Perceptual dimensions of short audio clips and corresponding timbre features

Jason Jiří Musil, Budr Elnusairi, and Daniel Müllensiefen

Goldsmiths, University of London
j.musil@gold.ac.uk

Abstract. This study applied a multi-dimensional scaling approach to isolating a number of perceptual dimensions from a dataset of human similarity judgements for 800ms excerpts of recorded popular music. These dimensions were mapped onto the 12 timbral coefficients from the Echo Nest's Analyzer. Two dimensions were identified by distinct coefficients, however a third dimension could not be mapped and may represent a musical feature other than timbre. Implications are discussed within the context of existing research into human musical cognition. Suggestions for further research are given, which may help to establish whether surface features are processed using a common feature set (as in many music information retrieval systems), or whether individuals use features idiosyncratically to quickly process surface features of music.

Keywords: Timbre perception, short audio clips, similarity perception, sorting paradigm, MDS

1 Introduction

Many application systems in music information retrieval rely on some kind of timbre representation of music [1, 2]. Timbre, or the surface quality of sound, seems to be a core aspect of computational systems which compare, classify, organise, search, and retrieve music. This dominance of timbre and sound representations in modern user-targeted audio application systems might be partially explained by the importance of the perceptual qualities of sound in popular music; writing about pop music in 1987, sociomusicologist Simon Frith already noted that "The interest today (...) is in constantly dealing with new textures" [3]. Whilst musical textures can contain a lot of musical structure, they also depend on surface features separate from any musical syntax or structure, such as the harmonicity of sound, the timbral and acoustical qualities of instruments and spaces, and recording or post-production methods. The precision with which many features of sound can be defined and implemented through modern signal processing has surely also contributed to their popularity in the information retrieval community. Acoustic and timbral features have been defined as part of the MPEG4 and MPEG7 standards and are easily implemented where not already available from one of many software libraries.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

Timbral features are popular in research and commercial music retrieval applications, yet there is surprisingly little rigorous research into perceptual principles explaining how certain timbral features can deliver results which are largely compatible with human music processing. Psychological and perceptual discourse around auditory processing often seems to be out of touch with parts of the audio engineering community. For example, an oft-cited validation of mel-frequency cepstral coefficients (MFCCs) as corresponding to human perceptual processing of sound is a brief engineering paper, rather than a psychological or psycho-acoustical study [4]. Conversely, some studies of human timbre perception (e.g. [5]) may have been unfairly overlooked by the psychological music research community due to their use of 'artificial' stimuli. Also, psychological studies of musical timbre have traditionally focused on the acoustics of musical instruments, or timbral qualities imparted by individual performers (e.g. vibrato, alteration of instrumental attack and decay). These are often studied in isolation and usually with reference to styles of Western art music (e.g. [6]; see [7] for an overview). Thus there is something of a discrepancy between the scope of psychological inquiries and the broader, data-driven goals of music information retrieval (MIR) as applied to finished recordings of popular music. This may exacerbate the relative ignorance between both fields.

The current study aims to bridge this gap to some extent by presenting data from a psychological experiment on human perception of timbral similarity, using short excerpts of Western commercial pop music as stimuli. In addition, this study also tries to identify the perceptual dimensions that Western listeners use when making similarity judgements based on timbre cues and to relate these to a set of timbral features that are well known to both music information researchers and software engineers: the 12 timbre feature coefficients provided through the Echo Nest Analyzer API[1]. As these involve considerable auditory modeling and dimensional reduction motivated to approximate human perception [8], we assume that the human and machine feature extractors under comparison are at least notionally parallel processes.

In this study, participants listen to very short excerpts of recorded commercial popular music and sort them into homogeneous groups. The paradigm is inspired by recent studies on genre [9] and song identification [10], which demonstrated that listeners are able to perform highly demanding tasks on the basis of musical information that is present in sub-second audio clips. Gjerdingen and Perrott found that 44% of participants' genre classifications of 250ms excerpts of commercially available music agreed with classifications they made of the same extracts when they were played for 3 seconds [9]. Krumhansl found that listeners could even identify the artists and titles of 25% of a series of 400ms clips of popular music spanning four decades [10]. At this timescale there are few, if any discernible melodic, rhythmic, harmonic or metric relationships to base judgements on. When musical-structural information is minimal, timbral information can be high; task performance also increased monotonically with longer exposures in both of the aforementioned studies.

