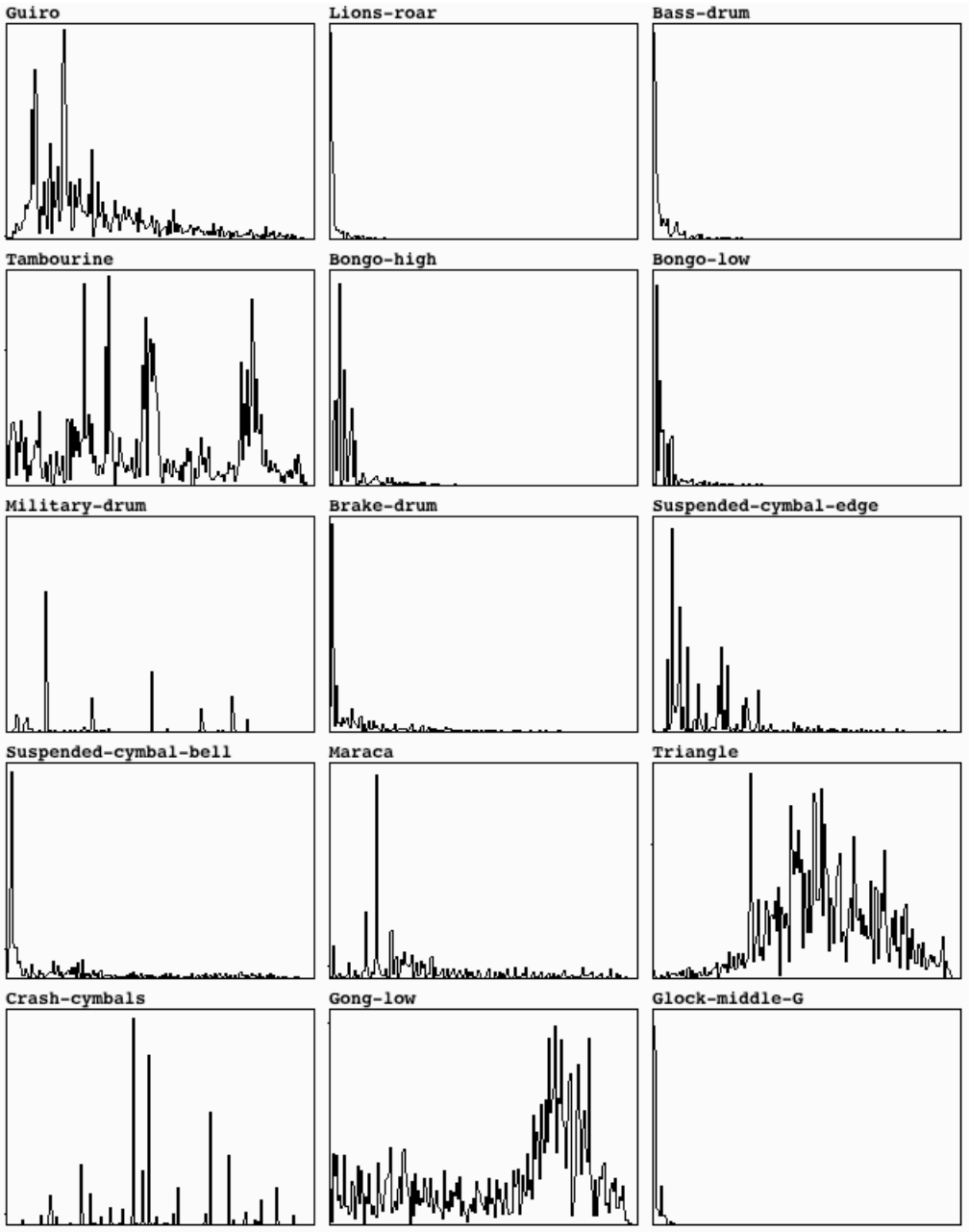


Figure A.2: Magnitude spectra for timbres 16—30 in the diverse set.

