Exploring Perceptual and Acoustical Correlates of Polyphonic Timbre

Vinoo Alluri and Petri Toiviainen
University of Jyväskylä, Jyväskylä, Finland

Polyphonic timbre has been demonstrated to be an important element for computational categorization according to genre, style, mood, and emotions, but its perceptual constituents have received less attention. The work presented here comprises two experiments, Experiment 1, to devise a framework of subjective rating scales for quantifying the perceptual qualities of polyphonic timbre and Experiment 2, to rate short excerpts of Indian popular music and correlate them with computationally extracted acoustic features. A factor analysis of the ratings suggested three perceptual dimensions: Activity, Brightness, and Fullness. The present findings imply that there may be regularities and patterns in the way people perceive polyphonic timbre. Furthermore, the perceptual dimensions can be predicted relatively well by the regression models. Spectrotemporal modulations were found to be most relevant, while the well known polyphonic timbre descriptors, the Mel-Frequency Cepstral Coefficients, did not contribute significantly to any of the perceptual dimensions.

Received June 26, 2009, accepted December 4, 2009.

Key words: polyphonic timbre, bipolar scales, computational analysis, acoustic features, modeling

Several definitions of timbre have been proposed over the years but there fails to exist one single, widely accepted definition. The multidimensional nature of timbre still renders it a difficult attribute of sound to deal with. Moreover, not much is known about the salient features of the overall emerging timbre of a jazz ensemble, a rock concert, or a symphony.

Previous Work on Timbre

Research on timbre has a long history, starting from the 70s, of perceptual similarity experiments of single instrument sounds, which concentrated on creating meaningful timbre spaces (Grey, 1977; Grey & Gordon, 1978; Iverson & Krumhansl, 1993; Lakatos, 2000; McAdams, Winsberg, de Soete, & Krimphoff, 1995). A timbre space can be thought of as a geometrical construct that aims at capturing a mental representation of the stimuli through the perceived similarity ratings of the sounds. This was achieved via multidimensional scaling of similarity ratings of the stimuli collected from a number of listeners. In addition, the aforementioned studies aimed at finding acoustic features that best correlate with the perceptual dimensions.

Several studies performed on monophonic timbre or the timbre of single isolated instrument sounds have reported that the spectral centroid, which can be conceptualized as the center of gravity of the spectrum, explains one of the timbre space dimensions (Grey & Gordon, 1978; Iverson & Krumhansl, 1993; McAdams et al., 1995) and is often referred to as a measure of perceived ‘brightness’ (Beauchamp, 1982; De Poli & Pradoni, 1997). However, interpretations of the other timbral space dimensions have lacked consensus, probably because the nature of the stimuli used in each experiment determines the perceptual and acoustic features that explain the resulting dimensions and thus cannot be generalized across other sets of stimuli. Nonetheless, in a meta-analysis of timbre spaces presented by McAdams (1999), the author argues that despite the diversity in the stimuli, the actual perceived similarity remains fairly constant irrespective of the context it is presented in. Various spectral and temporal features of the stimuli have been found to explain the dimensions of perceptual timbre spaces, such as the log-attack time, spectral flux, attack synchrony, and spectral irregularity, to name a few. In addition to the aforementioned features, the Mel-Frequency Cepstral Coefficients (MFCC), which are described in later sections, appear to be quite prominent in characterizing timbre (De Poli & Pradoni, 1997; Terasawa, Slaney, & Berger, 2005).

A few studies have aimed at creating ‘physical’ timbre spaces through self-organizing maps (SOM) from features of the audio in order to find correlations with perceptual timbre spaces (Cosi, De Poli, & Lauzzana, 1994; De Poli, Prandoni, & Tonella, 1993; Loureiro, de Paula,

Music Perception volume 27, issue 3, pp. 223–241, issn 0730-7829, electronic issn 1533-8312 © 2009 by the regents of the university of california. all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press’s rights and permissions website, http://www.ucpressjournals.com/reprintinfo.asp. DOI:10.1525/mp.2009.27.3.223

& Yehia, 2004; Toiviainen, Kaipainen, & Louhivuori, 1995). The clustering of sounds in these studies has been found to be comparable to human similarity judgments.

To add to this, a few neural studies on monophonic timbre have hinted at a high correlation between the amplitude of brain responses and the magnitude of changes in certain timbral features. These include brightness (Toiviainen et al., 1998), spectral center of gravity, and even harmonic attenuation (Caclin, McAdams, Smith, & Giard, 2008). These findings suggest that focusing on acoustic features such as brightness, attack time, and harmonic structure of the spectrum, and using computational modeling such as SOM, are effective ways to study timbre perception.

More recently, studies on monophonic timbre have moved towards finding semantic labels that best characterize timbre and their respective acoustic correlates. These studies have aimed at characterizing timbre using descriptors. Some studies have aimed at finding consistently used adjectives to describe timbre and acoustic features that can be associated with them, whereas other studies have required participants to verbally describe the timbre of the stimuli (Darke, 2005; Disley, Howard, & Hunt, 2006; Lukasik, 2005; Moravec & Stepanek, 2003; Nykänen & Johansson, 2003; Sarkar, Vercoe, & Yang, 2007).

Perceptual experiments that involve descriptions of timbral or emotional aspects of music have commonly used two rating methods: Verbal Attribute Magnitude Estimation (VAME) (Kendall & Carterette, 1993a) and Semantic Differentials (Osgood, Suci, & Tannenbaum, 1957). The VAME uses scales that quantify the applicability of each adjective or descriptor (e.g., *bright <-> not bright*) rather than using opposites or bipolar scales (e.g., *bright <-> dull*). The participants in these experiments are typically required to rate the applicability of these adjectives to monophonic instrument sounds on such scales. Some studies that used the VAME approach report significant negative correlation between descriptors such as *bright* and *dull* (Disley & Howard, 2004; Disley et al., 2006) as well as *bright* and *dark* (Lukasik, 2005).

Nevertheless, the use of semantic differentials, or bipolar scales, has been called into question, since the descriptors on each end of the scale are not necessarily opposites of each other (Darke, 2005; Kendall & Carterette, 1993a). However, this technique does enable controlled judgment of semantic concepts, and it is advocated since it proves to be quite adaptable in these kinds of perceptual tasks (Pratt & Doak, 1976).

Despite the fact that these studies have used very limited sets of instrument sounds—sometimes confined to a single instrument family—some adjectives have been found to be used repeatedly to describe timbre. These include, among others, *harsh*, *bright*, *full,* and *warm*. Following these results, Gounaropoulos & Johnson (2006) created an interface that allows users to synthesize new timbres through the manipulation of some existing feature that correlates to a perceptual quality of the sound. However, these studies may not be applicable to music, which is significantly more complex due to the interplay between its structure and texture.

Polyphonic Timbre

Polyphonic timbre refers to the overall timbral mixture in a music signal, or in simple words, the '*global sound*' of a piece of music (Aucouturier, 2006). The term 'polyphonic' in this context refers to the presence of more than one instrument and to the emerging timbral mixture found in music in general. It should not be confused with the music theoretical term of polyphony versus homophony or monophony.

Gjerdingen and Perrott (2008) refer to polyphonic timbre as an agglomerate of spectral and rapid time-domain variability in an acoustic signal formed in a manner comparable to the Gestalt effect that thereby enables listeners to identify, classify, and categorize the heard piece of music. For instance, the presence of high amounts of acoustic energy in the lower end of the spectrum points more towards rap and hip-hop rather than classical music (Gjerdingen & Perrott, 2008). Another analogy is that of the 'heaviness' of electric guitar sounds that is associated with heavy metal music (Berger & Fales, 2004).

While many studies have focused on understanding the perceptual and cognitive processes of higher-level features such as harmony, melody, and rhythm, the perceptual aspects of polyphonic timbre have received much less attention.

Polyphonic timbre has been found to be a significant perceptual component of music, especially in studies that involve tasks such as genre identification, categorization, or emotional affect attribution. The psychological study carried out by Gjerdingen & Perrott (2008) examined the time required for people to identify or classify into genres very short music excerpts. They reported that extracts as short as 250 ms were sufficient for genre identification. The authors also emphasized the importance of the overall timbre in the perceptual process of identification and categorization.

It also has been shown that adult participants can recognize the affective connotations of sad or happy musical excerpts taken from the Western classical repertoire, even when they are as short as half a second. This is most likely due to the overall timbral and spectro-temporal properties of the short sound signal (Peretz, Gagnon, & Bouchard, 1998).

On a more general note, an interesting development is that of contemporary music, which appears to be deviating from the well-known theories of Western melodic, harmonic, and rhythmic progressions. This music seems to move towards creating new sounds and textures by focusing on the blending of varied timbres. These results simply emphasize the importance of delving into the realm of polyphonic timbre perception.

Polyphonic Timbre in Music Computing

Features representing polyphonic timbre have been found to be important elements in the design of computational systems that categorize music according to genre, style, mood, and emotions. For example, many studies have focused on timbral and rhythmical features when designing such computational systems (e.g., Barbedo & Lopes, 2007; Jiang, Lu, Zhang, Tao, & Cai, 2002; Liu, Lu, & Zhang, 2003; Tzanetakis & Cook, 2002). The most common features used to represent the overall timbre of music signals are the MFCCs (Aucouturier & Pachet, 2003; Logan & Salomon, 2001; Lu, Liu, & Zhang, 2006; McKinney & Breebaart, 2003; Pampalk, Flexer, & Widmer, 2005; Pye, 2000; Zhu, Xue, & Lu, 2004). The MFCCs were first used in the domain of speech recognition (Davis & Mermelstein, 1980) and are derived by performing an operation similar to principal component analysis on spectra shaped by the Mel-Scale that approximates the human auditory response. They have often turned out to be the best predictors of class membership, which might suggest that they encapsulate some important properties of the overall timbre. Other features that are known to characterize polyphonic timbre are spectral flux, spectral roll-off, and spectral centroid (Barbedo & Lopes, 2007; Tzanetakis & Cook, 2002).