[1] `http://developer.echonest.com/`

Many kinds of timbral information can be extracted from musical excerpts. The presence of typical instrumental sounds can undoubtedly help to identify a particular genre [9] and perception of key spectral and dynamic features is robust even for incomplete instrumental tones [11]. However, if timbre is defined more broadly as the spectro-temporal quality of sound, many surface features of polyphonic music could potentially be seen as coefficients in a timbre space. Indeed, the expression of musical emotion can be ascertained from 250ms of exposure, and familiarity with a piece from 500ms [12]. Spectral coefficients also join metric cues as predictors of surface judgements of musical complexity [13]. Different recording and production techniques can give rise to a plethora of perceptual timbral dimensions [14, 15].

In this study, in order to establish how non-expert listeners make use of musical surface features in a similarity sorting task we first apply multi-dimensional scaling (MDS) to extract a small number of perceptual dimensions and then relate these to coefficients in a timbre space. The timbral coefficients returned by the Echo Nest's Analyze service were chosen as the initial pool, as they have been usefully applied in a number of real-world applications. This research paradigm was established by classic studies on timbral perceptual dimensions for instrumental tones [16, 17], and is sensitive to subtle processing differences not picked up by traditional discrimination paradigms [18].

2 Method

131 participants (59 male, with a mean age of 30.8, *SD*=11.8) sorted 16 randomly ordered excerpt test-items into four equally sized bins. Sorts were unconstrained (other than the need for solutions to have exactly four items per bin) and participants could audition items at will. The set contained four each of jazz, rock, pop and hip-hop items, taken from songs identified on the `http://www.allmusic.com` website as being genre-typical but not universally known (i.e. through not having achieved the highest pop chart ratings). Genres were chosen on the basis of Rentfrow and Gosling's high-level categories of musical genre: reflective/complex (jazz), energetic/rhythmic (hip-hop), upbeat/conventional (pop), and intense/aggressive (rock) [19]. Genre-category ratings for these are stable over time and appear to correlate somewhat with stable personality traits [20]. Participants could thus solve the task implicitly (by perceived similarity) even if they possessed no genre-specific knowledge. By focusing on these categories, we also avoided the inherent instability and fluidity of industry genre boundaries. Gjerdingen and Perrott also found that the presence of vocals in extracts reduced genre rating performance [9]. Although vocal features are important for recognising musical styles (and this is reflected in the technologies used in MIR) we chose stimuli without vocals to avoid making the already short excerpts too difficult to classify. Excerpts were representative of the typical instrumentation of the song. Several sets were tested, however results from only one of the 800ms item-sets are analysed here, following piloting which suggested this set to

have desirable psychometric properties[2]. Vectors of timbral features for the same items were extracted through the Echo Nest's Analyzer and used as predictors of item-placement on these dimensions.

3 Analyses and Results

Each possible pair of clips received a score based on the number of participants assigning both clips in the pair to the same group. The resulting distance matrix was taken as an input to the non-metric multi-dimensional scaling procedure as implemented in the R-function `isoMDS` (from package `MASS`). Computing a 2- as well as a 3-dimensional solution we obtained stress values of 12.05 and 6.52 respectively, indicating a much better fit of the 3-dimensional solution to the data, with the 3-dimensional solution also satisfying the elbow criterion in a stress plot (not reproduced here). As a rule of thumb, Kruskal considers MDS solutions with a stress of 5 or lower a good fit while solutions with a stress value of 10 are still fair [21]. Thus, it seems that 3 dimensions are sufficient to describe the participants' perceptual judgements. The 3-dimensional solution is shown in *Figure 1.* Clustering of clips by genre in the MDS space is clearly visible.

As a subsequent step we tried to identify the 3 perceptual dimensions identified by MDS with any of the Echo Nest's 12 timbre coefficients. The Echo Nest Analyzer divides audio into segments with stable tonal content, i.e. roughly per note or chord. For each audio clip we obtained 2 to 5 segments with 12 timbre coefficients each. In order to obtain a homogeneous set of timbral features to compare to the 3 MDS dimensions we used a simple first-order linear model of the time series values of each coefficient for each clip. From each linear model we used the intercept (mean value) and the variance across the number of segments as an indicator of variability of the coefficient in the given clip. In addition, we used the number of segments per coefficient and clip as another indicator of tonal variability.

