TYING SEMANTIC LABELS TO COMPUTATIONAL DESCRIPTORS OF SIMILAR TIMBRES

Rosemary A. Fitzgerald
Department of Music
Lancaster University,
Lancaster, LA1 4YW, UK

r.a.fitzgerald@lancaster.ac.uk

Adam T. Lindsay
Computing Department
Lancaster University
InfoLab21,
Lancaster, LA1 4WA, UK
atl@comp.lancs.ac.uk

ABSTRACT

This paper asserts the importance of using multiple representations when examining computational descriptions of any musical percept. We examine the existing state-of-the-art among the MPEG-7 timbre descriptors, and by choosing oboe timbre as our domain we investigate their application to timbres from the same instrument. Using a dual approach, correlating perceptual information with a wide range of computational descriptors, we propose potential extensions to MPEG-7 representations. By eliminating the need for generality across instruments, we believe we can reach a higher level of semantic representation within timbre domains from a single instrument class.

1. INTRODUCTION

Despite past efforts from standards bodies, MPEG-7, the first edition of which was published by ISO in 2002, standardises some aspects of musical instrument timbre. We look at the details of such an effort, examining the representation space that the standard spans. We then outline research to expand that representation space, based on psychological experiments and a wider range of signal processing. We expect to be able to reach a wider, more flexible range of representations than are currently available, and to allow for more sophisticated reasoning on musical timbre.

2. TIMBRE DESCRIPTION

Due to its complex multidimensional nature, timbre is still poorly understood and difficult to manipulate in a controlled, scientific way. The principal reason for this is that there is no universal definition of timbre. There are no standard units from which timbre may be quantified; any description given is verbal or phenomenological. Such qualitative descriptions have been used to compare the verbal attributes of orchestral instruments [1], [2] in attempts to gain greater understanding of timbre.

Many definitions of timbre have centred upon trying to construct a constitutive definition. This has been found as far back as early Chinese civilizations who developed sophisticated written definitions of timbre, based on a classification of material sources such as metal, stone, clay, skin, silk threads, wood, gourd and bamboo [3]. In a musical sense, the tonal quality characterising a sound can have many forms. Timbre may either refer to a specific family of tones (e.g., woodwind or double reed), or to an instrument (e.g., oboe). It may also be applicable to the perceived quality of a specific tone (e.g., a dark or a nasal oboe tone) and is sometimes used when describing different instrumental combinations when detailing the orchestration of a musical work.

The large number of variables involved in describing timbre leads to definitional vagueness. This is perhaps most evident in the frequently quoted statement by the American National Standards Institute (ANSI) in which timbre is described as "… an attribute of auditory sensation in terms of which a listener can judge that two sounds are similarly presented and having the same loudness and pitch are dissimilar" [4]. It has been suggested that the vagueness of the definition may, perhaps, be related to the multidimensionality of the phenomenon [5], which can perhaps be related to a number of perceptual attributes, described in the footnote to the ANSI definition. The footnote states that, "… timbre depends primarily upon the spectrum of the stimulus but it also depends upon the waveform, the sound pressure the frequency location of the spectrum and the temporal characteristics of the stimulus." [4].

3. A PROGRESSIVE APPROACH TO TIMBRE

3.1. Staircase model of perception

When studying the computational interpretation of multimedia, we have found it helpful to consider the progression of the computational process from expression (signal) to description (meaning) by imagining several steps on a staircase proceeding upwards from a signal towards meaning. This is very much inspired by Marr's [6] representational framework laid out in his seminal work. We abstract from his reliance on the primal and 2-½ D sketches to a more extensive survey of what may plausibly happen along the lines of human perception and understanding. One instance is illustrated in Figure 1.

We use this staircase as an instructional model to reveal multiple representations inherent in the computational analysis of timbre, the assumptions involved in doing such analysis, and to point the way forward for refined processing. The model, in brief, attempts to expose the steps – whether perceptual or computational, human or computer – taken in response to an external stimulus. Each step is conceivably a representation slightly more abstract than the previous one. At each

step, new knowledge, whether explicit, implicit, or algorithmic, is added to the previous representation. Once the representation reaches the top level, the model may be iterated, whether refining symbolic information (the jump to segmentation) or reinterpreting (potentially a segment of) the signal (the jump to the signal). It is hoped that the example below will further clarify the model.

