Looking to a Unified Theory of Timbre to Improve Timbral Similarity Systems in Music Information Retrieval

Rebecca Fiebrink
MUGS 695
Dr. Stephen McAdams
19 April 2005

Abstract

Recently, several groups of researchers in music information retrieval have proposed and implemented systems to compute the timbral similarity of pieces of music. Ratings of timbral similarity can be useful for tasks such as online music recommendation and playlist generation. Unfortunately, the creation and evaluation of these systems involves a host of thorny issues, and some researchers have recently postulated the existence of an upper limit on the performance possible using current approaches to design. A deeper understanding of timbre, especially regarding humans' perception and comparison, may allow for improved performance and greater usefulness of these systems. A set of questions is posed that, if answered by a unified theory of timbre, could lead to such improvements.

Introduction

Music information retrieval (MIR) encompasses a broad range of research endeavors and technological innovations. A relatively nascent discipline, it arises from the intersection of music, computer science, library and information science, business, and other fields (Downie 2003). MIR systems are quite diverse, but common goals include the facilitation of music searching, retrieval, distribution, and analysis (Downie 2003, Foote 1999).

Timbre has a great impact on MIR system goals, design, and performance. It may be explicitly related to system objectives and implementations, as is the case in the systems discussed in this paper. Or, timbre may be one of many features extracted from a musical signal and used toward other ends, such as classification based on genre or mood (e.g., Tzanetakis 2001). Even in systems where timbre is not explicitly measured or considered, such as beat-tracking and transcription (e.g., Pikrakis et al. 2004, Martin 1996), assumptions are made regarding the instrumentation and short- and long-term timbral variability. These assumptions impact systems' accuracy and their extensibility to a range of musical contexts. Timbre is an unavoidable consideration in any MIR system, just as timbre is an inextricable quality of music itself.

The success of current MIR systems, as well as the formulation and realization of new goals for the discipline, therefore hinges on a thorough understanding of timbre.

Acoustics, psychoacoustics, cognition, music theory, composition, cultural studies, and numerous other fields can all inform such an understanding. A unified theory of timbre relating all such facets of timbre research would therefore prove invaluable to the advancement of MIR research and development. While such a unified theory does not yet exist, a foundation of empirical, qualitative, and creative work on timbre has arisen in recent years.

This paper will highlight such questions that, if and when they are answered by a unified theory of timbre, may prove crucial to improvements and innovations in MIR technologies. Systems that judge timbral similarity are discussed specifically, though many of the questions raised in their design and implementation are likely similar to those raised elsewhere in MIR research. Before exploring these questions, however, such discussion is motivated by a critical review of the goals, implementation, and performance of existing timbral similarity systems.

Why Timbral Similarity?

Music similarity is an area of research within MIR that attempts to produce perceptually relevant ratings for the "similarity" of two or more segments or pieces of music. An often-cited application of these ratings is music recommendation systems, which suggest pieces of music that are similar to those in which a user has indicated an interest. Other applications, such as playlist generation systems that attempt to maintain a degree of uniformity among the pieces proposed, are certainly possible as well (Logan and Salomon 2001; Aucouturier and Pachet 2002a, 2002b).

Similarity could be judged based on metadata; for example, two pieces listed in a database as appearing on the same album by the same artist might likely be similar. However, this is obviously not always the case, and systems such as peer-to-peer file-sharing networks may contain music that is missing this information or is mislabeled. Another problem is that characteristics of a piece of music such as timbre, melody, or genre often lack standard representations (one notable exception is the MPEG-7 standard; see Herrera et al. 1999).

An alternative approach is the use of content-based techniques. The most widely-used of these is collaborative filtering, which is based on analysis of data collected from user behaviors (Aucouturier and Pachet 2002a). As a very simple example, if User X enjoys Songs A and B, and User Y enjoys Song A as well, a system might recommend that User Y listen to Song B. One problem with this approach is that recommendations may be only superficially relevant (Aucouturier and Pachet 2002); another is that such systems essentially ignore new music (Logan and Salomon 2001).

A second content-based technique, which has the potential to circumvent the above problems, focuses on analysis of the musical signal itself. This approach has received much attention from the MIR community in recent years. Foote (1997) and Wold et al. (1996) proposed two of the first automated systems to measure "musical similarity," without much elaboration of which musical characteristics were considered most important by their measures, and without a very thorough defense of their measures based on music perception or cognition. Nevertheless, according to the authors, these approaches produced promising results when incorporated into content-based retrieval systems. Welsh et al. (1999) created another early system, which extracted features intended to be specifically related to characteristics such as frequency content, volume, noise, rhythm and tempo, and "tonal transitions."

Since then, MIR researchers have proposed other content-based measurements of musical similarity (e.g., Herre et al. 2003), as well as measurements for specific characteristics such as melody (e.g., Grachten et al. 2004), rhythm (e.g., Foote et al. 2002; Paulus and Klapuri 2002), and timbre. The notion of timbral similarity measurement is particularly interesting, given that timbre itself is a complex phenomenon. There does not exist one standard method of quantifying the timbre of a musical signal. Studies on only single tones have shown timbre to be a perceptually multidimensional phenomenon, though the precise acoustic correlates proposed for its dimensions vary from study to study (see, for example, Plomp 1970; Wessel 1973; Grey 1977; Krumhansl 1989). Furthermore, there is no established definition of "timbral similarity," either among MIR researchers or the culture at large. The work discussed below tends to treat timbral similarity as something that makes songs "sound alike," based on some similarity in the spectral content of the music, but which is distinct from melodic and rhythmic similarity.

