A Comparison of Approaches to Timbre Descriptors in Music Information Retrieval and Music Psychology

Kai Siedenburg*, Ichiro Fujinaga, Stephen McAdams
Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT),
Schulich School of Music, McGill University
555 Sherbrooke Street West
Montreal, QC, Canada
email: kai.siedenburg@mail.mcgill.ca, ich@music.mcgill.ca, smc@music.mcgill.ca
*corresponding author

Acknowledgements: This work was supported by a grant from the Canadian Natural Sciences and Engineering Research Council (RGPIN 312774) and a Canada Research Chair (950-223484) awarded to Stephen McAdams; a Harman Scholarship from the Audio Engineering Society's Educational Foundation and a Québec International Merit Scholarship to Kai Siedenburg. The authors wish to thank the two anonymous reviewers for valuable comments on the manuscript.

Abstract

A curious divide characterizes the usage of audio descriptors for timbre research in music information research (MIR) and music psychology. While MIR uses a multitude of audio descriptors for tasks such as automatic instrument classification, only a highly constrained set is used to describe the physical correlates of timbre perception in parts of music psychology. We argue that this gap is not coincidental and results from the differences in the two fields' methodologies, their epistemic groundwork, and research goals. This paper lays out perspectives on the emergence of the divide and reviews studies in both fields with regards to divergences in research methods and goals. We discuss new representations for spectro-temporal modulations in MIR and psychology, and compare approaches to spectral envelope description in depth. Finally, we will propose that the interdisciplinary discourse on the computational modelling of music requires negotiations about the roles of scientific evaluation criteria.

Keywords: audio analysis, information retrieval, timbre perception, instrument classification, evaluation

A Comparison of Approaches to Timbre Descriptors in Music Information Retrieval and Music Psychology

1 Introduction

'Mel Cepstrum: *You're killing me. Are you seriously rejecting 10 years' worth of results as mere coincidences? Our findings that, say, taking the derivative of MFCCs improve genre classification by 10%, or that periodicities in the range 1–10 seconds (the rhythm fluctuation patterns [...]) are enough to account for timbre similarity, shouldn't that, at least, give you some sort of intuition about how these behaviours are cognitively produced?* Ann Ova: *In the MIR bestiary, we find, first, features deriving from traditional psychoacoustics [...], then, your field offers quite a lot of mathematical variants of these same characteristics, [...] which seem to be justified only by the fact that they are conceptually close [...] or even that they are easy enough to compute [...]; other features seem to start their career as intermediary steps in the processing chain of another feature, gain special status and then a name of their own [...]. And the list goes on, growing every year: the sole MIRtoolbox library offers more than 300 features, very few of which having a clear epistemological status. Now, I do not doubt they serve your purpose well, but I hope you see it is unclear whether they can serve ours.'* (Aucouturier & Bigand, 2012, p. 2-3).

In a fictive dialogue entitled 'Mel Cepstrum & Ann Ova: The Difficult Dialog Between MIR and Music Cognition' (Aucouturier & Bigand, 2012) and a subsequent journal publication (Aucouturier & Bigand, 2013), computer scientist Jean-Julien Aucouturier and psychologist Emmanuel Bigand identified a set of problems that hinders the academic disciplines of music information retrieval (MIR) and music cognition from productive collaboration. Today, the rhetorics of interdisciplinary collaboration have long become fashionable, yet still hard to set into practice. Position papers such as these are crucial because they clarify potential misunderstandings between researchers coming from different backgrounds and eventually guide collaborative research towards a fertile future.

Aucouturier and Bigand cover a variety of important aspects, including the usage of descriptors and algorithms and methods for scientific validation, as well as the status of limit cases in both fields. Their most fundamental claim is that parts of MIR are methodologically misguided because they do not respect the ways in which auditory information processing takes place in humans, but rather employ heuristics that have proven to be successful when evaluated algorithmically over large corpora of music. In the case of successful evaluation, these heuristics are then falsely interpreted as evidence of mechanisms of human auditory information processing. In their critique of facets of the MIR approach, Aucouturier and Bigand are not alone (cf., Wiggins, 2009; Marsden, 2012; Sturm, 2013, 2014). Bob Sturm, for instance, argued that relying on classification accuracy alone can be a misleading criterion in tasks such as musical genre classification, in some cases giving rise to operationalizations of musical genre that are implausible to any human listener.

In this review, we attempt to analyse the first aspect discussed by Aucouturier and Bigand (2012), namely the apparent divergence in the usage of audio descriptors in MIR and music psychology that quantify musical timbre. Timbre denotes the bundle of auditory attributes that endows musical sounds with their particular 'colour', 'shape' or 'texture' (which may covary with pitch, loudness, and duration) and that enables listeners to identify sound sources (McAdams, 2013). It thus comprises at least two partially separate perceptual facets: sound quality (e.g., 'colour') and sound source identity. Whereas the former can only be studied with subjective experimental tasks (such as dissimilarity rating), there is an objective 'ground truth' to the latter, in that the inference of a sounds' source-cause may be correct or incorrect. This distinction bears importance, because we will argue in Section 3 that one of the ways in which MIR and psychology diverge in their usage of descriptors may be related to exactly such task differences. The basic question underlying this paper then becomes 'How do MIR and psychology approach the quantitative modelling of musical timbre?' More particularly along the lines of Aucouturier

and Bigand (2012), 'Why does MIR use a multitude of timbre descriptors (≥ 20) when psychological research usually only identifies a few (≤ 5) to be relevant for the perception of a given set of timbres?'

For the purpose of disambiguation, we here we use the notion of audio *descriptors* to refer to those (usually continuously-valued) measures of audio signals that are more commonly referred to as audio *features* in MIR. In contrast, features in psychology generally denote stimulus characteristics, or, in a stricter sense, binary-valued properties that stimuli either do or do not possess. Although this usage may be unusual for some readers, distinguishing the stimulus (and its features) from the measurement (by means of descriptors) will become useful.

In what follows, we will portray the commonalities and differences of the two fields' approaches towards timbre description in terms of their scientific techniques, underlying experimental tasks, epistemic frameworks, and evaluation criteria. We begin by reviewing the techniques used for instrument classification in MIR and timbre similarity perception in psychology. The role of spectrotemporal fluctuations will be discussed in detail, as well as questions around spectral envelope description. Section 3 analyses implications of the prevalent task differences in the two fields, before we propose in Section 4 that the two fields do not necessarily share the same scientific objectives and epistemic framework. Along the way, we will discuss the ways in which Aucouturier and Bigand's critique may have neglected parts of the epistemic and methodological intricacies inherent to studies that appear under the umbrella of psychology or computational neuroscience. We will highlight the fact that contemporary models of complex auditory cognition, such as timbre perception, remain coarse approximations of the underlying psychological processes; such models can be evaluated according to multiple evaluation criteria that including goodness of fit, simplicity, physiological adequacy, or computational parsimony. Our point of arrival is that the computational modelling of music requires researchers—and in particular those involved in interdisciplinary research—to more explicitly negotiate the criteria according to

which scientific success should be evaluated.

2 Techniques and Results

This section reviews the scientific techniques that are used in MIR and music psychology for dealing with the parameter of timbre. For MIR, research on timbre mostly revolves around instrument classification, where computational systems represent audio signals via an ensemble of descriptors and use this representation to assign class (instrument category) memberships to instrument sounds. Music psychology has most often studied timbre perception by relying on dissimilarity ratings, that are correlated with descriptors in order to reveal the most salient physical parameters underlying subjective timbre perception. The two approaches have indeed led to substantial disagreement on the most suited set of audio descriptors.

2.1 Instrument Classification in MIR

For most tasks of audio-based MIR and instrument classification[1] in particular, algorithm design consists of two parts. The first concerns the representation of audio signals for which audio descriptors are chosen or are newly created, usually by mapping the signal's short-term Fourier transform (STFT) magnitude into a lower-dimensional domain that more clearly reveals the relevant signal characteristics. Oftentimes, this dimension-reduction phase is followed by a feature selection step that removes the least important descriptors from the model in order to reduce computational load, redundancy, and tendencies for overfitting. The second stage concerns the selection of the classification

[1]We use the term *classification* for what is often also called *recognition*, because the aim of the discussed tasks is to attach a class label (e.g., 'string instrument') to a given sound, and not to assess whether a sound has been encountered in the past. The latter is the central meaning of recognition in the psychology literature and does not require classification or identification (cf., McAdams, 1993; Berry, Shanks, Speekenbrink, & Henson, 2012). Accordingly, recognition is a necessary condition of classification, but not the other way around.

model. Popular supervised classifiers include k-nearest neighbours (k-NN), Gaussian mixture models (GMM), hidden Markov models (HMM), support vector machines (SVM) or neural networks (cf. Herrera-Boyer, Klapuri, & Davy, 2006; Fu, Lu, Ting, & Zhang, 2011). The resulting systems are then trained and evaluated, usually by employing cross-validation with regards to a 'ground-truth' defined a priori on annotated sets of audio data. Cross-validation denotes the partitioning of available data into training and test sets, which is repeated multiple times for different partitions ('k-fold'). The researcher then selects the model with the lowest average error on the test sets. This approach reduces the danger of classical overfitting, that is, the incorporation of too many predictors, which fit to the noise in the training data (alternative forms of overfitting may still exist, cf., Ng, 1997).

