Audio Engineering Society

Convention Paper

Presented at the 138th Convention
2015 May 7–10 Warsaw, Poland

Training-based Semantic Descriptors modeling for violin quality sound characterization

Massimiliano Zanoni[1], Francesco Setragno[1], Fabio Antonacci[1], Augusto Sarti[1], Gyorgy Fazekas [2], Mark Sandler[2] *

[1] *Politecnico di Milano, Milano, Italy*

[2] *Queen Mary University of London, London, UK*

Correspondence should be addressed to Massimiliano Zanoni (massimiliano.zanoni@polimi.it)

ABSTRACT

Violin makers and musicians describe the timbral qualities of violins using semantic terms coming from natural language. In this study we use regression techniques of machine intelligence and audio features to model in a training-based fashion a set of high-level (semantic) descriptors for the automatic annotation of musical instruments. The most relevant semantic descriptors are collected through interviews to violin makers. These descriptors are then correlated with objective features extracted from a set of violins from the historical and contemporary collections of the Museo del Violino and of the International School of Luthiery both in Cremona. As sound description can vary throughout a performance, our approach also enables the modelling of time-varying (evolutive) semantic annotations.

1. INTRODUCTION

The art of violin making begun in Cremona, Italy, five centuries ago and has grown to be what it is today thanks to the renowned families of Amati, Stradivari and Guarnieri. Cremona is currently home to over 150 violin makers, and thousands more have studied there and spread the tradition. In the year 2012 UNESCO crowned Cremona as a World Heritage Site for the art of lutherie confirming the leading role that this city has had for the tradition of violin making.

The study of the sound qualities of violins has been

*This research activity has been partially funded by the Cultural District of the province of Cremona, Italy, a *Fondazione CARIPLO* project, and by the Arvedi-Buschini Foundation

the subject of intense scientific investigation [1, 2] for decades. However, the physical phenomena that are involved in the characterization of their timbral quality are still far from being fully understood [3]. In past few years there has been a renewed frenzy in research, aimed at pushing the boundaries of our physical understanding of the quality of violin tone. This recently motivated a proliferation of research initiatives in the city of Cremona and the start of a new research projects with the Politecnico di Milano (for aspects of musical acoustics) and the University of Pavia (for aspects of material analysis), aimed at exploring new directions in contemporary lutherie.

Among the many goals of the projects are the investigation of the timbral quality of violins and, in particular, understanding the links that exist between objective and semantic descriptors related to such instruments. The former are geometric, vibro-acoustic, acoustic and timbral features; physical and chemical properties of materials, etc. The latter are the terms of natural language that are customarily used for describing qualities of the instrument.

In order to study the sound proprieties of musical instruments, one classical approach consists of extracting objective descriptors (*Low-Level Features - LLF*) [4, 5] and analyzing how such descriptors cluster up in feature space. As far as timbral characterization of violins based on low-level descriptors is concerned, some works have been presented in the literature. In [6, 7] the authors uses a set of MPEG spectral and harmonic descriptors for the characterisation of the violin sound quality. Whereas in [8], the author uses the long term cepstral coefficients. However, these descriptors are not semantically rich in nature, and do not match descriptions that are commonly used by violin makers and musicians (natural language). Examples of such terms are *warm* and *bright*, which are at a higher level of abstraction (*Semantic Descriptors or High-Level Features - HLF*). In the past decades, several studies have been presented in the literature [9, 10]. The main purpose of these studies is to build multi-dimensional perceptual spaces where semantic descriptors could be arranged. Similar approaches have been adopted also for the semantic description of the violin timbre [11, 12, 13, 14, 15].

Though our way of describing sounds is based on subjective Semantic Descriptors, there exists a strong connection between sound description, sound perception and physics. Our brain, in fact, processes stimuli from the auditory system in order to formulate a proper description. Understanding what aspects of the sound influence our perception [14] is not an easy task. For this reason, even if some remarkable work has been done [16, 17], this connection is still not fully understood. In the literature this is known as the *semantic gap between Low-Level and High-Level Features*.