The pair-wise distributions and correlations between each MDS-dimension and the means and variances of the 12 coefficients indicated that the relationships between the perceptual dimensions and the timbral coefficients are mainly non-linear and distributions are far from normal. We therefore chose a random forest as an analysis technique, as it is able to model non-linear relationships and can additionally deal with a relatively high number of predictors (means and variances for each of the 12 coefficients plus the number of segments resulted in 25 predictor variables) compared to the low number of observations (16 audio clips; for a discussion of random forests as a classification and regression technique see chapter 15 in [22]). More specifically, we chose the conditional random forest model as implemented in the `R` package `party` [23], which is assumed to deliver more reliable estimates of variable importance when predictors are highly correlated and represent different measurement levels [24].

[2] A floor effect for 400ms stimuli was significantly less pronounced for 800ms stimuli in a pilot dataset with 117 participants (800ms per-item successful pairs out of a maximum of 3: M=1.22, SD=0.44; 400ms: M=1.05, SD=0.37; $t_{(31)}$=4.87, p<.001).

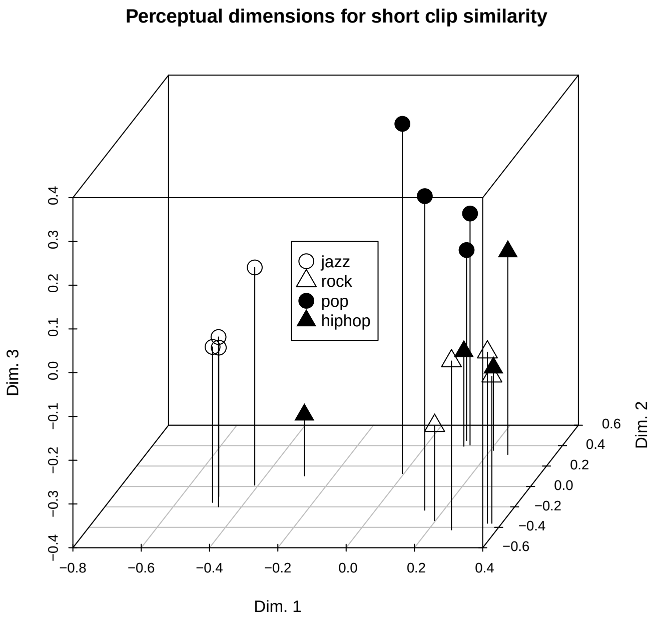

Fig. 1. The 3-dimensional solution of pairwise item distances. Points are differentiated by genre.

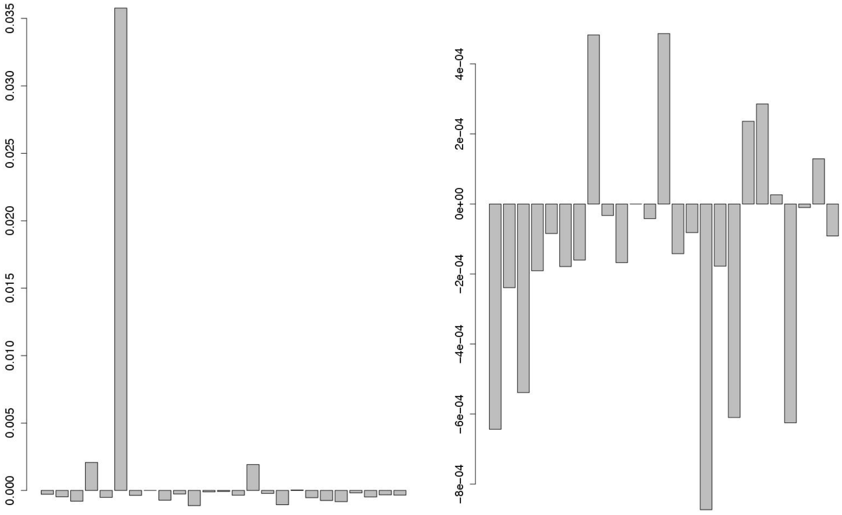

Fig. 2. Predictor importance for perceptual similarity dimensions 1 (*left*) and 3 (*right*). The tall bar for dimension 1 is the intercept of timbre coefficient 5. Note that the plots do not share a common y-axis.

Fitting a random forest model yielded a list of variable-importance values based on the usefulness of individual predictors for accurately predicting the so-called 'out-of-the-bag' (i.e. cross-validation) sample. The intercept (i.e. the mean) of the Echo Nest's timbral coefficient 5 was found to be of high importance as a predictor of perceptual dimension 1. A similarly clear picture was found for the intercept of coefficient 9, being highly important as a predictor of perceptual dimension 2. However, the picture was less clear for perceptual dimension 3, where all importance values for all variables remained within the margin of error around 0, indicating that perceptual dimension 3 cannot be closely associated with any (studied) timbral coefficient. Importance values of variables based on timbre coefficients are given in *Figure 2* for dimensions 1 and 3 for comparison.