The failure of timbral similarity systems to escape from a “wastebasket” definition of timbre as anything that is not pitch, rhythm, or loudness is only one problem of many contributing to the difficulty of creating such systems. However, because of timbre’s close relationship to genre, style, instrumentation, recording technology, expressiveness, and other interesting and meaningful characteristics of music, one might conclude that, if designed correctly, a “good” timbral similarity system could prove useful to music enthusiasts.

Overview of Timbral Similarity Systems

This section provides an overview of timbral similarity systems recently designed by the MIR and multimedia processing communities. Such systems tend to be similar in design: a song or clip is analyzed in some way to obtain a mathematical representation of its “timbre” or “sound,” and metrics are applied to calculate the distance between two representations and determine the similarity of the music. Often, these systems are tested by comparing their similarity ratings by those performed by a handful of human subjects. Alternatively, evaluation might be made using some assumption relating timbral similarity to qualities readily available in the metadata, such as genre: under the assumption that songs within a genre generally have similar timbres, a similarity measure is successful if it tends to rate same-genre songs as similar.

It is important to note that some of the systems here specifically aim to measure timbral similarity (namely, those by Aucouturier and Pachet, Pampalk et al., and Liu and Huang), while others are concerned with musical similarity in general. Systems in this latter group were nonetheless selected for study either because they incorporate timbral similarity as an explicit component of musical similarity or because they use essentially the same metrics as the timbral similarity systems. Knowledge gained through study of these systems is therefore directly applicable to systems attempting to measure timbral similarity alone. Likewise, if timbral similarity systems are to be improved, general musical similarity systems stand to benefit as well.

An in-depth discussion of an early system for the explicit measurement of timbral similarity will offer insight regarding how such a system might be designed and tested. Brief discussion of other systems will follow, with attention to the ways in which they differ from this example. The authors' findings on the performance of their systems will be described superficially here, and following sections will offer more critical analyses of these systems as a group.

An Example System: Aucouturier and Pachet

Aucouturier and Pachet's (2002a, 2002b) first work toward a timbral similarity system is motivated by the growth of Electronic Music Distribution (EMD) and the problem it presents of helping online music enthusiasts efficiently browse through online music collections. They argue that musical taste is often associated with timbre, so a timbral similarity measure is quite relevant to EMD systems.

Their approach involves the computation of a high-level timbral descriptor for each musical selection in a database. Based on the knowledge that a "large part" of the timbre of instruments is explained by their spectral envelope, they choose Mel Frequency Cepstral Coefficients (MFCC's) as the basis for these descriptors. Aucouturier and Pachet measure the cepstrum as the inverse Fourier transform of the log-spectrum:

$$c_n = \frac{1}{2\pi} \times \int_{\omega=-\pi}^{\omega=+\pi} \log(S(e^{j\omega})) \cdot e^{j\omega \cdot n} d\omega$$

The mel-cepstrum is computed via mapping linear frequencies to the psychoacoustically-based Mel scale. MFCC's are typically computed using an ordered sequence of these coefficients, (c_0, c_1, …, c_k), where *k* is chosen based on the desired resolution of the measure. Lower-ordered coefficients describe slow temporal changes in the spectral envelope, and higher-ordered coefficients describe increasingly fast variations. Higher-ordered coefficients are therefore increasingly dependent on the pitches present in a signal. Aucouturier and Pachet choose 8 coefficients for their 2002 work.

A set of MFCC sequences is computed for subsequent 50ms windows on each musical selection. The total number of coefficients for even a short song can become very numerous, however (in the order of tens of thousands), so it is necessary to devise a more compact representation of each song's set of MFCC's. Aucouturier and Pachet choose to model this set using a Gaussian Mixture Model (GMM), which in this work is composed of three component Gaussian probability distributions.

The similarity of two sets of MFCC's can be thought of as inversely related to the distance between their GMM's. This is computed via a sampling process, in which 100 random samples are taken from the GMM of one song (Song A), and the probability of these samples given the GMM of the other song (Song B) is calculated. Similarly, samples are taken from the GMM of Song B, and their probability is calculated given the GMM of Song A, resulting in a symmetric measure. The higher these probabilities, the more similar the songs are judged to be.

Aucouturier and Pachet evaluate the system by an informal analysis of the similarity matches it proposed. They note that many selections by the same artist or in the same genre were judged to be similar. They also note timbral matches such as a piece by Beethoven with a Beatles song. Rather than point to this as a failure of the measure, they describe these matches as "interesting," arguing that it is precisely those matches that cannot be made on the basis of artist or genre metadata alone wherein the system demonstrates its usefulness. They also compared the system to human listeners' judgments, and they found that the system agreed with the subjects 80% of the time.

Other Approaches

Liu and Huang (2000) also use MFCC's and GMM's to represent timbre, with the goal of facilitating indexing and retrieval based on audio content, though their focus is not restricted to music. A key difference with the work above is that they segment audio into relatively homogenous contiguous sections. This segmentation is performed automatically, and portions of the signal having low energy are considered as candidates for section boundaries. Rather than considering global timbre of entire audio stream or file, these segments are considered individually as potential timbral matches to a given query. In tests, this model tended to judge segments of recorded speech by one individual to be similar, regardless of the presence of background music or noise.

