Semantic Modelling for User Interaction with Sonic Content

António Sá Pinto[1*], Matthew E.P. Davies[2], Perfecto Herrera[3]

[1] Faculdade de Engenharia da Universidade do Porto
[2] Sound and Music Computing Group, INESC TEC, Porto
[3] Music Technology Group, Universitat Pompeu Fabra, Barcelona

Abstract

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.
The experimental evaluation was achieved with a listening test for the semantic characterization of drum samples, from which results will be presented.

Method

- Compute the audio features from the sonic samples;
- Model the relationship between the acoustic and psycho-acoustic features with the semantic descriptors, through regression analysis methods

$$S = f(F)$$

- Regressor *r* is defined as the function that minimizes the mean squared error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(s_i - r(f_i))^2$$

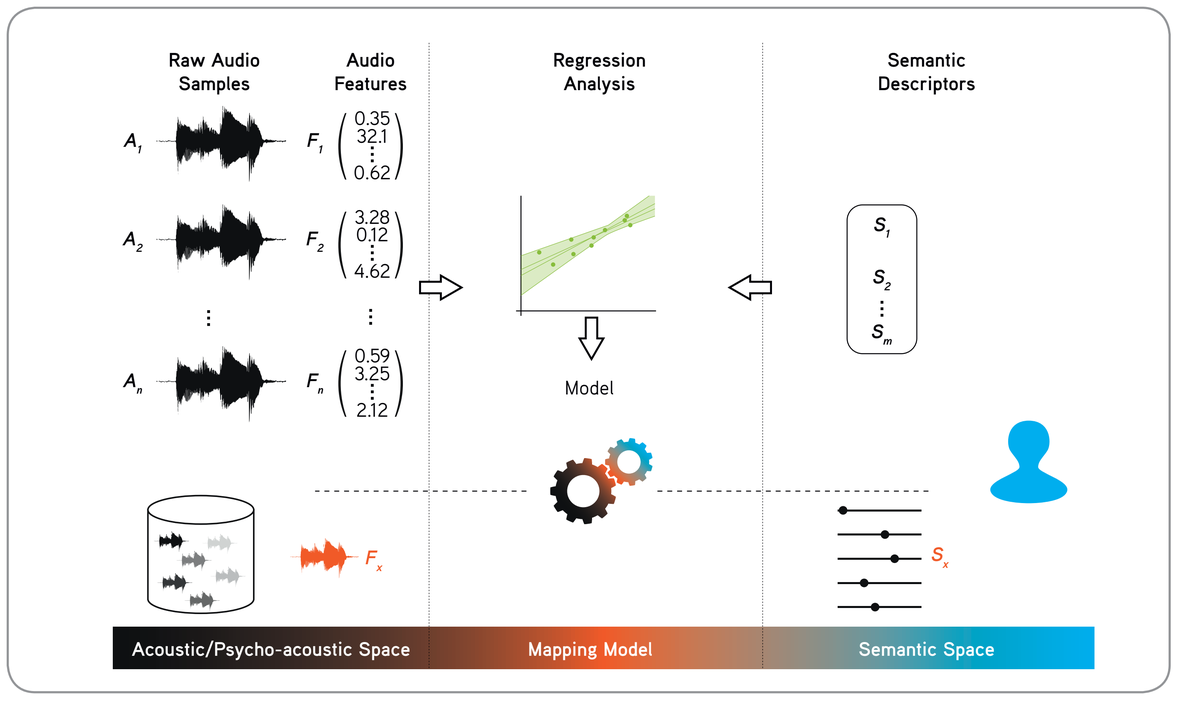
Implementation

- 48 audio features fitted to percussive instruments were computed, as defined in [1]. 4 global descriptors were added to the final set: the average and variance of the spectral centroid and spectral flux.
- Feature Transformation (e.g. *Log*) and Selection (e.g. Correlation-based feature selection) procedures were used to create different sets for evaluation. Instrument class and semantic descriptor information were added to further regression analysis.
- 4 regression techniques: Linear, Support Vector, Random Forest and Radial Basis Function (RBF). Following a 10-fold cross-validation procedure, performance was evaluated 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.