4 Discussion

Three perceptual dimensions explained listeners' similarity judgements of short musical clips. Two of these dimensions were predicted by distinct surface features. Mean values but not variances of coefficients were selected as important predictors, which is interesting because the excerpts were long enough to contain some note- and beat-like temporal variations. Unfortunately, only a few timbral features returned by the Echo Nest are publicly documented, so it is difficult to say what these correspond to. A scale-less spectrogram in the existing documentation[3] suggests that coefficient 5, which predicted the coordinates of the 16 clips in perceptual dimension 1, might be a kind of mid-range filter. This would not be surprising, as spectral and dynamic effects are used to add low-end power and high-end presence to recordings. This could reduce the amount of useful information contained in those frequency bands, whilst the mid-range could become the most informative for clip discrimination and classification. Indeed, the most distant cluster on this dimension was jazz, which tends towards conservative mastering and emphasises distinctive instrumental timbres.

The distribution of clips along perceptual dimension 2, as well as incomplete information from the Echo Nest documentation for coefficient 9, suggested that this dimension may represent a similar filtering function to coefficient 5, albeit shifted higher or polarised more to high and low frequency bands. Despite this evidence for possible commonality between the human and machine feature extractors under study, dimension 3 is not predicted by any of the 12 Echo Nest timbral coefficients. At 800ms, the stimuli we used contain rudimentary information about tempo, chord changes, and rhythm. It is possible that dimension 3 represents the influence of such abstracted structures. The results obtained from studies with shorter stimuli might not show these perceptual dimensions, or may indicate reliance on more than these timbral features if they were masked by the availability of musical structure information in the current stimuli. Additionally, the discrete sorting groups could invite top-down strategies based on retrieving explicit genre information from memory, and open subjective

[3] see `http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf`

experience responses will be taken in future studies to establish whether such information is cued by the clips. Nevertheless, the task is known to yield useful similarity data in a shorter and more easily administered experiment than would be possible with the more conventional pairwise similarity rating paradigm [25].

Scheirer and colleagues proposed that listeners may differ in the weight they give to a common set of perceived sound features when judging surface musical sound, or that different listeners may choose different features altogether [13]. Although they lacked enough data to explore these hypotheses, they were able to conclude that individual (participant) models explained complexity rating data better than a common model. Therefore, whilst we found some evidence of common feature-based perceptual dimensions, it is possible that further study with this paradigm will uncover individual strategy differences for this task. The INDSCAL variant of MDS may be helpful in exploring this hypothesis. The reverse is also possible, given that we used far shorter stimuli (800ms versus Scheirer et al.'s 5000) and may have measured a more constrained phenomenon. Individual differences are nonetheless plausible, as task-based measures of timbral perception can be improved by training [26, 27]. Indeed, because timbral perception does not require formalised musical knowledge, individuals could be expected to vary in the information they can access for this task purely on the basis of what they have previously listened to, and to what extent. We will look at three other datasets—including shorter, 400ms clips—and explore other features, for example those provided by Peeters and colleagues' recently published toolbox [28], as well as standard MFCC coefficients and spectral centroid-based measures.

References

1. Aucouturier, J., Pachet, F.: Improving timbre similarity: How high is the sky? In: Journal of Negative Results in Speech and Audio Sciences, vol. 1, pp. 1–13 (2004)
2. Pachet, F., Roy, P.: Exploring billions of audio features. In: Proceedings of CBMI 07, Eurasip, ed., pp. 227–235, Bordeaux, France (2007)
3. Frith, C., Horne, H.: Art into Pop. Methuen Young Books, London (1987)
4. Logan, B.: Mel frequency cepstral coefficients for music modeling. In: International Symposium on Music Information Retrieval, vol. 28, pp. 5–11 (2000)
5. Terasawa, H., Slaney, M., Berger, J.: A statistical model of timbre perception. In: ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA2006), pp. 18–23, Pittsburgh (2006)
6. Barthet, M., Depalle, P., Kronland-Martinet, R., Ystad, S.: Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance. In: Music Perception, vol. 28, pp. 265–278 (2011)
7. McAdams, S., Giordano, B.L.: The perception of musical timbre. In: The Oxford Handbook of Music Psychology, S. Hallam, I. Cross, M. Thaut, eds., pp. 72–80, Oxford University Press (2009)
8. Jehan, T.: Creating music by listening. Ph.D. thesis, Massachusetts Institute of Technology (2005)
9. Gjerdingen, R.O., Perrott, D.: Scanning the dial: The rapid recognition of music genres. In: Journal of New Music Research, vol. 37, pp. 93–100 (2008)

