AN INVESTIGATION OF MUSICAL TIMBRE: UNCOVERING SALIENT SEMANTIC DESCRIPTORS AND PERCEPTUAL DIMENSIONS.

Asteris Zacharakis
Queen Mary University of London,
Centre for Digital Music,
London, UK.
asteriosz@eecs.qmul.ac.uk

Kostantinos Pastiadis, Georgios Papadelis
Aristotle University of Thessaloniki,
School of Music Studies,
Thessaloniki, Greece
pastiadi@mus.auth.gr

Joshua D. Reiss
Queen Mary University of London,
Centre for Digital Music,
London, UK.
josh.reiss@eecs.qmul.ac.uk

ABSTRACT

A study on the verbal attributes of musical timbre was conducted in an effort to identify the most significant semantic descriptors and to quantify the association between prominent timbral aspects and several categorical properties of environmental entities. A verbal attribute magnitude estimation (VAME) type of listening test in which participants were asked to describe 23 musical sounds using 30 Greek adjectives together with verbal terms of their own choice was designed and conducted for this purpose. Factor and Cluster Analysis were performed on the subjective evaluation data in order to shed some light on the relationships between the adjectives that were proposed and to conclude to the number and quality of the salient perceptual dimensions required for the description of this set of sounds.

1. INTRODUCTION

Musical timbre perception and its acoustical correlates have been a subject of research since the late 19th century [15]. During the last decades numerous studies on musical timbre have tried to uncover the number of significant perceptual dimensions and their semantic associations. Having applied different techniques most of these studies have concluded to either 3 or 4 major perceptual dimensions for modelling timbres of monophonic acoustic instruments and have also proposed a wide range of verbal attributes to label them. Grey in his state-of-the-art study in 1977 proposed a 3-D space for musical timbre representation by applying Multidimensional Scaling techniques to pairwise dissimilarity rating data [3]. Krumhansl and McAdams have also proposed a 3-D space [8], [9] whose physical correlates vary compared to the ones proposed by Grey.

von Bismarck conducted a semantic differential listening test featuring 30 verbal scales in order to rate 35 speech sounds [14]. According to this study timbre would have four orthogonal dimensions. One of the four von Bismarck's dimensions is associated with volume (*full-empty*), another one is a blend of vision and texture (*dull-sharp*), the third is labelled *colourful-colourless* and the last one is labelled *compact-diffused*. Other related studies also revealed three or four perceptual axes. Pratt and Doak, working with simple synthetic tones have proposed a 3-D space featuring a vision (*bright-dull*), a temperature (*warm-cold*) and a wealth (*rich-pure*) axis [11]. Štěpánek's study in the Czech language [13] reveals one dimension associated with vision (*gloomy-clear*), another one with texture (*harsh-delicate*), a third one with volume (*full-narrow*) and a last one with hearing (*noisy/rustle-'undefined'*). Moravec's work again in Czech language has also resulted to four perceptual axes related to vision (*bright/clear-gloomy/dark*), texture (*hard/sharp-delicate/soft*), volume (*wide-narrow*) and temperature (*hot/hearty - 'undefined'*) [10]. Finally, Howard's study in the English language [6] has uncovered four salient dimensions the first of which is a mixture of vision, texture, volume and temperature (*bright/thin/harsh-dull/warm/gentle*). The second one is labelled *pure/percussive-nasal*, the third is associated with the material of the sound source (*metallic-wooden*) and the fourth is related to the evolution in time (*evolving*).

Although there seems to be some agreement concerning the number and attributes of the timbre dimensions, some differences between studies do exist. Such inconsistencies could be due to the different experimental protocol used each time and also due to generalization of the findings that resulted from a particular 'sampling' of the vast timbre space. Thus, the selection of an appropriate set of sounds that will represent as much of the variance of the existing musical timbres as possible and at the same time will keep the duration of a listening test relatively short is crucial. This work addressed this issue by including a wide range of musical timbres with high ecological validity drawn from acoustic

instruments, electric instruments and synthesisers.