In a previous work of ours [3], we studied the correlation between LLF and HLF using a set of correlation indices. In this study, we use machine learning techniques for modelling Semantic Descriptors using a large set of LLFs for automatic annotation and retrieval. In particular, we consider a generative approaches based on regression analysis, which was recently applied to Music Emotion Recognition [18, 19, 20] with very good results. In order to perform the mapping from LLF and HLF we explore parameter prediction using Multiple Linear Regression (MLR) [21], Ridge Regression [21], Polynomial Regression [21], Support Vector Regression (SVR) [22], Ada-boost Regression [23], Gradient Boost Regression[24].

In order to build the model for semantic descriptors we need to collect the low-level and the high-level representations of a large set of instruments. As far as the low-level representation is concerned we recorded thirteen historical violins (three Amati, two Guarnieri *del Gesù* and eight Stradivari) and fifteen modern violins from the collection of the “Museo del Violino” in Cremona and “International School of Lutherie” (Stradivari Institute) in Cremona, played by a professional musician according to a specific protocol. For each recording we extracted a large set of LLFs selected in order to capture timbral and harmonic proprieties of the instrument.

As far as HLFs are concerned, we collected the annotations by asking four professional violin makers to provide a description for each violin using a subser of the semantic descriptors presented in a previous work of ours [3]. In [3] we collected the set of most relevant terms used in lutherie to describe the sound of violins. In the listening test, each descriptor were presented along with its opposite (e.g. warm/not warm). The testers were asked to assign a graded annotation ranging from 0 to 1.

Although it is possible to provide an overall description of the sound quality of instruments, these proprieties tend to vary during a performance. Exploiting the short-time analysis, in this study we also use the regression approach in order to capture the evolution of the semantic descriptors over time.

2. LOW-LEVEL AUDIO FEATURES FOR MUSICAL INSTRUMENT CHARACTERIZATION

The study of timbral perception is still an open issue in music research. The ability of humans to discriminate, isolate and describe sounds has been subject of studies in many disciplines including psychology, sociology, acoustics, signal processing and music information retrieval. A comprehensive knowledge of the perceptual mechanisms involved in the human decision process is yet to achieve. However, many studies show how this tendency is mainly related to sets of simple acoustics and structural cues (LLF) [5, 25]. These cues are objective descriptors of sound that can be obtained by means of mathematical procedures. Each feature capture one specific aspect of the sound. In this study we are interested in understanding which cues are play a relevant role for each semantic descriptor.

The features that we select come from those extensively used in the music information retrieval field and exhaustively explained in [5, 25, 19].

In order to provide a measure of the noisiness of the sound the features that can be used are Zero Crossing Rate (ZCR), Spectral Flatness and Spectral Irregularity. The ZCR is defined as the normalized frequency at which the audio signal $s(n)$ crosses the zero axis. Spectral Flatness features are measures of the similarity between the spectral magnitude of the signal and the spectrum of a white noise signal (i.e. a flat spectrum).

As noisy signals tend to exhibit a weak correlation in the spectrum of successive temporal frame of analysis, Spectral Irregularity feature is used to capture the variation of the successive peaks of the spectrum, and it is defined as

$$F_{IR} = \frac{\sum_{k=1}^{K} (S_l(k) + S_l(k+1))^2}{\sum_{k=1}^{K} S_l(k)^2}, \tag{1}$$

where $S_l(k)$ is the magnitude spectrum at the l-th frame and the k-th frequency bin.

In order to provide a measure of the harmonicity we also consider Chromagram features. The Chromagram is a compact representation of the spectrum in the logarithmic scale. The spectrum is projected into 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.

Since part of the human perceptual process is still not well understood and since the process is mainly related to timbral characteristics, we include basic spectral descriptors to the set: Spectral Brightness, Roughness, Spectral Centroid, Spectral Kurtosis, Spectral Rolloff, Spectral Spread, Spectral Skewness, Mel-Frequency Cepstral Coefficients, Spectral Contrast. In particular, Spectral Roughness is an estimation of dissonance [26]. MFCC offer a compact representation of the spectrum, based on the human auditory model. They are obtained as the coefficients of the discrete cosine transform (DCT) applied to a reduced Power Spectrum. The reduced Power Spectrum derived as the log-energy of the spectrum is measured

$$c_i = \sum_{k=1}^{K_c} \log(E_k) \cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{K_c}\right] 1 \leq i \leq N_c, \tag{2}$$

where c_i is the ith MFCC component, E_k is the spectral energy measured in the critical band of the ith mel filter and N_c is the number of mel filters, K_c is the number of cepstral coefficients c_i extracted from each frame.