10. Krumhansl, C.L.: Plink: "Thin slices" of music. In: Music Perception: An Interdisciplinary Journal, vol. 27, pp. 337–354 (2010)
11. Iverson, P., Krumhansl, C.L.: Isolating the dynamic attributes of musical timbre. In: The Journal of the Acoustical Society of America, vol. 94, pp. 2595–2606 (1993)
12. Filipic, S., Tillmann, B., Bigand, E.: Judging familiarity and emotion from very brief musical excerpts. In: Psychonomic Bulletin & Review, vol. 17, pp. 335–341 (2010)
13. Scheirer, E.D., Watson, R.B., Vercoe, B.L.: On the perceived complexity of short musical segments. In: Proceedings of the 2000 International Conference on Music Perception and Cognition, Citeseer (2000)
14. Karadogan, C.: A Comparison of Kanun Recording Techniques as They Relate to Turkish Makam Music Perception. In: Proceedings of the 130th Audio Engineering Society Convention, Audio Engineering Society (2011)
15. Marui, A., Martens, W.L.: Timbre of nonlinear distortion effects: Perceptual attributes beyond sharpness. In: Proceedings of the Conference on Interdisciplinary Musicology (2005)
16. Wedin, L., Goude, G.: Dimension analysis of the perception of instrumental timbre. In: Scandinavian Journal of Psychology, vol. 13, pp. 228–240 (1972)
17. Grey, J.M.: Timbre discrimination in musical patterns. In: The Journal of the Acoustical Society of America, vol. 64, pp. 467–478 (1978)
18. Samson, S., Zatorre, R.J., Ramsay, J.O.: Deficits of musical timbre perception after unilateral temporal-lobe lesion revealed with multidimensional scaling. In: Brain, vol. 125, pp. 511–522 (2002)
19. Rentfrow, P.J., Gosling, S.D.: The do re mi's of everyday life: The structure and personality correlates of music preferences. In: Journal of Personality and Social Psychology, vol. 84, pp. 1236–1256 (2003)
20. Rentfrow, P.J., Gosling, S.D.: Message in a Ballad. In: Psychological Science, vol. 17, pp. 236–242 (2006)
21. Kruskal, J.: Nonmetric multidimensional scaling: A numerical method. In: Psychometrika, vol. 29, pp. 115–129 (1964)
22. Hastie, T., Tibshirani, R., Friedman, J.: Random Forests. In: The Elements of Statistical Learning, Springer Series in Statistics, pp. 1–18, Springer New York (2009)
23. Hothorn, T., Hornik, K., Zeileis, A.: Model-based recursive partitioning. In: Journal of Computational and Graphical Statistics, vol. 17, pp. 492–514 (2008)
24. Strobl, C., Boulesteiz, A., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for Random Forests. In: Bioinformatics, vol. 9, pp. 307–327 (2008)
25. Müllensiefen, D., Gingras, B., Stewart, L., Musil, J.J.: Goldsmiths Musical Sophistication Index (Gold-MSI) v0.9: Technical Report and Documentation Revision 0.2. Tech. rep., Goldsmiths, University of London, London (2012), URL `http://www.gold.ac.uk/music-mind-brain/gold-msi`
26. Shahin, A.J., Roberts, L.E., Chau, W., Trainor, L.J., Miller, L.M.: Music training leads to the development of timbre-specific gamma band activity. In: NeuroImage, vol. 41, pp. 113–122 (2008)
27. Gfeller, K., Witt, S., Adamek, M., Mehr, M., Rogers, J., Stordahl, J., Ringgenberg, S.: Effects of training on timbre recognition and appraisal by postlingually deafened cochlear implant recipients. In: Journal of the American Academy of Audiology, vol. 13, pp. 132–145 (2002)
28. Peeters, G., Giordano, B.L., Susini, P., Misdariis, N., McAdams, S.: The Timbre Toolbox: Extracting audio descriptors from musical signals. In: Journal of the Acoustical Society of America, vol. 130, pp. 2902–2915 (2011)