All of the cited studies have applied Factor Analysis and Cluster Analysis techniques in order to achieve dimension reduction of their multidimensional perceptual data. Factor analysis is a multivariate statistical technique that is used to uncover the latent structure of a set of inter-correlated variables [4]. It is widely applied in musical timbre research in order to reduce a large number of semantic descriptions to a smaller number of interpretable factors. Cluster Analysis is another statistical technique that seeks to identify homogeneous subgroups within a larger set of observations [12]. In the research on timbre perception it can indicate groups of semantically related verbal descriptors.

The current work has also made use of these data analysis techniques seeking for more definitive conclusions concerning the nature of the significant verbal descriptors of musical timbre. Overall, it aims at yielding a content analysis framework based on extramusical semantics.

2. METHOD

For the purpose of this study a listening test exploiting a variation of the Verbal Attribute Magnitude Estimation (VAME) [7] method was designed and conducted. The subjects were provided with a pool of 30 Greek verbal descriptors and were asked to describe timbral attributes of 23 sound stimuli by choosing the adjectives they believed that were more appropriate for each case. Once a subject chose a descriptor he was further asked to insert its amount of relevance on a scale anchored by the verbal attribute and its negation, such as " not brilliant - very brilliant". This rating was performed by a horizontal slider with a hidden continuous scale ranging from 0 to 100. The verbal descriptors used, were English language equivalents that are commonly found in timbre perception literature [1], [14], [2], [5] and are depicted in Table 1. The subjects were also free to insert up to three adjectives of their own choice for describing each stimuli in case they felt that the provided terms were inadequate.

2.1 Stimuli - Material

A set of 23 sounds of high ecological validity (acoustic instruments, electric instruments and state-of-the-art synthesisers) was selected. The following 14 instrument tones come from the MUMS (McGill University Master Samples) library: *violin, sitar, trumpet, clarinet, piano* at A3 (220 Hz), *double bass pizzicato, Les Paul Gibson guitar, baritone saxophone B flat* at A2 (110 Hz), *oboe* at A4 (440 Hz), *Gibson guitar, pipe organ, marimba, harpsichord* at G3 (196 Hz) and *french horn* at A3# (233 Hz). A *flute* recording at A4 was also used along with a set of 8 synthesiser sounds: *Acid, Hammond, Moog, Rhodes piano* at A2, *electric piano (rhodes), Wurltitzer, Farfisa* at A3 and *Bowedpad* at A4. The samples were loudness equalised with an informal listening test within the research team. The playback level was set between 65 and 75 A weighted dB SPL rms. 83% of the subjects found that level comfortable and 78% reported that loudness was perceived as being constant across stimuli.

The listening test was conducted in an acoustically isolated listening room. Sound stimuli were presented through the use of a desktop computer (Intel pentium 2.8 GHz, 1 GB Ram, WinXP(SP3)), with an M-Audio (Firewire 410) external audio interface, and a pair of Sennheiser HD60 ovation circumaural headphones. The interface of the experiment was built in Max/MSP.

2.2 Listening Panel

Forty one subjects (aged 19-55, mean age 23.3, 13 male) participated in the listening test. None of them reported any hearing loss and all of them were critical listeners and had been practising music for 13.5 years on average (ranging from 5 to 35). The majority of subjects were students at the Department of Music Studies of the Aristotle University of Thessaloniki. Course credit was offered as a reward for their participation.

2.3 Procedure

Initially the listeners were presented with a familiarisation stage which consisted of the random presentation of the stimuli set in order for them to get a feel of the timbral range of the experiment. For the main part of the experiment the playback of each sound was allowed as many times as needed prior to submitting a rating. The sounds were presented in a random order for each listener in order to minimize bias to the responses. Subjects were advised to use as many of the terms as they felt were necessary for an accurate description of each different timbre and also to take a break in case they felt signs of fatigue. They were also free to withdraw at any point. The overall listening test procedure, including instructions, lasted around 40 minutes for the majority of the subjects. The wide majority of subjects rated the above procedure as easy to follow, clear and meaningful.