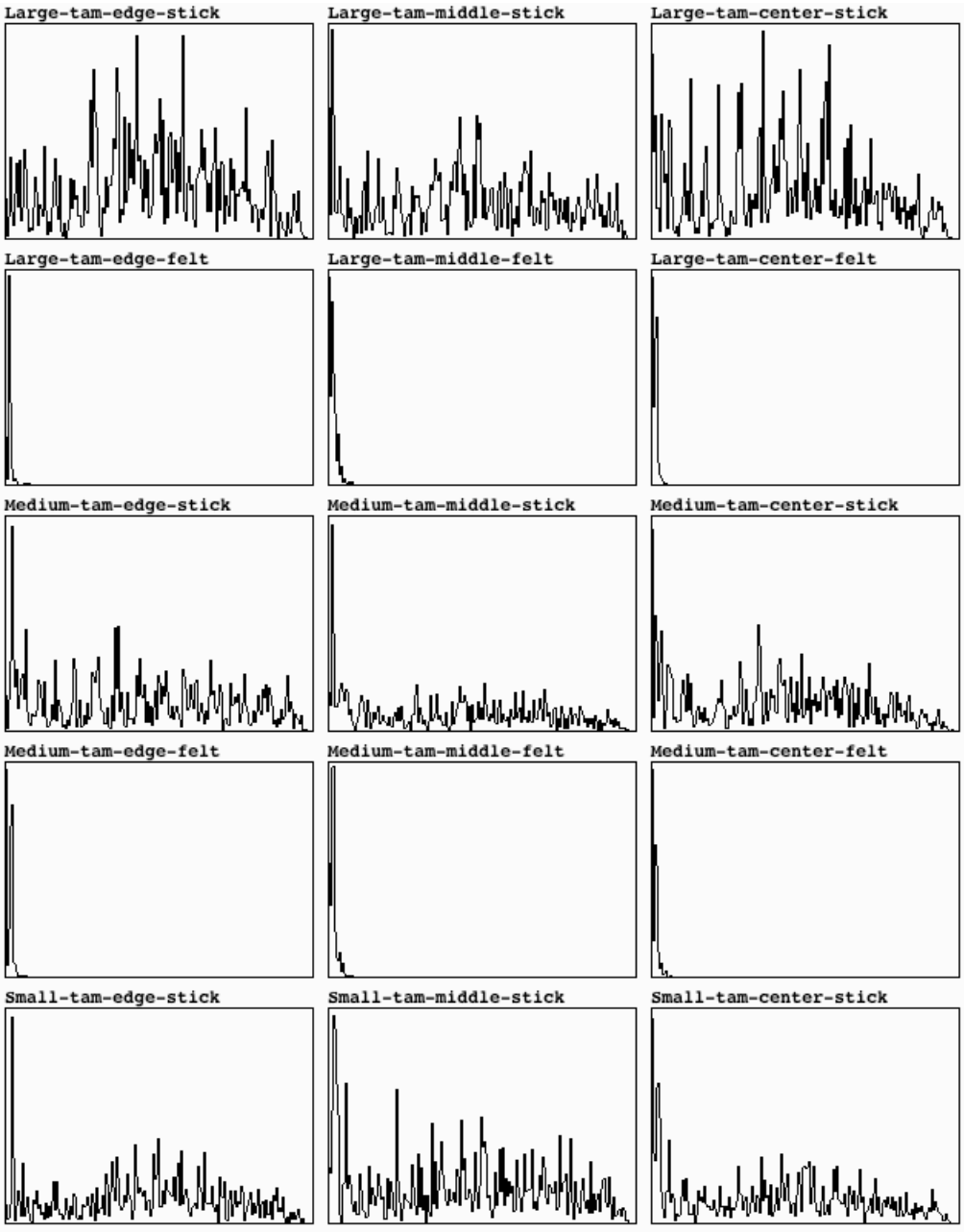


Figure A.3: Magnitude spectra for timbres 1—15 in the similar set.

