Semantic Modelling for User Interaction with Sonic Content

António Sá Pinto*
antoniosapinto@gmail.com
Faculdade de Engenharia da Universidade do Porto
Porto

Matthew E.P. Davies
mdavies@inesctec.pt
Sound and Music Computing Group, INESC TEC
Porto

Perfecto Herrera
perfecto.herrera@upf.edu
Music Technology Group, Universitat Pompeu Fabra
Barcelona

Abstract

In this paper we present a methodology for converting semantic descriptions of sounds into computable audio features. This process aims to enable the use of commonly used notions of timbre in an audio engineering context where the user interacts (e.g. searches for sounds in large digital collections) with sonic content, bridging the gap between the high-level perceptual sound notions and low-level machine-ready descriptors. Although our focus is the description of the constituent blocks of a general-purpose semantic framework, examples from an experimental test for the semantic characterization of drum samples will be given to illustrate the process.

1 Introduction

The semantic gap that separates human descriptions of sounds from computable definitions is a recognized topic in the Music Information Retrieval (MIR) field of research [9]. Due to escalating data availability, and the potential and natural appeal of using common vocabulary for interacting with this abundance of sound resources, semantic audio studies have been increasingly addressed by the MIR scientific community. Accordingly, multiple applications have been proposed, such as active listening [11], music recommendation [2], sound retrieval [12] or intelligent audio production [10].

Our focus is the use of semantic descriptors (adjectives) for user interaction with sonic content, in a music production environment, where tasks such as browsing, retrieval or identification of samples are typical. Rather than detailing a specific system or application, we aim at identifying the building blocks required to enable the use of (high-level) human vocabulary by the user, dismissing the need for the extensive training required for interaction with sonic content in the (low-level) machine-extractable acoustic space.

2 Methodology

In this section we explain the proposed method for mapping semantic notions of sound into acoustic-based computable parameters, and the underlying procedures of extracting the audio signals features and creation of a semantic lexicon (a three-tier scheme is presented in Figure 1). Our focus lies on the description of a general methodology, but in this paper we illustrate the approach via the semantic characterization of drum samples.

2.1 The Acoustic (and Psycho-Acoustical) Space

Audio signals are computationally described in terms of audio features. In the case of isolated notes of musical instruments, for each of a set of samples ($A_1,...,A_n$), a vector of attributes ($F_1,...,F_n$) is obtained, which provides a compact representation of each of the sonic instances. Audio feature extraction is a cornerstone of audio signal processing, thus a consolidated body of work has been produced in this area (e.g. [8]). This process is achieved through the application of digital signal processing techniques directly to the raw audio signal, whether in the time-domain, wavelet, constant-Q or other spectral domains. There are countless audio features, several domain-oriented taxonomies, which group descriptors by their nature under distinct categories or related standpoints: spectral vs temporal, attack vs sustained vs decay, energy-related, perceptual, etc. Feature Selection and Transformation are normally the following processing blocks, aiming to refine and adapt the feature extraction process to the subsequent tasks in a specific target-application.

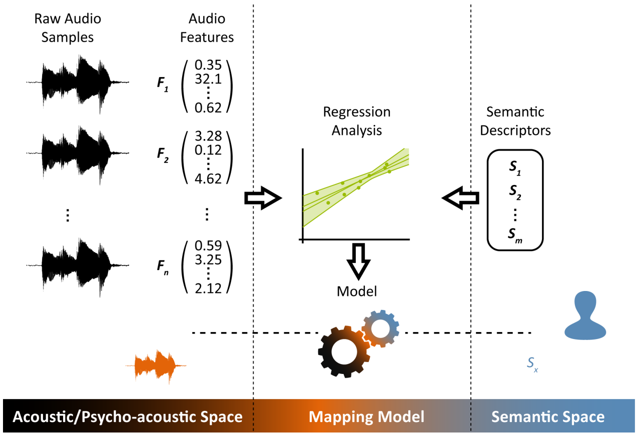

Figure 1: General Methodology Scheme

In our study, we use the 48 features which are fitted to percussive instruments [5]. In addition, four global descriptors were added to our final set: the average and variance of the Spectral Centroid and Spectral Flux, due to their reported relevance in the perceptual (and consequently semantic) space [7].