2.4 Factor Analysis

Although the choice between Exploratory Factor Analysis (FA) or Principal Components Analysis (PCA) for data reduction has long been debated, we believe that FA is the appropriate choice for our investigation, as we focus on the identification of potential underlying structures that shall describe and justify the semantic representation of listeners' timbral experiences and judgements, across different musical sounds.

The basic FA model is described as:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jn}F_n + U_j = \sum_{i=1}^{n} a_{ji}F_i + U_j \quad (1)$$

where $j = 1 \ldots m$ or in matrix notation,

$$\mathbf{Z = A \cdot F + U} \quad (2)$$

where

$$\mathbf{Z}^T = \begin{bmatrix} z_1 & \cdots & z_m \end{bmatrix}$$

is the array of m analysed variables

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

is the matrix of *factor loadings* to be estimated from the data,

$$\mathbf{F}^T = \begin{bmatrix} F_1 & \cdots & F_n \end{bmatrix}$$

is the array of *n Common Factors*, and

$$\mathbf{U}^T = \begin{bmatrix} U_1 & \cdots & U_m \end{bmatrix}$$

is the array of *m Unique Factors*.

Actually, the problem and methodology of FA is to try to create, from a set of original variables, a new set of constructs (the common factors, with n < m) that will compactly describe the correlations between the original variables. Unique factors add to the versatility of the solution, as they account for that part of the original variance that cannot be attributed or modelled by the common factors.

3. RESULTS

The listeners' responses were analysed employing Cluster Analysis and Factor Analysis (FA). For this reason the quantity estimations on each verbal descriptor and each musical timbre were averaged over the 41 subjects of the test. Basic statistics for each descriptor are shown in Table 1.

Only 37% of the subjects inserted at least one extra verbal descriptor thus providing 36 additional terms. However, only 9 of them where mentioned more than once and only 4 were mentioned by more than one subject. This sparsity and inconsistency of the findings implies that our proposed set of 30 adjectives was adequate for describing this particular set of musical timbres.

As the distributions for most descriptors showed excessive positive skewness, a square root monotonic transformation was applied. Initially, the terms *empty, distinct, nasal* were removed following a bivariate correlation analysis over the 30 descriptors that was employed to identify and remove

Table 1. Basic statistics for each verbal descriptor.

Descriptor	Range	Mean	Descriptor	Range	Mean
Brilliant	25.68	8.63	Deep	59.93	10.82
Hollow	17.43	6.08	Distinct	34.34	11.65
Clear	48.39	8.76	Dry	24.00	8.13
Rough	33.45	8.47	Light	25.54	4.76
Metallic	39.17	14.02	Messy	39.73	4.90
Warm	23.66	9.01	Empty	36.80	6.93
Smooth	19.24	5.05	Dirty	41.51	8.60
Thick	47.32	8.26	Compact	17.22	7.91
Rounded	26.10	11.22	Dark	23.95	7.81
Harsh	25.88	9.48	Soft	34.32	6.14
Dull	30.41	10.93	Nasal	33.07	9.30
Thin	18.76	5.61	Full	35.90	13.50
Shrill	55.37	17.90	Dense	20.07	8.89
Cold	13.33	6.59	Bright	16.95	5.44
Sharp	36.31	10.96	Rich	20.49	6.68

those with several instances of low correlation coefficients (absolute value < 0.2), which could potentially reduce the validity of further dimensionality reduction analysis. A centroid Hierarchical Cluster Analysis based on squared Euclidean distances over the remaining 27 descriptors (Figure 1) identified 3 major clusters of descriptors, namely Cluster 1: *soft, light, warm, smooth, rounded, dull, rich, full, thick, deep, dense, dark, compact, hollow,* Cluster 2: *bright, brilliant, thin, clear,* Cluster 3: *shrill, sharp, rough, harsh, dirty, messy, dry, cold, metallic.* In order to further reduce the number of verbal descriptors, a preliminary Factor Analysis was performed within each cluster and those with absolute factor loadings [1] > 0.7 were selected for the subsequent final Factor Analysis.