Spectral Contrast coefficients, which have been used in many MIR applications [18, 27], attempt to capture the relative distribution of the harmonic and non-harmonic components in the spectrum. The spectrum is divided in sub-bands, and the samples from each subb-and are sorted in descending order. At this point the peaks and spectral valleys of the i-th can be calculated as follow:

$$\mathrm{P}_i = \log\left\{\frac{1}{\alpha N_i}\sum_{j=1}^{\alpha N_i} s'_{i,j}\right\}, \tag{3}$$

$$\mathrm{V}_i = \log\left\{\frac{1}{\alpha N_i}\sum_{j=1}^{\alpha N_i} s'_{i,N_i-j+1}\right\}. \tag{4}$$

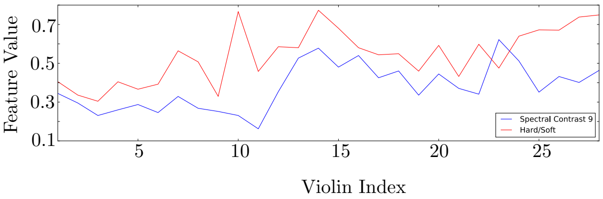

Fig. 1: Comparison of the distribution of the first sub-bands of SC feature and the Hard/Soft descriptor.

Finally, the Spectral Contrast can be calculated as their difference:

$$\mathrm{SC}_i = \mathrm{Peak}_i - \mathrm{Valley}_i, \tag{5}$$

where *alpha* is a corrective factor used in order to ensure the steadiness of the feature, $s'_{i,j}$ is the j-th sample of the sorted i-th sub-band and N_i is total number of samples in the j-th sub-band. In this study we keep both peaks, valleys and SCs as low-level descriptors (29 descriptors). Fig. 1 depicts the distribution of the first sub-bands of SC feature and the correspondent Hard/Soft descriptor for each instrument. The figure outlines the SC highly descriptive attitude for the Hard/Soft modeling since values of the two features has similar distribution.

The total number of LLFs that we use in this study is 59.

3. REGRESSION APPROACH

The goal of regression analysis is to model the relationship between a dependent variable and a set of independent variables of a formulated problem. From a different perspective, regression analysis includes a set of methods for discovering the set of coefficients for a function that best fits predefined data observations. According to the latter formulation, regressors have been recently widely applied as predictors in machine learning applications [18]. Indeed, they can be used to predict a real value from a set of observed variable by projecting a multidimensional feature space into a novel continuous space with a limited number of dimensions. In our case, for each semantic descriptor, the LLF space is mapped into a novel conceptual one-dimensional space of real values (HLF).

Formally, given $(\mathbf{x}_i, y_i)$, $i \in \{1, ..., N\}$ a set of N pairs, where $\mathbf{x}_i$ is a $1 \times M$ feature vector and y_i is the real HLF value to predict, a regressor $r(\cdot)$ is defined as the function that minimize the mean squared error (MSE) ϵ:

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}(y_i - r(\mathbf{x}_i))^2 \tag{6}$$

Based on this idea, several regression methods have been presented in the past few years. Since it is not clear the correlation between LLF and HLF, in order to discover the most appropriate method, in this study we use a set of regression functions resulted to be effective in many MIR applications [18, 27]: Multiple Linear Regression (MLR) [21], Polynomial Regression [21], Ridge Regression [21], Polynomial Regression [21], Support Vector Regression (SVR) [22], Ada-boost Regression [23], Gradient Boost Regression[24].

4. METHODOLOGY

The overall scheme of the method is depicted in Fig. 2. The figure shows the approach adopted for a single HLF and it follows a classic schema of a training-based technique. As described so far in this study, human attitude to sound discrimination and description is mainly based on acoustic cues and it is performed through spectral analysis. For this reason, the low-level characterization of each recording is provided through the extraction of the set of low-level features described in section 2. Each recording is then represented by a feature vector $\mathbf{x}_i \in \mathbb{R}^D$ where D is the number of features. In the training phase, the generative models (regressors) are trained on the high dimensional feature space computed on a training dataset of recordings. At this end, the regressors take as input a set of pairs $\langle \mathbf{x}_i; y_i \rangle$, where $y_i \in \mathbb{R}$ is the real value subjective annotation for the recording.