Studies performed using the query-by-semantic-description paradigm have focused on creating a meaningful vocabulary to describe aspects of music such as instrumentation, emotional content, genre, song concepts, and usage terms. For instance, recent work by Turnbull and colleagues (2006, 2007, 2008) has dealt with semantic annotation and retrieval of music. They used the delta-MFCCs (ΔMFCCs), which capture the temporal fluctuation of the MFCCs, to represent features of the music tracks.

Whitman and Ellis (2004) created an automatic record review system that annotated music excerpts based on web-reviews of artists. They made use of the features derived from the MFCCs to characterize short music excerpts. By way of machine learning, the system then learned to create reviews based on acoustic properties and their associated semantic descriptions. For generalizability, the authors emphasized the need to use features that capture the overall perceptual quality of the music excerpts rather than use features such as beat and pitch, that require previous knowledge. Although not explicitly mentioned in their study, one might speculate that it is the overall timbre of the music that causes listeners to identify and describe its emotive, situational, or structural aspects.

Acoustic Feature Selection and Modeling Strategies

When attempting to link semantics of polyphonic timbre to acoustic content, selection of acoustic features is a crucial step. Feature selection varies depending on the task at hand and the modeling approach used. Most often features are selected based on existing theoretical knowledge or other criteria such as interpretability. Often automatic feature selection is subsequently performed by employing methods such as step-wise regression or sequential feature selection. These take in all the features as input to produce a desired output, and then through iterative processing select the features that best help predict or provide better results. These approaches commonly are found in a few studies that focus on mood, emotion or expression prediction using audio features (Eerola, Alluri, & Ferrer, 2008; Leman, Vermeulen, De Voogdt, Moelants, & Lesaffre, 2005; Mion & De Poli, 2008; Yang, Lin, Su, & Chen, 2008). In addition there exist methods for automatic generation of acoustic features (Li & Stern, 2004; Schuller, Wallhoff, Arsic, & Rigoll, 2006). However, this is not the approach used in the present study.

Along with spectral structure, one could assume that the temporal evolution of the features plays a role in the perception of polyphonic timbre. The most common modeling techniques employed in computational systems involve static approaches that model the global distribution of features while ignoring their temporal evolution. In their study regarding static and dynamic modeling[1] of polyphonic timbre, Flexer, Pampalk, and Widmer (2005) report that dynamic modeling does not

[1] Static modeling does not take into account the temporal ordering of the features. Rather, it uses them in parallel to estimate the probability distribution. Dynamic modeling, however, is based on the statistical and temporal evolution of the features.

significantly improve the accuracy of such categorization systems at least in the frame-based approach of polyphonic timbre modeling. Moreover, Aucouturier & Pachet (2007) found that the use of dynamic modeling, such as Hidden Markov Models, did not result in any significant increase in the performance of timbre-based classifier systems. Thus it seems that the static approach of modeling is able to capture essential attributes of polyphonic timbre perception.

Aim of the Study

To summarize, there have been a large number of perceptual studies on monophonic timbre. Polyphonic timbre research, however, has mainly confirmed its importance for computational purposes, while its perceptual constituents have been investigated far less. Thus, there is a need for controlled studies that focus on the perceptual aspect of polyphonic timbre.

The aim of the work presented here was to investigate the consistency and predictability of semantic associations of listeners to polyphonic timbre, and to determine the most salient features of polyphonic timbre perception by investigating the descriptive auditory qualities of music and mapping acoustic features to these descriptors. To this end, two experiments were carried out. The first experiment included both a survey and listening test, and helped develop a framework for acquiring quantitative assessments of polyphonic timbre. The second experiment comprised a listening test and computational modeling. In the latter, acoustic features were extracted from the stimuli and were used to predict the ratings through statistical modeling techniques. This process is explained in detail in the following sections.

Experiment 1

Method

The purpose of Experiment 1 was to investigate the descriptive qualities of polyphonic timbre and devise a set of Semantic Differentials or bipolar scales to measure quantitatively its perceptual attributes. It comprised two stages, a survey and Listening Test 1.

STAGE 1: SURVEY

The purpose of the survey was to assess the applicability of rating scales as descriptors of polyphonic timbre and to select a representative subset of scales for subsequent listening tests.

Choice of perceptual scales. Developing perceptual scales that are better suited for a rating task is heavily dependent on the nature of the study (Osgood et al., 1957). Of the two previously mentioned methods of rating, namely VAME and Semantic Differentials, we chose to use the latter in our study. Although the VAME approach is quite popular for perceptual rating tasks, the Semantic Differential approach allows for more concepts to be covered with the same number of ratings, which makes the task less cumbersome especially if the experiment requires the rating of a large number of stimuli for statistical significance. Additionally it has been suggested that the appropriateness of using bipolar scales in a study is determined by the goal of the task at hand, and the choice of semantic concepts should be based on the purpose of the research and need to be relevant to the task (Osgood et al., 1957; Pratt & Doak, 1976).

Consequently, we decided to use bipolar scales starting with those found in the previous literature describing monophonic timbre (Pratt & Doak, 1976; Sethares, 1998). In addition, bipolar scales were formed using timbre descriptors from previous work on monophonic timbre (Helmholtz, 1885/1954) and other studies (Darke, 2005; Disley et al., 2006; Fizgerald & Lindsay, 2004; Kendall & Carterette, 1993a; Moravec & Stepanek, 2003; Sarkar et al., 2007). A few additional scales that were not part of any published studies were included (See Appendix 1 for an overview).

To address the criticism put forth by Kendall and Carterette (1993a) about the potential problem of bipolar opposites not being antipodes, the participants of this survey first were asked whether the suggested pairs forming each scale were opposites before rating their applicability as polyphonic timbre descriptors.

Participants. A total of 71 persons (32 males, age M = 24.8, SD = 6.8) participated in the survey. The majority (96%) of them were either students or staff members at the department of music.

Procedure. The participants were instructed to rate the applicability of 36 bipolar scales on a five-point scale with “1” representing not applicable and “5” most applicable. In addition they were asked to suggest words that they would use to describe polyphonic timbre.

Results. For the subsequent listening tests, we decided to reduce the number of scales to 12 to make the rating task manageable. Among the 12, 11 were chosen based on the applicability ratings and one was chosen based on the suggestions given by the participants. To this end, the ratings were first rank ordered based on their mean applicability scores. If two scales shared

TABLE 1. Mean Applicability Scores of Scales that Scored the Highest.

	Mean Applicability
Heavy-Light	4.27
Cold-Warm	4.24
Soft-Hard	4.23
Dark-Bright	4.16
Simple-Complex	4.00
Colorful-Colorless	3.94
Wide-Narrow	3.93
Strong-Weak	3.89
Acoustic-Synthetic	3.82
Full-Empty	3.66
Gentle-Harsh	3.66

a common descriptor, such as 'Dark-Bright' and 'Dark-Light,' the scale with lower mean applicability was eliminated. From the rank ordered scales, the top 11 scales were included in the subsequent stage. These scales are displayed in Table 1.

Among the additional descriptors suggested by the participants, words related to energy were most frequently mentioned. Hence, an additional bipolar scale, 'High Energy-Low Energy,' was included, resulting in a set of 12 scales for the successive stage.

STAGE 2: LISTENING TEST 1

The aim of Listening Test 1, was to ascertain whether the process of rating the overall timbre of music using bipolar scales was a viable task and to further reduce the number of scales for Experiment 2 to make it more manageable.

Stimuli. One hundred musical excerpts (each with a duration of 1.5 s) of Indian popular music were selected. The reason for choosing Indian music was to reduce the influence of familiarity with the excerpts and bias. The stimulus set was manually chosen to encompass a wide range of genres such as pop, rock, disco, electronic, and contained various instrument combinations including those commonly used in Western music, such as piano, violin, drums, and guitar. All the excerpts were converted to mono files in wav-format (44.1kHz, 16 bit) and were equalized in level by RMS value normalization. All of the excerpts can be found at http://users.jyu.fi/~vialluri/PolyphonicTimbre/.

Participants. Seven persons (5 males, age M = 35.7, SD = 8.2) participated in rating the music excerpts. Almost everybody reported as having very little or no familiarity with Indian popular music. Two reported having no formal music education and the rest had a mean of 13.2 years of formal music training. One reported having no formal music theory education and the rest had a mean of ten years. All of the participants reported listening to music as a hobby for more than half of their lives (M = 26.0 years), with a mean of 15.9 hrs/week. One reported having absolute pitch. None reported any hearing problems. Three of the participants of the survey phase participated in this listening test.

Procedure. The listening experiment took place in a silent room and the participants were given written instructions before the experiment. To present the stimuli and obtain the ratings, a graphical interface was developed in PureData[2] that displayed the eight bipolar rating scales. The music examples were presented via headphones and in random order for each participant. Each scale was divided into nine levels from which the participant could choose the level that best described the music excerpt presented. The interface also had three buttons: one that allowed the participants to play the excerpt as many times as they wish, one to play the music excerpt in a continuous loop with a 400 ms silence between every repetition, and one that returned the ratings to the neutral position and played the next excerpt. The actual experiment was preceded by a trial session in order to familiarize participants with the music examples and the rating process. The excerpts used in the trial session were different for each participant and were chosen randomly from the same set of stimuli used in the actual experiment. Participants were able to view their progress on the left upper corner of the interface. The experiment lasted an hour on average including a small break halfway through.