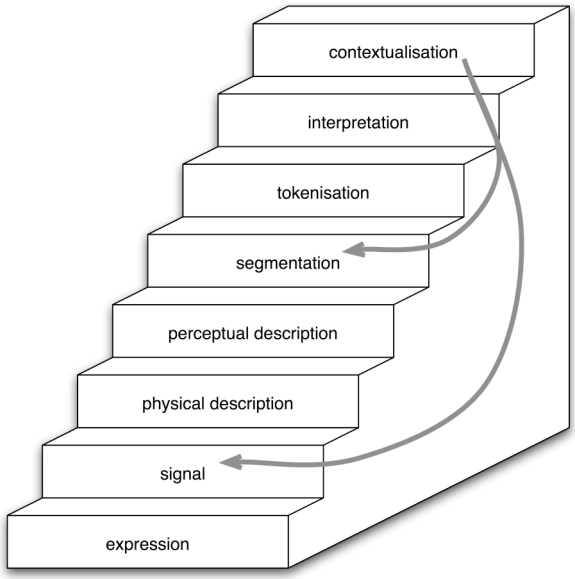

Figure 1. A generic series of steps taken to extract "meaning" from a signal. An interpretation may be iteratively re-segmented and "chunked" into larger items of meaning.

3.2. Musical Instrument Timbre Comparisons in MPEG-7

The MPEG-7 activities represent an effort to establish a standard for computational descriptions of multimedia content. For audio in particular, there is an aspiration to derive meaning via signal processing. Although the official stance throughout the developing standard is that there is no preference as to whence a description arises (e.g., hand-annotated or computationally derived) it is clear throughout the Audio and Visual parts of the standard that the chosen representation favours a signal-processing method of feature extraction. We examine the timbre descriptors from the audio part of the standard to see where the processing assumptions are made, and how they fit the above staircase model of content understanding.

The application-oriented description schemes concerned with musical instrument timbre within the Audio part of the MPEG-7 standard draw upon research in musical perception and psychophysics that attempts to determine what features of a given musical sound distinguish it from other sounds at the same loudness and pitch [8]. The scheme draws upon low-level descriptors that have direct analogues within signal processing. The relationship between these schemes and the description output from a system implementing the descriptors is worthy of comment, but it is more suitable to begin by examining of the types of multimedia content that are expected to be input to the system, and what that the implications of these are.

The MPEG-7 Timbre tools describe perceptual features of “monophonic, non-mixed, non-layered instrument sounds" [8]. This places an explicit limit on the types of signal they may describe. The range of all possible signals is first limited to musical instrument sounds, and is further constrained to be of a solo instrument playing a note in isolation. Thus, a perceiver (or possibly another computer system) is required to ascend the staircase via the physical and perceptual steps to determine that this is indeed a monophonic signal. Then, that note must be either isolated from its neighbours in a temporal stream, or it must be determined that this has already been done (segmentation). At this point, the instrumental note will have been labelled with a token indicating that it is a sound suitable for timbre description.

Following the initial assumptions discussed immediately above, the clause in the standard posits four classes of musical instrumental sounds that may be described: non-sustained sounds; sustained, harmonic, coherent sounds; sustained, non-harmonic, coherent sounds; and sustained non-coherent sounds. The standard currently accommodates only the first two of these classes of sounds, which then form perceptual “spaces” in which sounds are compared. Thus, the sounds that have been segmented are interpreted and placed into a context (and reaching the top of the staircase) before they are even analysed by a computer for their intended use – “low-level” timbral descriptors.

By deciding which class of musical sound the signal belongs to, one decides which of the seven possible low-level temporal and spectral features are used to describe the sound. Two of the possible descriptors applicable to sustained, harmonic, coherent sounds are log attack time and harmonic spectral centroid. These are physical features of signals that are proxies for the perceptual features of “attack” and “brightness”, respectively. The computational details are irrelevant here, but it can be observed that both features rely on another ascent up the staircase (e.g., segment through the temporal signal to note the beginning and the loudest initial part of the sound) and then yield a physical measure that finally approximates a perceptual comparison. The final representation consists of four or five quantitative values. These values may then be compared in a perceptually scaled space to judge the perceptual similarity between two sounds.

The above description is not intended to denigrate the technology behind the MPEG-7 Timbre descriptors; those descriptors are the state of the art, backed up by experimental evidence. The intention here is rather to expose the series of assumptions that underpin the simple application of a pre-defined group of descriptors for a sound. A supervisory system (a human, in most cases) applies various analytical processes at each step up the staircase, to arrive at a classified segmented sound. In

other words, the computational techniques that purport simply to traverse the space between signal and perception actually embody techniques that incorporate segmentation and tokenisation as intermediate steps in order to compute their corresponding features.