During the training, the regression processes aims at finding the hypersurface that best fits the data in order minimize the error in eq. 6. Whereas, in test phase, generated models are used to predict the real value label on a set of previously unseen recording.

Moreover, since some features are not informative for all the HLFs, feature selection methods can be

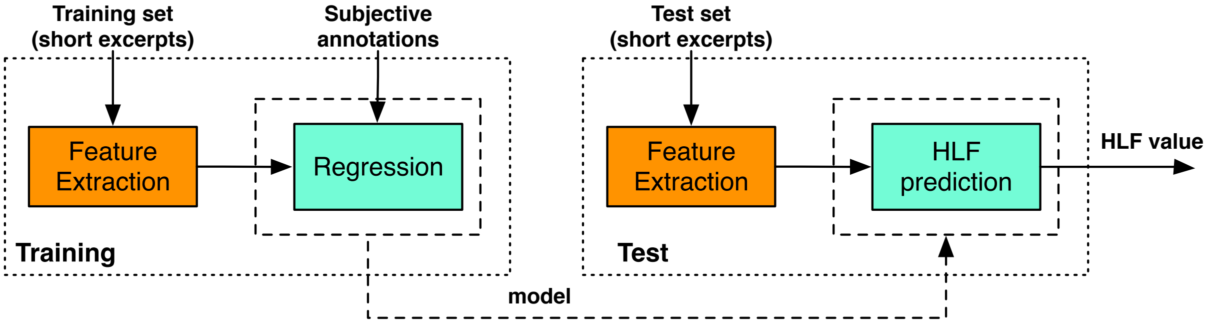

Fig. 2: General example-based regression learning schema. Models are the result of the training phase, performed over low-level features extracted by the excerpts in the training dataset and using the subjective annotation as the ground truth. Models are then used in the testing phase in order to analyze a previously unseen audio excerpt.

applied. To this end, in this study we used the Univariate Feature Selection algorithm that resulted to be very effective in music classification applications in the literature [28].

4.1. Data collection and Feature Extraction

The set of semantic descriptors used in this work represents the most used set of terms described in [3], which it has been obtained by several interviews to professional violin makers. The list of terms used in this study is shown in table 1.

Bright	Dark
Warm	Not Warm
Sweet	Harsh
Full	Not Full
Soft	Hard
Deep	Not Deep

Table 1: List of terms related to timbre used in this work. Terms in the same row and the same column are considered synonyms; terms in the same row but in different column are considered opposites.

With the intent to validate our method, a dataset of recordings has been conveniently collected. We recorded 28 violins of different qualities and ages: thirteen historical violins (three Amati, two Guarnieri del Gesu‘ and eight Stradivari) and fifteen modern violins from the collection of the Museo del Violino in Cremona and Scuola di Liuteria Istituto Stradivari in Cremona. Recordings have been performed in a semi-anechoic room using high-quality recording system and 44100 Hz as sample rate. A unique professional musician were performing for all the the recordings. In order to best emphasize the timbre characteristics of the instruments, the musician were asked to play a set short pieces of songs.

We collected the subjective annotation for each instrument through a listening test to 4 professional violin makers. For each pair of Semantic Descriptor in table 1, testers were asked to place the instruments on a mono-dimensional space. The position in the space represents how the violin is described by the two terms and corresponds to a real value ranging from 0 to 10. As an example, in figure 3, the violin 2 has been placed very close to Dark. This means that the timbre of the instrument is quite dark, it is darker than the violin 5 and it has assigned the value 1.1. The tester were allowed to listen the recordings of all the instruments. We computed the average of the annotations in order to obtain a single HLF value for each violin.

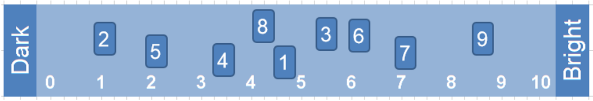

Fig. 3: Screenshot of the listening test related to a single HLF.

In order to enrich the dataset, we segmented each recording by extracting segments each 5 seconds with an overlap of the 60%. We considered each segment as an independent recording. The final dataset is composed by 500 segments, 70% used for compose the training dataset and 30% used for test dataset. The train dataset and the test dataset have been populated by randomly chosen segments.

The features have been extracted from each segment using the MIR toolbox [25].