Logan and Salomon (2001) propose to measure music similarity using MFCC's on subsequent frames. They use 19 MFCC's, resulting in a measure that considered relatively fine detail compared to Aucouturier and Pachet's (2002a, 2002b) measure. Instead of using a GMM, they represent the collection of MFCC's using standard K-means clustering, wherein each song is summarized by a "spectral signature" of 16 typical MFCC sequences. Signatures are compared using Earth Mover’s Distance, a metric that calculates the amount of “work” needed to convert one song’s spectral signature into the other’s. Logan and Salomon conducted tests with human users, and they found that users tended to agree that, of the top five songs the system judged as similar to a seed, 2.5 songs were indeed similar. Baumann and Pohle (2003) implemented a very similar system, and they found loosely comparable performance among their system, Logan and Salomon’s, and Aucouturier and Pachet’s.

Pampalk, Dixon, and Widmer (2003a) present several measures for musical similarity in the context of a music browsing system that allows a user to interactively decide which of the measures should be used. The measure that specifically addresses timbre employs a spectrum histogram, which is a relatively straightforward technique in comparison to Logan and Salomon (2001) and Aucouturier and Pachet (2002a, 2002b) use of MFCC’s and representative models. To create this histogram, a piece of music first undergoes psychoacoustically-informed preprocessing to convert it to a Sone/Bark representation, account for the effects of the outer and middle ear, and calculate the effects of masking. Each “bin” of the resulting histogram corresponds to a critical band, and it may take on one of fifty values representing the loudness in that band. This process is performed over a windowed version of the signal, and for each band, a count is taken of how many times some specified loudness threshold is exceeded. To compute the similarity between two songs, a simple Euclidean-distance measure is performed on the histograms.

Herre et al. (2003) investigate a variety of frequency- and time-domain analysis techniques for musical similarity analysis, many of which are closely related to timbral qualities. They found that combining spectral flatness measure, normalized loudness, MFCC’s, Real Cepstral Coefficients (RCC’s), and temporal log-loudness derivatives yielded the best results in a small-scale music recommendation system. Herre et al. are the only group of researchers discussed here who explored such a wide range of signal metrics. They analyzed each metric individually for its correlation to listener judgments of similarity, and they also compared listener judgments with a final system that incorporated several measures. However, they did not investigate what perceptual qualities were actually evaluated by each measure in the given implementation or postulate perceptually-based reasons for the success of some measures over others.

Berenzweig, Ellis, and Lawrence (2003) use MFCC's and their first-order differences to define an "anchor space." Pieces of music are mapped to clusters in this space, and clusters are represented as GMM's. The system measures similarity with respect to selected "anchor" points in this space, based on the recognition that people often describe music via its similarity to well-known "anchor" or "canonical" artists and genres.

Performance of Existing Timbral Similarity Systems

Several groups of researchers have undertaken larger-scale studies comparing the performance of the above systems and other variants on them. Pampalk et al. (2003b) reviewed work by Logan and Salomon (2001) and Aucouturier and Pachet (2002a, 2002b) with their own spectrum histogram (as well as two other measures that deal more with rhythmic similarity than timbral similarity). They compared the systems based on their degree of match to album, artist, genre, style, and "tones" from the All Music Guide (All Media Guide 2005), using a sample database containing a variety of popular and classical music. The spectral histogram correlated with these characteristics the best, followed by Aucouturier and Pachet's system, then by Logan and Salomon's. Their findings suggested poorer performance of these other systems than their authors had reported. However, Pampalk's measure was highly dependent on the music present in their own database, which was different from the test databases of the other researchers, so no firm conclusions regarding performance could be drawn.

Berenzweig and Logan et al. (2003) performed a study evaluating the relative performance of MFCC's (as used by Logan and Salomon 2001, for example) with anchor space (as used by Berenzweig, Ellis, and Lawrence 2003) and the effectiveness of various means of modeling and comparing feature distributions. To obtain ground truth data, they performed a user survey, consulted the All Music Guide (All Media Guide 2005), and obtained information from the internet on user playlist co-occurrence, user collection co-occurrence, and textual data from documents describing music. They found that MFCC's and anchor space techniques performed similarly and that the choice of modeling scheme and comparison method was somewhat dependent on the feature space being used. They also suggested personal music collection co-occurrence as the best ground truth measure (notably, however, they were interested in evaluating measurements of musical similarity in general, not timbral similarity only).

Aucouturier and Pachet (2004a, 2004b) performed an extensive, systematic study on the optimization of their 2002 timbral similarity system, incorporating techniques used by other researchers from MIR and speech processing. They experimented with changing parameters such as the audio sample rate, number of MFCC’s, number of GMM components, and window size, and they picked optimal parameters for each based on this experimentation. They also examined the use of Earth Mover’s Distance as an alternative to sampling for similarity calculation, tried using Hidden Markov Models instead of GMM’s, and experimented with a variety of front-end processing techniques. Using genre metadata for an intentionally simplified and conservative similarity ground truth, they were able to improve upon their original system to some extent. Most of this improvement arose from parameter fine-tuning; they found that substantial changes to their initial approach (e.g., the use of GMM’s) failed to improve performance substantially. After all improvements were implemented, Aucouturier and Pachet still found the system’s performance to be unsatisfactory, and they posited the existence of a ceiling on performance that might not be overcome using this basic system structure.