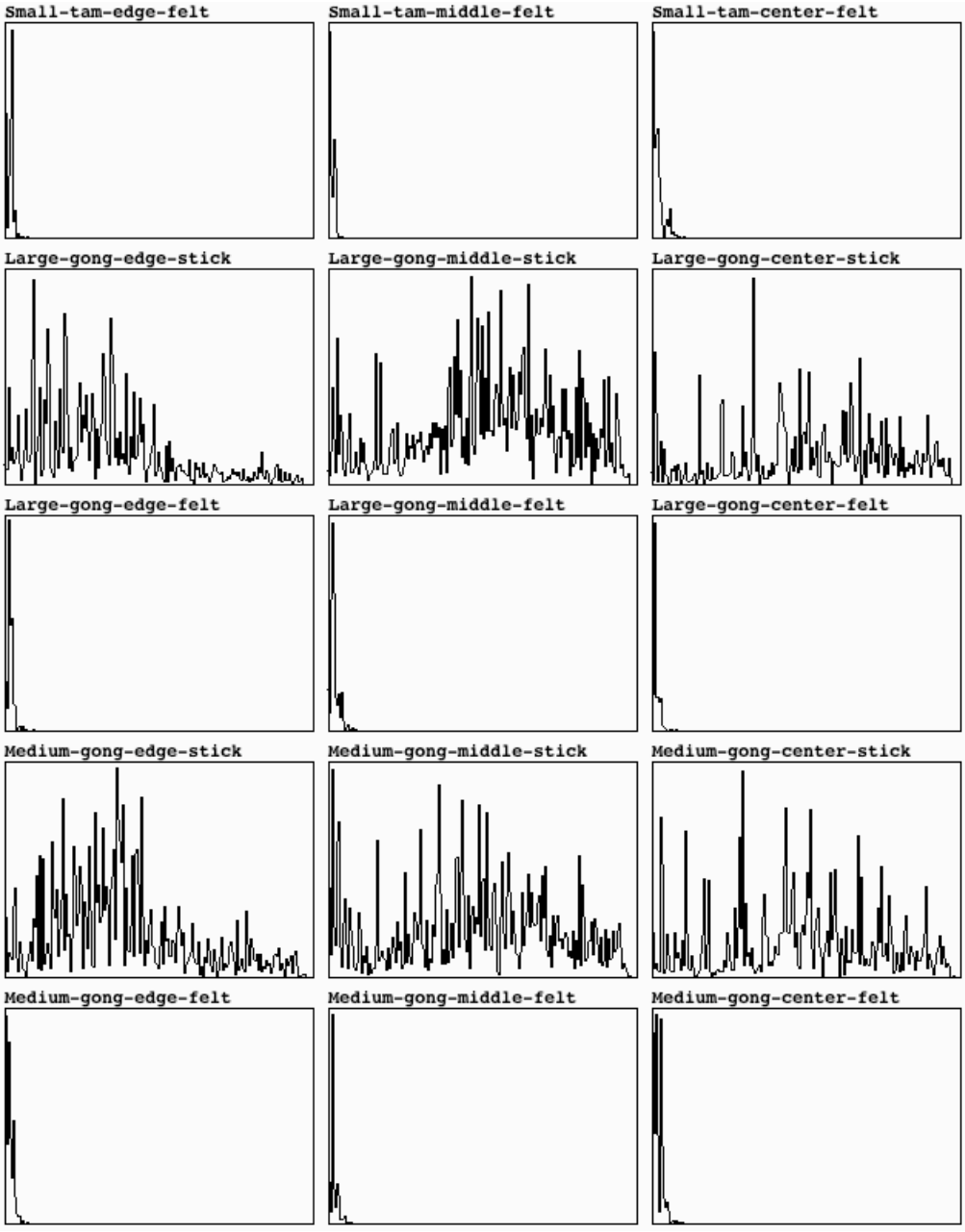


Figure A.4: Magnitude spectra for timbres 16—30 in the similar set.

Appendix B

100 Adjectives

Table B.1 contains the original list of 100 adjectives from which the final 15 adjectives were selected by a participant in the experiment described in Chapter 6. Most of the adjectives were drawn from the collection of 61 terms in [KC93b, p. 501]. Additional terms were taken from [vB74b]. The 15 selected adjectives are highlighted in blue.

Table B.1: One hundred adjectives from which the final set of 15 were drawn.

Obtrusive	Hard	Open	Tremulous	Ringing
Clear	Hoarse	Buoyant	Velvety	Scattered
Crisp	Light	Voluminous	Muffled	Artificial
Shuffling	Piercing	Even	Balanced	Silvery
Gentle	Dry	Delicate	Smooth	Subdued
Choked	Muted	Broad	Reedy	Sweet
Sparkling	Clicking	Fused	Dead	High
Nasal	Deep	Dirty	Pulsating	Sharp
Strained	Rustling	Pleasant	Brilliant	Violent
Penetrating	Rich	Assertive	Loud	Beautiful
Sensuous	Blurred	Dark	Edgy	Raspy
Quivering	Brilliant	Shrill	Pure	Coarse
Mellow	Forceful	Pale	Cold	Warm
Bright	Sustained	Weak	Tight	Hollow
Soft	Strong	Biting	Thin	Rough
Ominous	Full	Evocative	Clean	Tense
Pungent	Incisive	Cutting	Round	Throbbing
Complex	Noisy	Colorful	Damped	Intense
Prominent	Dramatic	Dull	Brittle	Resonant
Heavy	Luminous	Thick	Dynamic	Strident

Bibliography

[Ber49] L.L. Beranek. *Acoustic Measurements*. Wiley, New York, 1949.