5. EXPERIMENTAL RESULTS AND EVALUATIONS

Since we proposed to study the relation existing between acoustic cues and semantic descriptors, we are also interested in studying the contribute of different feature sets. More specifically, we performed the evaluation using the following groups: MFCC, Spectral Contrast, Chromagram, All (all the features), All+FS (use of a feature selection procedure applied to the whole set of features).

We evaluate the performance of the proposed regression approach in terms of R^2 index [21], which is a standard metric for measuring the accuracy of the fitting of regression models and in terms of Mean Squared Error (MSE). Let us notice that a negative value of R^2 means the prediction model is worse than simply taking the sample mean, whereas the value of R^2 represents the best performance.

The evaluation are collected in table 2. Let us notice that the feature selection procedure is not applied to ADABoost and GradientBoost cases, since they already include a feature selection method.

As shown in table 2, the overall performance is very prominent. The best results ($R^2 = 0.763$) are obtained combining the feature selection procedure applied on the whole set of features and the Linear Regression for the Hard/Soft descriptor. In general the overall accuracy is prominent (R^2 over 0.4). For the Dark/Bright descriptor the best result ($R^2 = 0.507$) is obtained computing the Polynomial regression using the feature selection procedure applied to the whole set of features. Feature selection results to be effective also for Hard/Soft descriptors where the best score is obtained using the Linear regression ($R^2 = 0.763$), which is the overall best result. For the Warm/Not Warm, Harsh/Sweet and Full/Not Full descriptors, the best score is obtained using Spectral Contrast features respectively using Ridge regression ($R^2 = 0.405$), ADABoost regression ($R^2 = 0.560$) and Polynomial regression ($R^2 = 0.594$). The MFCC features result to be the best solution only for the Deep/Not Deep descriptor by means of the SVR regression with the RBF kernel ($R^2 = 0.428$).

Let us provide some general consideration. Since less informative features can produce noise in the classification process, feature selection resulted to be very effective on almost all the cases. Moreover, Spectral Contrast features are very discriminant since obtained high score for all the HLFs. This confirms that the human ability to recognize bootlegs mainly relies on spectral cues.

In Fig. 5 and in Fig. 4 we present a pair of examples of the prediction for an historical violin and for a modern violin. The plots provide an intuitive description of the overall sound quality of the instrument. The annotations and the predictions are represented as curves in order to better outline the similarities.

The use of short segments for the training makes our method valid also for short-time analysis to capture the evolution of semantic descriptors along the performance. Fig. 6 shows that the method is effective also for small segments (1s).

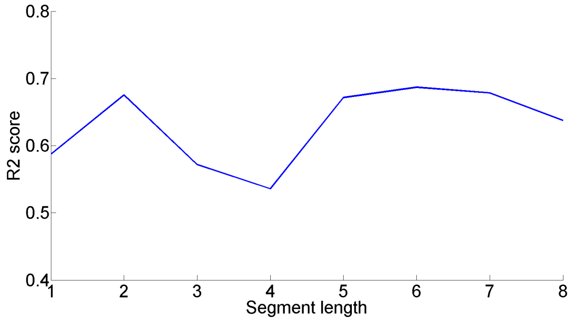

Fig. 6: R^2 score varying the length of segments for training and test dataset in the case of the Harsh/Sweet descriptor using ADABoost regression and Spectral Contrast.