Three main conclusions can be drawn from the above studies. First, any meaningful comparison of timbral or musical similarity measures is wrought with practical difficulties. Second, most of the similarity systems discussed above can perform reasonably well in certain contexts. Third, none of the approaches greatly outperforms the others, and it appears that there is an upper limit on the performance achievable using the general framework shared by these approaches to measure similarity. This observation is indicative of larger problems inherent to the construction and evaluation of timbral similarity systems as they have been conceived thus far.

Larger Issues in the Construction and Evaluation of Timbral Similarity Systems

One serious issue that complicates the construction and evaluation of musical similarity systems is the lack of an operational definition of what is being measured and compared. For timbral similarity systems, there seems to be agreement among researchers that the “timbre” of a piece relates to its spectral content, and pieces whose spectra are somehow similar are more likely to sound similar than pieces whose spectra are different. This is a problem among more general music similarity systems as well; though there may be countless factors that contribute to perceptual similarity between two pieces of music, most researchers have not attempted to elaborate specifically what these factors are. Even when they incorporate spectral measures such as MFCC’s into their general similarity systems, they offer little discussion regarding why. The problem of definition is not a trivial one; studies on the perception of the timbre of single notes reveal its multidimensional perceptual nature (e.g., Grey 1977), so it is likely that the perception of the timbre of an entire piece of music is even more complex.

For that matter, most of these systems also assume that it is perceptually meaningful to assign an entire piece of music *one* representation of timbral similarity. Even considering that this representation might be a statistical distribution rather than a single point in a feature space, and even considering that much of the music of interest is only a few minutes long, this is quite problematic. A few researchers, such as Liu and Huang, do segment an audio signal into sections, but even they do not demonstrate firm evidence that the resulting segmentation is relevant to human perception.

Another drawback of these systems is their limited musical focus. Most systems are evaluated using Western popular music, and some incorporate Western common practice period music as well. The above difficulties regarding a lack of understanding of what the perceived timbre of a piece of music entails and whether it even make sense to discuss the timbre of a whole piece of music are problematic enough for this limited set of genres. However, they present even more challenges when non-Western music, electroacoustic music, and all other possible varieties are considered as candidates for timbral similarity measurements (as they should be).

A fourth, very serious problem of timbral similarity systems is the lack of ground truth for objective evaluation. This problem exists in part because of the lack of a definition for timbral similarity. It is exacerbated by the fact that human similarity judgments of music are likely influenced by a variety of musical and non-musical factors. For example, when user tests are conducted to assess the performance of a system, their similarity ratings might be highly influenced by whether they are familiar with the music played, but this is not accounted for in any of the studies. Furthermore, researchers have at some times pointed toward the convergence of timbral similarity measures with artist and genre similarity as a marker of success, and at other times they have claimed that the *divergence* of these measures indicates success (see specifically Aucouturier and Pachet 2002a, 2002b). Even when artist and genre similarity is established as a baseline to which comparisons are made, bias is introduced by the fact that the system creators themselves have usually put together the databases for testing and have chosen which artists and genres to include.

This is not to say that these systems are not measuring anything of relevance; many of the user tests indicate that these systems' similarity judgments often correlate with human listeners'. But it is not clear whether this correlation is actually a result of the systems successfully measuring timbre or timbral similarity in a meaningful way. Without an understanding of which perceptual qualities are being measured and how measurements for different songs should be compared to best mimic human similarity judgments, "timbral similarity" measures can only claim to be heuristics of questionable applicability.

Many, if not most, of these issues arise from a lack of understanding of perceptual, cognitive, and cultural aspects of timbre perception. This may come at little surprise, given the general lack of understanding of these issues by the wider academic music community itself. Fortunately, recent decades have seen an increase in the rigorous study of timbre in disciplines such as psychology, music theory, composition, and many others. One need only glance through the proceedings of the recent CIM Conference (Traube and Lacasse 2005) to appreciate that timbre research is an active and vibrantly diverse endeavor. A unified theory of timbre might elegantly link components from these different fields. Such a theory would have immense potential to guide MIR research around the above-mentioned pitfalls. It would open avenues for improved conceptualization and implementation of timbral similarity systems, as well as benefit broader MIR research.

The Role of a Unified Theory of Timbre in Improving Timbral Similarity Systems

The quest for improved, perceptually relevant timbral similarity systems might begin with six questions, each of which can be addressed to a large degree by a unified theory of timbre:

1. *What do we want to measure? (What aspects of timbre are we concerned with?)* Existing timbral similarity work is vague with regard to the perceptual qualities of music it hopes to capture. A working definition of timbre as it is perceived in the context of entire musical selections is a necessary first step.
2. *What metrics are most appropriate to measure the qualities/quantities from Question 1?* We must identify acoustic correlates to perceptually relevant measures and choose or invent means to compute and represent them in the context of these systems.
3. *How do humans perform timbral similarity comparisons?* Given a mechanism for perceiving the timbre of a segment of music, we must understand which facets of the perceived timbre contribute to a judgment of similarity, and how.
4. *What approaches are most appropriate for performing human-like comparisons, using the metrics from Question 2?*
5. *How can we best evaluate timbral similarity systems?* We must be able to clearly identify shortcomings in systems in order to remedy them, and we must be able to evaluate systems against each other.
6. *How can we best incorporate timbral similarity measures into larger EMD systems so that they are useful and relevant to music enthusiasts?*

Of course, it is possible that the even the structure of these questions makes incorrect assumptions about human perception of timbre and timbral similarity. For instance, the act of perceiving or recognizing the timbre of a sound might itself involve a comparison with prototypical models. The above outline will nevertheless be used as a loose structure to organize the following discussion, and possible deviations from the model the outline presupposes will also be pointed out.

