Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009) Jyväskylä, Finland
Jukka Louhivuori, Tuomas Eerola, Suvi Saarikallio, Tommi Himberg, Päivi-Sisko Eerola (Editors)

# In Search of Perceptual and Acoustical Correlates of Polyphonic Timbre

Vinoo Alluri & Petri Toiviainen

*Department of Music, Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland*
vialluri@jyu.fi, petri.toiviainen@jyu.fi

## ABSTRACT

Polyphonic timbre refers to the overall timbre mixture of a music signal, or in simple words, the 'global sound' of any piece of music. It has been proven to be an important element for computational categorization according to genre, style, mood, and emotions, but its perceptual constituents have been less investigated. The aim of the study is 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. Descriptors of monophonic timbre taken from previous literature were used as a starting point. Based on three pilot studies, eight scales were chosen for the actual experiment. Short musical excerpts from Indian popular music were rated on these scales. Relatively high agreement between the participants' ratings was observed. A factor analysis of the scales suggested three perceptual dimensions. Acoustic descriptors were computationally extracted from each stimulus using signal processing and correlated with the perceptual dimensions. The present findings imply that there may be regularities and patterns in the way people perceive polyphonic timbre. Furthermore, most of the descriptors can be predicted relatively well by the acoustical features of the music. Finally the results suggest that spectrotemporal modulations are most relevant in the perception of polyphonic timbre.

## I. INTRODUCTION

Polyphonic timbre refers to the overall timbral mixture in a music signal, or in simple words, the '*global sound*' of any piece of music (Aucouturier, 2006)[1]. While many studies have focused on understanding the perceptive and cognitive processes of higher-level features such as harmony, melody and rhythm, the perceptual aspects of polyphonic timbre have been less studied and understood. Polyphonic timbre has been found to be a significant component especially in studies that involve perceptual tasks such as genre identification, categorization or emotional affect attribution. The psychological study done by Gjerdingen & Perrott in 1999 examined the time required for people to identify or classify very short music excerpts into genres. They reported that extracts as short as 250 ms were sufficient for genre identification thereby emphasizing the importance of the overall timbre in the perceptual process of identification and categorization. Similarly, it has been shown that adult subjects could recognize the sad or happy affective connotations of musical excerpts taken from the Western classical repertoire even when they were as short as half a second, most likely relying on the overall timbral characteristics of the short sound signal (Peretz et al., 1998). On a more general note, an interesting development is that of modern music, which appears to be deviating from the well-known theories of Western melodic, harmonic and rhythmic progressions. It apparently seems to move towards creating new sounds and textures by focusing on blending of varied timbres and concentrating less on structures and forms. This development simply emphasizes the importance of delving into the realm of polyphonic timbre perception. Besides its perceptual relevance, polyphonic timbre has been found to be an important element for computational categorization according to genre (Tzanetakis et al., 2001; Aucouturier & Pachet., 2003), artist (Berenzwig et al., 2002), mood (Lu et al., 2006), semantics (Slaney, 2002; Turnbull et al., 2008) and emotions (Leman et al., 2005; Trohidis et al., 2008; Yang et al., 2008). It has been often reported that commonly occurring features in the design of such models include timbral features such as brightness, roughness, spectral flux and spectral rolloff, to name a few. Considerable advances have been made in the area of polyphonic timbre modeling for computational purposes but its perceptual constituents have been less investigated. This calls for controlled perceptual studies that focus on polyphonic timbre.

[1] 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 and should not be confused with the music theoretical term of polyphony vs. homophony or monophony

## II. AIM OF THE STUDY

The aim of the work presented here is to investigate the consistency and predictability of semantic associations of listeners to polyphonic timbre; also 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, a listening test was carried out where the participants were required to rate the overall timbre of short music excerpts using the semantic differential approach (Osgood et al., 1957). Feature extracted from the music stimuli were correlated with the perceptual dimensions. This process will be explained in detail in the following sections.