Results

None of the participants reported any major difficulties in performing the rating task, which suggests that this approach to rating polyphonic timbre is feasible. Table 2 displays the mean inter-subject correlation and Cronbach alphas for each of the perceptual scales. As can be seen, acceptable internal consistency was observed for most of the scales (Cronbach $\alpha \geq .72$ for all except four scales).

Based on the participants' feedback, we decided to reduce the number of scales for Experiment 2 from twelve to eight to reduce the length of the experiment. First, the two scales with the lowest inter-subject

[2] The graphical programming environment is available at http://www.crca.ucsd.edu/~msp/software.html.

TABLE 2. Mean Inter-Subject Correlations and Corresponding Cronbach Alphas for Each of the Listening Test 1 Bipolar Scales.

	Mean Inter-Subject *r*	Cronbach Alpha
Soft-Hard	.50	.88
Colorless-Colorful	.24	.62
Heavy-Light	.41	.83
Warm-Cold	.28	.72
Dark-Bright	.32	.76
Wide-Narrow	.13	.46
Simple-Complex	.20	.61
Acoustic-Synthetic	.55	.90
Gentle-Harsh	.52	.87
Strong-Weak	.20	.62
Empty-Full	.22	.62
High Energy-Low Energy	.39	.82

consistency—'Wide-Narrow' and 'Simple-Complex'—were eliminated, mean inter-subject correlation ≤.21; Cronbach alpha ≤ .61. Correlations among the remaining ten scales are shown in Table 3.

As can be seen, some scales are collinear, suggesting that they might be measuring similar perceptual properties of the music stimuli. For instance, correlations ranging from .71 to .90 (all $p < .001$) were found between the scales 'Gentle-Rough,' 'Soft-Hard,' 'Heavy-Light,' and 'High Energy-Low Energy,' as well as between 'Colorless-Colorful' and 'Acoustic-Synthetic, $r(98) = .79$, $p < .001$.

Elimination of collinear variables was based on squared multiple correlation (SMC). SMC is known to represent the maximal proportion of variance in each variable that can be explained by a linear combination of the other variables and is usually considered the best estimate of communality (Harris, 2001). SMC is inversely related to the uniqueness of the variable. In other words, lower SMC indicates that the variable is unique and vice versa. The SMC values of each of the scales can be seen in Table 4.

As can be seen, the scales 'Heavy-Light,' 'Soft-Hard,' and 'Gentle-Harsh' have the highest SMCs and are thus least unique. The SMCs of these three scales are almost equal and are significantly larger than those of the other scales. Of these three scales, 'Soft-Hard' has a high Cronbach alpha, a high mean applicability score and was thus retained; the other two scales were eliminated.

As a result, the following eight scales were selected for inclusion in Experiment 2: 'Colorful-Colorless,' 'Warm-Cold,' 'Dark-Bright,' 'Acoustic-Synthetic,' 'Soft-Hard,' 'Strong-Weak,' Empty-Full,' and 'High Energy-Low Energy.'

TABLE 3. Inter-Scale Correlations.

	Soft Hard	Colorless Colorful	Light Heavy	Warm Cold	Dark Bright	Acoustic Synthetic	Strong Weak	Gentle Harsh	Empty Full
Colorless Colorful	−.35***								
Heavy Light	−.75***	.39***							
Warm Cold	.64***	−.69***	−.50***						
Dark Bright	−.17	.38***	.64***	−.30**					
Acoustic Synthetic	.48***	−.79***	−.51***	.66***	−.27**				
Gentle Harsh	.90***	−.39***	−.74***	.70***	−.22*	.52***			
Strong Weak	−.58***	−.04	.70***	−.17	.38***	−.10	−.53***		
Empty Full	.22*	.21*	−.36***	−.23*	−.03	−.11	.14	−.53***	
High Energy Low Energy	−.75***	.19	.56***	−.39***	−.00	−.33***	−.71***	.61***	−.36***

*$p < .05$, ** $p < .01$, *** $p < .001$

TABLE 4. Squared Multiple Correlation of All Scales.

Bipolar Scales	Squared Multiple Correlation
Soft-Hard	.88
Colorless-Colorful	.75
Heavy-Light	.90
Warm-Cold	.75
Dark-Bright	.78
Acoustic-Synthetic	.72
Gentle-Harsh	.86
Strong-Weak	.71
Empty-Full	.51
High Energy-Low Energy	.69

TABLE 5. Mean Inter-Subject Correlations and Corresponding Cronbach Alphas for Each of the Listening Test 2 Bipolar Scales.

	Mean Inter-Subject *r*	Cronbach Alpha
Colorless-Colorful	.25	.91
Warm-Cold	.21	.90
Dark-Bright	.40	.96
Acoustic-Synthetic	.43	.96
Soft-Hard	.41	.96
Strong-Weak	.24	.91
Empty-Full	.15	.84
High Energy-Low Energy	.41	.96

Experiment 2

Experiment 2 comprised Listening Test 2 followed by analysis of the behavioral data, extraction of acoustic features, correlation analysis, and regression modeling.

LISTENING TEST 2

For Listening Test 2, the scales selected as a result of Experiment 1 were used to obtain perceptual ratings of music excerpts.

Stimuli. The stimuli comprised the same 100 excerpts used in Listening Test 1.

Participants. Thirty-five persons (20 males, age M = 25.3, SD = 4.1) participated in the experiment. Almost everyone reported having very little or no familiarity with Indian popular music, except for two who reported having taken a one-month course in Indian music. Twenty-five percent of participants reported having no formal music education and the rest had a mean of 9.4 years of formal music training and 6.9 years of music theory. All but one of the participants reported listening to music as a hobby for more than half of their lives (M = 14.1 years), with a mean of 15 hrs/week. Only one participant reported having absolute pitch. Three reported occasional tinnitus and one of a small loss of high frequency response due to aircraft exposure. None of the participants from Listening Test 1 participated in Listening Test 2.

Procedure. The same procedure used in the Listening Test 1 was employed.

Results. The behavioral data were initially checked for inconsistencies and outliers. For each scale, two to three participants were eliminated owing to their mean inter-subject correlation being 2 SDs below the overall mean inter-subject correlation.

Table 5 displays the mean inter-subject correlation and Cronbach alphas for each of the perceptual scales. High agreement between the participants' ratings was observed (Cronbach $\alpha > .90$ for all scales, except for one with α =.84). These findings suggest the presence of fairly consistent opinions among listeners with respect to these scales. For subsequent analysis, the individual ratings for each concept were averaged across all participants.

Table 6 shows the correlations among the scales. As can be seen, many of the scale pairs revealed high mutual correlations. This suggests that they might be associated with the same perceptual dimension. High correlations (greater than .82, p < .001) were found between the scales 'Strong-Weak,' 'Soft-Hard,' and 'High Energy-Low Energy,' and relatively high correlations (greater than .69, p < .001) between the scales 'Colorless-Colorful,' 'Warm-Cold,' and 'Acoustic-Synthetic.' Interestingly, 'Colorless-Colorful' correlated highly with 'Dark-Bright,' $r(98) = .82$, p < .001.

ANALYSIS OF PERCEPTUAL RATINGS

To investigate the underlying structure of the perceptual dimensions, a factor analysis of the ratings was carried out. Various rotations were performed on the factor space to identify the best arrangement of the factors. Based on the Kaiser (1960) criterion, most of the rotations suggested the presence of three or four factors in the data. We found that three factors provided a good compromise between accuracy and interpretability. Table 7 summarizes the three factors obtained as a result of varimax rotation and the corresponding loadings of each of the scales. These findings are consistent with the inter-scale correlations shown in Table 6.

The first factor had high loadings from the scales 'Strong-Weak,' 'Soft-Hard,' and 'High Energy-Low Energy.' This factor appears to describe the overall activity present in the musical excerpt. The scale 'Warm-Cold' appears to play an equal role in the first

TABLE 6. Inter-Scale Correlations.

	Colorless	Warm	Dark	Acoustic	Soft	Strong	Empty
	Colorful	Cold	Bright	Synthetic	Hard	Weak	Full
Warm-Cold	−.69***						
Dark-Bright	.82***	.41***					
Acoustic-Synthetic	−.71***	.74***	.44***				
Soft-Hard	−.25***	.69***	.00	.53***			
Strong-Weak	.04	.48***	−.08	−.38***	.89***		
Empty-Full	.58***	−.30**	.43***	−.21*	.23*	.52***	
High Energy-Low Energy	−.19	−.32**	.35***	−.23*	.82***	.90***	.64***

*p < .05, **p < .01, ***p < .001

two factors but in this case is associated with the first factor owing to its higher loadings. The scales 'Colorless-Colorful,' 'Dark-Bright,' and 'Acoustic-Synthetic,' which seem to represent the perceptual brightness or colorfulness, largely influence the second factor. The third factor relates to the fullness or conversely the sparseness of the music excerpt, owing to the high contribution from the scale 'Empty-Full.' The corresponding factor scores of the above-mentioned perceptual dimensions were used for subsequent analysis and will be referred to from now on as Activity, Brightness, and Fullness.