What do we want to measure?

The first issue that a unified theory of timbre could address is whether, and for which types of music, it even makes sense to discuss a high-level, “global” timbre. Perhaps the idea of a global timbre can make sense when conceptualized as a range or distribution within the space of perceptible timbres rather than a single representative point. Perhaps the idea of a global timbre can make sense for a piece of music with little variation in instrumentation, texture, pitch, and loudness, If this is the case, the degree of variation of these parameters that is allowable before a piece of music is perceived as having multiple distinct high-level timbres should be understood. If a global timbre representation does have a perceptual basis, perhaps it is linked to timbre memory: the most distinct, characteristic, or otherwise memorable timbres would then be the ones a timbral similarity measure would need to capture and represent somehow. Or, perhaps one global timbre is too restrictive a notion for most pieces of music; in this case, it should be understood what constitutes a timbrally homogenous perceptual segment, and how the *general* timbral characteristics of diverse segments could be combined (or not) into a global average or distribution.

Existing timbre research has already demonstrated people’s ability to discriminate among and compare the similarity of the timbre of single notes (e.g., Plomp 1970, Grey 1977, Krumhansl 1989). We already have some understanding of the acoustic correlates of similarity measures at this very primitive level. A perceptually-informed timbral similarity measure would ideally mirror human performance on single-note similarity judgments. A music file containing one note played by a trombone should be rated as more similar to a music file containing one note on a trumpet than a music file containing one note on a kazoo. In order to answer the above questions regarding high-level timbre perception, perhaps the work on single-note timbre perception could be extended to consider musical segments of increasingly greater complexity. A unified theory of timbre could answer, for instance: What are the acoustic correlates of the perceived timbre of a chord or tone cluster, played with one instrument or many? How do the perceived timbres of temporally contiguous chords affect the overall perceived timbre of the group of chords?

At increasingly higher temporal levels, the questions of which acoustic properties might be important to the notion of a “general” timbre (that is, “global” within the bounds of the segment), and whether a general timbre may exist at this level, becomes increasingly complicated. Temporal variations in timbre at the note level (such as occurs in hocket, for example) are likely to have different effects on perception than the microtemporal variations that occur within a single note. Temporal variations in timbre at the measure level are likely to have another effect. A theory of timbre informed by physiology, music perception, and cognition is necessary to predict the effect on any perceived general timbre as the temporal scale of spectral variation changes from low to high.

Even considering relatively short and homogenous segments of music, the question of the contribution of timbral variance to a general timbre (or its preclusion of the perception of a general timbre) is complicated by the question of auditory streaming and its impact on timbre perception. A very short music clip might be perceived as containing multiple streams if sufficient variation in pitch, localization, dynamics, timbre, or other characteristics are present. Supposing there is indeed a way to represent the general timbre of a stream of music given constraints on time and timbral variability, what happens when two or more streams combine? The complexity of the perceptual mechanisms that operate in auditory stream segregation and auditory scene analysis (Bregman 1990) and the multidimensional nature of timbre perception at the note level both suggest that any perceived general timbre of a sound containing multiple streams might not be computed by the sum of its parts. Perhaps it no longer makes sense at all to discuss any sort of general timbre at the level of complexity where multiple streams are perceived. In any case, a theory of timbre informed by an understanding of human auditory scene analysis might further describe how distinct component streams contribute to a perceived general timbre and delimit when the perception of a general timbre might occur at all.

Knowledge regarding the factors that contribute to some timbres being more salient than others would also benefit timbral similarity systems. In popular music, for example, the timbre of the singer’s voice might have a greater effect on the perceived general timbre than the background instrumentation. Sounds that are surprising or seem out of place might also have effects on the perceived general timbre that are out of proportion to their relative temporal or dynamic prominence in the music. Of course, in the event that there is no perceptual basis for a general or global timbre at all, a timbral similarity system that considered the individual timbres of greatest salience, such as a singer's voice, might still be very useful.

The perception of any general timbre and the perception of component timbres as more salient than others are also likely dependent on genre. Timbre plays very different roles in the organization of punk rock, Javanese gamelan, and electroacoustic music, for example. It might make more sense to consider a general timbre of a short rock piece with static instrumentation than a modern Western art music piece wherein timbre is a primary bearer of form. A unified theory of timbre that explained the function of timbre in such different contexts and its impact on perception would be necessary for a timbral similarity tool to handle a wide variety of music.

What metrics are appropriate?

After it is clear what aspects of the signal should be measured, that is, what is perceptually relevant to the formation of a general timbre for a segment of music, it is necessary to choose a set of metrics for computing these from the audio signal. Current systems that use MFCC's only measure stationary spectral envelopes computed over short, successive time windows. Information we already know to be important to timbre perception, such as attack time and spectral fluctuation characteristics, is not captured well by this metric. In order to take into consideration the myriad of factors that influence the perception of a general or global timbre, it is likely that multiple metrics will be necessary.

One approach to combining metrics is to take measurements of different signal qualities (for instance, MFCC's 1-10, spectral flux, and attack time) and concatenate them in a "feature vector" describing a timbre from several semi-independent perspectives. A vector of *n* measurements would describe a timbre point in an *n*-dimensional space. This type of data representation is common to many pattern recognition systems. However, its link to human perception is questionable.