3.3. Multiple representations are key

We see that although the final product is rather terse, typically with five parameters describing a note, there are many points along the way to that product that are plausible representations. We believe that further representations may be derived from the MPEG-7 timbre descriptors, both supplementing them at the same level and building upon them at higher levels. This is entirely consistent with MPEG-7 audio, trading off generality for descriptive power. With a wider range of salient representations, there is more flexibility in processing, as well as the possibility of a more intuitive interface for a user navigating timbre-space than navigating along axes such as "Harmonic Spectral Variation."

4. A PROGRAMME FOR EXPANDED TIMBRE DESCRIPTION

In order to prise open the series of perceptual and computational assumptions for ourselves we examined how musically-trained listeners tie perceptual similarities to descriptive labels. We now describe a perceptual experiment that explores perceptual similarities and verbal attribute magnitude estimation (VAME) of oboe tones from two different performers.

4.1. Psychoacoustic experiments

To obtain the perceptual dissimilarity, 32 musically-trained subjects were asked to make judgements of dissimilarity on a scale from 0–100. Twenty-four isolated tones digitally recorded at six different pitches (C4, F4, A4, C♯5, A♯5, F6) and two different dynamic levels *mf*, *ff*) by two oboists (A and B) from different schools of playing (British and American) were used as a primary data set. The recordings were premised on the use of real-world sounds, which involves treating the oboist, reed and instrument as a whole mechanism and recording the tones in an acoustically live room. The tones in the data set were equalised for duration (by adding a false decay) and amplitude as they would be compared against each other experimental conditions.

Eight adjectives, taken from the principal components analysis by Kendall and Carterette [1] in their studies on wind instrument dyads, were used as a basis to assess their suitability for describing oboe timbre. Kendall and Carterette [2] also used the same group of eight adjectives to assess the verbal characteristics of natural and synthetic single instrument tones. In their study these adjectives were found to describe four different factors/dimensions of the wind timbres: strong; tremulous; light (factor 1); nasal; rich (factor 2); brilliant; ringing (factor 3); reedy (factor 4).

To collect the VAME data for the study of oboe timbre subjects were asked to rate the magnitude of the verbal attributes for each tone after hearing it played once. They achieved this in the same manner as the perceptual similarity scaling by means of a computer-based moving slider, which converted the positioning by the subject to a value scale of 0–100. The poles of the rating scale were labelled "not adjective" – "adjective". The order of the presentation of the tones within each set of verbal attributes and dissimilarity comparison was randomly assigned.

4.1.1. Results

Perceptual differences are revealed between most of the tones played by the two oboists as shown by the Multi-dimensional Scaling (MDS) representation in figure 2. There are significant differences between tones for different performers across both the same loudness level (although some confusion occurs for pitch C4) and different loudness levels. Perceptual differences are also revealed within a performer: there are significant differences between tones across different loudness levels (although some confusion occurs for pitch C♯5).

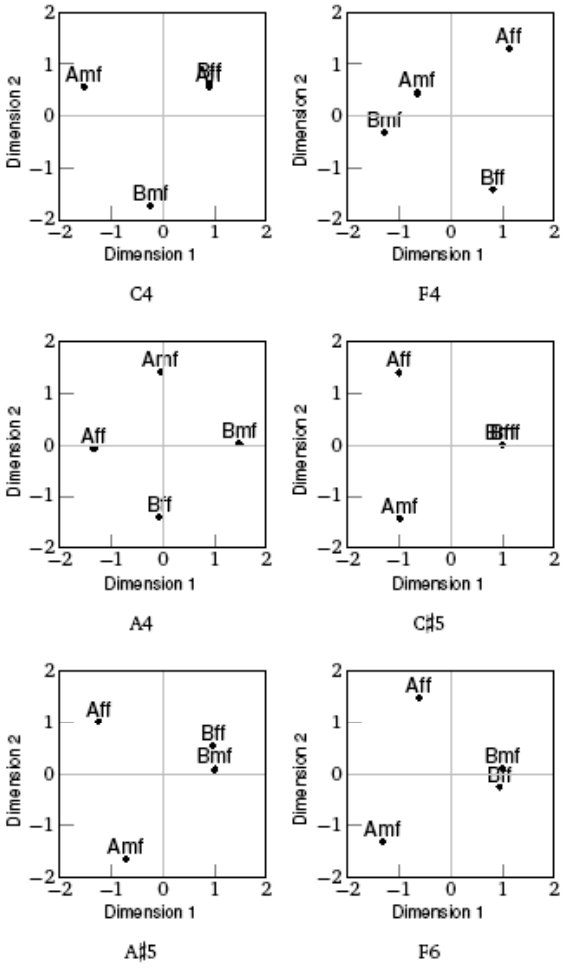

Figure 2. Two-dimensional MDS solution for the oboist/dynamic tones (*by pitch*)