## III. EXPERIMENT

### A. Experiment Details

*1) Choice of Perceptual Scales*. Timbre, both monophonic and polyphonic, is known to be an abstract concept, and selecting a set of descriptors that best characterize it can be a debatable task. The notions of descriptors may vary from person to person and have different mental representations. Nevertheless, the choice of semantic concepts should be based on the purpose of research and need to be relevant to the task at hand (Osgood et al., 1957; Kendall & Carterette, 1991). Descriptors of monophonic timbre taken from previous literature (Sethares, 1999; Pratt & Doak, 1975) were used as a starting point. Common words from surveys and studies (Disley et al., 2006; Darke, 2005; Moravec & Stepanek, 2006, Sarkar et al., 2007; Kendall & Carterette, 1991) were chosen

and then used to form bipolar scales. As a result of three pilot studies, the following eight scales were finally chosen: 'Colourful-Colourless', 'Warm-Cold', 'Dark-Bright', 'Acoustic-Synthetic', 'Soft-Hard', 'Strong-Weak, 'Empty-Ful l', and 'High Energy-Low Energy'.

*2) Stimuli*. One hundred music excerpts with a duration of 1.5 seconds each were chosen form Indian popular music. The duration of the stimuli was so chosen in order to increase the focus of the participant on the global sound rather than higher musical structures such as harmony, rhythm and melody. The stimuli encompassed a wide range of instrument combinations including those commonly used in Western music. All the excerpts were converted to mono files in wav-format (44.1kHz, 16 bit) and were equalized in terms of loudness by RMS value normalization. A 23ms fade-out at the end of each sample was done to prevent any abrupt termination of the sound.

*3) Participants*. Thirty-five persons participated (20 males, age M=25.29, SD=4.05) in the rating test. Almost everybody reported as having very little or no familiarity with Indian music except two who reported as having taken a one-month course in Indian music. 25% reported having no formal music education and the rest had a mean of nine (9.38) years of formal musical training and seven years (6.88) of theory. All but one of the participants reported as having had music listening as a hobby for more than half of their lives (M = 14.06 years), with a mean of 15 hrs/week. Only one reported having absolute pitch. Three reported occasional tinnitus and one of small loss of high frequency response due to aircraft exposure.

*4) Procedure*. The listening experiment took place in a silent room and the participants were given written instructions before the experiment. The music examples were presented via headphones and presented in random order for each participant to avoid any bias in rating. To present the stimuli and obtain the ratings, an interface developed using PureData[2] was used. The interface displayed the eight bipolar rating scales with each scale divided into 9 levels from which the subject could choose the level that best described the heard music excerpt. The interface had three buttons, one that allowed the subjects to play the excerpt as many times as they wish, the other to play the music excerpt in a continuous loop with a 400ms silence between every repetition; and another that played the next excerpt and returned the ratings to the neutral position. The actual experiment was preceded by a training session in order to familiarize the participants with few of the music examples and the rating process. Participants were able to view their progress on the left upper corner of the interface. The experiment lasted on average for an hour including a small break midway.

## B. Acoustic Descriptors

Parameterization of audio is an important step in computational modelling. A plethora of features exist in literature that define spectral, temporal and spectrotemporal aspects of audio. In addition to the commonly used timbre descriptors from previous studies such as zero-crossing rate, spectral centroid, spectral flux, roughness, spectral roll-off to name a few, (Tzanetakis and Cook, 2002; Aucouturier and Pachet, 2003; Aucouturier, 2006; McAdams et al., 1995), a new set of features is proposed, namely the sub-band flux or octave-based flux. The octave-based sub-band flux represents the fluctuation of frequency content in octave-scaled bands of the spectrum. Each stimulus was subjected to frame-by-frame analysis and a total of nineteen features (including ten of the octave-based flux features) were computationally extracted (See Appendix for an overview). The feature space consisted of the means and standard deviations of all nine features across all frames except the octave-based features as they are represented only by their mean fluctuation in each band, thereby resulting in a 28-dimensional feature vector. The entire analysis was carried out in the MATLAB environment with the MIRToolbox (Lartillot & Toiviainen, 2007) for feature extraction.

## C. Results

The behavioural data was 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 1. displays the mean inter-subject correlation and croncbach alphas for each of the perceptual scales. As can be seen, high agreement between the participants' ratings was observed (Cronbach $\alpha > 0.9$ for all scales except one scale with $\alpha=0.84$). These findings suggest the presence of fairly mutually consistent opinions among listeners in the perception of polyphonic timbre. For subsequent analysis, the individual ratings for each concept were averaged across all participants and then used.