A unified theory of timbre that accounted for the interaction of many levels of human sensing, perception, and cognition could result in a more informed approach to signal analysis. Perceptually-informed blackboard systems have been used with some success in the transcription of polyphonic music (e.g., Martin 1996, Bello and Sandler 2000). These systems attempt to combine bottom-up and top-down processing in meaningful ways. In polyphonic transcription, for example, the presence of energy in a critical band exerts bottom-up pressure for the perception of a note or partial with a corresponding frequency, and the recognition of a chord exerts top-down pressure for the perception of a particular frequency that fits into that chord. Temporal information can also be incorporated into these systems: for example, a blackboard system for Western tonal music could "expect" to find a tonic chord following a dominant chord and therefore exert top-down pressure to seek out the tonic if it exists. One might imagine a blackboard model of timbre perception that used understanding of low-level sensing of the inner ear all the way up to cognitive models of expectation to "hone in" on the aspects of a signal that are most important to timbre perception at each particular moment in a piece.

Any representation of timbre constructed by a system might also need to be compacted to make storage manageable (if it is to be stored in a database rather than computed on-the-fly). For instance, many existing systems use GMM's to approximate large sets of MFCC's instead of storing all the MFCC's explicitly. The choice of how the complexity is to be reduced should be as informed by human perception as possible. If work in timbre perception does suggest that a general timbre arises from some distribution of temporally local timbres, than the use of GMM's makes some sense. Otherwise, alternative representations should be explored. If research indicates that longer or more timbrally heterogeneous pieces are perceived has having distinct segments, each with a general timbre, perhaps more states could be used for their GMM's than for more homogenous pieces. Notably, Toiviainen et al. (1995) suggest that the representation of timbre (or of other phenomena) in the brain is essentially a projection from a multidimensional to a less-dimensional space. This suggests that a timbral similarity system might also perform reductions of dimensionality without a degradation of performance.

How do humans perform timbral similarity comparisons?

Similarity comparisons of the timbre of single notes have been helped to form a foundation for timbre perception research since Plomp’s work (1970). User studies done by MIR timbral similarity researchers (e.g., Logan and Salomon 2001) indicate that humans do tend to agree on what sounds similar at the song level, as well. Assuming, therefore, that there is some perceptual and cognitive mechanism for timbre comparison, a unified theory of timbre should explain this mechanism in greater detail.

One fundamental question regarding timbral similarity judgments of musical selections is symmetry. Aucouturier and Pachet’s 2002 system treats timbral similarity as symmetrical: the system will respond identically to “How similar is Song A to Song B?” and “How similar is Song B to Song A?” A unified theory of timbre should address whether this is always in fact the case for human judgments.

A related issue is the very relevance of any question formulated as “How similar is Song A to Song B?” In their later work (2004a), Aucouturier and Pachet note studies indicating that human comparisons involve a choice between two models (“A sounds like B” and “A does not sound like B”) rather than by testing the significance of a hypothesis (“A sounds like B”). A unified theory of timbre perception would elucidate how this finding applies to timbral comparisons.

Berenzweig, Ellis, and Lawrence (2003) note that the All Music Guide (All Media Guide 2005) describes the “sound” of artists in relation to standard “anchor” artists. Perhaps high-level timbral similarity is assessed with respect to “anchor” timbres that are common (e.g., “full string orchestra” or “jazz combo”) or distinctive (e.g., “Louis Armstrong playing the trumpet”). The similarity of two non-anchor timbres might be mediated with respect to their similarity to an anchor timbre. Perhaps timbre perception and recognition is directly linked to similarity judgments with an anchor timbre or a timbral/temporal template, as is suggested for sound recognition by the information processing approach to psychology (described in McAdams 1993).

A practical issue that must be addressed is the effect of audio compression, sampling rates, noise, and other factors that can alter the perception of a signal without changing the music itself. Perhaps it is a safe assumption to say that, if Song A and Song

B sound very similar, then down-sampled or noisy versions of Song A and Song B will still sound similar, as long as the signals are not degraded too much. But, what if Song A is down-sampled, compressed, transmitted over a telephone line, or otherwise modified, and Song B is not? How robust is human timbre perception to these modifications? In large music databases where the music is likely to come from a variety of different sources, might be encoded in several different formats, and is certainly recorded and mixed using a variety of technologies, this question is quite relevant. A unified theory of timbre should account for how such changes to audio signals affect similarity judgments.

A unified theory of timbre should also elucidate the role of culture and familiarity in timbral similarity judgments. One might imagine a Western listener judging all Indian classical music to be very timbrally similar, while a trained Indian sitar player considers this music to be quite heterogeneous based on variants in instrumentation, playing style, etc. These two listeners may use quite different strategies to characterize the timbre of this music and compare the "sound" of pieces within this genre.

What similarity metrics can be used?

The similarity metrics used should reflect, to the greatest extent possible, the strategy used by humans in similarity judgments of large-scale timbre. Aucouturier and Pachet's method of sampling using GMM's might be appropriate, for instance, when the similarity judgment is assumed to be symmetrical. Berenzweig et al.'s anchor model is more appropriate for anchor-based judgments. Still other metrics may be suggested by further research into timbre perception and timbral similarity.

How can timbral similarity systems be evaluated?