Using analysis of variance (ANOVA), individual VAME relationships are revealed for each performer as there are significant differences in the data between the two oboists. This is evident for tones at the same loudness level and different loudness levels. VAME relationships are revealed within a performer as there are differences in judgments between tones for the same performer across different loudness levels, (except for pitch F6 for oboist B where VAME judgments were almost identical.)

Principal components analysis (PCA) with Varimax rotation (Kaiser normalisation) was performed on the VAME ratings for all tones. Figure 3 shows the three-dimensional solution for the PCA loadings for the mean verbal attribute ratings across pitches. Three factors, those with eigenvalues over 1, account for 54.055% of the variance. Factor 1, which accounts for 20.586% of the variance could, perhaps, be named the 'Power' factor as the attributes Strong, Rich and Brilliant rate the most positively, whilst the attributes Light and Nasal load negatively. The 'Power' factor was found by Kendall and Carterette [1] onto which the attribute Strong loaded positively. Factor 2 accounts for 17.025% of the variance and could, perhaps, be named as the 'Vibrancy' factor for the attributes Tremulous, Ringing and Brilliant rate the most positively, whilst Rich and Reedy load negatively. Factor 3 accounts for 16.444% of the variance and could, perhaps, be labelled the 'Pinched' factor. Attributes Nasal and Reedy rate the most positively, whilst Light and Brilliant load negatively. The averaged VAME ratings suggest that subjects are not confusing the attributes Nasal and Reedy and that the negative loadings of Light and Brilliant are being judged almost as opposites.

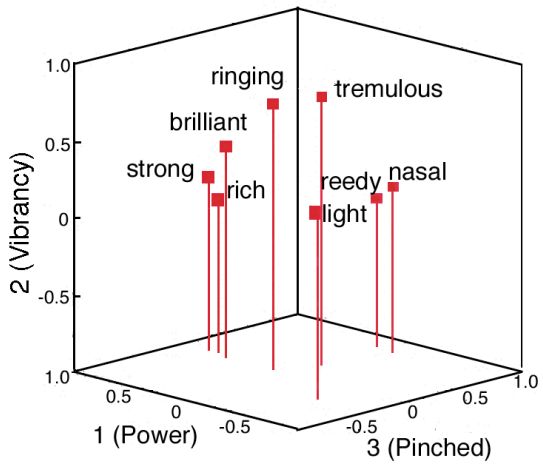

Figure 3. Three-dimensional configuration of the verbal attributes across all tones and pitches for the principal components analysis.

Although only two dimensions are found for the dissimilarity scaling solutions (see figure 2), it is suggested that the first two factors of 'Power' and 'Vibrancy' may account for these dimensions. It is possible that the 'Power' factor could be used as the label on the dimension differentiating the tones by dynamic, whereas the 'Vibrancy' factor is used to differentiate between oboist. This is reflected in the results of the ANOVA (as oboist A's tones were judged to be higher over all pitches on ratings of Tremulous and Ringing than those from oboist B).

The Kendall and Carterette adjectives are suitable for describing oboe timbre as subjects seem to be using each adjective differently. One drawback to having a limited number of results for the dissimilarity rating experiment is that only a two-dimensional solution could be plotted, whereas the PCA analysis of the VAME data suggested that three or four dimensions are needed to differentiate results. In summary, VAME ratings for each attribute generally distinguish between oboists at all dynamic levels. At extremes of pitch the VAME ratings are more similar for each performer suggesting that their tones are being perceived as being more alike.

4.2. Signal Processing

To quantitatively examine the dimensions of the timbre space obtained from the perceptual experiments, we have developed a timbre analysis toolbox to extract spectral and temporal features of tones. (For details of the descriptors please see [9].) The toolbox is implemented in Scilab (an open source signal processing environment) [10]. The analysis functions that extract both the spectral and temporal timbral features have their origins in many previous studies on timbre. Vibrato tracking features using a method of autocorrelation are also implemented. Also included in the toolbox are implementations of the timbre descriptors used in the MPEG-7 specification [7] and the Kendall and Carterette studies [1][2].