Instrument classification has been an important task since the beginnings of MIR. Fujinaga (1998) and Fujinaga and MacMillan (2000) classified steady-state portions of musical instrument tones in an exemplar-learning-based approach. Their system relied on spectral descriptors such as higher-order moments and amplitudes of spectral peaks. Classification was realized using the non-parametric k-NN scheme, which assigns an item to a class based on the majority vote of the item's k nearest neighbours. Classification performance was around 50% with more than 60 spectral descriptors and dropped to 42% using only four descriptors (fundamental frequency and the first three spectral moments). Impulsively excited sounds were classified considerably worse, which seems natural because only manually selected steady-state portions of the sounds were considered in the analysis.

Martin (1999) modelled perceptual sound source recognition inspired by the hierarchy of perceptual processing proposed by McAdams (1993). The idea was to model source recognition as a process that incrementally accumulates information at multiple, increasingly fine-grained levels of abstraction. Using a three-dimensional audio representation based on auditory-filterbank autocorrelation (Ellis, 1999), a large number of descriptors were computed relating to spectral, attack, pitch, vibrato, and tremolo characteristics. Instrument prototypes were accumulated over several instances of the same

instrument. The classification scheme comprised a hierarchical Bayesian decision tree with three levels: all instruments, instrumental families, and specific instruments. Classification was realized via a log-likelihood decision that ruled out alternative categories at every level of the tree. Interestingly, a context-dependent descriptor weighting was implemented, selecting those descriptors that best separated the remaining categories at each level. With around 75% classification accuracy for instruments, the system performed better than human subjects for specific instruments, whereas humans performed better at the family level.

Eronen (2001) used the nowadays ubiquitous *mel-frequency cepstral coefficients* (MFCC), which originated from speech processing. They are obtained by computing the logarithm of the power of a Mel-scale-warped STFT before applying a discrete cosine transform (DCT), which yields a representation of the spectral shape (i.e., lower order coefficients represent coarse spectral variability, higher order coefficients represent increasingly finer spectral detail). For capturing spectral envelope information (and not pitch), only the first few (e.g., 13) coefficients are used. As the DCT has a de-correlating effect, it also helps to remove the redundancy that plagues the first moments of the spectral envelope, thus facilitating classification.

Using a collection of 160 descriptors including MFCCs and newly proposed octave-band signal intensities and octave-band signal-intensity ratios, Essid, Richard, and David (2006) evaluated different descriptor-selection and classification strategies for solo-instrument signals of 0.5 s duration. Their first experiment compared feature-selection strategies based on so-called 'genetic algorithms' and 'inertia ratio maximization'. Genetic algorithms implement a randomly initialized, iterative search process in which subsets of features are encoded as 'chromosomes' and evaluated according to their 'fitness' (Siedlecki & Sklansky, 1989). The fitness function was implemented by Essid et al. (2006) as the mutual separability of class probability densities for a given chromosome. Even higher classification accuracy was achieved by using intertia ratio maximation (Peeters, 2003),

which iteratively adds descriptors with a maximal ratio of between-class variance to the overall variance (Fisher discriminant), followed by a descriptor orthogonalization step for the removal of descriptor redundancies. A second experiment compared different kernel functions for an SVM classifier (which, in the case of two-classes and a linear kernel, yields a hyperplane that maximally separates the feature values of both classes' items as a decision boundary). Using a radial basis function kernel instead (thus non-linearly transforming the descriptor-space) improved classification accuracy from 81% to 87%. The last experiment tested the influence of the signal duration on accuracy where the increase from 0.5 to 3 s yielded an increase to 93% classification accuracy.

Joder, Essid, and Richard (2009) considered temporal integration in the descriptor and classification stage. Temporal integration involves the combination of descriptor observations over successive time frames. The list of low-level descriptors that was suggested by descriptor selection included the first three spectral moments, two sets of 13 MFCCs with 11 or 30 Mel subbands, 6 octave-band spectral intensities and 5 ratios of such coefficients, 5 wavelet transform coefficients, 3 spectral irregularity descriptors, spectral roll-off, spectral flatness, 2 zero-crossing rates over different time windows, and amplitude modulation strength between 10 Hz and 40Hz. These descriptors underwent early and late temporal integration. In early integration, new descriptor vectors are computed that characterize the signal at a higher time scale by summing local descriptors extracted from a sequence of analysis frames. Late integration does not attempt to extract descriptor dynamics, but either combines successive primary decisions of the classifier or uses a classifier that can deal with sequences. Suited classifiers include HMMs (which track transition probabilities between a system's states) or SVMs with alignment kernels (which dynamically align sequences). These authors showed that including both early and late temporal integration yielded small improvements of classification accuracy compared to static reference systems, i.e., GMM and SVM (with Gaussian kernels) in conjunction with non-integrated features.

An important insight on the interrelation of descriptors was provided by Peeters, Giordano, Susini, Misdariis, and McAdams (2011). Based on an analysis of a large corpus (> 6000) of instrument sounds, they found that their initial 164 audio descriptors from the Timbre Toolbox fell into 10 fairly independent classes, but descriptors within each class were highly collinear. Classes included the same descriptors computed on the basis of different time-frequency representations (e.g., linear STFT vs. Gammatone-filterbank), but also different types of descriptors (such as the ratio of levels of od and even harmonics, noisiness, inharmonicity) regardless of their initial audio representation. This implies that in many situations there may indeed be a much smaller number of substantially independent variables present than raw numbers of descriptors suggest.

In summary, instrument classification has reached an impressive accuracy during a relatively short period of research. However, the psychoacoustic meaning of many of the descriptors used in these systems, as exemplified by the aforementioned list that is representative of many studies in the field, is hard to decipher and expresses a considerable level of eclecticism. Zero-crossing rates per se (a waveform's number of sign-shifts in a given duration), for instance, do not relate to perceptual processing in a straight-forward fashion, but they somehow contribute to computational classification accuracy.

An alternative example of a high-dimensional representation derived from a formal standpoint (thus perhaps less idiosyncratic than some of the collections of descriptors mentioned above) was presented by Mallat (2012). His nonlinear *scattering transform*, if applied to sound, provides a mathematical representation tailored towards spectrotemporal modulation analysis. The formal requirement imposed on the representation is invariance to operations such as small time-shifts and log-frequency shifts, a useful property (shared by MFCCs) for classification tasks, and it is also of potential perceptual relevance. The resulting transform iteratively decomposes a signal into layers of coefficients by cascading wavelet transforms on the low-pass filtered modulus (i.e., envelope) of the previous layer. Its first layer of coefficients encodes a signal's frequency content, whereas the second and

further layers mainly capture the temporal evolution. The approach thereby generalizes MFCCs in that the first layer of scattering coefficients is comparable to MFCCs, but the further layers yield temporal details that are not considered in MFCCs. Andén and Mallat (2014) showed that using this representation improves genre classification over a set of MFCCs and their first order differences (deltas). Note that in contrast to the aforementioned audio analysis features, this method can also be used for signal reconstruction. Bruna and Mallat (2013) applied the method to an analysis-resynthesis of environmental sound textures and found that the scattering transform requires around half the number of coefficients compared to the results by McDermott and Simoncelli (2011) who had used a modulation representation similar to the one proposed by Dau, Kollmeier, and Kohlrausch (1997b). Although the scattering representation has not been used for instrument recognition specifically, it seems to be well suited for this task.

Similar examples of high-dimensional signal representations that encode spectrotemporal modulations have been used for tasks such as genre classification (Sturm & Noorzad, 2012). Interestingly, the authors also observed that representations that mimicked basic aspects of cochlear processing did not necessarily improve classification results, potentially due to the nature of the genre classification task. The importance of considering perceptual and computational tasks in detail will be further discussed in Section 4.