Currently there is no established method of formally evaluating timbral similarity systems. However, it is necessary to gain an understanding of how closely a system mimics human perception and where its shortcomings are, if one is to be able to improve upon it and predict the contexts in which it would be a useful tool. A unified theory of timbre could point toward the best approaches to evaluating timbral similarity systems.

Most basically, a theory of timbre could offer understanding regarding the relative importance and interdependence of timbre, melody, pitch, rhythm, familiarity with an artist or song, etc. in humans’ similarity judgments. Any test comparing system similarity ratings with human judgments that uses real music should consider which of these other factors should be held constant. An understanding of the nature of similarity judgments would also aid in test design; for instance, if people make timbral similarity judgments based on two competing models (i.e., “This sounds like X” and “This doesn’t sound like X”), this should be reflected in any written surveys used for testing.

A unified theory of timbre could also point to other means of analyzing human similarity judgments. Toiviainen et al. (1998) have shown that measured electrical brain activity correlates well with human similarity judgments on single tones. Perhaps it is possible to analyze similarity perception of larger segments of audio via direct physical measurements as well.

How should timbral similarity systems be embedded in larger EMD systems?

If the above work is to yield fruitful results, we must gain an understanding of the most effective ways of incorporating timbral similarity measures into electronic music distributions so that they are actually helpful to music enthusiasts. A unified theory of timbre might predict aspects of user behavior, or it could aid in the design of usability experiments. Some key questions that must be answered are, How would people use timbral similarity-based queries for electronic music databases, if such queries were possible? Would people be interested in using them only to find pieces of music with similar instrumentation, or would they come up with more novel applications? If the system operated by returning a list of closest matches to a seed query (as most existing systems do), does the user expect the returned matches to also have similar mood or genre to the seed? Will the user be intrigued by matches that are from very different genres, or will she or he feel that the system failed? These questions address very complex issues regarding how people think of timbre as a useful way to describe music and whether they view it as something that can truly be separated from other characteristics of music.

A unified theory of timbre that lent insight into such complexities of the social aspect of timbre could also lead to the design of new interfaces for timbral query systems. Instead of having to provide a timbral similarity-based recommendation system with a seed song, perhaps people would find it useful to vocally imitate some timbre. The system could record their voice and return audio samples based on knowledge of how people tend to mimic certain non-vocal sounds. Or, an understanding of how people use language to describe timbre could lead to a system that would allow queries as simple as, “Find me music with a rough sound,” or “Find me music with a trumpet-like synthesized sound.”

Aucouturier and Pachet (2002b) recognized that users might want different recommendation strategies depending on various factors. In their 2002 system, they therefore included a feature called an “AHA slider,” which gave the user a degree of control over the novelty or “interestingness” of the results returned. The user could request to see only timbral matches that also correlated strongly with a seed’s recorded metadata, or they could request that the system return more interesting results. Deeper knowledge of how people view timbre as relevant to their music browsing and musical taste could lead to the implementation of other such controls.

New Avenues in MIR

It is clear that a deeper understanding of timbre could lead to numerous improvements in timbral similarity systems and their integration into electronic music distribution systems. Improvements in automated, perceptually-informed measurements of timbre and timbral similarity could also be applied elsewhere in MIR. Knowledge regarding auditory stream segregation and what contributes to the salience of certain timbres in a complex sound would be very useful for polyphonic transcription. McKay (2004) has shown that timbral information—specifically instrumentation—can be very useful in MIDI genre classification; however, currently the technology does not exist to extract such detailed timbral information from audio signals. Answers to many of the questions above would be helpful in the design of such a system. An understanding of the role of timbral variance in creating perceived sectional boundaries of a piece of music, as well as knowledge regarding the salience of certain timbres, could be employed for automatic theme extraction. New research on timbre could undoubtedly also spark innovative music information retrieval technologies that have not yet been imagined.

Conclusions

The above discussion has highlighted the motivation behind construction of timbral similarity systems in MIR, described implementations of existing systems and their shortcomings, and posed key questions answerable by a unified theory of timbre. A more thorough knowledge of timbre informed by research in music perception, cognition, usability, linguistics, cultural studies, and other areas can lead to more useful, perceptually-based measurements for timbre and timbral similarity. It can also point toward the best means of integrating timbral similarity measures into electronic music distribution systems to make online music most easily accessible to music lovers.

Currently, timbre research cannot answer many of the questions posed here. Several of these questions are so complex that it is unlikely they will be answered soon. However, one should view the illumination of these holes in timbre research as a starting point for future work rather than a proclamation of the inadequacy of existing work. As ongoing timbre research continues to incrementally improve our understanding of the above issues, music information retrieval stands to benefit from each new insight. MIR researchers should therefore keep abreast of, support, and participate in timbre research to the best of their abilities.

References

All Media Guide. 2005. *All Music Guide*. <http://www.allmusic.com>.

Aucouturier, J., and F. Pachet. 2002a. Finding songs that sound the same. *Proceedings of the First IEEE Benelux Workshop on Model Based Processing and Coding of Audio*, 91–8.

———. 2002b. Music similarity measures: What's the use? *Proceedings of the International Conference on Music Information Retrieval*.

———. 2004a. Improving timbral similarity: How high's the sky? *Journal of Negative Results in Speech and Audio Sciences,* 1 (1).