**Table 1. Mean inter-subject correlation and corresponding Cronbach's alphas**

| | Mean inter-subject correlation | Cronbach's alpha |
|---|---|---|
| **Colorless Colorful** | 0.25 | 0.91 |
| **Warm Cold** | 0.21 | 0.90 |
| **Dark Bright** | 0.40 | 0.96 |
| **Acoustic Synthetic** | 0.43 | 0.96 |
| **Soft Hard** | 0.41 | 0.96 |
| **Strong Weak** | 0.24 | 0.91 |
| **Empty Full** | 0.15 | 0.84 |
| **High Energy Low Energy** | 0.41 | 0.96 |

[2] A graphical programming environment http://www-crca.ucsd.edu/~msp/software.html

**Table 2. Inter-scale correlation**

| | Colourless Colourful | Warm Cold | Dark Bright | Acoustic Synthetic | Soft Hard | Strong Weak | Empty Full |
|---|---|---|---|---|---|---|---|
| **Warm Cold** | -0.69*** | -- | | | | | |
| **Dark Bright** | 0.82*** | -0.41*** | -- | | | | |
| **Acoustic Synthetic** | -0.71*** | 0.74*** | -0.44*** | -- | | | |
| **Soft Hard** | -0.25* | 0.69*** | 0.00 | -0.53*** | -- | | |
| **Strong Weak** | 0.04 | -0.48*** | -0.08 | -0.38*** | -0.89*** | -- | |
| **Empty Full** | 0.58*** | 0.30*** | 0.43*** | -0.21* | 0.23* | -0.52*** | -- |
| **High Energy Low Energy** | -0.19 | -0.32** | -0.35*** | -0.23* | -0.82*** | 0.90*** | -0.64*** |

*$p$<0.05, **$p$<0.01, ***$p$<0.001

Many of the scales revealed high mutual correlation exceeding 0.6 in absolute value, as can be seen in Table 2. This suggests that they might be associated with the same perceptual dimension. Significantly high correlation was found between the scales 'Strong-Weak', 'Soft-Hard', and 'High Energy-Low Energy' ($|r| \geq 0.82$) and relatively high correlation between the scales 'Colourless-Colourful', 'Warm-Cold', and 'Acoustic-Synthetic' ($|r| \geq 0.69$).

*1) 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. The scree plots obtained with 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 3 summarizes the three factors obtained as a result of varimax rotation and the corresponding loadings of each of the scales.

**Table 3. Factor Loadings**

| | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| **Colourless Colourful** | -0.07 | 0.94 | 0.31 |
| **Warm Cold** | 0.59 | -0.58 | -0.34 |
| **Dark Bright** | 0.17 | 0.86 | 0.07 |
| **Acoustic Synthetic** | 0.43 | -0.67 | -0.13 |
| **Soft Hard** | 0.96 | -0.18 | -0.03 |
| **Strong Weak** | -0.91 | 0.06 | -0.27 |
| **Empty Full** | 0.33 | 0.36 | 0.87 |
| **High Energy Low Energy** | -0.90 | -0.16 | -0.33 |

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 two factors but in this case is associated with the first factor owing to its higher loadings. The scales 'Colourful-Colourless', 'Dark-Bright' and 'Acoustic-Synthetic' that seem to represent the perceptual brightness or colourfulness largely influence the second factor. The third factor relates to the 'fullness' or '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 describing the 'activity', 'brightness' and 'fullness' were used for subsequent analysis.

*2) Correlation between acoustical cues and perceptual dimensions*.

Next we investigated the correlation between factor scores and acoustic features. Correlation revealed $|r| = 0.75$ between 'activity' and spectral flux, $|r|=0.40$ between 'brightness' and zero-crossing rate, and $|r|=0.58$ between 'fullness' and sub-band-2 flux (all $p < 0.001$). As an example, figure 1 displays the scatter plot of spectral flux and 'activity'.