For each cluster FA, Maximum Likelihood (ML) factor extraction with Oblimin rotation was employed. Maximum Likelihood estimation of factor loadings allows for sufficient, consistent and efficient representation of the FA's pattern matrix, under the provision of multivariate normality of the data, a condition for which special steps have been taken in this work (e.g. variable transformation).Traditionally, FA results in a reduced size description of correlations between the subjected variables using new 'combined' variables (the factors) which are designed and computed as mutually orthogonal. However, in several cases, orthogonality of factors could impede the interpretability of results by constituting an unexpectedly strict and excluding possibility. We believe that in this work we should relax the factors' or-

[1] Factor loadings are the correlation coefficients between variables and factors. The values of the factor loadings indicate how well a certain variable is represented by a particular factor and are crucial for the labelling and interpretation of the factors.

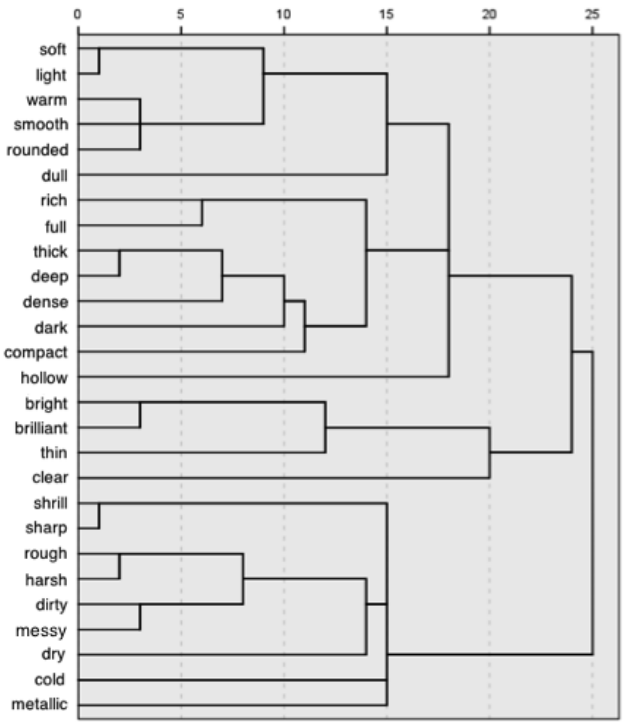

Figure 1. Dendrogram of the Hierarchical Cluster Analysis over the 27 descriptors.

thogonality requirement, and follow a conceptually 'wider' approach, by employing a non-orthogonal (oblique) rotation of the initial orthogonal solution. Later on, as it is usually preferred, it will be possible to check and justify the necessity for such a divergence from orthogonality requirements, by considering inter-factor correlations. The Direct Oblimin method (among others) is considered as a viable approach to the problem of oblique factor rotation.

Principal components extraction was used prior to factor extraction in order to determine the number of factors and ensure absence of multicollinearity. The Kaiser-Meyer-Olkin (KMO) [2] measure of sampling adequacy was for all three clusters bigger than 0.6 (Cluster 1: 0.672, Cluster2: 0.69, Cluster 3: 0.76), and the Bartlett's test of sphericity [3] also showed statistical significance. For each cluster, the first 3 factors were decided to be retained from the initial eigenvalues and the scree plots, accounting for more than 79% of cumulative variance. After factor extraction, the selected factors based on communalities [4] bigger than 0.6 were: Cluster 1: *soft, light, warm, smooth, rounded, rich, full, thick, deep, dense*, Cluster 3: *shrill, sharp, rough, harsh, dirty, messy, dry*. However, for the second cluster, a 3-factor solution could not be obtained and we decided to reduce the number of factors to 1, leading to retained descriptors as Cluster 2: *bright, brilliant*. In all 3 cases all eigenvalues were > 0.014, avoiding singularity.