COMPUTATIONAL ANALYSIS

Parameterization of audio is an important step in computational modeling. A plethora of features exist in literature defining spectral, temporal, and spectrotemporal aspects of audio. In the present study, timbre-related computational features commonly used in previous work (Aucouturier, 2006; Aucouturier & Pachet, 2003; McAdams, Winsberg, de Soete, & Krimphoff, 1995; Tzanetakis & Cook, 2002) were used as a starting point. From these features we included those that are easily interpretable from a perceptual point of view. This led us to exclude features such as spectral kurtosis, skew, irregularity, spread, and flatness. The initial set of features comprised the zero-crossing rate, spectral centroid, high energy-low energy ratio, entropy, spectral roll-off, Mel-Frequency Cepstral Coefficients (13 coefficients), and roughness (see Appendix 2 for an overview).

Additionally, we introduce a new set of features, namely the Sub-Band Flux. The Sub-Band Flux represents the fluctuation of frequency content in ten octave-scaled bands of the spectrum. Figure 1 displays schematically the calculation of the Sub-Band Fluxes.

The division into Sub-Bands was obtained using a 10-channel filterbank of octave-scaled second-order elliptical filters. The frequency response of the filterbank can be seen in Figure 2.

For each of the ten channels the spectral flux was calculated as the Euclidean distance between successive amplitude spectra obtained using Short-Time Fourier

TABLE 7. Loadings of Perceptual Scales onto Each of the Three Factors.

	Factor 1	**Factor 2**	**Factor 3**
	Variance Explained	Variance Explained	Variance Explained
	(41%)	(33%)	(15%)
Colorless-Colorful	−.07	.94	.31
Warm-Cold	.59	−.58	−.34
Dark-Bright	.17	.86	.07
Acoustic-Synthetic	.43	−.67	−.13
Soft-Hard	.96	−.18	−.03
Strong-Weak	−.91	.06	−.27
Empty-Full	.33	.36	.87
High Energy-Low Energy	−.90	−.16	−.33

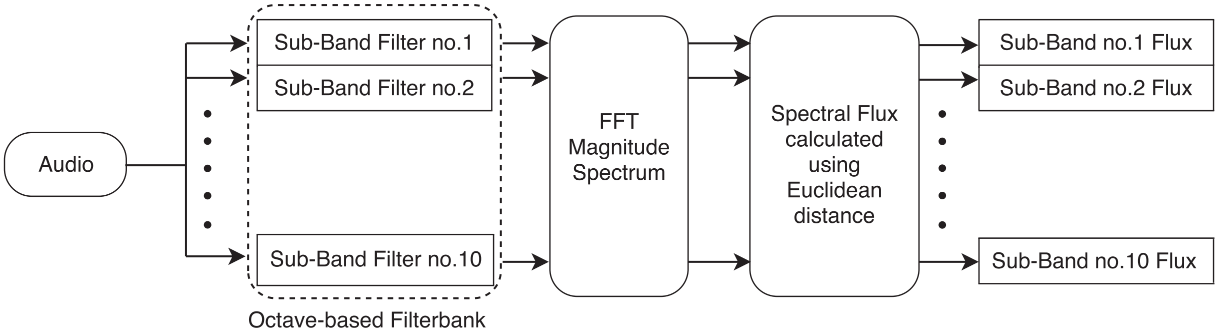

FIGURE 1. Procedure for calculating Sub-Band fluxes.

Transform. The Euclidean distance d is calculated as shown in Equation 1.

$$d = \sqrt{\sum_{n=1}^{N} (A_t[n] - A_{t-1}[n])^2}$$

where A_t and A_{t-1} are magnitude spectra of the audio frames at time t and time t–1, normalized to have unit Euclidean norm, i.e., $\sum A[n]^2 = 1$.

Table 8 displays the boundaries of each of the Sub-Bands.

For many features, the obtained values depend on the length of the analysis window used in feature extraction. For example, roughness increases as a function of the window size. Similarly, the standard deviation for a feature such as entropy, which measures the distribution of spectral energy, can decrease as a function of the window size. In addition, the optimal window size can be dependent on the nature of the stimuli.

To determine the optimal window size for this experiment, we correlated the perceptual ratings with the feature values extracted using various window lengths. For most of the features, we did not observe any significant changes in the correlation values when the window length was changed. Hence we decided to extract all the features using a window length of 25 ms and an overlap of 50% between successive windows. Window lengths of this magnitude often are used for feature extraction in Music Information Retrieval (Barbedo & Lopes, 2007; Lu et al., 2006; Pampalk et al., 2005; Tzanetakis & Cook, 2002).

The feature set consisted of the mean of each acoustic feature across all frames. Hence, a total of 29 acoustic features were extracted per stimulus. The entire analysis was carried out in the MATLAB environment using

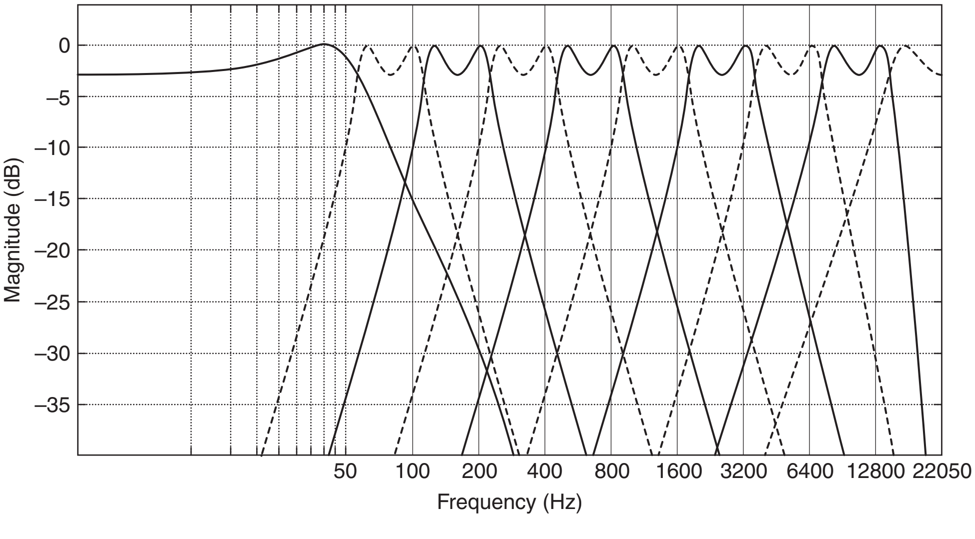

FIGURE 2. Frequency response of the 10-channel filterbank of octave-scaled second-order elliptical filters used for obtaining the Sub-Band fluxes.

TABLE 8. Frequency Ranges of the Sub-Bands.

	Frequency Range
Sub-Band No. 1	0 – 50 Hz
Sub-Band No. 2	50 – 100 Hz
Sub-Band No. 3	100 – 200 Hz
Sub-Band No. 4	200 – 400 Hz
Sub-Band No. 5	400 – 800 Hz
Sub-Band No. 6	800 – 1600 Hz
Sub-Band No. 7	1600 – 3200 Hz
Sub-Band No. 8	3200 – 6400 Hz
Sub-Band No. 9	6400 – 12800 Hz
Sub-Band No. 10	12800 – 22050 Hz

the MIRToolbox (Lartillot & Toiviainen, 2007) for feature extraction.

CORRELATION BETWEEN ACOUSTIC CUES AND PERCEPTUAL DIMENSIONS

Next, we investigated the correlation between factor scores and computationally extracted acoustic features. A Lilliefors test was used to check for the normality of the distribution of all the acoustic features and the perceptual data. As a result, six out of the 29 features were transformed using the Box-Cox transformation (Box & Cox, 1964). Multivariate outliers were screened and constrained to ± 2 *SD*. However, all of the perceptual dimensions passed the normality test and hence did not require any transformation.

A number of acoustic features were found to correlate highly with the perceptual dimensions. For instance, a correlation of $r(98) = .76, p < .001$ was found between Activity and Sub-Band No. 7 Flux, $r(98) = .44$, $p < .001$ between Brightness and zero-crossing rate, and $r(98) = .60$, $p < .001$ between Fullness and Sub-Band No. 2 Flux. As an example, Figure 3 displays the scatter plot of Activity and Sub-Band No. 7 Flux.

Table 9 displays the five acoustic features having the highest correlation values with each of the three perceptual dimensions.

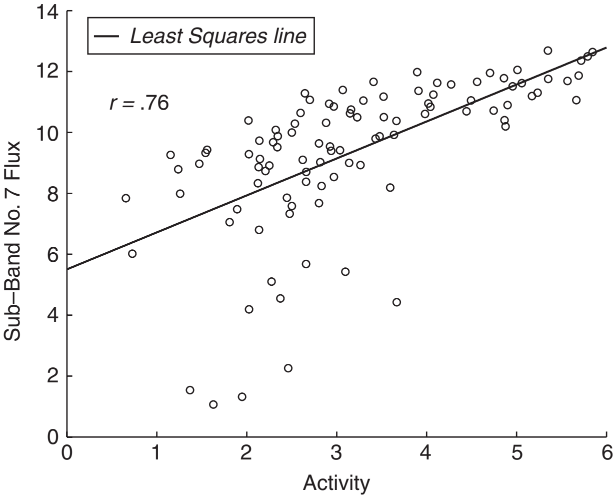

FIGURE 3. Scatter plot of Activity and Sub-Band No. 7 Flux.

As can be seen, the stimuli with higher Activity scores tended to be associated with high flux in the higher end of the spectrum, above 1600 Hz (Sub-Band No. 7, Sub-Band No. 8), more energy in the higher frequency bands when compared to the lower frequency bands (high energy-low energy ratio, zero-crossing rate), and a more even spectral distribution (entropy).