2.2 The Semantic Space

Terms such as *bright* or *warm* are common to experts when describing the sound of an instrument, and are among reported collections of verbalizations for the description of timbre of musical instruments [4]. However in common usage, such terms are less precise in their meaning, and while this may not prejudice the comprehensibility of a conversation about timbre, it disrupts the development of a computational semantic-based system. Therefore, to accomplish this goal, a consistent lexicon of semantic descriptors must be developed, in order to establish what will become the interface "language" between the user and the system: expert verbalizations of sound descriptions must be collected, and carefully reduced to a consistent subset, whose reliability must be confirmed through appropriate statistical validation techniques (e.g. the Chronbach's Alpha Coefficient).

Despite the existence of several instrument-specific studies, a significant gap still exists in the literature for percussive instruments. Only two systematic studies addressed the constitution of a drum timbre lexicon have been made [1, 3]; nevertheless, they did not establish a body of percussive semantic adjectives commonly accepted by the community as a reference lexicon. Given this absence, an initial collection of timbral adjectives was built upon the referred percussive lexicons, complemented by terms collected from other sources (e.g. e-commerce drum samples websites and an inquiry addressed to percussionists). From this large collection (more than 700 terms), a final subset of five sonic attributes was obtained as a "common denominator" from these several sources (preceding the reliability measurement, the adopted criteria was the unambiguity of their meanings and the coverage of salient sonic perception aspects, balanced with the additional statistical processing effort): (1) Brightness: the quality in sound of being clear, vibrant, and typically high-pitched; (2) Hardness: the quality in sound of being firm, rigid, stiff; (3) Tone (Sensation): the sound provokes a tonal sensation (pitch); (4) Size: the apparent external size, form of the sound source; (5) Ambiance: the conditions or atmosphere in which the sound was produced are explicit (e.g. reverb).

*The reported work was part of the Master's thesis done in the MTG-UPF.

2.3 Mapping the Semantic and Acoustic Space

The final component for a semantic-based system is the one that maps the two parameter spaces: the semantic space, which reflects the user-perceived prominent timbral aspects, and the underlying acoustic and psycho-acoustic features extracted from the sound samples. Given this pivotal role, global design decisions are hereby assembled and disclosed by a thorough analysis: e.g. the suitability of the set of features chosen to characterize the acoustic space or the reliability of the group of semantic descriptors. From its examination, we may draw key findings, namely which sonic cues play a relevant role for each semantic descriptor.

At this stage, we aim to model the relationship between a dependent variable (each of the semantic descriptors) and a group of independent variables (the audio descriptors):

$$S = f(A) \tag{1}$$

For this purpose, a regression analysis has been applied. These techniques provide a set of coefficients for a function that best fits predefined data observations, thus mapping acoustic features into the semantic space. Formally, given $(x_i, y_i), i \in 1,...,N$ a set of N pairs, where x_i is a $1 \times M$ feature vector and y_i is the real semantic rate value to predict, a regressor r is defined as the function that minimize the mean squared error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - r(x_i))^2 \tag{2}$$

We tested four different state-of-the-art regression techniques: Linear, Support Vector, Random Forest and Radial Basis Function (RBF). A brief performance analysis will be presented in the following sub-section. As an example of the mapping model (eq. 1), we highlight the five most significant terms of the Linear Regressor for the Brightness descriptor.

$$\begin{aligned} Brightness =& 3.6 * dSpecCentroid + 2.9 * dSkewness + \\ & 2.3 * dKurtosis - 1.8 * mfcc_{ave05} - 1.8 * RED5 \end{aligned} \tag{3}$$

where $dSpecCentroid$ represents the audio descriptor decay spectral centroid, $dSkewness$ the decay skewness, $dKurtosis$ the decay kurtosis, $mfcc_{ave05}$ the average of the 5th band (of 13) of the mel-frequency cepstrum coefficients (MFCC) representation of the signal spectral envelope, and $RED5$ represents the 5th (of 8) band energy relative percent of the signal, as defined in [5].

2.4 Experimental Results

In order to validate the proposed methodology, a semantic listening experiment applied to drum samples was undertaken, by means of a verbal attribute magnitude estimation (VAME) questionnaire [6], in which 47 (expert) subjects were asked to rate a set of 30 percussive samples (A_1,...,A_{50}), using a 6-point ordinal scale for each of the 5 semantic verbalizations (S_1,...,S_5). In parallel, the aforementioned set of acoustic descriptors (F_1,...,F_{52}) were computed for each audio signal, and after feature transformation and selection, different sets were included for evaluation. Following a 10-fold cross-validation procedure, the regressor's performance evaluation was obtained in terms of R^2 index (the squared correlation coefficient, a standard metric for measuring the accuracy of the fitting of the regression models) and the correspondent MSE. The accuracy was promising in many cases (reaching values of 0.695 for the R^2 score); yet, distinct improvement paths pave the way for a robust model to be built (e.g. coping with the significant inter-rater disagreement). After an informal comparison to other related studies [13], an auspicious overall performance was confirmed.