2.2 Timbre Similarity in Music Psychology

Timbre similarity and multidimensional scaling. A shared point of departure for most psychological studies of timbre that make use of audio descriptors is to circumvent verbal description and semantics by probing timbral *similarity*, assumed not to require language. Note that this approach probes the qualitative facet of timbre; audio features for perceptual instrument identification, on the contrary, are only beginning to be studied empirically (see e.g., Agus, Suied, Thorpe, & Pressnitzer, 2012; Patil, Pressnitzer,

Shamma, & Elhilali, 2012). Multidimensional scaling (MDS) (R. Shepard, 1962; Kruskal, 1964) has been a pivotal tool for the study of timbre dissimilarity. MDS generates a spatial configuration of points whose pairwise distances approximate the original perceptual dissimilarity data. An important variant of the algorithm is CLASCAL (Winsberg & De Soete, 1993), which includes latent classes of subjects weighting the obtained dimensions differently, as well as so-called *specificities*, which provide additional distance values to account for perceptual features that are specific to individual items. The specificities take up as much unexplained variance as possible but do not make any assumptions about the relationships among timbres and turn the model indeed into a compromise between strictly spatial models (R. N. Shepard, 1987) of similarity and non-spatial tree-based approaches to similarity (Sattath & Tversky, 1977). It is common practice to select the scaling model that minimizes possibly both (or either of) the Bayesian (BIC) (Schwarz et al., 1978) and Akaike (AIC) (Bozdogan, 1987) information criteria in order to avoid overfitting by adding too many dimensions (Winsberg & De Soete, 1993).

For applying MDS to timbre similarity perception, first conducted by Plomp (1970), interference from other perceptual parameters must be avoided. Stimuli are thus subjectively equalized in pitch, loudness, and duration before participants are asked to judge dissimilarity of subsequently presented pairs of timbres. It is then up to the researcher to search for physical correlates of the obtained spatial dimensions. Due to the constrained duration of experiments and the fact that an increase in the number of items results in a quadratic increase in the number of pairs to be compared, experiments commonly use small sets of timbres, usually in the range of 10–20. A recent exception is Elliott, Hamilton, and Theunissen (2013) who used 42 tones by allocating different sparse subgroups of sounds, tested by different groups of subjects. The full dissimilarity matrix was obtained by averaging across groups. This small number of distinct sounds used in perceptual dissimilarity studies therefore stands in sharp contrast to the hundreds of samples used in MIR studies.

Acoustic Interpretation of Timbre Spaces. In his seminal work on timbre (Grey, 1975, 1977; Grey & Gordon, 1978), John Grey used emulations of orchestral tones, generated by line-segment-approximated amplitude and frequency trajectories of partials. He settled on a three-dimensional MDS solution (dimensions are referred to as D1/D2/D3 in the following). The physical correlates were interpreted in terms of properties of the spectral energy distribution for D1. D2 was related to the attack synchronicity of partials, but he also noticed that this dimension was related to the amount of spectral fluctuation. D3 was related to spectral balance during the attack portion of tones. Using a set of FM-synthesized sounds from Wessel, Bristow, and Settel (1987, August), Krumhansl (1989) was the first to present a timbre space including specificities using CLASCAL (Winsberg & De Soete, 1993). MDS dimension D1 was interpreted qualitatively as corresponding to rapidity of attack, D2 to centre of gravity of the spectrum, D3 to spectral fluctuations over time. Iverson and Krumhansl (1993) used recorded instrumental sounds and studied the influence of attack portions on similarity judgements. For all three sets of stimuli (full tones, transients only, sustained parts only), similarity judgements correlated with spectral centroid frequency (first moment of the spectral distribution and the centre of gravity of the spectrum, correlates with subjective *brightness*, cf., Schubert & Wolfe, 2006) and amplitude envelope shape.

McAdams, Winsberg, Donnadieu, De Soete, and Krimphoff (1995) synthesized many of the previously mentioned possibilities of MDS, including specificities plus latent classes of subjects using CLASCAL, as well as rigorous quantification of physical correlates of MDS dimensions, and used a subset of 18 tones from Krumhansl (1989). The audio descriptors log-rise time (logarithm of duration from start of tone to amplitude maximum), spectral centroid, spectral flux (average of correlations between adjacent short-time amplitude spectra), and spectral irregularity (log of the standard deviation of component amplitudes of a tone's spectral envelope, derived from a running average of the amplitudes of three adjacent harmonics) were considered for an interpretation of a CLASCAL-based

MDS model. The best model fit was obtained by a six-dimensional solution without specificities that yielded an ambiguous acoustic interpretation, however. The authors thus settled on a three-dimensional solution that was easier to interpret psychophysically. Here, D1 and rise time and D2 and spectral centroid both had correlations of .94. D3 had a correlation of .54 with spectral flux. The three stimuli with highest specificity values were the harpsichord, clarinet and *vibrone* (vibraphone/trombone hybrid). This study confirmed the salience of the spectral centroid and amplitude envelope properties, but it also highlighted the interpretative role of the researcher using MDS. The relevance of spectral flux remained somewhat more vague due to its relatively low correlation with D3, and non-correspondence with earlier findings. For example Krimphoff, McAdams, and Winsberg (1994) found that the third dimension of the Krumhansl (1989) space correlated strongly with spectral deviation (the irregularity or jaggedness of the spectral envelope). Lakatos (2000) confirmed the relevance of spectral centroid and rise time for a large set of recorded timbres comprising both harmonic and nonharmonic percussive timbres, but did not further report investigations of the role of spectral flux and irregularity.

Choosing a somewhat different focus, there is also research that considers the acoustical features underlying timbral differences within instruments, such as the distinct sound qualities of sounds played in different pitch registers (Marozeau, de Cheveigné, McAdams, & Winsberg, 2003), produced with different playing efforts or dynamics (Gadermaier & Reuter, 2014), coming from natural acoustic sources or their synthetic emulations (Kendall, Carterette, & Hajda, 1999), or differences resulting from expressive intent (Barthet, Depalle, Kronland-Martinet, & Ystad, 2010). Zooming into the sound of one instrument, Barthet, Guillemain, Kronland-Martinet, and Ystad (2010) studied perceptual dissimilarity between clarinet tones synthesized from a physical model that varied in bowing pressure and lip pressure on the reed. They obtained a three-dimensional MDS solution with dimensions that correlated with attack time and spectral centroid (D1), the ratio of the energy of the partials no. 2, 3, and 4 compared to the overall harmonic

energy (i.e., the 'tri-stimulus' coefficient no. 2; D2), and the odd-even ratio of harmonics (D3). Barthet, Depalle, et al. (2010) further found systematic effects of expressive clarinet performance on timbre descriptors such as the spectral centroid or the odd-even ratio.

Considering even more homogeneous sets of sounds by holding playing-related aspects constant, there may still remain timbral differences between exemplars of the same instrument or object type, differences that underly the general question of instrument sound quality. Although we do not attempt to review this large field here, suffice it to state that research on this issue has mostly sought to correlate verbal descriptors of quality with timbre descriptors. For instance, C. Fritz, Blackwell, Cross, Woodhouse, and Moore (2012) found correlations between verbal descriptors of violin sounds and sound energy in different octave bands (also cf., Saitis, Scavone, Fritz, & Giordano, 2015; Štěpánek, Syrový, Otčenášek, Taesch, & Angster, 2005; Štěpánek & Otčenášek, 2004).

The prevalent differences in stimulus heterogeneity raise the question whether there is an upper limit in heterogeneity beyond which single audio descriptors may loose their usefulness. One could suppose, for instance, that ratings between very different types of sounds may be driven by cognitive factors instead of low-level acoustic features (Susini, Lemaitre, & McAdams, 2012). Misdariis et al. (2010) therefore proposed a two-layered quantitative description for various sets of environmental sounds. It included a broad categorization step into sounds stemming from similar sound sources, followed by a subordinate model of within-category dissimilarity based on continuous dimensions. Compared to this variety of environmental sounds, however, the subsets of orchestral instrumental timbres used in the dissimilarity studies reviewed above were rather homogeneous. Moreover, similarity studies that used categorically different subsets of sounds, such as harmonic and percussive timbres (Lakatos, 2000) or acoustic and synthetic timbres (Kendall et al., 1999), did not yield a categorical separation of ratings between subsets. Regarding the similarity structures in sets of musical instruments, it is important to note that the particularly salient dimensions, such as attack time and spectral centroid,

traverse instrument categories (McAdams et al., 1995; Kendall et al., 1999; Lakatos, 2000; Elliott et al., 2013). This does not, however, exclude the possibility that there may be acoustic dimensions that only become prevalent for within-category comparisons, and that are therefore not revealed as (global) latent dimensions by the MDS approach.

Spectro-Temporal Cues. Coming back to the parameter of spectral flux, Caclin, McAdams, Smith, and Winsberg (2005) addressed the issue of correlation vs. causation in timbre-space studies. Tones varying along the parameters spectral centroid, rise time and spectral flux were synthesized. The latter was operationalized as variation of spectral centroid within the first 100 ms of the tone. Contrary to spectral deviation, which was confirmed to be perceptually salient in another experiment in that paper, the obtained timbre spaces suggested that spectral flux is unlikely to serve as a salient perceptual dimension of timbre, at least in the parameterization used for the experiment. This parametrization—well suited to describe an instrument's 'brassiness' and coherent with the interpretation of D2 in Grey (1977)—nonetheless diverged from what had been measured as spectral flux in McAdams et al. (1995), namely the average correlation of adjacent short-time spectral magnitudes over the full signal. A parameterization that may better fit this latter measure of flux was used by Golubock and Janata (2013). Here, flux was parameterized as joint AM and FM variation of *individual* partials, thus leading to a sensation of roughness rather than tremolo or vibrato. Measuring discriminability of timbre dimensions, they found that thresholds for this kind of flux were stable over time. Studying various simplifications of time-varying parameters of re-synthesized instrument tones, McAdams, Beauchamp, and Meneguzzi (1999) also observed that spectral flux was among the most salient parameters which allowed listeners to discriminate between the full and the simplified re-synthesis.