4.2.1. Results

The spectral descriptors correlate the above perceptual results with the results of the toolbox, thereby evaluating the performance of the MPEG-7 descriptors. Three main results can be summarised thus: spectral centroid and spectral deviation (from MPEG-7 descriptors) correlated best with the 'Power' factor, spectral flux and centroid variability (from Kendall and Carterette descriptors) having less significant correlations; spectral spread and spectral variation (from MPEG-7 descriptors) with the 'Vibrancy' factor, and spectral centroid and spectral variation with the 'Pinched' factor. Significant attribute correlations are Rich with spectral centroid and spectral deviation (positive), Ringing and Tremulous with spectral spread (negative), Light for spectral flux (positive), Rich and Reedy for spectral flux (negative), Rich for the centroid variability (negative).

Further investigation is recommended using both oboe-specific verbal and computational descriptors, e.g., vibrato variations, and formant structure, to obtain more accurate features of timbre relating to the oboe that correlate with instrument-specific verbal descriptors. Future work, leading to expanding the MPEG-7 timbre descrip-

tor set, needs to examine timbres from a number of different instrumen.

5. CONCLUSION

We have seen that multiple representations of timbre are extremely helpful in terms of analysis, and that one should be aware of all of the assumptions going into a computational signal processing system. The semantic descriptions in this current work were used to describe all instruments and were taken, by Kendal and Carterette [1], from Piston's work on Orchestration [11]. We have seen a way to move beyond the current way of thinking within the MPEG-7 standard, by providing more semantic 'axes' for navigating within the representational space formed by computational timbre descriptors. Most significantly this research highlights the importance of exploring timbres from the same-instrument class to further examine relationships between perceptual and computational descriptors.

We believe that, with further verification, the 'Power,' 'Vibrancy,' and 'Pinched' groupings of descriptors could be used as a higher-level, oboe-specific description, atop the existing MPEG-7 timbre descriptors. This layering of descriptions is entirely consistent with both the MPEG-7 approach and the staircase model described above. The adjective groupings could be used to make a user interface more intuitive.

5.1. Future Work

With the continuation of the psychological statistics and signal processing analyses, there are many potential directions this research could take. The techniques used in this very restricted, single-instrument domain could be examined to see if they could be transferred to other instruments, or made more general again. The MPEG-7 standard may be enhanced with this further research, potentially included in a second version of the standard. In any case, we believe this research to add to the repertoire of representations for anyone attempting to work with timbre – whether perceptually or computationally.

6. REFERENCES

[1] R. A. Kendall and E. C. Carterette, "Verbal attributes of simultaneous wind instrument timbres: II. Adjectives induced from Piston's Orchestration," Music Perception,vol. 10, no. 4, pp. 469 – 502, 1993.

[2] R. A. Kendall and E. C. Carterette, "Perceptual and acoustical features of natural and synthetic orchestral instrument tones," Music Perception, vol. 16, no. 3, pp. 327-364, 1999.

[3] E. M. von Hornbostel and C. Sachs, "Classification of musical instruments," Journal of the Galpin Society, vol. 14, pp. 3 – 29, 1961. (trans. by A. Baines and K. P.Wachsmann. Original work published in 1914).

[4] ANSI, "American national standard: Psychoacoustical terminology. timbre," Tech. Rep. ANSI S3.20-1973, American National Standards Institute, 1973.

[5] R. Plomp, *Aspects of Tone Sensation: A Psychophysical Study*. London: Academic Press, 1976.

[6] D. Marr, *Vision*. San Francisco: Freeman, 1982.

[7] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of MPEG-7," in Proceedings of the ICMC 2000, (Berlin), International Computer Music Conference, August 2000.

[8] ISO/IEC JTC1/SC29/WG11 (MPEG), "Multimedia Content Description Interface Part 4: Audio," International Standard 15938-4, ISO/IEC, 2002.

[9] R. A. Fitzgerald *Performer-dependent dimensions of timbre: identifying acoustic cues for oboe tone discrimination* PhD Thesis, School of Music, University of Leeds, UK 2003

[10] http://scilabsoft.inria.fr/

[11] W. Piston, *Orchestration*. London: Gollancz. 1991