The descriptors selected in the preliminary stage were then subjected to a final FA, again using ML and Oblimin rotation. The KMO measure was 0.654 and the Bartletts test of sphericity also showed statistical significance. Although singularity was again avoided, extreme multicollinearity was present leading to removal of 'culprit' descriptors. Next, the FA was repeated with a reduced set of 15 remaining descriptors. Again, 3 factors were extracted, accounting for more than 85% of initial variance. Although only messy and dirty had extracted communality < 0.6, for reasons of parsimony we additionally posed a criterion of absolute factor loading > 0.75 as a final step to data reduction. Maximum correlation between rotated factors was 0.249. The prominent descriptors over the three factors are shown in Table 2. Factor scores coefficients are given in Table 3. Multiplied by a sample's standardized measured score on the corresponding variables, these coefficients will sum to the score of a given sample on a given factor.

Table 2. Factor Loadings.

	Factor		
	1	**2**	**3**
Brilliant		-0.885	
Deep		0.824	
Soft			0.881
Full	0.851		
Bright		-0.946	
Rich	0.993		
Harsh			-0.861
Rounded			0.904
Thick		0.798	
Warm			0.787
Sharp			-0.779

Factor loading values are the basis for inputting a label to each of the different factors. A high factor loading indicates that a particular variable is expressed strongly by a certain factor. Based on Table 2, the three factors could be identified as Factor 1 *volume/wealth*, Factor 2 *brightness* and *density*, and Factor 3 *texture* and *temperature(warmth)*. Thus, it would seem possible to address musical timbre with semantic associations to material objects properties. It also seems, based on indications from the extracted variances, and since the oblique rotation results in relatively low levels

[2] The KMO assesses the sample size (i.e. cases/variables) and predicts if data are likely to factor well based on correlation and partial correlation. The KMO can be calculated for individual and multiple variables. KMO varies from 0 to 1.0 and KMO overall should be .60 or higher to proceed with factor analysis.

[3] Bartlett's test concerns whether correlations between variables are overall significantly different from zero.

[4] The communality measures the percent of variance in a given variable explained by all the factors jointly.

Table 3. Factor Scores Coefficients.

	Factor		
	1	**2**	**3**
Brilliant	-0.17	-0.121	0.020
Deep	-0.057	0.266	0.079
Soft	-0.035	0.098	0.160
Full	0.065	0.103	-0.022
Bright	-0.051	-0.286	0.079
Rich	0.898	-0.186	-0.099
Harsh	0.003	0.006	-0.106
Rounded	0.011	0.006	0.588
Thick	0.076	0.258	0.009
Warm	-0.000	0.006	0.065
Dense	0.18	0.052	0.003
Dry	-0.005	0.018	-0.057
Sharp	0.003	-0.043	-0.095

of correlation between factors, that all factors share some common and balanced portion (23%, 34% and 24% correspondingly) of the total explained variance ($\sim$ 82%), which by turn reveals a relatively equal importance of descriptors upon the timbral targets.

The low correlation between factors implies the existence of a nearly orthogonal perceptual space, thus a positioning of the 23 sound stimuli into a euclidean 3-D space seems justified and is shown in Figures 2, 3 and 4. Figures 3 and 4 reveal a noticeable influence of fundamental frequency on the brightness axis, as higher pitched sounds tend to be rated as brighter than lower pitched ones. A potential similar influence on the other two axis cannot be supported by these depictions.

4. DISCUSSION

The above findings share many things in common with results of previous studies -as presented in the introduction- both on the number and on the attributes of the uncovered timbre space dimensions. Indeed, *volume*, *wealth*, *texture*, *temperature* and *vision* related terms have also been attributed as labels to timbre space dimensions from previous research. Furthermore, most of the past studies result in perceptual spaces of either three or four dimensions for musical timbre representation. This agreement is present even among studies that apply different experimental protocols and methods for the creation of timbre spaces such as Multidimensional Scaling on data from pairwise dissimilarity listening tests or Principal Component Analysis for dimension reduction among perceptual variables. It is important, however, to emphasize the fact that the Factor Analysis applied on the variables (i.e adjectives) of this experiment was based on strictly mathematical criteria avoiding any bias from past studies results.