Evaluation

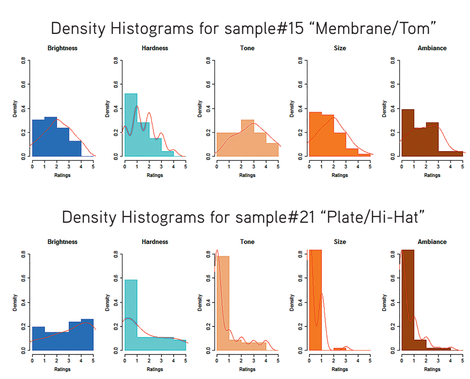
Verbal Attribute Magnitude Estimation (VAME) [2] Listening Test, where 47 musically trained participants (80.4% musicians, 54.3% percussionists, 37.0% experienced with DAW's) rated 30 percussive samples (2 types: membranes and plates; 7 classes: Kick, Snare, Tom, China, Hi-Hat, Crash, Ride) in a 6-point ordinal scale for the chosen semantic lexicon:

S_1 - **Brightness**: the quality in sound of being clear, vibrant, and typically high-pitched;
S_2 - **Hardness**: the quality in sound of being firm, rigid, stiff;
S_3 - **Tone** (Sensation): the sound provokes a tonal sensation (pitch);
S_4 - **Size**: the apparent external size, form of the sound source;
S_5 - **Ambiance**: the environment in which the sound was produced is explicit (e.g. reverb).



Results

- Slight variations in regressor's accuracy, but not significant enough to define a method that well suits all the descriptors (across different audio feature sets);
- Encouraging accuracy in comparison to related studies [4], reinforced by hampering differences in experimental conditions (e.g. number of raters, diversity of samples, overall complexity);
- High improvement of results (doubling accuracy in a case) when best subjects where selected;
- Confirmation of reported findings in timbre studies [3]: the relevance of spectral centroid to the perception of Brightness, the association between MFCC and the sensation of Tone, and the relevance of attack energy and log-attack time to the Hardness descriptor;
- Significant inter-rater disagreement, as well as intra-class divergence (e.g. for Hardness and Tone descriptors), which hinders firm inferences.



Descriptor	Regressor	All Subjects All Features (AF) R^2		Top-5 Subjects AF + Class R^2
Brightness	Linear	0.389	0.654	**0.695**
	RBF	0.388	0.677	0.688
Hardness	Linear	0.072	0	0.003
	RBF	**0.074**	0.003	0.005
Tone	Linear	0.181	0.395	**0.407**
	RBF	0.182	0.386	0.397
Size	Linear	0.393	0.509	**0.522**
	RBF	0.397	0.514	0.512
Ambiance	Linear	0.405	0.238	0.231
	RBF	**0.408**	0.25	0.256

Conclusions

Our studies demonstrated the adequacy of the semantic approach for the description of sonic content (namely for percussive instruments). Despite the promising results, statistical validity could not be assured due to the reported inconsistencies in ratings.
In this regard, we could ascertain the relevance of experimental design-related aspects to the overall performance and statistical validity: semantic descriptors must be unequivocal and suitable to a consistent set of sonic samples, the quality and reliability of the ratings; these are key factors for overall performance and statistical validity, assuming a pivotal role on a semantic framework.
In spite of the limitations, we were able to infer reported relationships between acoustic/psycho-acoustic descriptors and semantic attributes, confirming reference literature, while validating our framework and pave the way for future work.

References

[1] P. Herrera, A. Dehamel, F. Gouyon. Automatic labeling of unpitched percussion sounds. In *Audio Engineering Society Convention 114*. Audio Engineering Society, 2003.
[2] R. Kendall and E. Carterette. Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. *Music Perception: An Interdisciplinary Journal*, 10(4):445–467, 1993.
[3] S. Lakatos. A common perceptual space for harmonic and percussive timbres. *Perception & psychophysics*, 62(7):1426–1439, 2000.
[4] 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.

* The reported work was part of the Sound and Music Computing Master thesis at the Universitat Pompeu Fabra, Barcelona. antonio.sa.pinto@fe.up.pt