[Ber64] K. Berger. Some factors in the recognition of timbre. *Journal of the Acoustical Society of America*, 36(10):1888–1891, 1964.

[BHT63] B. Bogert, M.J.R. Healy, and J.W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis*, 1963.

[Bou87] P. Boulez. Timbre and composition - timbre and language. *Contemporary Music Review*, 2:161–171, 1987.

[Bul07] J. Bullock. Libxtract: A lightweight library for audio feature extraction. In *Proceedings of the International Computer Music Conference*, 2007.

[CG07] M. Casey and M. Grierson. Soundspotter/remix-tv: Fast approximate matching for audio and video performance. In *Proceedings of the International Computer Music Conference*, 2007.

[CH78] W. C. Campbell and J. J. Heller. The contribution of the legato transient to instrument identification. In *Proceedings of the Research symposium on the psychology and acoustics of music*, pages 30–44, University of Kansas, 1978.

[CLA+63] M. Clark, D. Luce, R. Abrams, H. Schlossberg, and J. Rome. Preliminary experiments on the aural significane of parts of tones of orchestral instruments and choral tones. *Journal of the Audio Engineering Society*, 11(1):45–54, 1963.

[CMSW05] A. Caclin, S. McAdams, B. Smith, and S. Winsberg. Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, 118(1):471–482, 2005.

[DAC07a] S. Dubnov, G. Assayag, and A. Cont. Audio oracle: A new algorithm for fast learning of audio structures. In *Proceedings of the International Computer Music Conference*, Copenhagen, Denmark, 2007.

[DAC07b] S. Dubnov, G. Assayag, and A. Cont. Guidage: A fast query guided assemblage. In *Proceedings of the International Computer Music Conference*, Copenhagen, Denmark, 2007.

[Dar05] G. Darke. Assessment of timbre using verbal attributes. In *Proceedings of the conference on interdisciplinary musicology*, Montreal, 2005.

[DM80] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-4(4):357–366, 1980.

[Don07] S. Donnadieu. Mental representation of the timbre of complex sounds. In J. Beauchamp, editor, *Analysis, Synthesis, and perception of musical sounds: the sound of music*, pages 251–271. Springer, 2007.

[Fre90] D. Freed. Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *Journal of the Acoustical Society of America*, 87(1):311–322, 1990.

[Fuj98] I. Fujinaga. Machine recognition of timbre using steady-state tone of acoustic instruments. In *Proceedings of the International Computer Music Conference*, pages 207–210, 1998.

[Gas54] G. Gassler. Ueber die hörschwelle für schallereignisse mit verschiedenbreitem frequenzspektrum. *Acustica*, 4:408–414, 1954.

[GG78] J. Grey and J. Gordon. Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, 63(5):1493–1500, 1978.

[Gib66] J. J. Gibson. *The senses considered as perceptual systems*. Houghton Mifflin, Boston, 1966.

[Gre75] J. Grey. *An Exploration of Musical Timbre Using Computer-based Techniques for Analysis, Synthesis and Perceptual Scaling*. PhD thesis, Stanford University, 1975.

[Gre77] J. Grey. Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

[Had07] J. Hadja. The effect of dynamic acoustical features on musical timbre. In J. Beauchamp, editor, *Analysis, Synthesis, and perception of musical sounds: the sound of music*, pages 251–271. Springer, 2007.

[Han95] S. Handel. Timbre perception and auditory object identification. In B. Moore, editor, *Hearing*, pages 425–461. Academic Press, 1995.

[HBPD03] P. Herrera-Boyer, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21, 2003.

[HDW94] G. Holmes, A. Donkin, and I.H. Witten. Weka: a machine learning workbench. In *Proceedings of the second Australia and New Zealand Conference on Intelligent Information Systems*, pages 357–361, Brisbane, Australia, 1994.

[HE01] S. Handel and M. Erickson. A rule of thumb: the bandwidth for timbre invariance is one octave. *Music Perception*, 19(1):121–126, 2001.

[IK93] P. Iverson and C. Krumhansl. Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, 94(5):2595–2603, 1993.

[Jen99] K. Jensen. *Timbre Models of Musical Sounds*. PhD thesis, University of Copenhagen, 1999.

[Jos67] E. Jost. *Akustische und psychometrische Untersuchungen an Klarinettenklängen*. Arno Volk Verlag, Köln, 1967.