One other important outcome of the current work is that inter-dimension correlation is low. Consequently, even though the orthogonality requirement was not initially followed, as in most previous works, the result is still a nearly orthogonal space with independent dimensions.

A confirmatory study for examining the adequacy of the extracted perceptual dimensions regarding timbre description will be the next step for reaching the desired content analysis framework. The definition of such a framework will contribute towards a better understanding of musical timbre and can be used for the development of perceptual driven applications on musical sound modification and synthesis.

Finally, this study also positively adds to the concept of inter-linguistic agreement regarding musical timbre verbalization and proposes a certain rationale for the interpretation of the salient musical timbre space dimensions. The notion of timbre perception as being projected on other less abstract senses in order to facilitate expression and communication could in a sense justify the inter-linguistic agreement. The orientation of the human mind towards decoding and categorizing all incoming information to familiar entities could be responsible for the semantic associations to material objects that were revealed in this study.

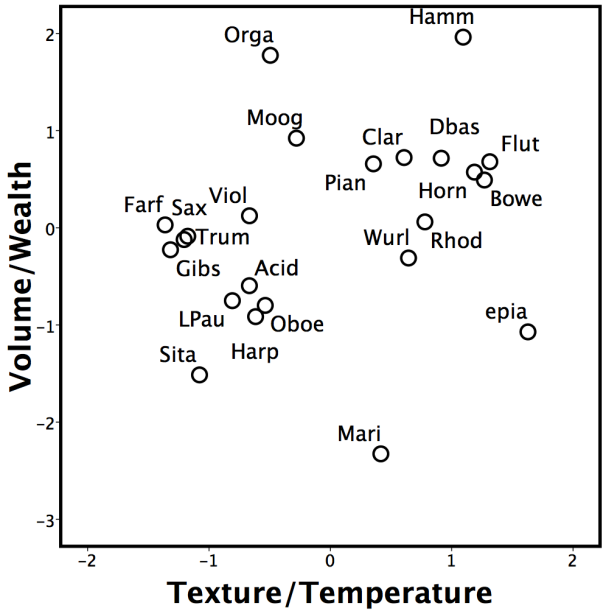

Figure 2. Volume/Wealth vs Texture/Temperature

5. CONCLUSION

In this paper, we have conducted an initial exploration of the possible underlying semantic structure of adjective timbral descriptors for musical sounds. Factor and Cluster Analysis applied on the subjective evaluation responses revealed

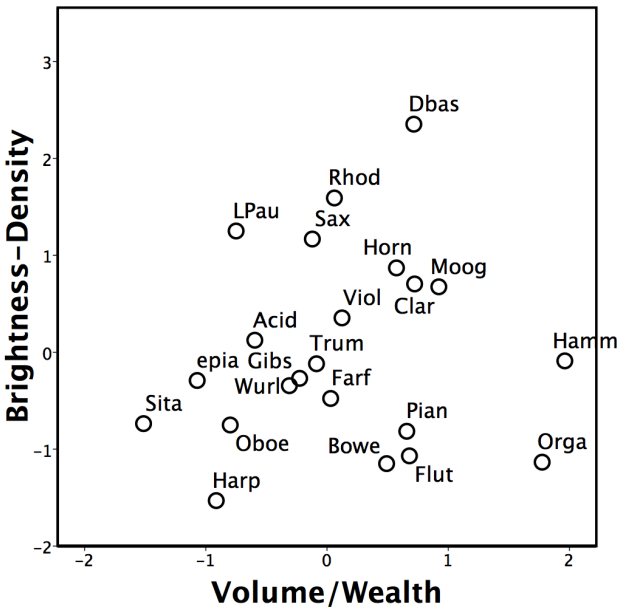

Figure 3. Brightness-Density vs Volume/Wealth.