6. CONCLUSIONS

		Chromagram		Spectral Contrast		MFCC		All		All+Selection	
Descriptor	Regressor	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2
Dark/Bright	Linear	1.712	0.075	1.090	0.411	1.175	0.365	0.950	0.486	1.031	0.443
	Ridge	1.703	0.080	1.276	0.310	1.177	0.364	0.928	0.498	1.046	0.435
	Polynomial	1.488	0.196	1.116	0.396	1.131	0.388	1.021	0.448	**0.912**	**0.507**
	SVR	1.406	0.240	1.385	0.251	1.303	0.296	1.003	0.458	1.436	0.224
	GradBoost	1.596	0.138	1.184	0.360	1.299	0.298	1.198	0.353	-	-
	ADABoost	1.715	0.073	1.347	0.272	1.554	0.160	1.360	0.265	-	-
Warm	Linear	0.820	-0.095	0.538	0.281	1.215	-0.624	0.762	-0.018	0.604	0.191
	Ridge	0.846	-0.131	**0.445**	**0.405**	1.078	-0.441	0.522	0.301	0.554	0.259
	Polynomial	0.759	-0.014	0.646	0.135	1.292	-0.727	0.807	-0.079	0.835	-0.116
	SVR	0.759	-0.014	0.647	0.134	0.997	-0.333	0.544	0.271	0.552	0.261
	GradBoost	0.890	-0.189	0.649	0.131	1.262	-0.686	0.592	0.208	-	-
	ADABoost	0.711	0.0492	0.532	0.289	1.429	-0.909	0.531	0.289	-	-
Harsh/Sweet	Linear	1.224	0.269	0.653	0.610	1.789	-0.067	1.230	0.265	0.820	0.510
	Ridge	1.232	0.264	0.845	0.495	1.711	-0.021	0.880	0.474	0.880	0.474
	Polynomial	1.185	0.292	0.873	0.478	1.670	0.003	1.348	0.195	0.911	0.456
	SVR	2.037	-0.215	1.514	0.096	1.503	0.103	1.312	0.217	1.384	0.173
	GradBoost	1.441	0.140	1.107	0.339	1.559	0.069	0.922	0.449	-	-
	ADABoost	0.926	0.447	**0.736**	**0.560**	1.862	-0.111	0.818	0.511	-	-
Full	Linear	1.687	0.327	1.024	0.591	2.124	0.153	2.071	0.174	1.033	0.588
	Ridge	1.762	0.297	1.197	0.522	2.016	0.196	1.266	0.495	1.186	0.527
	Polynomial	1.503	0.400	**1.017**	**0.594**	2.012	0.198	1.959	0.219	1.074	0.571
	SVR	2.627	-0.047	2.217	0.116	1.874	0.252	1.701	0.321	1.222	0.512
	GradBoost	1.643	0.344	1.392	0.445	1.914	0.236	1.359	0.458	-	-
	ADABoost	1.451	0.421	1.289	0.486	2.220	0.114	1.249	0.501	-	-
Hard/Soft	Linear	0.696	0.417	0.306	0.743	0.872	0.269	1.230	0.265	**0.281**	**0.763**
	Ridge	0.735	0.383	0.428	0.641	0.912	0.236	0.880	0.474	0.450	0.623
	Polynomial	0.607	0.493	0.329	0.724	0.837	0.298	1.348	0.195	0.318	0.734
	SVR	0.641	0.463	0.422	0.646	0.841	0.294	1.312	0.217	0.526	0.560
	GradBoost	0.723	0.395	0.566	0.526	0.830	0.305	0.922	0.449	-	-
	ADABoost	0.570	0.523	0.541	0.546	1.151	0.035	0.818	0.512	-	-
Deep	Linear	1.773	0.131	1.257	0.384	1.144	0.292	1.737	0.148	1.415	0.306
	Ridge	1.803	0.116	1.182	0.421	1.523	0.253	1.198	0.412	1.202	0.411
	Polynomial	1.679	0.177	1.288	0.369	1.380	0.323	2.012	0.016	1.370	0.328
	SVR	2.234	-0.094	1.756	0.139	**1.167**	**0.428**	1.902	0.067	1.592	0.220
	GradBoost	1.926	0.055	1.707	0.163	1.746	0.144	1.552	0.239	-	-
	ADABoost	1.826	0.105	1.658	0.187	1.962	0.038	1.494	0.267	-	-

Table 2: Performance for each regressor expressed with the R^2 score and the MSE

In this work we modeled a set of high-level descriptors for violin timbre, employing regression techniques typically used in machine learning and low-level audio features. The descriptors have been collected by means of interviews to violin makers and the ground truth came from a listening test where the subjects had to annotate every violin with the collected descriptors.

The results highlighted important aspects of timbre perception. As we imagined, only features related to spectral components achieved good performances (regression scores obtained using the Chromagram were low). Moreover, the use of feature selection techniques improved the results, since the presence of useless features made the data noisier. The accuracy was satisfying in many cases, reaching values of 0.76 for the R^2 score and 0.28 for the MSE. Finally, it is not possible to define a regression method that well suits all the high-level descriptors: each descriptor needs a specific method to be designed and tuned.

With our model it is possible to predict the highl-level timbral description of an instrument, starting from a recording. We also showed that with the right setting we can perform a time-varying prediction, by segmenting the audio file and processing each segment separately.