This question is closely related to the latest published timbre space study (Elliott et al., 2013), which explicitly used modulation power spectra (MS) to quantify timbre dimensions, potentially suited to better capture spectral fluctuations. They obtained a

five-dimensional MDS space for dissimilarity judgements of 42 timbres. Dimensions were then quantified by projecting the MSs into a 20-dimensional vector space with Principal Components Analysis (PCA; 20 dimensions were optimal in cross-validation), and then using the corresponding scores of these first 20 principle components as independent variables in regularized regression. Explained variance was 0.73 for D1; 0.59 for D2; 0.60 for D4; and 0.10 for D5. D3 did not correlate significantly with any of the given principle components. However, the areas within the MS that corresponded to the correlation with the five MDS dimensions seemed not as straight-forward as for speech signals (Elliott & Theunissen, 2009). For comparison with previous research (which the authors incorrectly characterize as *exclusively* focused on spectral *or* temporal measures), moment statistics and entropies of the spectral and temporal envelopes were computed. Regression of MDS dimensions yielded explained variances that had surprisingly similar magnitudes compared to those reported for the MS approach beforehand. In sum, although certainly a valuable way of representing the information content of audio signals, the novel MS approach did not yield substantial improvements in the fit of acoustic features to MDS dimensions when compared to the classical, audio-descriptor-based approach.

This again highlights the difficulties of developing physical interpretations of MDS spaces for recorded instrumental timbres. Moreover, the MS might not be the most suitable tool for modelling timbre space dimensions from MDS in the first place. MDS assumes the existence of a few latent, perceptually orthogonal dimensions. The MS is high-dimensional and redundant, without having explicit 'regions' that represent spectral centroid or rise time. But in 40 years of MDS of orchestral instrument timbres, the latter two descriptors have usually been represented as two separate MDS dimensions. If it is impossible to construct single audio descriptors accounting for the multifaceted manifestations of spectrotemporal modulations, as is assumed by the MS approach, one alternative would be to perhaps construct a hybrid framework, in which traditional audio descriptors would account for a low-dimensional MDS space, whereas MS could be used to

represent potential specificities in a CLASCAL model.

A more radical alternative would be to discard MDS all together and model dissimilarity judgements directly, a pathway that is taken by Patil et al. (2012). Dissimilarities were modelled by fitting a Gaussian kernel distance on a high-dimensional time×frequency×modulation-rate×modulation-scale representation based on estimates of spectrotemporal receptive fields (STRF) modelled on auditory cortical neurons. They obtained impressive correlations of $r = .94$ with human perceptual judgements after a complex kernel optimization was performed. Using simple Euclidean norms instead (i.e., the equivalent of the simple geometric distance for spaces of arbitrary dimension), correlation with dissimilarity judgements reduced to $r = .61$. At the same time, instrument classification accuracies using SVM were above 95% (however, see also Patil & Elhilali, 2013; Giannoulis et al., 2013, a classification challenge for environmental acoustic scenes and events where the same model did not yield convincing results). This demonstrates that there is enough information in STRFs to accomplish the tasks of timbre classification and dissimilarity prediction.

2.3 The example of spectral envelopes

Spectral envelope descriptors may serve as a particularly good example of the disparities between MIR and perception research. The relative amplitudes of a tone's partials have long been considered as the primary determinants of tone colour (von Helmholtz, 1875). A modern version of this 'spectral view' on timbre is to consider the spectral envelope (i.e., the smoothed spectral distribution) which is to some extent invariant across pitch for many musical instruments (Reuter, 2002; Patterson, Gaudrain, & Walters, 2010; Lembke, 2014). Spectral envelopes are most often described with a set of 13–24 MFCCs in MIR (with increasing number describing increasingly fine-grained spectral detail), but in psychology, often only a small number of descriptors are used (e.g., the spectral centroid and the even-odd ratio). One should note that the centroid is usually

highly correlated with the first MFCC which captures the coarsest portion of spectral variability by definition. For that reason, a core question regarding the value of MFCCs for psychological purposes thus concerns the amount of spectral detail that listeners rely on.

MFCCs were designed for speech classification, where the cepstral analysis is assumed to deconvolve information from vocal source and filter into additive components; it is interesting to note that this assumption would not hold for musical tones which encompass an extended pitch range (Richard, Sundaram, & Narayanan, 2013). A formal interpretation of MFCCs states that they yield invariance to small signal perturbations such as small translations in time or transpositions in frequency (Andén & Mallat, 2014); this is one reason why they are valuable in classification. Yet another view assumes that they model aspects of perceptual processing such as compression and non-linear frequency resolution in the cochlea, because they are derived from the logarithm of the instantaneous energy of a Mel-scale filterbank. Therefore, it is sometimes assumed that *'a small (resp. large) numerical change in the MFCC coefficients corresponds to a small (resp. large) perceptual change'*(Müller, Ellis, Klapuri, & Richard, 2011, pp. 1097–1098).

However, only a few studies have actually tested how well MFCCs are suited for the prediction of perceptual data. Terasawa, Berger, and Makino (2012) had subjects rate the pairwise dissimilarity of synthesized timbres whose MFCCs were precisely varied. The design of their first experiment implied that MFCCs perfectly correlated with spectral centroids, unfortunately prohibiting a comparison of their predictive powers (i.e., prediction from centroids were identical to MFCCs). In Exp. 2, two MFCCs were varied at the same time. Notably, for one out of their five experimental conditions, the centroid predicted perceived dissimilarity significantly better than MFCCs, whereas no other significant differences between descriptors were observed. A parsimonious interpretation of these results would suggest to reject MFCCs as descriptions of spectral envelope perception in dissimilarity judgements. Nevertheless, the opposite direction is preferred by the authors: *'Experiments [...] suggest that an MFCC-based description holds a similar degree*

of linearity in predicting spectral envelope perception to a spectral centroid-based description. Yet the spectral centroid is essentially a single-dimensional descriptor and does not describe the complex shapes of the spectral envelope itself.' (Terasawa et al., 2012, p. 682) Even more surprising, the lack of gain in predictive power of MFCC coefficients compared to the spectral centroid seems to go unnoticed in MIR. Supposedly reflecting the MIR position on the issue, Aucouturier and Bigand (2012) note: *'How about that study by Terasawa, Slaney and their colleagues at Stanford: they resynthesized sounds from MFCCs and showed that human timbre dissimilarity ratings between sounds correlated exactly with the MFCCs. Doesn't that prove something?'*

Horner, Beauchamp, and So (2011) compared different error metrics to predict listeners' spectral discrimination performance. Tones were additively re-synthesized such that the amplitudes of selected time-varying partial trajectories could be altered. Otherwise, the stimuli were matched with regard to subjective loudness and spectral centroid (F0 = 311 Hz). Predicting discrimination data with a fit of $R^2 = .85$ required to include the relative amplitude error of the first five harmonics in the regression model (the full comparison that took into account all 30 partials obtained a fit of $R^2 = .91$). A metric that measured Mel-band errors required ten bands to achieve the same fit of $R^2 = .85$. Similarly, ten MFCC coefficients were required. This suggests that for centroid-matched tones, subjects focus on the first few harmonics in discrimination, and these are only resolved with a higher number of Mel bands or MFCC coefficients. In a similar study, McAdams et al. (1999) had shown that listeners can discriminate well relatively fine-grained modifications of spectral envelope fine structure and spectral flux. Differences of the partials' amplitude envelope values and differences in spectral centroids both predicted well discrimination performance across instruments and the different re-synthesis manipulations (but sounds were not matched in centroid). Leaving the realm of controlled synthesized tones in a study of timbral features in polyphonic Indian popular music, Alluri and Toiviainen (2009), however, found only one MFCC (no. 13) to correlate weakly

($r = -.32$) with one of the extracted semantic factors ('fullness').

Turning towards the centroid, Schubert and Wolfe (2006) addressed the question of whether sound brightness is better predicted by the absolute spectral centroid or the (supposedly pitch invariant) centroid rank, i.e., the centroid divided by the fundamental frequency. The latter predictor, however, failed to correlate significantly with subjective brightness, whereas the regular centroid did (r = .53). Explicitly varying spectral centroids of tones additively re-synthesized from orchestral instrument samples, Wun, Horner, and Wu (2014) found good discrimination performance for centroid deviations from ±8 to ±40%. A second experiment showed that pairs of tones were still identified as originating from the same instruments with centroid changes within ±32%. Further, changes within ±64% were still judged to not alter instrument family identity.