[Jus00] P. Juslin. Cue utilization in communication of emotion in music performance: relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(6):1797–1813, 2000.

[KC93a] R. Kendall and E. Carterette. Verbal attributes of simultaneous wind instrument timbres: I. von bismarck's adjectives. *Music Perception*, 10(4):445–468, 1993.

[KC93b] R. Kendall and E. Carterette. Verbal attributes of simultaneous wind instrument timbres: Ii. adjectives induced from piston's *Orchestration*. *Music Perception*, 10(4):469–502, 1993.

[KI92] C. Krumhansl and P. Iverson. Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology*, 18(3):739–751, 1992.

[KMW94] J. Krimphoff, S. McAdams, and S. Winsberg. Caractérisation du timbre des sons complexes. ii: Analyses acoustiques et quantification psychophysique. *Journal de Physique*, 4(C5):625–628, 1994.

[Kru64] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[Kru89] C. Krumhansl. Why is musical timbre so hard to understand? In S. Nielzen and O. Olsson, editors, *Structure and Perception of Electroacoustic Sound and Music*. Elsevier, 1989.

[Lak00] S. Lakatos. A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, 62(7):1426–1439, 2000.

[LEB03] B. Logan, D.P.W. Ellis, and A. Berenzweig. Toward evaluation techniques for music similarity. In *Proceedings of the 4^{th} International Symposium on Music Information Retrieval*, pages 81–85, 2003.

[Ler87] F. Lerdahl. Timbral hierarchies. *Contemporary Music Review*, 2:135–160, 1987.

[Lic41] W. Lichte. Attributes of complex tones. *Journal of experimental psychology*, 28:455–481, 1941.

[LJB05] P. Laukka, P.N. Juslin, and R. Bresin. A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, 19:633–653, 2005.

[LMC97] S. Lakatos, S. McAdams, and R. Caussé. The representation of auditory source characteristics: simple geometric form. *Perception & Psychophysics*, 59(8):1180–1190, 1997.

[Log00] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the International Symposium on Music Information Retrieval*, 2000.

[LT07] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10^{th} International Conference on Digital Audio Effects*, Bordeaux, France, 2007.

[MC92] S. McAdams and J.C. Cunible. Perception of timbral analogies. *Philosophical Transactions: Biological Sciences*, 336(1278):383–389, 1992.

[MWD+95] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58:177–192, 1995.

[OJ08] J. Oliver and M. Jenkins. The silent drum: A new percussive gestural interface. In *Proceedings of the International Computer Music Conference*, 2008.

[Oli10] J. Oliver. The mano controller: A video-based hand-tracking system. In *Proceedings of the International Computer Music Conference*, 2010.

[OS89] A. Oppenheim and R.W. Schafre. *Discrete-Time Signal Processing*. Prentice Hall, New Jersey, 1989.

[OST57] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The measurement of meaning*. University of Illinois Press, 1957.

[PAZ98] M. Puckette, T. Apel, and D. Zicarelli. Real-time audio analysis tools for pd and msp. In *Proceedings of the International Computer Music Conference*, pages 109–112, 1998.

[PD76] R. L. Pratt and P. E. Doak. A subjective rating scale for timbre. *Journal of sound and vibration*, 45(3):317–328, 1976.

[PJ82] H.F. Pollard and E.V. Jansson. A tristimulus method for the specification of musical timbre. *Acustica*, 51, 1982.

[Plo70] R. Plomp. Timbre as a multidimensional attribute of complex tones. In R. Plomp and G. Smoorenburg, editors, *Frequency Analysis and Periodicity Detection in Hearing*, pages 397–414. Sijthoff, 1970.

[Plo76] R. Plomp. *Aspects of Tone Sensation*. Academic Press, New York, 1976.

[PMH00] G. Peeters, S. McAdams, and P. Herrera. Instrument sound description in the context of mpeg-7. In *Proceedings of the International Computer Music Conference*, Berlin, Germany, 2000.

[QT79] T.F Quatieri and J.M. Tribolet. Computation of the real cepstrum and minimum phase reconstruction. In *Programs for Digital Signal Processing*, pages 7.2–1–7.2–6, New York: IEEE Press, 1979.

[Rah66] V. Rahlfs. *Psychometrische Untersuchungen zur Wahrnehmung musikalischer Klänge*. PhD thesis, University of Hamburg, 1966.

[Rey87] C. Reynolds. Flocks, herds, and schools: A distributed behavioral model. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, pages 25–34, 1987.