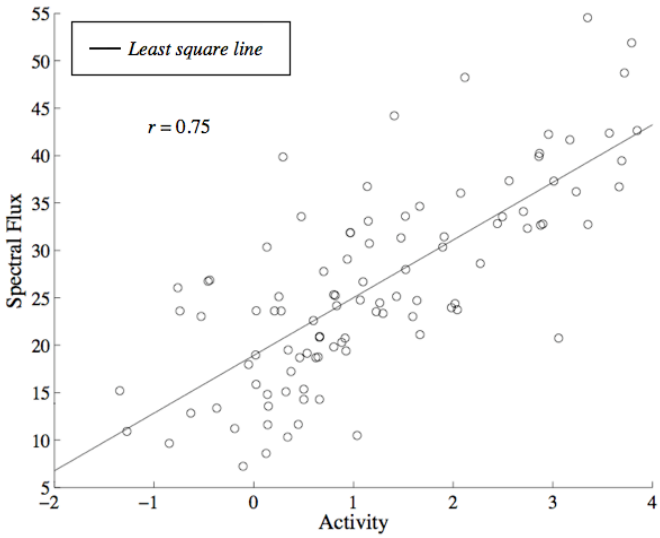


**Figure 1. Scatter plot of spectral flux and the 'activity' dimension.**

Table 4 displays the acoustic features having the highest correlation values with the three perceptual dimensions

**Table 4. Correlation results between acoustic features and the perceptual dimensions.**

| Activity | Brightness | Fullness |
|---|---|---|
| Spectral Flux 0.75*** | Zero-crossing Rate 0.40*** | Sub-Band No.2 Flux 0.58*** |
| High Energy-Low Energy ratio 0.74*** | Sub-Band No.1 Flux -0.40*** | Sub-Band No.1 Flux 0.45*** |
| Sub-Band No.8 Flux 0.65*** | High Energy-Low Energy ratio 0.36*** | Zero-crossing Rate -0.43*** |
| Spectral Centroid 0.63*** | Sub-Band No.6 Flux 0.34*** | Sub-Band No.3 Flux 0.40*** |
| Sub-Band No.7 Flux 0.62*** | Sub-Band No.7 Flux 0.32** | Irregularity(std) -0.36*** |

**$p$<0.01, ***$p$<0.001

## IV. DISCUSSION

Recent studies on timbre semantics have mainly focused on single sound sources aiming at finding semantic labels that best characterize timbre and possibly their respective acoustical correlates. Commonly cited is the work of von Bismarck (1974) regarding adjectives describing timbre (Pratt & Doak, 1975; Disley et al., 2006; Movarec & Stepanek, 2003; Darke, 2005; Disley & Howard, 2004; Nykanen & Johansson, 2003). He suggests a subset of four scales (dull-sharp, compact-scattered, full-empty, and colourless-colourful) 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 the 'brightness' (representing dull-sharp, and colourless-colourful) and 'fullness' (representing compact-scattered, and full-empty). Fitzgerald & Lindsay (2004) in their study on semantic labels and acoustical correlates of oboe report, as a result of principal component analysis, '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 (1993) report the same finding of the word 'Strong' being associated with the 'Power' dimension. A similar finding is observed in our study wherein the 'Activity' dimension has a relatively high loading of the scale 'strong-weak' (see Table 3).

In several studies, the most common descriptor of monophonic timbre has been reported to be brightness. Either the word 'bright' (Darke, 2005; Disely et al., 2006; Disley & Howard, 2004; Johnson & Gounaropoulos, 2006) or bipolar scales such as 'dark-bright' (Lukasik, 2005, Sethares, 1998) or variants of it such as 'brilliant-dull' (Pratt & Doak, 1975) have been used to describe monophonic timbre. In our study, the second factor seems 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 (Darke 2005; Von Bismarck, 1974). Additionally, the term 'thin', which can be regarded as an antonym for full, appears in some studies (Darke, 2005; Disley et al., 2006; Disley & Howard, 2004).

Though the semantic labels discussed above have been used for monophonic instrument sounds, it is surprising 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 results of the analysis of 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. A plausible reason for not finding such high correlations for 'brightness' with the spectral centroid or any other of the other acoustic features can be explained in relation to the cognitive listening process of people as described by Aucouturier (2006). He suggests 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. For example, the stimuli containing instruments belonging to the brass family had higher scores for 'brightness' although this is not reflected in the computational measure of the spectral centroid. In addition, the presence of high pitch in the stimuli may render it perceptually 'brighter'. Nevertheless, the zero-crossing rate, which correlates highly with the spectral centroid (r = 0.89), was found to be the feature that most correlated with 'brightness' dimension. Additionally, the High energy-Low energy ratio, which again correlates with the spectral centroid (r = 0.88) also appears to correlate in a similar fashion with 'brightness'.