This overview corroborates the idea that the centroid parameterises the most salient perceptual feature of spectral envelopes, and does so in a way that may predict perceptual results somewhat better than MFCCs (Terasawa et al., 2012). At the same time, the review has shown that there are more fine-grained spectral envelope features that are easily discriminable (Horner et al., 2011; McAdams et al., 1999) and thus part of the 'perceptual repertoire'. This lends empirical support to the intuition that there are many ways in which spectral envelopes can vary that cannot be captured by the centroid. As observed by McAdams et al. (1999), however, when multiple acoustic cues allow for discrimination, perceptual performance may rely on the most salient cue alone. Similar processes of perceptual 'feature selection' might be at play in dissimilarity ratings. Conclusively, the reason underlying the centroid's success in perceptual studies may be that in situations with variability in multiple (envelope) features, variability in the coarsest spectral portion (i.e., the centre of gravity that is measured by the centroid and the *first* MFCC) may be the most salient one and therefore may override other cues.

3 Tasks

Overall, the observed discrepancies in techniques and results seem striking. MIR systems have been using more and more descriptors, which has improved classification performance significantly. Yet in music psychology, only two descriptors have proven to be robust enough to reappear across a number of different studies. How can research yield such different results?

So far, the review has concentrated on *instrument classification* in MIR and *similarity perception* in psychology. Two comments must be made on these two tasks. First, MIR-systems for instrument classification do not necessarily model perception. Instrument classification is a well-defined task, fairly independent of human judgement. There are certainly many discriminating acoustic features of acoustic instruments that perceptual systems may not take into account, and they can be efficiently used for automatic classification, as, for example, employed by Barbedo and Tzanetakis (2011).

Perception and computational modelling are not independent, however, as soon as the modelled phenomena are of inherently psychological nature. This has largely been ignored in MIR, a field as hesitant as many parts of signal processing to integrate systematic perceptual evaluations into their methods, even though many of their systems are targeting human users. That might be one reason why researchers have commented on the existence of a 'glass ceiling' for the performance of music similarity algorithms (Pachet & Aucouturier, 2004; Pampalk, Flexer, & Widmer, 2005): despite hard work on descriptors and classifiers, retrieval performance as measured by precision and recall scores does not seem to have improved significantly. This may be a natural consequence of neglecting the inherently psychological nature of the notion of music similarity. More recently, Sturm (2013, 2012) has demonstrated that three state-of-the-art systems for genre classification do not recognize genre, if anything, despite yielding high classification accuracy scores. For example, he modified the spectral weighting of songs by mere spectral filtering, with the result that some were classified radically differently afterwards. Most humans would not

agree with this notion of genre, yet the classifiers obtained high accuracy scores overall. Sturm thus argued for a less narrow notion of evaluation in MIR, which often relies exclusively on classification accuracy scores. A broader conception of evaluation would also be better able to cope with data sets in which the variable of interest, genre for instance, is confounded by other irrelevant variables, such as long-term spectral distribution or reverberation. If a classifier distinguishes jazz from classical music because it has become sensitive to room-acoustical cues that confound the training exemplars, it has certainly learned a conception of genre that does not correspond with a human definition (cf., Sturm, 2014). With perhaps even more sobering results, Flexer (2014) reconsidered the 2006–2013 *Music Information Retrieval Evaluation eXchange* (MIREX) audio music similarity and retrieval task. The task randomly drew M query songs from a database (M=60 in 2006, M=100 in 2007–2011, M=50 in 2012–2013), selected the 5 (10 in 2012–2013) most similar candidate songs for each query as ranked by the algorithms, and had human listeners judge the subjective similarity of each query-candidate pair. During the years 2007–2013, the evaluation was based upon one human rating per query-candidate pair. In 2006, when any pair was rated by three human listeners, their intercorrelation ranged only between .37–.43. This implies that the 'ground truth' itself was highly inconsistent throughout 2006–2013, questioning the very purpose of that MIREX task.

Overall, these results suggest the obvious: as soon as machine learning attempts to model generically psychological constructs, such as similarity or genre, it would be short-sighted not to thoroughly consider the cognitive processes involved in their formation (also see Griffiths, 2015, a manifesto for a new cognitive revolution in the computational sciences). For MIR, it will not be enough to switch to physiologically inspired audio representations, but classification and similarity modelling will need to incorporate knowledge about processes in music perception and cognition, as well as factors of personal experience and human memory. This does not question MIR's achievements in tasks that do not require a psychological foundation, automatic instrument classification is one

example. In that case, the falling tree does make a sound, even if no human is there to listen.

Secondly, we wish to highlight the idea that even human subjects may rely on different acoustic features in categorization and similarity judgements. As made clear by Tversky (1977), different perceptual features may contribute to different tasks: *'Our total data base concerning a particular object (e.g., a person, a country, or a piece of furniture) is generally rich in content and complex in form. It includes appearance, function, relation to other objects, and any other property of the object that can be deduced from our general knowledge of the world. When faced with a particular task (e.g., identification or similarity assessment) we extract and compile from our data base a limited list of relevant features on the basis of which we perform the required task.'*(p. 329) Similar to furniture, musical timbre varies along a variety of features that enable identification and discrimination. Yet, if confronted with the task of similarity assessment, subjects could rely on the perceptually most salient properties according to which musical tones can be most easily compared with one another. Spectral centre of gravity and attack time may be dimensions well suited for such comparative tasks. For absolute identification or classification, however, it would be a non-optimal strategy to not make use of all other available features that reduce ambiguity between stimuli. This may be particularly important for a perceptual parameter such as timbre, for which the *'perceptual consequences of the multiplicity of cues created by the sound production process are varied. [...] Any single cue will provide some level of identification performance, and combinations of cues usually will produce better performance than a single one. Moreover, the effectiveness of any cue will vary across contexts.'*(Handel, 1995, p. 433) Recent empirical evidence supports the idea that cues for identification and similarity assessment may differ. Agus et al. (2012) demonstrated that neither solely spectral nor solely temporal properties can account for speeded perceptual classification of timbre, although these properties usually play a salient role in dissimilarity judgements. Task sensitivity also seems to be an important component in reconciling

apparent divergences between studies showing that listeners implicitly recognize the subtlest variations in frozen noise sounds (Agus, Thorpe, & Pressnitzer, 2010), or discriminate between subtle changes in spectrotemporal behaviour (McAdams et al., 1999), but appear to be insensitive to such subtleties in timbre dissimilarity ratings. Or listeners may be sensitive to certain mechanical properties of sounding objects carried by acoustic cues when making dissimilarity ratings, but will use only a subset of those for source material identification, i.e., the ones that are most reliable for the task according to listeners' past auditory experience in the world (McAdams, Roussarie, Chaigne, & Giordano, 2010).

Neurophysiological studies on the task-sensitivity of sensory representations begin to shed light on the neural mechanisms potentially underlying these phenomena. Measuring frequency selectivity of auditory cortical neurons in behaving ferrets, J. Fritz, Shamma, Elhilali, and Klein (2003) found that neurons exhibited facilitated response properties in a tone detection task compared to a passive listening condition: When ferrets had to detect tones of a specific frequency (a task on which they were trained in advance) neurons in primary auditory cortex adapted within seconds and showed greater sensitivity to the target frequency. J. Fritz, David, and Shamma (2013) conclude, *'RFs [receptive fields] of A1 neurons in the adult are in a state of rapid flux that is modulated by continuous "top-down" biasing as a function of changing salience and task-relevance of auditory stimuli.'*(p.83) This type of neurophysiological evidence even questions the very existence of a task-independent, comprehensive (i.e.,'platonic') representation of perceptual sound features in auditory cortex. What has been hypothetically described above as a 'feature selection' process, might indeed turn out to be closer to a cortical process of task-sensitive 'feature generation'.

This perspective also calls into question the psychological status of the 'spatial metaphor' of timbre (Wessel, 1973). An 'emphatic' interpretation of the timbre space studies reviewed above would assert that there are a few (say, less than or equal to five, cf.,

Elliott et al., 2013) perceptual dimensions that constitute the timbre of musical instruments, regardless of perceptual task or stimulus context that only need to be revealed by an appropriate experimental methodology. Coherent with General Recognition Theory (Ashby, 1992), this would imply that timbre categorization could be implemented via category boundaries in timbre space (cf., Giordano & McAdams, 2010). A more 'liberal' interpretation would interpret the obtained MDS dimensions as the most salient dimensions for the given stimulus set and rating task. Such a view would also acknowledge that the resulting low-dimensional description of timbre may be incomplete in regards to other tasks such as discrimination or identification. As should be clear by now, the above discussion leans towards a liberal view, suggesting a view of timbre space as a powerful exploratory tool of dissimilarity data, rather than as a 'hard-wired' and comprehensive, 'perceptual coordinate system' of timbre.