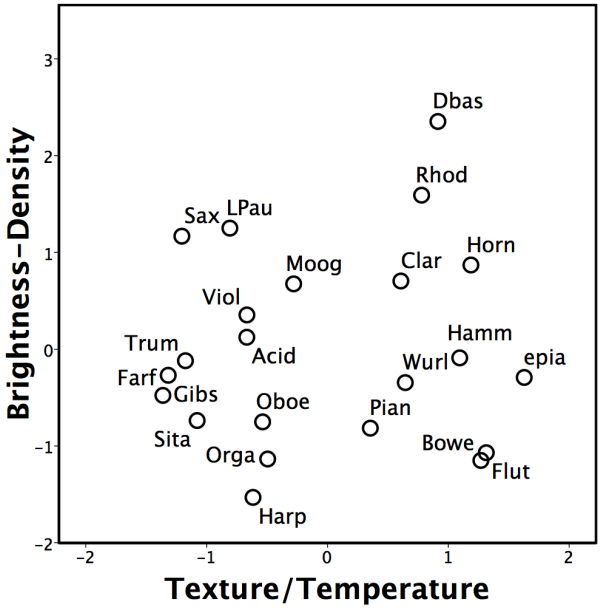

Figure 4. Brightness-Density vs Texture/Temperature.

three perceptual dimensions with high degree of independence that explained over 80% of the total variance. These dimensions are associated with material object properties such as *volume*, *brightness-density* and *texture-temperature* and constitute a framework for the semantic description of this particular set of sound stimuli. A further challenging issue is the conduction of confirmatory structural analysis (e.g. Confirmatory Factor Analysis) along different groups of sounds and/or different groups of listeners, since all aesthetic, stylistic and cultural factors could possibly affect the validity of the hereby developed semantic model. Subsequently, such a developed semantic framework could be deployed in a semantically driven framework of audio signal processing with application in musical sound synthesis, audio post-production or other similar fields.

6. REFERENCES

[1] R. Ethington and B. Punch. Seawave: A system for musical timbre description. *Computer Music Journal*, 18(1):30–39, 1994.

[2] A. Faure, S. McAdams, and V. Nosulenko. Verbal correlates of perceptual dimensions of timbre. In *Proc. 1996 Int. Conf. on Music Perception and Cognition*, pages 79–84, 1996.

[3] J.M. Grey. Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61:1270–1277, 1977.

[4] H. H. Harman. *Modern Factor Analysis*. University of Chicago Press, 3 edition, 1976.

[5] D. Howard, A. Disley, and A. Hunt. Timbral adjectives for the control of a music synthesizer. In *19th International Congress on Acoustics*, Madrid, 2-7 September 2007.

[6] D. Howard and A. Tyrrell. Psychoacoustically informed spectrography and timbre. *Organised Sound*, 2(2):65–76, 1997.

[7] R. A. Kendall and E. C. Carterette. Verbal attributes of simultaneous wind instrument timbres: I. von bismarck's adjectives. *Music Perception*, 4(10):445–468, 1993a.

[8] C. L. Krumhansl. Why is musical timbre so hard to understand? In S. Nielzén and O. Olsson, editors, *Structure and Perception of Electroacoustic Sound and Music: Proc. Marcus Wallenberg Symposium*, pages 43–53, Lund, Sweden, August 1988. Excerpta Medica, Amsterdam.

[9] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes. *Psychol. Res.*, 58:177–192, 1995.

[10] O. Moravec and J. Štěpánek. Verbal description of musical sound timbre in czech language. In *Proceedings of the Stockholm Music Acoustics Conference (SMAC03)*, pages 643–645, Stockholm, Sweden, 4-5 September 2003.

[11] R.L Pratt and P.E. Doak. A subjective rating scale for timbre. *Journal of Sound and Vibration*, 45, 1976.

[12] C. Romesburg. *Cluster Analysis for Researchers*. Lulu.com, 2004.

[13] J. Štěpánek. Musical sound timbre: Verbal descriptions and dimensions. In *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06)*, Monteral, Canada, 18-20 September 2006.

[14] G. von Bismarck. Timbre of steady tones: A factorial investigation of its verbal attributes. *Acustica*, 30:146–159, 1974.

[15] H. L. F. von Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. New York: Dover (1954), 4 edition, 1877.