In previous studies, interpretations of timbre space dimensions other than brightness, 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).

The 'Activity' dimension correlated highly with spectral flux and high energy-low energy ratio. Fitzgerald & Lindsay (2004) report strong correlation of the 'Power' dimension with the spectral centroid and less significant correlation with the spectral flux. On the other hand, the second dimension reported in their study, that is, 'Vibrancy' was also 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 high the correlation found with the spectral flux and spectral centroid.

In the early study performed by Grey in 1977, the author suggests spectral fluctuation as a possible physical interpretation of one of the dimensions of the perceptual space. McAdams et al. (1995), add to this by reporting spectral flux as a measure that quantified of one of the perceptual

dimensions. Specifically, as can be observed from Table 4, the flux in the frequency range of around 1600 Hz ~ 6400 Hz, represented by sub-bands 6 and 7, correlates significantly with 'activity'. Interestingly enough, this frequency region corresponds to the region in the spectrum that the ear is most sensitive to (Fletcher & Munson, 1933). Further controlled experiments need to be performed to better investigate the implications of this finding.

The third dimension, that is, the 'fullness' correlates 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.

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.

## V. CONCLUSION

The aim of this study was to explore the perceptual components of polyphonic timbre and look for, if any, regularities in the process of perception. 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, we found that spectrotemporal modulations play a vital role in the perception of polyphonic timbre.

In future, to predict the perceptual dimensions we plan to carry out regression analysis using the acoustic features as predictors. The predicted perceptual dimensions could then be used as predictors for higher-level concepts such as emotions and preference. A possible extension to this study would be to conduct a cross-cultural study by comparing the ratings of people who are similar with Indian music to the data collected here. If similar patterns can be found from that data, it could allow for generalizing the process 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 it to the neural domain might provide valuable insights into the processing of complex sound dimensions.

## ACKNOWLEDGMENT

We thank Olivier Lartillot for his help in computational analysis of the stimuli.

## REFERENCES

Aucouturier, J.J. (2006) Dix Expériences sur la Modélisation du Timbre Polyphonique, PhD Dissertation.

Aucouturier, J.J. & Pachet, F. (2003). Representing Musical Genre: A State of the Art. *Journal of New Music Research*, 32(1), (pp. 83-93).

Beauchamp, J. W. (1982) Synthesis by spectral amplitude and "brightness" matching of analyzed musical instrument tones. *The Journal of Audio Engineering Society,* 30(6), (pp. 396-406).

Berenzweig, A., Ellis, D. and Lawrence, S. (2002). Using voice segments to improve artist classification of music. *Proceedings of the twenty-second International Conference of Audio Engineering Society,* Espoo, Finland, (pp. 15–17).

Darke. G. (2005). Assessment of timbre using verbal attributes. In *Proceedings of the 2005 Conference on Interdisciplinary Musicology*. http://www.oicm.umontreal.ca/cim05/.

Disley, A., Howard, D. & Hunt, A. (2006). Timbral description of musical Instruments. *Proceedings of International Conference of Music Perception and Cognition*. Bologna: Bologna Univ. Press.

Disley, A.C. and Howard, D.M. (2004). Spectral correlates of timbral semantics relating to the pipe organ. *Speech, Music and Hearing,* 46.

Fitzgerald, R. A., and Lindsay, A. T. (2004) Tying semantic labels to computational descriptors of similar timbres. *Proceedings of Sound and Music Computing*.

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

Grey, J. M. (1977) Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustic. Society of America.,* 61(5), (pp. 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(5), (pp. 1493-1500).

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

Johnson, C. G. and Gounaropoulos, (2006) A. Timbre interfaces using adjectives and adverbs. *NIME* (pp. 101–102).

Kendall,R.A., & Carterette, E.C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Percetpion,* 8, (pp. 369-404).

Lakatos, S (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, 62(7) (pp. 426–1439).

Lartillot, O., & Toiviainen, P. (2007). MIR in Matlab (II): A toolbox for musical feature extraction from audio, *Proceedings of the 8th International Conference on Music Information Retrieval* (pp. 237-244).Vienna,AT:Österreichische Computer Gesellschaft.