Conceptually, the hypothesis of distinct cues for timbre similarity judgements and identification brings us back to machine learning where it is common practice to use high-dimensional descriptor spaces in order to achieve high discriminatory power and to facilitate classification. If similarity and classification rely on partially distinct features, it would only be appropriate if computational models of these tasks require different compilations of descriptors. This is what was observed for modelling in MIR and psychology.

4 Objectives and Evaluation Criteria

Yet another factor (if not a 'meta-factor') plays a substantial role in the divide of the disciplines. It is related to the fact that MIR and psychoacoustics originate from disparate scientific traditions with different objectives. MIR has its roots in applied computer science and machine learning and is therefore primarily interested in the question of how to build robust systems. This implies that it is important *whether* a descriptor has predictive power, rather than *what* exact acoustic properties it encodes. Aucouturier and Bigand

(2013) polemically summarize this by stating, *'MIR research practices are notoriously goal-oriented.'*(p. 488) Being goal-oriented enables MIR, on the other hand, to be innovative in terms of design of descriptors, still the core of much MIR research. More recently, researchers have nonetheless argued that heuristic descriptor design should be replaced by deep neural network architectures that enable the fully-automated joint optimization of the descriptor extraction and classification steps (Humphrey, Bello, & LeCun, 2012). Following the success of deep learning algorithms in many other applications, this would leave the classical two-level architectures (descriptors + classifiers) behind, whose optimization might be tedious, but still easy to interpret globally. From the machine-learning perspective, this does not seem to be unconvincing, particularly because there are well-defined criteria of achievement, for instance by comparing systems with the current *state of the art*. Characteristic of this competitive approach are the annual MIREX competitions mentioned above, which evaluate systems on pre-specified tasks. Scholarly achievement is evaluated in a strict manner, similar to sports, by definite rankings.

Despite the success of such systems in some applications, the machine-learning approach has been harshly criticized. For instance, cognitive science icon Noam Chomsky ironically envisioned what it would mean to study gravity by means of machine learning: *'If you took tons of video tapes of what's happening outside my office window, leaves flying and various things and if you did an extensive analysis of them, you would get some kind of prediction of what's likely to happen next, certainly way better than anybody in the physics department could give.'*[2] Chomsky here promotes the old distinction between *knowledge-how* and *knowledge-that*, practical and theoretical knowledge, between engineering and science. Instead of defining scholarly success by the precision with which an a priori defined ground-truth can be approximated, he endorses the pursuit of the

[2]His statement can be retrieved under `http://techtv.mit.edu/videos/13200-keynote-panel-the-golden-age-a-look-at-the-original-roots-of-artificial-intelligence-cognitive-science-and-neuroscience-`. The quote stems from the last minute of the recording.

underlying principles. This is a research project that cannot be measured with precision scores, because there is no ground for the ground truth.

Psychoacoustics and music psychology have traditionally attempted to position themselves in that scientific realm. The attempt to look for the most parsimonious explanation of a given set of phenomena—to keep things as simple as possible but not simpler—is supposed to be at the heart of the disciplines' methodological approach, a principle often referred to as *Ockham's razor*. This principle indeed mediates between different criteria for model selection, in particular *goodness of fit* on the one hand, and *simplicity* on the other (cf., Myung, Tang, & Pitt, 2009). Whereas fit can be strictly quantified, e.g., by the prediction error, the R^2 coefficient, or maximum likelihood coefficients that measure the probability of the data given the best fitting model parameters, simplicity can only be assessed in a straight-forward manner for mathematical models that possess a directly accessible number of freely varying parameters. The AIC and BIC criteria introduced above in the context of MDS are examples that combine these two criteria: Both consist of a negative log-likelihood term (fit) plus a measure of simplicity (AIC: $2k$, BIC: $k \ln(n)$, k being the number of parameters, n the sample size). Minimizing AIC or BIC thus implies finding a viable compromise between goodness of fit and simplicity, which, from a statistical standpoint, also serves to counteract overfitting (Myung et al., 2009).

When a direct quantification of model parameters becomes inviable and data sets are sufficiently large, cross-validation becomes the standard evaluation method. Note that although cross-validation delivers robust means to prevent overfitting, it does not necessarily promote parsimonious descriptions of the data, thus allowing for accounts that fit well and generalize, but are overly complex and might be confounded (as discussed above). A different approach thus seeks parsimony in the modelling process itself. To give an example, Dau, Püschel, and Kohlrausch (1996) presented a quantitative model of auditory perception. Subsequently, a modulation filterbank was added in order to account

for experimental data on modulation detection and modulation masking (Dau, Kollmeier, & Kohlrausch, 1997a; Dau et al., 1997b). This means, model *generality* was improved, i.e., the breadth of phenomena that the model accounted for. Later on, Jepsen, Ewert, and Dau (2008) replaced the original linear gammatone filterbank by an outer- and middle ear transformation plus a nonlinear cochlear filtering stage, in order to further explain phenomena such as spectral masking. Here processing modules were not added in order to let the model serve as an 'archive' of known functional principles, but on the contrary, yield baselines for the *minimal* amount of information processing required to reproduce empirical results with a desired goodness of fit and degree of generality. In other words, *'the approach is to focus on the "effective" signal processing which uses as little physiological and physical parameters as necessary, but tries to predict as many perceptual data as possible. On the one hand, such a modelling strategy will never allow conclusions about the details of signal processing at a neuronal or single-unit level. On the other hand, if the effective model correctly describes the transformation of physical parameters in a large variety of experimental conditions, this strongly suggests certain general processing principles.'*(Dau, 2008, p. 180)

To further complicate things, there are other, not strictly quantifiable criteria, which are indispensable for the scientific discourse (Jacobs & Grainger, 1994; Myung et al., 2009): *Explanatory adequacy* refers to whether a model minimizes the number of ad-hoc assumptions necessary to account for sets of data by relying on widely accepted theoretical principles and empirical observations. *Interpretability* denotes whether the model components or parameters are theoretically transparent and tied to established theoretical principles. The well-known criterion of *falsifiability* (Popper, 1963) denotes whether a theory or model can be proven wrong by empirical observation.

Coming back to psychological models of timbre perception, these considerations suggest that one remain sceptical about the utility of any additional variable that does not add a significant amount of predictive power to a model (i.e., to respect parsimony), in

particular if its acoustic meaning is hard to decipher (i.e., to ensure interpretability). Correspondingly, many of the reviewed MDS studies on similarity perception achieved impressively compact explanations, relating the acoustical basis of timbre similarity perception to only a handful of relevant dimensions. Interestingly, more recent studies seem not to adhere to these principles so strictly any more. Published in a classical journal of psychoacoustics, Elliott et al. (2013) compare the predictions provided by their modulation power spectrum (MPS) with those of traditional audio descriptors, both of which explain similar proportions of variance in timbre dissimilarity ratings. They conclude, *'When both give positive correlation results, the MPS analysis has the advantage of being more quantitative and detailed.'*(p. 400) Nonetheless, from the perspective of dimensionality, modulation spectra are a less parsimonious description compared to the few audio descriptors that the study identified as yielding similar correlations. As mentioned above, the very same scepticism towards low-dimensional representations of envelopes appear to motivate Terasawa et al. (2012) in their work on MFCCs. And even if the representation from Patil et al. (2012) is physiologically plausible (yielding explanatory adequacy), the computations that enable classification, for example, tensor singular value decomposition for reducing dimensionality (e.g., from 4224 to 420 per timbre), place this approach in a different league of computational complexity compared to earlier MDS models. Although this approach shows that there is enough information provided by STRFs in order to account for timbre classification and dissimilarity in principle, the computational opacity of the subsequent machine-learning steps do not yield specific hypotheses on information processing beyond the representation stage. The authors conclude, *'Timbre percepts can be effectively explained by the joint spectro-temporal analysis performed at the level of mammalian auditory cortex. However, unlike the small number of spectral or temporal dimensions that have been traditionally considered in the timbre literature, we cannot highlight a simple set of neural dimensions subserving timbre perception.'*(Patil et al., 2012, p. 11). However, with this conclusion there may even be the danger of proposing a theory

of timbre that is too unspecific to be falsifiable.

This complex interplay of different modelling stages and corresponding evaluation criteria seems to go unnoticed by Aucouturier and Bigand (2013). They note: *'But should one conclude that the brain implements a discrete cosine transform? Probably not. It would be like concluding that jet planes demonstrate how birds fly just because they both move in air. [...] When such physiologically and psychologically validated alternatives exist [STRFs], it is [...] increasingly difficult to justify the use of features like MFCCs in MIR studies pretending to have any relevance for cognition.'*(Aucouturier & Bigand, 2013, p. 487) They hence question the explanatory adequacy of discrete cosine transforms in the case of MFCCs, but endorse Patil et al.'s work as a 'valid' alternative. This raises doubts about whether they consider processing steps beyond the initial audio representation as important at all (one could ask alternatively, 'Does the brain implement tensor singular value decomposition?'). To be clear, we do not wish to argue for or against any audio representation in particular at this point, but rather discuss the ways in which we draw conclusions about representations as part of cognitive models. We suspect that considering one part of the puzzle through the lense of one evaluation criterion is not enough; a biologically plausible audio representation is not the only valuable property of a sensory-cognitive auditory model. To use the words of David Marr, *"Trying to understand perception by only studying neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structures of the feathers and the different shapes of birds' wings make sense."*(Marr, 2010, p. 27) It should become part of the interdisciplinary discourse to negotiate how different criteria of scientific evaluation may frame the indispensable modelling process.