[RJ93] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, 1993.

[RMW02] T. Rossing, F.R. Moore, and P. Wheeler. *The Science of Sound*. Addison Wesley, New York, 2002.

[Roa96] C. Roads. *The Computer Music Tutorial*. MIT Press, Cambridge, 1996.

[SBVB06] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton. Real-time corpus-based concatenative synthesis with catart. In *Proceedings of the 9th International Conference on Digital Audio Effects*, Montréal, Canada, 2006.

[Ser95] S. Serafini. Timbre judgments of javanese gamelan instruments by trained and untrained adults. *Pscyhomusicology*, 14:137–153, 1995.

[SF64] E. L. Saldanha and Corso J. F. Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America*, 36:2021–2026, 1964.

[Sib65] F. Sibley. Aesthetic and non-aesthetic. *Philosophical Review*, 74:135–159, 1965.

[SMN96] E. Samoylenko, S. McAdams, and V. Nosulenko. Systematic analysis of verbalizations produced in comparing musical timbres. *International Journal of Psychology*, 31(6):225–278, 1996.

[Sol58] L. Solomon. Semantic approach to the perception of complex sounds. *Journal of the Acoustical Society of America*, 30(5):421–425, 1958.

[Sol59] L. Solomon. Search for physical correlates to psychological dimensions of sounds. *Journal of the Acoustical Society of America*, 31(4):492–497, 1959.

[Sta83] J. P. Stautner. Analysis and synthesis of music using the auditory transform. Master's thesis, Massachusetts Institute of Technology, 1983.

[SVN37] S.S. Stevens, J. Volkman, and E.B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8:185–190, 1937.

[TC99] G. Tzanetakis and P. Cook. Marsyas: a framework for audio analysis. *Organised Sound*, 4(3):169–175, 1999.

[TC02] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[Ter70] E. Terhardt. Frequency analysis and periodicity detection in the sensations of roughness and periodicity pitch. In R. Plomp and G. Smoorenburg, editors, *Frequency Analysis and Periodicity Detection in Hearing*, pages 278–291. Sijthoff, 1970.

[Tra90] H. Traunmüller. Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America*, 88(1):97–100, 1990.

[Tru71] B. S. Truax. *Development of a verbal technique for identifying children's musical concepts of instrumental timbre*. PhD thesis, The Pennsylvania State University, 1971.

[Tza02] G. Tzanetakis. *Manipulation, Analysis and Retrieval Systems for Audio Signals*. PhD thesis, Princeton University, 2002.

[vB74a] G. von Bismarck. Sharpness as an attribute of the timbre of steady sounds. *Acustica*, 30:160–172, 1974.

[vB74b] G. von Bismarck. Timbre of steady sounds: a factorial investigation of its verbal attributes. *Acustica*, 30:146–159, 1974.

[WBS87] D. Wessel, D. Bristow, and Z. Settel. Control of phrasing and articulation in synthesis. In *Proceedings of the 1987 International Computer Music Conference*, pages 108–116, San Francisco, 1987.

[Wes73] D. Wessel. Psychoacoustics and music. *Bulletin of the Computer Arts Society*, 30:1–2, 1973.

[Wes79] D. Wessel. Timbre space as a musical control structure. *Computer Music Journal*, 3(2):45–52, 1979.

[WG72] L. Wedin and G. Goude. Dimension analysis of the perception of instrumental timbre. *scandinavian journal of psychology*, 13:228–240, 1972.

[YJO+00] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book*. Entropic Labs and Cambridge University, 2000.

[ZF90] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. Springer Verlag, Berlin, 1990.

[ZFS57] E. Zwicker, G. Flottorp, and S.S. Stevens. Critical bandwidth in loudness summation. *Journal of the Acoustical Society of America*, 29:548–557, 1957.

[ZR07] X. Zhang and Z.W. Ras. Analysis of sound features for music timbre recognition. In *Proceedings of the IEEE CS International Conference on Multimedia and Ubiquitous Engineering*, pages 3–8, 2007.

[ZS65] E. Zwicker and B. Scharf. A model of loudness summation. *Psychological Review*, 72:3–26, 1965.

[Zwi52] E. Zwicker. Die grenzen der hörbarkeit der amplitudenmodulation und der frequenzmodulation eines tones. *Acustica*, 2:125–133, 1952.

[Zwi60] E. Zwicker. Subdivision of the audible range into critical bands (frequenzgruppen). *Journal of the Acoustical Society of America*, 33(2):248, 1960.