The correlations between the acoustic features and Brightness were lower than for the other two perceptual dimensions. However, the correlations suggest that perceptually bright stimuli possess relatively more high frequency energy (zero-crossing rate, high energy-low energy ratio) than less bright stimuli. In addition, Bright stimuli tended to have less fluctuation in the lower frequency regions (Sub-Band No. 1) and more fluctuation in the higher frequency regions (Sub-Band no. 6 and Sub-Band No. 7).

The perceptual dimension of Fullness exhibits moderately high correlation with the acoustic features. According to the correlations, fuller sounds exhibit

TABLE 9. Acoustic Features Exhibiting the Highest Correlation with the Perceptual Dimensions.

Activity		Brightness		Fullness	
Sub-Band No. 7 Flux	.76***	Zero-crossing Rate	.44***	Sub-Band No. 2 Flux	.60***
Entropy	.73***	Sub-Band No. 1 Flux	−.40***	Sub-Band No. 1 Flux	.47***
High Energy-Low Energy ratio	.71***	Sub-Band No. 6 Flux	.38***	Sub-Band No. 3 Flux	.41***
Zero-crossing Rate	.67***	Sub-Band No. 7 Flux	.36***	High Energy-Low Energy ratio	−.32**
Sub-Band No. 8 Flux	.65***	High Energy-Low Energy ratio	.34***	MFCC13	−.32**

$^{**}p < .01$, $^{***}p < .001$

higher fluctuation in the lower end of the spectrum (Sub-Band No. 1, Sub-Band No. 2, Sub-Band No. 3).

It is noteworthy that the MFCCs did not correlate highly with any of the perceptual dimensions, with the exception of the thirteenth MFCC, which exhibited a moderate correlation, $r(98) = -.32$, $p < .01$, with the Fullness dimension. However, the perceptual relevance of this finding is not straightforward to explain.

Moreover, given the ubiquity of the spectral centroid in previous timbre studies, it was interesting to observe that it was not among the most highly correlated features with any of the perceptual dimensions. The spectral centroid displayed relatively high correlation with Activity, $r(98) = .63, p < .001$, and low correlation with Brightness, $r(98) = .27, p < .01$, and Fullness, $r(98) = -.18$, *n.s.*

REGRESSION ANALYSIS

As a next step, we performed two kinds of regression analyses—stepwise regression and principal components regression (PCR)—to investigate the extent to which the three perceptual dimensions could be predicted by the acoustic features. The MFCCs were excluded from these analyses for two reasons. First, they failed to correlate highly with any of the perceptual dimensions. Second, we found that their inclusion also deteriorated the predictive power of the principal component regression models.

First, we employed stepwise regression analyses with an inclusion criterion of $p < .05$ and an exclusion criterion of $p < .10$. Table 10 summarizes the results of these regression analyses.

As can be seen, a relatively high proportion of the Activity ratings could be explained by four acoustic features, $R^2 = .70, F(4, 95) = 57.73, p < .001$. These, in order of inclusion are Sub-Band No. 6 Flux, Sub-Band No. 3 Flux, High energy–Low energy ratio, and Sub-Band No. 8 Flux. For the Brightness dimension, 31% of the variance could be explained by acoustic features zero-crossing rate, Sub-Band No. 1 Flux, and Sub-Band No. 3 Flux, $F(3, 96) = 15.46, p < .001$. A moderate proportion of the variance, $R^2 = .51, F(4, 95) = 26.93, p < .001$, of the Fullness dimension can be explained using four acoustic features, namely Sub-Band No. 2 Flux, zero-crossing rate, High energy-Low energy ratio, and entropy.

TABLE 10. Beta Coefficients as a Result of Stepwise Regression

	Regression Coefficients (ß)		
Acoustic Feature	Activity $R^2 = .70$	Brightness $R^2 = .31$	Fullness $R^2 = .51$
Zero-crossing Rate	—	.48	−.57
High Energy-Low Energy ratio	.68	—	−.37
Entropy	—	—	.67
Sub-Band No. 1 Flux	—	−.44	—
Sub-Band No. 2 Flux	—	—	.44
Sub-Band No. 3 Flux	.28	−.31	—
Sub-Band No. 6 Flux	.49	—	—
Sub-Band No. 8 Flux	.20	—	—

To reduce the number of independent variables and avoid unwanted effects in regression analysis, such as suppression, due to their collinearity (Tabachnick & Fidell, 2001), a principal component analysis was subsequently performed on the acoustic features. The result of this operation gives rise to principal components that are linear combinations of the acoustic features. They explain maximal variance in the data and are mutually orthogonal. The first seven principal component projections were chosen as the independent variables for a linear regression analysis. This number conforms with the maximal predictor-to-observation ratio suggested by Green (1991). The models thus created were able to explain 67%, $F(7, 92) = 29.44$, $p < .001$, 26%, $F(7, 92) = 6.00$, $p < .001$, and 41%, $F(7, 92) = 10.64$, $p < .001$, of the variance in the Activity, Brightness, and Fullness factor scores, respectively.

CROSS-VALIDATION OF THE MODEL

To screen for eventual overfitting of the data, we performed a cross-validation operation on the models. To this end, we partitioned the data randomly into two subsets consisting of 80 and 20 samples, created a regression model with the larger set, and tested it with the smaller set. This was repeated 1000 times for each model type and each perceptual dimension. The mean of the variance explained across the 1000 runs was taken as the performance measure in the cross-validation. The results can be seen in Table 11.

As can be seen, the cross-validation result for Activity is comparable to that of the original regression model. However, for Brightness and Fullness, the cross-validated values of R^2 are considerably lower than the respective original values. These results suggest that the regression

TABLE 11. Cross-Validation Results of the Regression Models.

	Variance Explained		
	Activity	Brightness	Fullness
Stepwise Regression	70%	31%	51%
Cross-Validation of Stepwise Regression	61%	21%	37%
PCR	67%	26%	41%
Cross-Validation of PCR	64%	22%	36%

model for Activity is more robust in comparison to the Brightness and Fullness models.

Discussion

The work presented here comprises two studies exploring verbal and acoustic correlates of polyphonic timbre. Experiment 1 investigated the semantic space concerning polyphonic timbre from which a subset of eight bipolar scales were selected for quantitative measurements in Experiment 2. The second experiment allowed us to identify acoustic correlates of the obtained underlying perceptual dimensions via correlation analysis. Subsequently, we built models that predicted the perceptual dimensions from acoustic features.

Recent studies on timbre semantics mainly have focused on single sound sources and have aimed at finding semantic labels that best characterize timbre and possibly their respective acoustic correlates. Commonly cited is the work of Von Bismarck (1974) regarding adjectives that describe timbre (Darke, 2005; Disley & Howard, 2004; Disley et al., 2006; Movarec & Stepanek, 2003; Nykänen & Johansson, 2003; Pratt & Doak, 1976). Bismarck suggested a subset of four scales (dull-sharp, compact-scattered, full-empty, and colorless-colorful) to describe the timbre of single instrument sounds. It is interesting to see that the result of the present factor analysis revealed similar dimensions, namely Brightness (representing dull-sharp, and colorless-colorful) and Fullness (representing compact-scattered, and full-empty). In their study on semantic labels and acoustic correlates of oboe and as a result of principal component analysis, Fitzgerald & Lindsay (2004) reported 'Power' and 'Vibrancy' as two of their three perceptual dimensions. 'Power' was found to have high loadings of the semantic label 'Strong.' Kendall and Carterette (1993b) reported the same finding of the word 'Strong' being associated with the 'Power' dimension. In the present study, the Activity dimension appears to measure a similar perceptual quality as the 'Power' dimension in the above mentioned studies.

In several studies, the most common descriptor of monophonic timbre has been reported to be brightness. Helmholtz (1885/1954) uses the term 'bright' or 'brilliant' to classify the musical quality of a tone. In more recent studies, either the word 'bright' (Darke, 2005; Disley & Howard, 2004; Disley et al., 2006; Gounaropoulus & Johnson, 2006) or bipolar scales such as 'dark-bright' (Lukasik, 2005; Sethares, 1998) or variants of it such as 'brilliant-dull' (Pratt & Doak, 1976) have been used to describe monophonic timbre. In our study, the second factor seemed to capture this aspect of perceptual brightness.

Similarly, Fullness, which was the third perceptual factor found in this study, has also been used to describe a perceptual aspect of timbre (Von Bismarck, 1974; Darke 2005). The term 'thin,' which can be regarded as an antonym for full, appears in some studies (Darke, 2005; Disley & Howard, 2004; Disley et al., 2006). Additionally, Helmholtz (1885/1954) used the term 'full' as being descriptive of the musical quality of a tone, or in a simpler word, timbre.

Though the semantic labels discussed above have been used for monophonic instrument sounds, it is interesting to see similar patterns in describing polyphonic timbre. It can be inferred that the semantic associations of monophonic timbre could be extended to polyphonic timbre as well. This suggests common perceptual mechanisms in play while processing timbre, be it monophonic or polyphonic.

The correlations between the perceptual dimensions and acoustic features are interesting in terms of previous studies on timbre spaces of single instrument sounds. Several studies have reported the spectral centroid as an acoustic correlate that explains one of the perceptual dimensions, that is, the perceived brightness (Beauchamp, 1982; Grey, 1977). However, in our study the spectral centroid did not highly correlate with the Brightness dimension, $r(98) = .27$, $p < .01$. A plausible reason for not finding such high correlations for Brightness with the spectral centroid or any of the other acoustic features can be explained in relation to the cognitive listening process of people as described by Aucouturier (2006). Aucouturier suggested that music listeners may hear elements contained in the music that might not be statistically or computationally significant and hence leads to discrepancies between the computational and perceptual data.