Although a general performance trend could be identified (a slight pre-eminence of the Random Forest, followed by the RBF), it is not yet possible to define a regression method that well suits all the high-level descriptors, given its dependency on several concurrent issues, such as the audio features set or the selected semantic descriptors. This initial analysis highlights several important issues for consideration in future work, namely the cross-dependency on other experiment design-related factors: the unambiguity of the adjectives (e.g. Hardness and Tone codings were inconsistent among raters), as well as the quality (or expertise) of the ratings (a selection of the top-5 raters, selected under a thorough correlation analysis, approximately doubled one regressor's accuracy). On the other hand, our results infer the suitability of the set of features used to characterize the sound samples. Addressing a main goal of our experimental study (mapping the acoustic and the semantic space), our framework provided some expected results in line with the literature: the relevance of the spectral centroid to Brightness perception, the association between the descriptor Tone and MFCC representation, and the relevance of attack energy and log-attack time to the Hardness descriptor.

3 Conclusions

In this paper we have presented in the context of a practical application, a methodology for converting expert sound notions into computable definitions. This enables a semantic approach for the description of sonic content, that tries to approximate human perception when describing timbre using common adjectives. A general framework architecture has been proposed, and their processing blocks have been characterized. Some experimental examples were given to illustrate the methodology, and aided the discussion of some prominent issues; related both to procedure and implementation viewpoints. We were able to infer some relationships between semantic and acoustic descriptors, confirming some findings reported in reference timbre studies, while validating the presented methodology. Over our study span (here resumed), some interesting paths to further research have been identified (e.g. the use of onomatopoeias to describe percussive instruments, or the application of deep learning methods for our framework). In conclusion, the “semantic gap” represents an important obstacle to overcome, thereupon we foresee a key role for semantic approaches that target a wide range of relevant applications for interaction with sonic content.

References

[1] R. Bell. *PAL : The Percussive Audio Lexicon*. Doctoral dissertation, Swinburne University of Technology, Melbourne, Australia, 2015.

[2] D. Bogdanov, M. Haro, F. Fuhrmann, A. Xambó, E. Gómez, and P. Herrera. Semantic audio content-based music recommendation and visualization based on user preference examples. *Information Processing and Management*, 49(1):13–33, 2013.

[3] W. Brent. *Physical and Perceptual Aspects of Percussive Timbre*. Doctoral dissertation, University of California, San Diego, 2009.

[4] A. C. Disley, D. M. Howard, and A. D. Hunt. Timbral description of musical instruments. In *International Conference on Music Perception and Cognition*, pages 61–68, 2006.

[5] P. Herrera, J. P. Bello, G. Widmer, M. Sandler, O. Celma, F. Vignoli, E. Pampalk, P. Cano, S. Pauws, and X. Serra. SIMAC: Semantic Interaction with Musical Audio Content. In *The 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT)*, pages 399–406, 2005.

[6] R. Kendall and E. Carterette. Verbal attributes of simultaneous wind instrument timbres: I. von bismarck's adjectives. *Music Perception*, 10(4):445–467, 1993.

[7] S. Lakatos. A common perceptual space for harmonic and percussive timbres. *Perception & psychophysics*, 62(7):1426–1439, 2000.

[8] G. Peeters. A large set of Audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, 2004.

[9] X. Serra, M. Leman, and G. Widmer. A Roadmap for Sound and Music Computing. Technical report, 2007.

[10] R. Stables, S. Enderby, B. De Man, G. Fazekas, and J. Reiss. SAFE: A System for the Extraction and Retrieval of Semantic Audio Descriptors. In *Late Breaking Demo Session, 15th ISMIR Conference*, 2014.

[11] F. Thalmann, A. P. Carillo, G. Fazekas, G. A. Wiggins, and M. B. Sandler. The Semantic Music Player: A Smart Mobile Player Based on Ontological Structures and Analytical Feature Metadata. *Proceedings of the 2nd Web Audio Conference (WAC-2016)*, 2016.

[12] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Trans. Audio, Speech, Language Process.*, 16(2):467–476, 2008.

[13] M. Zanoni, F. Setragno, F. Antonnaci, A. Sarti, G. Fazekas, and M. Sandler. Training-based Semantic Descriptors modeling for violin quality sound characterization. In *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.