———. 2004b. Tools and architecture for the evaluation of similarity measures: Case study of timbral similarity. *Proceedings of the International Conference on Music Information Retrieval*.

Baumann, S., and T. Pohle. 2003. A comparison of music similarity measures for a P2P application. *Proceedings of the Sixth International Conference on Digital Audio Effects (DAFX).*

Bello, J., and M. Sandler. 2000. Blackboard system and top-down processing for the transcription of simple polyphonic music. *Proceedings of the COST G-6 Conference on Digital Audio Effects*.

Berenzweig, A., D. P. W. Ellis, and S. Lawrence. 2003. Anchor space for classification and similarity measurement of music. *Proceedings of the IEEE International Conference on Multimedia and Expo*.

Berenzweig, A., B. Logan, D. P. W. Ellis, and B. Whitman. 2003. A large-scale evaluation of acoustic and subjective music similarity measures. *Proceedings of the International Conference on Music Information Retrieval.*

Bregman, A. 1990. *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: Bradford Books, MIT Press.

Downie, J. S. 2003. Music information retrieval. *Annual review of information science and technology,* 37: 295–340.

Foote, J. 1997. Content-based retrieval of music and audio. In C.-C. J. Kuo et al., eds. *Multimedia storage and archiving systems II, Proceedings of SPIE,* 3329: 138–47.

———. 1999. An overview of audio information retrieval. *Multimedia systems,* 7(1): 2–10.

Foote, J., Cooper, M., and Nam, U. 2002. Audio retrieval by rhythmic similarity. *Proceedings of the International Conference on Music Information Retrieval*.

Grachten, M., J. Arcos, and R. de Mantaras. 2004. Melodic similarity: Looking for a good abstraction level. *Proceedings of the International Conference on Music Information Retrieval*.

Grey, J. 1977. Multidimensional perceptual scaling of musical timbres. *Journal of the acoustical society of America,* 82: 88–105.

Herre, J., E. Allamanche, and C. Ertel. 2003. How similar do songs sound? Towards modeling human perception of musical similarity. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

Herrera, P., X. Serra, and G. Peeters. 1999. Audio descriptors and descriptors schemes in the context of MPEG-7. *Proceedings of the International Computer Music Conference*.

Krumhansl, C. 1989. Why is musical timbre so hard to understand? In S. Nielzen and O. Olsson, eds. *Structure and perception of electroacoustic sound and music*. Amsterdam: Excerpta Medica.

Liu, Z., and Q. Huang. 2000. Content-based indexing and retrieval by example in audio. *IEEE International Conference on Multimedia and Expo,* 2: 877–80.

Logan, B., and A. Salomon. 2001. A music similarity function based on signal analysis. *Proceedings of the IEEE International Conference on Multimedia and Expo*, 745–8.

Martin, K. 1996. A blackboard system for automatic transcription of simple polyphonic music. *MIT Media Laboratory Perceptual Computing Section Technical Report No. 385*.

McAdams, S. 1993. Recognition of sound sources and events. In S. McAdams and E. Bigand, eds. *Thinking in sound: The cognitive psychology of human audition*, 146–98. Oxford: Oxford University Press.

McKay, C. 2004. *Automatic genre classification of MIDI recordings*. M.A. Thesis. McGill University, Canada.

Pampalk, E., S. Dixon, and G. Widmer. 2003a. Exploring music collections by browsing different views. *Proceedings of the International Conference on Music Information Retrieval.*

———. 2003b. On the evaluation of perceptual similarity measures for music. *Proceedings of the 6th International Conference on Digital Audio Effects*, 6–12.

Paulus, J., and A. Klapuri. 2002. Measuring the similarity of rhythmic patterns. *Proceedings of the International Conference on Music Information Retrieval*.

Pikrakis, A., I. Antonopoulos, and S. Theodoridis. 2004. Music meter and tempo tracking from raw polyphonic audio. *Proceedings of the International Conference on Music Information Retrieval.*

Plomp, R. 1970. Timbre as a multidimensional attribute of complex tones. In R. Plomp and G. F. Smoorenburg, eds. *Frequency analysis and periodicity detection in hearing*. Leiden: Sijthoff.

Toiviainen, P., M. Kaipainen, and J. Louhivuori. 1995. Musical timbre: Similarity ratings correlate with computational feature space distances. *Journal of new music research,* 24: 282–98.

Toiviainen, P., M. Tervaniemi, J. Louhivuori, M. Saher, M. Huotilainen, and R. Näätänen. 1998. Timbre similarity: Convergence of neural, behavioral, and computational approaches. *Music perception,* 16(2): 223–41.

Traube, C., and S. Lacasse, eds. 2005. *Proceedings of the Conference on Interdisciplinary Musicology: Timbre in composition, performance, perception, and reception of music.*

Tzanetakis, G. 2001. Automatic musical genre classification of audio signals. *Proceedings of the International Symposium on Music Information Retrieval*.

Welsh, M, N. Borisov, J. Hill, R. von Behren, and A. Woo. 1999. Querying large collections of music for similarity. *Technical Report UCB/CSD-00-1096, U.C. Berkeley Computer Science Division.*

Wessel, D. 1973. Psychoacoustics and music: A report from Michigan State University. *PACE: Bulletin of the Computer Arts Society* 30: 1–2.

Wold, E., T. Blum, D. Keslar, and J. Wheaton. 1996. Content-based classification, search, and retrieval of audio. *IEEE multimedia*, Fall 1996: 27–36.