In the present study, it appears that pitch content is a crucial component in judgments concerning perceived brightness. For example, the stimuli containing brass instruments had higher scores for Brightness, although this is not reflected in the computational measure of the average spectral centroid. In addition, the presence of high pitch in the stimuli may render it perceptually brighter. Conversely, the absence of high pitch rendered some of the stimuli as perceptually less bright even if they contain higher amounts of energy in the higher frequency range. For example, some of the stimuli contained rapidly repeating percussive sounds with an absence of a prominent pitch in the higher registers. These were rated as less perceptually bright despite the presence of the rapidly varying percussive sounds causing an increase in the energy contained in the higher end of the spectrum.

Nevertheless, the zero-crossing rate and the entropy, which correlated highly with the spectral centroid, $r(98) = .89$, $p < .001$ and $r(98) = .87$, $p < .001$, respectively, were found to be the features that correlated highest with Brightness. Additionally, the high energy-low energy ratio, which again correlated with spectral centroid, $r(98) = .84$, $p < .001$, also appears to correlate in a similar fashion with Brightness.

In previous studies, interpretations of timbre space dimensions other than brightness—indirectly referring to high energy-low energy ratio—have lacked consensus, although various spectral and temporal features have been suggested such as the log-attack time, spectral flux, attack synchrony, spectral irregularity, to name a few (Grey, 1977; Grey & Gordon, 1978; Iverson & Krumhansl 1993; Lakatos, 2000; McAdams et al., 1995). However, Caclin, McAdams, Smith, and Winsberg (2005) were unable to confirm the importance of spectral flux in synthesized isolated tones. In previous studies, log-attack time has been reported as another commonly occurring feature that allows monophonic timbre identification. However, we did not include it in our analysis, because estimating the log-attack time computationally in a polyphonic mixture might not be sufficiently reliable due to the superposition and interleaving of various timbres.

The Activity dimension correlated highly with the flux in Sub-Bands 7 & 8, and the high energy-low energy ratio. Grey (1977) suggested spectral fluctuation as a possible physical interpretation of one of the dimensions of the perceptual space. McAdams et al. (1995) added to this by reporting spectral flux as the acoustic correlate of one of the perceptual dimensions in their timbre space. Specifically, as can be observed from Table 4, the flux in the frequency range of around 1600 ~ 6400 Hz, represented by Sub-Bands 7 and 8, correlated significantly with Activity. Interestingly enough, this frequency region corresponds to the region in the spectrum to which the ear is most sensitive (Fletcher & Munson, 1933). Moreover, Leman et al. (2004) reported the 'activity' factor in their study to be highly related to the manually annotated loudness measure. Further experiments need to be performed to investigate the implications of this finding.

Fitzgerald & Lindsay (2004) reported a strong correlation of the 'Power' dimension with the spectral centroid and a less significant correlation with the spectral flux. On the other hand, their second dimension, 'Vibrancy' also was found to correlate strongly with spectral centroid and spectral variation. The Activity dimension in our study may be thought of as a conglomerate of 'Power' and 'Vibrancy' owing to its high correlation found with Sub-Band Flux and the high energy-low energy ratio.

The Activity or Arousal dimension often is used in studies dealing with emotion or affect evaluation and prediction in speech and music. For example, in their study on emotional connotations of monophonic timbre, Eerola et al. (2008) reported the high energy-low energy ratio as being positively linked to the arousal or energy dimension. In their study the term 'brightness' was used to represent the high energy-low energy ratio. Leman et al. (2005) reported similar finding of the Activity dimension being associated with the same acoustic feature. In the speech literature, Laukka, Juslin, and Bresin (2005) associated high activation to the presence of more high frequency energy, which is measured here by the high energy-low energy ratio. Scherer and Oshinsky (1977) also reported that the presence of high pitch and many harmonics, contributed significantly to the Activity present in electronically synthesized tone sequences, again corresponding to the same acoustic feature.

The third dimension, Fullness, correlated strongly with the fluctuation of the lower end of the spectrum. It is interesting to note that the word 'thin,' which in this case can be regarded as a lack of fullness, has been associated with a reduction in lower frequency components (Disley & Howard, 2004). This association can be seen in the correlation between fluctuation in the lower end of the frequency spectrum—that is below 200 Hz—and the factor scores of the Fullness dimension. This finding is in accordance with Helmholtz's (1885/1954) description of a sound as 'full' due to the outweighing of the prime tone over the upper partials. It can be argued that Fullness could be predicted by mere energy content in the lower frequencies. To test this hypothesis we calculated sub-band energies and correlated them with the Fullness dimension. However, they were rejected owing to their low correlation when compared to the Sub-Band Fluxes with the perceptual dimensions.

As a general remark, an interesting observation is that all the perceptual dimensions seem to correlate significantly with fluctuations in the spectrum. This suggests that spectrotemporal modulations could be most relevant in the perception of polyphonic timbre. Unexpectedly, the MFCCs failed to show high correlations with any of the perceptual dimensions, despite their widespread use as predictors of perceived similarity of timbre (Aucouturier & Pachet, 2003; Logan & Salomon, 2001; Lu et al., 2006;

McKinney & Breebaart, 2003; Pampalk et al., 2005; Pye, 2000; Zhu et al., 2004). One possible explanation for this could be that the transformation used to obtain the MFCCs renders a single MFCC less meaningful, with the exception of the first MFCC, which indicates the low or high pass nature of the sound. To this end, we also tried delta-MFCCs (ΔMFCCs) and Mel-flux as alternative measures of spectral fluctuation. However the correlations obtained between the ΔMFCCs and the perceptual dimensions were much lower (all $\leq .32$).

Mel-flux, which indicates the temporal fluctuations of the spectrum warped on the Mel-scale, was found to correlate highly with the Sub-Band flux. This finding is not surprising as they measure spectral fluctuation on a non-linear scale. However, the correlations obtained between Mel-flux and the perceptual dimensions (all $\leq .67$), was found to be lower than those obtained using Sub-Band flux (all $\leq .75$). The Sub-Band flux hence proved to be a more useful and compact acoustic representation of spectrotemporal modulations.

In the regression analysis, the models performed better for Activity ($R^2 = .70$ in stepwise regression and .67 in PC regression) than for the other two perceptual dimensions. Interestingly, an increase in window size during feature extraction caused a decrease in the variance explained to 58% (stepwise regression) and 56% (PC regression). The model of the Brightness dimension performed inadequately by explaining only 31% and 26% of the variance in the stepwise and PC regression models, respectively. We observed, however, that for this dimension the regression models could explain a higher proportion of variance if the window length used in the feature was increased. For example, 50% (stepwise regression) of the variance can be explained by seven acoustic features with a window size of 700 ms. Similarly, the variance explained as result of PC regression increased to 44%.

The regression model for Fullness performs moderately by explaining 51% (stepwise regression) and 41% (PC regression) of the variance. The variance explained drops to 34% (stepwise regression) and 24% (PC regression) for Fullness as the window size for feature extraction increases to 700 ms. These trends seem to suggest that different perceptual dimensions operate on different timescales. This is a potentially interesting finding that deserves further investigation. It is not clear if this multiscale aspect could be attributed to some characteristics distinctive to the stimulus set we used or if the same trends could be observed across different stimulus sets.

Conclusion

The aim of this study was to explore the perceptual components of polyphonic timbre and look for eventual regularities. The main findings can be summarized as follows. First, there seem to exist mutual consistencies in the ratings, suggesting regularities in the perception of polyphonic timbre across individuals. Second, semantic associations with polyphonic timbre appear to be similar to those of monophonic timbre, suggesting that these two phenomena share common underlying perceptual mechanisms. Third, the results of correlation between acoustic features and perceptual dimensions seem to depend on the length of the window used for feature analysis, suggesting that the perceptual dimensions operate on mutually different time scales. This result calls for further investigation. Fourth, the MFCCs do not seem to contribute significantly to any of the perceptual dimensions. Fifth, spectrotemporal modulations seem to play a vital role in the perception of polyphonic timbre. Finally, the perceptual dimension of Activity can be predicted to a high degree of accuracy using linear regression on acoustic features extracted from the sound.

A possible extension of the work presented here would be to conduct a cross-cultural study in which Indian Popular music and Western Popular music is rated by listeners with varying degrees of familiarity to these musical styles. If similar patterns can be found from those data, it could allow for generalizations about the process of polyphonic timbre perception. Moreover this setting would permit a deeper insight into the aforementioned seemingly multiscale aspect of polyphonic timbre perception.

As mentioned earlier, polyphonic timbre has been a perceptually and computationally important attribute of sound and music. Conducting further controlled experiments and possibly extending them to the neural domain might provide valuable insights into the processing of complex sound dimensions.

Author Note

The authors wish to thank Olivier Lartillot for his help in the computational analysis of the stimuli. This research was supported by the Academy of Finland (project number 7118616).

Correspondence concerning this article should be addressed to Vinoo Alluri, Department of Music, PL 35(M), 40014 University of Jyväskylä, Jyväskylä, Finland. E-MAIL: Vinoo.Alluri@campus.jyu.fi

References

AUCOUTURIER, J. J. (2006). *Dix expériences sur la modélisation du timbre polyphonique* [Ten experiments on the modelling of polyphonic timbre]. Unpublished doctoral dissertation, Universite Paris 6, France.