5 Conclusion

In this article, we discussed the divide between the computational descriptions of musical timbre in MIR and music psychology. It turned out that a seemingly narrow question on descriptors not only raised issues around differences in the two fields' technical approaches, but also highlighted differences in underlying tasks and evaluation criteria. Specifically, we discussed new approaches to represent spectro-temporal modulations, both for MIR and psychology, and compared approaches to spectral envelope description. We outlined task differences and epistemological foundations of the two fields. They mostly deal with different facets of timbre-related tasks: MIR most often considers instrument *classification*, psychology has predominantly dealt with timbre *dissimilarity* perception (in most studies where audio descriptors were taken into account). We argued that as cognitive phenomena, it is by no means clear that classification (or identification) and similarity assessment rely on the same compilation of perceptual features when dealing with high-dimensional perceptual objects such as timbre.

In closing, we would like to suggest three basic questions whose discussion could benefit both researchers in MIR and music psychology (in particular before embarking on joint projects):

1. What kind of knowledge is pursued? (Computational models of) perceptual principles or computational application?

2. What kind of task does the research focus on? Does it involve a cognitive component? If yes, is it based on subjective judgements (e.g., similarity, affect, etc.) or does it involve absolute identification or classification scores? If not, are you sure?

3. What kind of evaluation criteria apply? Biological plausibility, computational parsimony, robustness in applications, conceptual simplicity, etc.?

Having answers to such questions would help frame the work to be done more clearly and rigourously.

Acknowledgements

This work was supported by a grant from the Canadian Natural Sciences and Engineering Research Council (RGPIN 312774) and a Canada Research Chair (950-223484) awarded to Stephen McAdams; a Harman Scholarship from the Audio Engineering Society's Educational Foundation and a Québec International Merit Scholarship to Kai Siedenburg. The authors wish to thank the two anonymous reviewers for valuable comments on the manuscript.

References

Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *The Journal of the Acoustical Society of America*, *131*(5), 4124–4133.

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. *Neuron*, *66*, 610–618.

Alluri, V., & Toiviainen, P. (2009). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, *27*(3), 223–241.

Andén, J., & Mallat, S. (2014). Deep scattering spectrum. *IEEE Transactions on Signal Processing*, *62*(16), 4114–4128.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Lawrence Erlbaum Associates.

Aucouturier, J.-J., & Bigand, E. (2012). Mel Cepstrum & Ann Ova: The difficult dialog between MIR and Music Cognition. In F. Gouyon, P. Herrera, L. G. Martins, & M. Müller (Eds.), *Proceedings of the 13th International Society for Music Information Retrieval Conference, Porto, Portugal, Oct 8–12, 2012* (pp. 397–402). Porto, Portugal: FEUP Edições.

Aucouturier, J.-J., & Bigand, E. (2013). Seven problems that keep MIR from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems*, *41*(3), 483–497.

Barbedo, J. G. A., & Tzanetakis, G. (2011). Musical instrument classification using individual partials. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(1), 111–122.

Barthet, M., Depalle, P., Kronland-Martinet, R., & Ystad, S. (2010). Acoustical correlates of timbre and expressiveness in clarinet performance. *Music Perception*, *28*(2), 135–153.

Barthet, M., Guillemain, P., Kronland-Martinet, R., & Ystad, S. (2010). From clarinet control to timbre perception. *Acta Acustica united with Acustica*, *96*(4), 678–689.

Berry, C. J., Shanks, D. R., Speekenbrink, M., & Henson, R. N. (2012). Models of recognition, repetition priming, and fluency: exploring a new framework. *Psychological Review*, *119*(1), 40.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Bruna, J., & Mallat, S. (2013). Audio texture synthesis with scattering moments. *arXiv preprint arXiv:1311.0407*.

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

Dau, T. (2008). Auditory processing models. In D. Havelock, S. Kuwano, & M. Vorländer (Eds.), *Handbook of signal processing in acoustics* (Vol. 1, pp. 175–196). New York, NY: Springer.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997a). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, *102*(5), 2892–2905.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997b). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *The Journal of the Acoustical Society of America*, *102*(5), 2906–2919.

Dau, T., Püschel, D., & Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *The Journal of the Acoustical Society of America*, *99*(6), 3615–3622.

Elliott, T., Hamilton, L., & Theunissen, F. (2013). Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *The Journal of the Acoustical Society of America*, *133*(1), 389–404.

Elliott, T., & Theunissen, F. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, *5*(3), e1000302.

Ellis, D. P. (1999). Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Communication*, *27*(3), 281–298.

Eronen, A. (2001). Comparison of features for musical instrument recognition. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, Oct 21–24, 2001* (pp. 19–22). Piscataway, NJ: IEEE.

Essid, S., Richard, G., & David, B. (2006). Musical instrument recognition by pairwise classification strategies. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(4), 1401–1412.

Flexer, A. (2014). On inter-rater agreement in audio music similarity. In H. Wang, Y. Yang, & J. H. Lee (Eds.), *Proceedings of the 15th International Society for Music Information Retrieval Conference, Taipei, Taiwan, Oct 27–31, 2014* (pp. 245–250).

Fritz, C., Blackwell, A. F., Cross, I., Woodhouse, J., & Moore, B. C. (2012). Exploring violin sound quality: Investigating English timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties. *The Journal of the Acoustical Society of America*, *131*(1), 783–794.

Fritz, J., David, S., & Shamma, S. (2013). Attention and dynamic, task-related receptive field plasticity in adult auditory cortex. In Y. E. Cohen, A. N. Popper, & R. R. Fay (Eds.), *Neural correlates of auditory cognition* (Vol. 45, pp. 251–291). New York, NY: Springer.

Fritz, J., Shamma, S., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, *6*(11), 1216–1223.

Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A survey of audio-based music

classification and annotation. *IEEE Transactions on Multimedia*, *13*(2), 303–319.

Fujinaga, I. (1998). Machine recognition of timbre using steady-state tone of acoustic musical instruments. In M. Simoni (Ed.), *Proceedings of the International Computer Music Conference, Ann Arbor, MI, USA, Oct 1–6, 1998* (pp. 207–210). San Francisco, CA: International Computer Music Association.

Fujinaga, I., & MacMillan, K. (2000). Realtime recognition of orchestral instruments. In I. Zannos (Ed.), *Proceedings of the International Computer Music Conference, Berlin, Germany, Aug 27–Sep 1, 2000* (pp. 141–143). San Francisco, CA: International Computer Music Association.

Gadermaier, T., & Reuter, C. (2014). Strukturelle Merkmale von Blasinstrumentenspektren–die Schumannschen Klangfarbengesetze aus heutiger Sicht [Structural features of brass instruments – Schumann's laws of tone colour from a contemporary perspective]. In *Fortschritte der Akustik – Tagungsband der 40. DAGA, Oldenburg, Germany, Mar 10–13, 2014* (pp. 48–49).

Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., & Plumbley, M. D. (2013). Detection and classification of acoustic scenes and events: An IEEE AASP challenge. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz NY, USA, Oct 20–23, 2013* (pp. 1–4). Piscataway, NJ: IEEE.

Giordano, B. L., & McAdams, S. (2010). Sound source mechanics and musical timbre perception: Evidence from previous studies. *Music Perception*, *28*(2), 155–168.

Golubock, J. L., & Janata, P. (2013). Keeping timbre in mind: Working memory for complex sounds that can't be verbalized. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 399–412.

Grey, J. M. (1975). *An exploration of musical timbre*. Unpublished doctoral dissertation, CCRMA, Stanford University.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal*

of the Acoustical Society of America, *61*(5), 1270–1277.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, *63*(5), 1493–1500.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Handel, S. (1995). Timbre perception and auditory object identification. In B. C. Moore (Ed.), *Hearing* (Vol. 2, pp. 425–461). San Diego, CA: Academic Press.

Herrera-Boyer, P., Klapuri, A., & Davy, M. (2006). Automatic classification of pitched musical instrument sounds. In A. Klapuri & M. Davy (Eds.), *Signal processing methods for music transcription* (pp. 163–200). New York, NY: Springer.

Horner, A. B., Beauchamp, J. W., & So, R. H. (2011). Evaluation of mel-band and MFCC-based error metrics for correspondence to discrimination of spectrally altered musical instrument sounds. *Journal of the Audio Engineering Society*, *59*(5), 290–303.