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*, Vol. 34(1), (pp. 39-67).

Lu, L., Liu, D., Zhang, H.J. (2006). Automatic Mood Detection and Tracking of Music Audio Signals, *IEEE transactions on audio, speech, and language processing,* Vol. 14(1), (pp. 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).

Malcolm Slaney.(2002). Semantic–Audio Retrieval. *Proceedings of the 2002 IEEE, ICASSP*, Orlando, FL.

McAdams, S., Winsberg, S., de Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes, *Psychological Research*, 58, (pp. 177-192).

Moravec, O., Stepanek, J. (2003). Verbal description of musical sound timbre in Czech language. *Proceedings of the Stockholm Music Acoustics Conference*, Stockholm (pp. 643-645).

Nykänen A & Johansson Ö (2003). Development of a language for specifying saxophone timbre, *Proceedings of the Stockholm Music Acoustic Conference,* (pp. 647-650).

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana, Ill: University of Illinois Press.

Peretz, I., Gagnon, L., & Bouchard, B. (1998). Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. *Cognition,* 68(2), (pp. 111–141).

Pratt, R. L. & Doak P. E. (1976). A subjective rating scale for timbre, *Journal of Sound and Vibration*,43(3), (pp. 317-328).

Rosemary A. Fitzgerald and Adam T. Lindsay. (2004) Tying semantic labels to computational descriptors of similar timbres. *Proceedings of Sound and Music Computing*.

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 (Hg.), *Language and Music as Cognitive Systems Conference*, Conference Booklet (S. 37-38). Cambridge.
http://web.media.mit.edu/~mihir/documents/mihir_sound_survey_poster.pdf
Sethares, W. A. (1998). *Tuning, Timbre, Spectrum, Scale*, Springer-Verlag
Trohidis, K., Tsoumakas, G., Kalliris, G. and Vlahavas, I. (2008) Multilabel classification of music into emotions. *Proceedings of the ninth International Conference on Music Information Retrieval,* Philadelphia, PA, USA.
Turnbull, D., Barrington, L., Torres, D. and Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects, *IEEE Transactions on Audio, Speech and Language Processing,* 16, (pp. 467-476).
Tzanetakis, G., Essl, G. and Cook, P. (2001) Automatic musical genre classification of audio signals. *Proceedings of the International Symposium on Music Information Retrieval*, (pp. 205–210).
von Bismarck G (1974). Sharpness as an attribute of the timbre of steady sounds, *Acustica* 30(3), (pp. 159-172).
Yang, Y.-H., Lin, Y.-C., Su, Y.-F., and Chen. H.-H. (2008) A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, 16(2) (pp. 448–457).

# APPENDIX

**Table A. Acoustic features and their descriptions**

| **Feature** | **Description** |
|---|---|
| *Temporal* | |
| Zero Crossing Rate | Number of time-domain zero-crossings of the signal per time unit |
| *Spectral* | |
| Centroid | Geometric centre of the amplitude spectrum |
| High Energy – Low Energy Ratio | Ratio of energy content below and above 1500 Hz |
| Spread | Standard Deviation of the amplitude spectrum |
| Skewness | Asymmetry measure of the amplitude spectrum |
| Rolloff 95 | Frequency below which 95% of the total energy exists |
| Irregularity | Measure of jaggedness of spectrum |
| *Spectrotemporal* | |
| Spectral Flux | Measure of change in the spectrum over time |
| Roughness | Estimate of Sensory dissonance |
| Sub-band flux | Measure of fluctuation of frequency content in octave-scaled sub-bands of the spectrum<br><br>The octave-based sub-band flux is obtained using octave-scaled second-order elliptical filters. The frequency domain is first divided into ten octave-scaled sub-bands (0-50 Hz, 50 Hz-100 Hz, 100 Hz – 200 Hz, 200 Hz-400 Hz, 400-800 Hz, 800 Hz – 1600 Hz, 1600 Hz-3200 Hz, 3200 Hz-6400 Hz, 6400 Hz – 13200 Hz, 13200 Hz - 22050 Hz). The mean value of spectral flux of the output of each filterbank constitutes the feature vector. |