AUCOUTURIER, J. J., & PACHET, F. (2003). Representing musical genre: A state of the art. *Journal of New Music Research*, *32*, 83-93.

AUCOUTURIER, J. J., & PACHET, F. (2007). The influence of polyphony on the dynamical modeling of musical timbre. *Pattern Recognition Letters, 28*, 654-661.

BARBEDO, J. G. A., & LOPES, A. (2007). Automatic genre classification of musical signals. *The European Association for Signal Processing Journal on Advances in Signal Processing, 2007,* 157-168.

BEAUCHAMP, J. W. (1982). Synthesis by spectral amplitude and "brightness" matching of analyzed musical instrument tones. *The Journal of Audio Engineering Society, 30*, 396-406.

BERGER, H. M., & FALES, C. (2004). The match of perceptual and acoustic features over time. In P. D. Greene & T. Porcello (Eds.), *Wired for Sound: Engineering and Technologies in Sonic Cultures* (p. 181). Middletown, CT: Wesleyan University Press.

BOX, G., & COX, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society.Series B (Methodological), 26*, 211-252.

CACLIN, A., MCADAMS, S., SMITH, B. K., & GIARD, M. H. (2008). Interactive processing of timbre dimensions: An exploration with event-related potentials. *Journal of Cognitive Neuroscience, 20*, 49-64.

CACLIN, A., MCADAMS, S., SMITH, B. K., & WINSBERG, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America, 118*, 471-482.

COSI, P., DE POLI, G., & LAUZZANA, G. (1994). Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research, 23*, 71-98.

DARKE, G. (2005). Assessment of timbre using verbal attributes. In C. Traube & S. Lacasse (Eds.)*, Proceedings of the Second Conference on Interdisciplinary Musicology* (pp. 1-12). Montreal, Quebec, Canada: Université de Montréal.

DAVIS, S. B., & MERMELSTEIN, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *Institute of Electrical and Electronics Engineers, Transactions on Acoustics, Speech, and Signal Processing, 28*, 357-366.

DE POLI, G., & PRANDONI, P. (1997). Sonological models for timbre characterization. *Journal of New Music Research*, *26*, 170-197.

DE POLI, G., PRANDONI, P., & TONELLA, P. (1993) Timbre clustering by self-organizing neural networks. In G. Haus & I. Pighi (Eds.), *Proceedings of X Colloquium on Musical Informatics* (pp. 102-108). Milan: Associazione di Informatica Musicale Italiana.

DISLEY, A. C., & HOWARD, D. M. (2004). Spectral correlates of timbral semantics relating to the pipe organ. *Proceedings of the Joint Baltic-Nordic Acoustics Meeting* (pp. 1-12). Mariehamn, Åland.

DISLEY, A. C., HOWARD, D. M., & HUNT, A. D. (2006). Timbral description of musical instruments. In M. Baroni, A. R. Addessi, R. Caterina, & M. Costa (Eds.), *International Conference on Music Perception and Cognition,* 61-68. Bologna, Italy: Society for Music Perception and Cognition (SMPC) and European Society for the Cognitive Sciences of Music (ESCOM).

EEROLA, T., ALLURI,V., & FERRER, R. (2008). Emotional connotations of isolated instruments sounds. In S. W. Yi (Ed.), *Proceedings of the 10th International Conference on Music Perception and Cognition* (pp. 483-489). Sapporo, Japan: University of Hokkaido.

FITZGERALD, R. A., & LINDSAY, A. T. (2004). Tying semantic labels to computational descriptors of similar timbres. *Proceedings of Sound and Music Computing.* Paris: IRCAM. Retrieved December 7, 2009, from http://smcnetwork.org/files/proceedings/2004/P45.pdf

FLETCHER, H., & MUNSON, W. A. (1933). Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, *5*, 82-108.

FLEXER, A., PAMPALK, E., & WIDMER, G. (2005). Hidden Markov models for spectral similarity of songs. *Proceedings of the 8th International Conference on Digital Audio Effects* (pp. 131-136). Madrid, Spain: International Conference on Digital Audio Effects.

GJERDINGEN, R, O., & PERROTT, D. (2008). Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research, 37*, 93-100.

GOUNAROPOULOS, A., & JOHNSON, C. G. (2006). Synthesising timbres and timbre-changes from adjectives/adverbs. *Lecture Notes in Computer Science, 3907,* 664-675.

GREY, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America. 61*, 1270-1277.

GREY, J. M., & GORDON, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, *63*, 1493-1500.

HARRIS, R. J. (2001). *A primer of multivariate statistics.* New York: Academic Press.

HELMHOLTZ , H. L. F. VON (1885/1954). *On the sensations of tone as a physiological basis for the theory of music* (A. J. Ellis, Trans. of the 1877 German edition). New York: Dover.

IVERSON, P., & KRUMHANSL, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, *94*, 2595-2603.

JIANG, D. N., LU, L., ZHANG, H. J., TAO, J. H., & CAI, L. H. (2002). Music type classification by spectral contrast feature. *Proceedings of the IEEE International Conference on Multimedia and Expo, Vol. 1* (pp. 113-116). Lausanne, Switzerland: IEEE Computer Society.

KAISER, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

KENDALL, R. A., & CARTERETTE, E. C. (1993a). Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck adjectives. *Music Perception*, *10*, 445-467.

KENDALL, R. A., & CARTERETTE, E.C. (1993b). Verbal attributes of simultaneous wind instrument timbres: II. Adjectives induced from Piston's Orchestration. *Music Perception*, *10*, 469-502.

LAKATOS, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception and Psychophysics*, *62*, 426-1439.

LARTILLOT, O., & TOIVIAINEN, P. (2007). MIR in Matlab (II): A toolbox for musical feature extraction from audio, In S. Dixon, D. Bainbridge, & Rainer Typke (Eds.), *Proceedings of the 8th International Conference on Music Information Retrieval* (pp. 237-244). Vienna, Austria: Austrian Computer Society.

LAUKKA, P., JUSLIN, P. N., & BRESIN, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion, 19*, 633-653.

LEMAN, M., VERMEULEN, V., DE VOOGDT, L., MOELANTS, D., & LESAFFRE, M. (2005) Prediction of musical affect using a combination of acoustic structural cues, *Journal of New Music Research*, *34*, 39-67.

LEMAN, M., VERMEULEN, V., DE VOOGDT, L., TAELMAN, J., MOELANTS, D. & LESAFFRE, M. (2004). Correlation of gestural music audio cues and perceived expressive qualities. In R. Goebel, J. Siekmann, & W. Wahlster (Eds.), *Lecture Notes in Artificial Intelligence* (pp. 40-54). Heidelberg, Germany: Springer Verlag.

LI, X., & STERN, R. M. (2004). Feature generation based on maximum normalized acoustic likelihood for improved speech recognition, *Proceedings of Institute of Electrical and Electronics Engineers Transactions on Speech and Audio Processing* (pp. 545-548). Piscataway, NJ: IEEE.

LIU, D., LU, L., & ZHANG, H. J. (2003). Automatic mood detection from acoustic music data. In H. H. Hoos & D. Bainbridge (Eds.), *Proceedings of the 4th International Conference on Music Information Retrieval* (pp. 81-87). Baltimore, Maryland: Johns Hopkins University.

LOGAN, B., & SALOMON, A. (August, 2001). A music similarity function based on signal analysis. *Proceedings of the Institute of Electrical and Electronics Engineers International Conference on Multimedia and Expo* (pp. 745-748). Tokyo, Japan: IEEE Computer Society.

LOUREIRO, M. A., DE PAULA, H. B., & YEHIA, H. C. (2004). Timbre classification of a single musical instrument. *Proceedings of the 5th International Conference on Music Information Retrieval* (pp. 546-549). Barcelona: Universitat Pompeu Fabra.

LU, L., LIU, D., & ZHANG, H. J. (2006). Automatic mood detection and tracking of music audio signals. *Institute of Electrical and Electronics Engineers Transactions on Audio, Speech, and Language Processing*, *14*, 5-18.

LUKASIK, E., (2005). Towards timbre-driven semantic retrieval of violins. *Proceedings of the Fifth International Conference on Intelligent Systems Design and Applications* (pp. 55-60). Wroclaw, Poland: EEE Computer Society.

MCADAMS, S. (1999). Perspectives on the contribution of timbre to musical structure. *Computer Music Journal*, *23*, 85-102.

MCADAMS, S., WINSBERG, S., DE SOETE, G., & KRIMPHOFF, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes, *Psychological Research*, *58*, 177-192.

MCKINNEY, M. F., & BREEBAART, J. (2003). Features for audio and music classification. In H. H. Hoos & D. Bainbridge (Eds.), *Proceedings of the 4th International Conference on Music Information Retrieval* (pp. 151-158). Paris: IRCAM.

MION, L., & DE POLI, G. (2008). Score-independent audio features for description of music expression. *Institute of Electrical and Electronics Engineers, Transactions on Audio Speech and Language Processing,16*, 458-466.

MORAVEC, O., & STEPANEK, J. (2003). Verbal description of musical sound timbre in Czech language. In A. Askenfelt, S. Felicette, E. Jansson, & J. Sundberg (Eds.), *Proceedings of the Stockholm Music Acoustics Conference* (pp. 643-645). Stockholm: University of Stockholm.