Humphrey, E. J., Bello, J. P., & LeCun, Y. (2012). Moving beyond feature design: Deep architectures and automatic feature learning in music informatics. In F. Gouyon, P. Herrera, L. G. Martins, & M. Müller (Eds.), *Proceedings of the 13th International Society for Music Information Retrieval Conference, Porto, Portugal, Oct 8–12, 2012* (pp. 403–408). Porto, Portugal: FEUP Edições.

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, *94*(5), 2595–2603.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human perception and performance*, *20*(6), 1311–1334.

Jepsen, M., Ewert, S., & Dau, T. (2008). A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America*, *124*(1), 422.

Joder, C., Essid, S., & Richard, G. (2009). Temporal integration for audio classification with application to musical instrument classification. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(1), 174–186.

Kendall, R. A., Carterette, E. C., & Hajda, J. M. (1999). Perceptual and acoustical features of natural and synthetic orchestral instrument tones. *Music Perception*, *16*(3), 327–363.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique. *Le Journal de Physique IV*, *4*(C5), 625–628.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzén & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (Vol. 846, pp. 43–53). Amsterdam, The Netherlands: Excerpta Medica.

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

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

Lembke, S.-A. (2014). *When timbre blends musically: perception and acoustics underlying orchestration and performance*. Unpublished doctoral dissertation, McGill University.

Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics*, *65*(10), 1331–1398.

Marozeau, J., de Cheveigné, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *The Journal of the Acoustical Society of America*, *114*(5), 2946–2957.

Marr, D. (2010). *Vision: A computational approach*. Cambridge, MA: MIT Press.

Marsden, A. (2012). Interrogating melodic similarity: a definitive phenomenon or the product of interpretation? *Journal of New Music Research*, *41*(4), 323–335.

Martin, K. D. (1999). *Sound-source recognition: A theory and computational model*.

Unpublished doctoral dissertation, Massachusetts Institute of Technology.

McAdams, S. (1993). Recognition of sound sources and events. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: The cognitive psychology of human audition* (pp. 146–198). Oxford, UK: Oxford University Press.

McAdams, S. (2013). Musical timbre perception. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 35–67). San Diego, CA: Academic Press.

McAdams, S., Beauchamp, J. W., & Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *The Journal of the Acoustical Society of America*, *105*(2), 882–897.

McAdams, S., Roussarie, V., Chaigne, A., & Giordano, B. L. (2010). The psychomechanics of simulated sound sources: Material properties of impacted thin plates. *The Journal of the Acoustical Society of America*, *128*(3), 1401–1413.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*(3), 177–192.

McDermott, J., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*, 926–940.

Misdariis, N., Minard, A., Susini, P., Lemaitre, G., McAdams, S., & Parizet, E. (2010). Environmental sound perception: Metadescription and modeling based on independent primary studies. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*(1), Article ID 362013.

Müller, M., Ellis, D. P., Klapuri, A., & Richard, G. (2011). Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, *5*(6), 1088–1110.

Myung, J. I., Tang, Y., & Pitt, M. A. (2009). Evaluation and comparison of computational models. *Methods in Enzymology*, *454*, 287–304.

Ng, A. Y. (1997). Preventing "overfitting" of cross-validation data. In D. H. Fisher (Ed.), *Proceedings of the 14th International Conference on Machine Learning, Nashville,*

TN, USA, July 8–12, 1997 (pp. 245–253). San Francisco, CA: Morgan Kaufmann Publishers.

Pachet, F., & Aucouturier, J.-J. (2004). Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, *1*(1), 1–13.

Pampalk, E., Flexer, A., & Widmer, G. (2005). Improvements of audio-based music similarity and genre classification. In *Proceedings of the 6th International Society for Music Information Retrieval Conference, London, UK, Sep 11–15, 2005* (pp. 628–633).

Patil, K., & Elhilali, M. (2013). Task-driven attentional mechanisms for auditory scene recognition. In *Proceedings of the 2013 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, Mar 26–31, 2013* (pp. 828–832). Piscataway, NJ: IEEE.

Patil, K., Pressnitzer, D., Shamma, S., & Elhilali, M. (2012). Music in our ears: The biological bases of musical timbre perception. *PLOS Computational Biology*, *8*(11), e1002759.

Patterson, R. D., Gaudrain, E., & Walters, T. C. (2010). The perception of family and register in musical tones. In M. Riess Jones, R. R. Fay, & A. Popper (Eds.), *Music perception* (pp. 13–50). New York, NY: Springer.

Peeters, G. (2003). Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *Proceedings of the 115th Audio Engineering Society Convention, New York, NY, USA, Oct 10–13, 2003* (pp. 1–14). New York, NY: Audio Engineering Society.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, *130*(5), 2902–2916.

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. F. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing*

(pp. 397–414). Leiden, The Netherlands: Suithoff.

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London, UK: Routledge and Kegan Paul.

Reuter, C. (2002). *Klangfarbe und Instrumentation*. Peter Lang.

Richard, G., Sundaram, S., & Narayanan, S. (2013). An overview on perceptually motivated audio indexing and classification. *Proceedings of the IEEE*, *101*(9), 1939–1954.

Saitis, C., Scavone, G. P., Fritz, C., & Giordano, B. L. (2015). Effect of task constraints on the perceptual evaluation of violins. *Acta Acustica united with Acustica*, *101*(2), 382–393.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*(3), 319–345.

Schubert, E., & Wolfe, J. (2006). Does timbral brightness scale with frequency and spectral centroid? *Acta Acustica united with Acustica*, *92*(5), 820–825.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shepard, R. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Siedlecki, W., & Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, *10*(5), 335–347.

Štěpánek, J., & Otčenášek, Z. (2004). Interpretation of violin spectrum using psychoacoustic experiments. In *Proceedings of the International Symposium on Musical Acoustics (ISMA2004), Nara, Japan, Mar 31–Apr 3, 2004* (pp. 324–327).

Štěpánek, J., Syrový, V., Otčenášek, Z., Taesch, C., & Angster, J. (2005). Spectral features influencing perception of pipe organ sounds. In F. Augusztinovicz, A. B. Nagy, &

Z. Hunyadi (Eds.), *Proceedings of the Forum Acusticum, Budapest, Hungary, Aug 29–Sep 2, 2005* (pp. 465–469).

Sturm, B. L. (2012). Two systems for automatic music genre recognition: What are they really recognizing? In C. C. S. Liem (Ed.), *Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies, Nara, Japan, Oct 29 – Nov 2, 2012* (pp. 69–74). New York, NY.

Sturm, B. L. (2013). Classification accuracy is not enough. *Journal of Intelligent Information Systems*, *41*(3), 371–406.

Sturm, B. L. (2014). A simple method to determine if a music information retrieval system is a horse. *IEEE Transactions on Multimedia*, *16*(6), 1636–1644.

Sturm, B. L., & Noorzad, P. (2012). On automatic music genre recognition by sparse representation classification using auditory temporal modulations. In M. Aramaki, M. Barthet, R. Kronland-Martinet, & S. Ystad (Eds.), *Proceedings of the 9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012), London, UK, Jun 19–22, 2012* (pp. 379–394). Heidelberg, Germany: Springer.

Susini, P., Lemaitre, G., & McAdams, S. (2012). Psychological measurement for sound description and evaluation. In B. Berglund, G. B. Rossi, J. T. Townsend, & L. R. Pendrill (Eds.), *Measurement with persons: Theory, methods, and implementation areas* (pp. 227–253). New York, NY: Psychology Press.

Terasawa, H., Berger, J., & Makino, S. (2012). In search of a perceptual metric for timbre: Dissimilarity judgments among synthetic sounds with MFCC-derived spectral envelopes. *Journal of the Audio Engineering Society*, *60*(9), 674–685.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

von Helmholtz, H. (1875). *On the sensations of tone as a physiological basis for the study of music.* London, UK: Longmans, Green.

Wessel, D. L. (1973). Psychoacoustics and music: A report from Michigan State

University. *PACE: Bulletin of the Computer Arts Society*, *30*, 1–2.

Wessel, D. L., Bristow, D., & Settel, Z. (1987, August). Control of phrasing and articulation in synthesis. In J. Beauchamp (Ed.), *Proceedings of the 1987 International Computer Music Conference, Urbana-Champaign, IL, USA* (pp. 108–116). San Francisco, CA: Computer Music Association.

Wiggins, G. A. (2009). Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music. In *Proceedings of the 11th IEEE International Symposium on Multimedia, San Diego, CA, USA, Dec 14–16, 2009* (pp. 447–482). Piscataway, NJ: IEEE.

Winsberg, S., & De Soete, G. (1993). A latent class approach to fitting the weighted Euclidean model, CLASCAL. *Psychometrika*, *58*(2), 315–330.

Wun, S., Horner, A., & Wu, B. (2014). Effect of spectral centroid manipulation on discrimination and identification of instrument timbres. *Journal of the Audio Engineering Society*, *62*(9), 575–583.