NYKÄNEN, A., & JOHANSSON, Ö. (2003). Development of a language for specifying saxophone timbre. In A. Askenfelt, S. Felicette, E. Jansson, & J. Sundberg (Eds.), *Proceedings of the Stockholm Music Acoustic Conference* (pp. 647-650). Stockholm: University of Stockholm.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. (1957). The measurement of meaning. Urbana, IL: University of Illinois Press.

PAMPALK, E., FLEXER, A. & WIDMER, G. (2005). Improvements of audio-based music similarity and genre classification. *Proceedings of Sixth International Symposium on Music Information Retrieval* (pp. 628-633). London: Queen Mary, University of London.

PERETZ, I., GAGNON, L., & BOUCHARD, B. (1998). Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. *Cognition, 68*, 111-141.

PRATT, R. L., & DOAK P. E. (1976). A subjective rating scale for timbre, *Journal of Sound and Vibration*, *43*, 317-328.

Pye, D. (2000). Content-based methods for the management of digital music. *Proceedings of the 2000 Institute of Electrical and Electronics Engineers, Inc. International Conference on Acoustic Speech and Signal Processing* (pp. 2437-2440). Los Alamitos, CA: IEEE Computer Society.

Sarkar, M., Vercoe, B., & Yang, Y. (2007). Words that describe timbre: A study of auditory perception through language. In J. Cross, J. Hawkins, P. Rebuschat, & M. Rohrmeier (Eds.), *Language and Music as Cognitive Systems Conference* (pp. 37-38). Cambridge, UK: University of Cambridge

Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion, 1*, 331-346.

Schuller, B., Wallhoff, F., Arsic, D., & Rigoll, G. (2006). Musical signal type discrimination based on large open feature sets. *Proceedings of Institute of Electrical and Electronics Engineers. International Conference on Multimedia and Expo* (pp. 1089-1092). Los Alamitos, CA: IEEE Computer Society.

Sethares, W. A. (1998). *Tuning, timbre, spectrum, scale.* Berlin: Springer-Verlag.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal, 27*, 379-423.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn and Bacon.

Terasawa, H., Slaney, M., & Berger, J. (2005). Perceptual distance in timbre space. *Proceedings of International Conference on Auditory Display* (pp. 61 - 68). Limerick: International Community for Auditory Display.

Toiviainen, P., Kaipainen, M., & Louhivuori, J. (1995). Musical timbre: Similarity ratings correlate with computational feature space distances. *Journal of New Music Research, 24*, 282-298.

Toiviainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huotilainen, M., & Näätänen, R. (1998). Timbre similarity: Convergence of neural, behavioral, and computational approaches. *Music Perception*, *16*, 223-242.

Turnbull, D., Barrington, L., & Lanckriet, G. (2006). Modeling music and words using a multi-class naive bayes approach. *Proceedings of the International Symposium on Music Information Retrieval* (pp. 254-259). Victoria, Canada: University of Victoria.

Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2007). Towards musical query-by-semantic-description using the cal500 data set. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, & N. Kando (Eds.), *Proceedings of the 30th Annual International Association for Computing Macinery Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval* (pp. 439-446). Amsterdam: ACM SIGIR.

Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. *Institute of Electrical and Electronics Engineers. Transactions on Audio, Speech and Language Processing, 16*, 467-476.

Tzanetakis, G., & Cook, P. (2002). Music genre classification of audio signals. *Institute of Electrical and Electronics Engineers. Transactions on Speech and Audio Processing, 10*, 293-302.

Von Bismarck, G. (1974). Sharpness as an attribute of the timbre of steady sounds. *Acustica 30*, 159-172.

Whitman, B., & Ellis, D. (2004). Automatic record reviews. *Proceedings of the Fifth International Symposium on Music Information Retrieval* (pp. 470-477). Barcelona: Universitat Pompeu Fabra.

Yang, Y. H., Lin, Y. C., Su, Y. F., & Chen, H. H. (2008). A regression approach to music emotion recognition. *Institute of Electrical and Electronics Engineers. Transactions on Audio, Speech and Language Processing*, *16*, 448-457.

Zhu, J., Xue, X., & Lu, H. (2004). Musical genre classification by instrumental features. *Proceedings of the International Computer Music Conference* (pp. 580-583). Miami: International Computer Music Association.

APPENDIX 1. List of bipolar scales used for the survey in English (and Finnish translation).

		Mean Applicability Rating
Dull (Tylppä, Tylsä)	Sharp (Terävä)	3.04
Blunt (Tylppä)	Sharp (Terävä)	3.13
Dull (Tylppä, Tylsä)	Bright (Kirkas)	2.58
Heavy (Painava)	Light (Kevyt)	4.27
Dark (Tumma)	Light (Kevyt)	3.62
Dark (Tumma)	Bright (Kirkas)	4.16
Cold (Kylmä)	Warm (Lämmin)	4.24
Cold (Kylmä)	Hot (Kuuma)	2.86
Soft (Pehmeä)	Hard (Kova)	4.23
Full (Täysi)	Empty (Tyhjä)	3.67
Pure (Puhdas, Paljas)	Rich (Rikas)	2.73
Colorful (Värikäs)	Colorless (Väritön)	3.94
Compact (Tiivis)	Scattered (Hajaantunut)	3.31
Dense (Tiheä, Tiivis)	Scattered (Hajaantunut)	2.83
Dense (Tiheä, Tiivis)	Thin (Ohut)	2.07
Thick (Paksu)	Thin (Ohut)	3.58
Fat (Tukeva)	Thin (Ohut)	2.56
Acoustic (Akustinen)	Synthetic (Synteettinen)	3.82
Metallic (Metallinen)	Wooden (Puinen)	3.61
Resonant (Sointuva, Täyteläinen)	Ringing (Soiva, Kilisevä)	2.45
Strong (Voimakas)	Faint (Heikko)	3.52
Strong (Voimakas)	Weak (Heikko)	3.89
Gentle (Hillitty, Hellä)	Harsh (Karkea, Kova)	3.66
Gentle (Hillitty, Hellä)	Rough (Karkea)	3.45
Smooth (Pehmeä)	Rough (Karkea)	3.65
Crunchy (Rapea)	Smooth (Pehmeä)	2.31
Clean (Puhdas)	Gritty (Rakeinen)	2.38
Clear (Kirkas, Kuulas)	Muted (Vaimennettu)	3.23
Clear (Kirkas, Kuulas)	Hazy (Sumea)	3.34
Nasal (Nasaali)	Clear (Kirkas)	2.83
Nasal Nasaali)	Solid (Kiinteä)	1.99
Wide (Leveä)	Narrow (Kapea)	3.93
Deep (Syvä)	Shallow (Matala)	2.93
Simple (Yksinkertainen)	Complex (Monitahoinen)	4.00
Crispy (Murea, Rapea)	Damp (Kostea)	1.83
Steady (Tasainen)	Shaky (Huojuva)	3.42

Appendix 2

All audio files have been subjected to frame-by-frame analysis with a frame size of 25 ms and an overlap of 50% between successive frames. For the spectral and spectrotemporal features, each frame was multiplied with the Hamming window before applying the Discrete Fourier Transform. A brief description of each of the acoustic features is presented below. A detailed explanation can be found in the user manual of the MIRToolbox (Lartillot & Toiviainen, 2007).

Temporal

ZERO CROSSING RATE:

Number of time-domain zero-crossings of the signal per time unit

Spectral

CENTROID:

Geometric center of the amplitude spectrum

HIGH ENERGY–LOW ENERGY RATIO:

Ratio of energy content below and above 1500 Hz

ENTROPY:

The relative Shannon (1948) entropy is calculated using the equation given below.

$$H_t = \frac{\sum_{n=1}^{N} A_t[n] \log A_t[n]}{\log N}$$

where A_t is the amplitude spectrum of audio frame at time t and N the number of frequency bins in the amplitude spectrum.

The relative Shannon entropy indicates whether the spectrum contains predominant peaks or not. For example, a single sine tone has minimal entropy and white noise maximal.

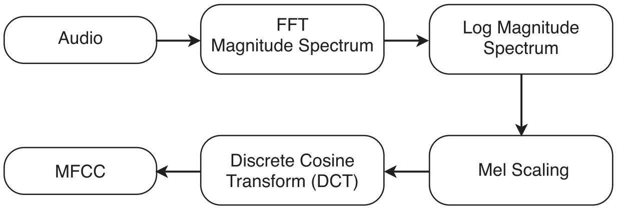

FIGURE 4. Procedure for calculating Mel-Frequency Cepstral Coefficients.

ROLL-OFF 85:

Frequency below which 85% of the total energy exists

MEL-FREQUENCY CEPSTRAL COEFFICIENTS (13 FEATURES IN TOTAL):

MFCCs offer a description of the spectral shape of the sound. MFCCs are a result of the discrete Cosine Transform (DCT) of the log amplitude of the Mel-frequency spectrum of an audio signal.

Figure 4 illustrates the process of obtaining the MFCCs.

As can be seen, the magnitude spectrum of the audio is first positioned on the Mel scale (which approximates the human auditory system's response more closely than the linearly-spaced frequency bands). This weighted spectrum is then subjected to a log operation followed by a discrete cosine transform (DCT), which is a Fourier-related transform similar to the discrete Fourier transform (DFT), but using only real numbers. Most of the signal information tends to be concentrated in the first few components of the DCT. Following the convention we used the first 13 components.

Spectrotemporal

SUB-BAND FLUX (10 FEATURES IN TOTAL):

Measure of fluctuation of frequency content in ten octave-scaled sub-bands of the spectrum (See main text for details).

ROUGHNESS:

Estimate of sensory dissonance (Sethares, 1998)