In future studies, new low-level features, specifically designed for violin sound analysis, will be tested.

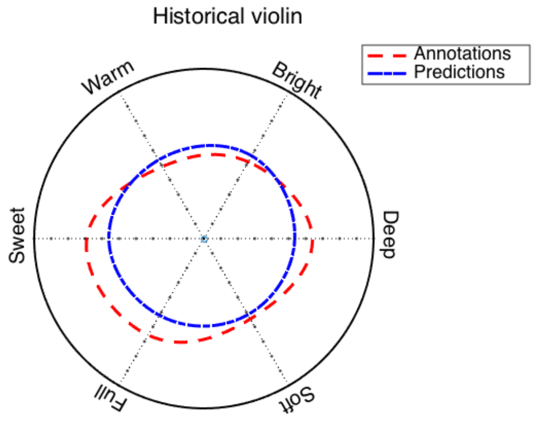

Fig. 4: Circular HLF description for an historical violin

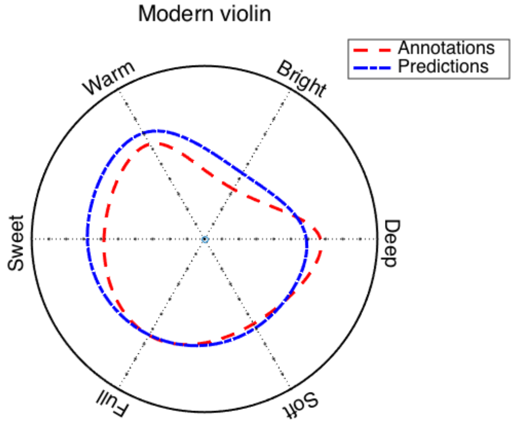

Fig. 5: Circular HLF description for an modern violin

Moreover, since the feature selection process is very complex and important, we want to test other selection algorithms.

The semantic gap represents an arduous obstacle in the study of sound perception. Nevertheless, this work can be considered a further step toward the comprehension of the relations that exist between physical attributes of violin sounds and the description of its timbre.

7. ACKNOWLEDGEMENTS

The authors are grateful to the Violin Museum Foundation, Cremona, Italy, for supporting the acquisitions activities on historic violins. We are also grateful to the Stradivari International School of Lutherie (particularly to Prof. Alessandro Voltini) for their continuous support with timbral acquisitions of their violins. We would also like to thank the violin players that helped us produce the audio data for the analysis and, in particular, the extraordinary violinist Anastasiya Petryshak for her patient work with us.

8. REFERENCES

[1] C.M. Hutchins. A history of violin research. *The Journal of the Acoustical Society of America*, (73):1421–1432, 1983.

[2] Jim Woodhouse. The acoustics of the violin: a review. *Reports on progress in physics. Physical Society (Great Britain)*, 77(11):115901, November 2014.

[3] M. Zanoni, F. Setragno, and A. Sarti. The violin ontology. In *In proceedings of the 9th Conference on Interdisciplinary Musicology (CIM14)*, Berlin, Germany,, 2014.

[4] M. Casey. Mpeg-7 sound recognition tools. In *IEEE Transactions on Circuits and Systems for Video Technology,*, volume 11, pages 737–747, 2001.

[5] T. Sikora H.G. Kim, N. Moreau. *MPEG-7 Audio and Beyond. Audio Content Indexing and Retrieval,*. John Wiley & Sons Ltd, 2005.

[6] A. Kaminiarz and E. Lukasik. Mpeg-7 audio spectrum basis as a signature of violin sound. In *In proceedings of the European Signal Processing Conference (EUSIPCO)*, 2007.

[7] J A Charles, D Fitzgerald, and E Coyleo. Violin Timbre Space Features. In *Irish Signals and Systems Conference, 2006. IET*, pages 471–476, 2006.

[8] E. Lukasik. Long term cepstral coefficients for violin identification. In *In proceedings of the Audio Engineering Society Convention 128 (AES128)*, 2010.

[9] Asterios Zacharakis, Konstantinos Pastiadis, and Joshua D. Reiss. An investigation of musi-

cal timbre: uncovering salient semantic descriptors and perceptual dimensiopns. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, 2011.

[10] Alastair C. Disley, David M. Howard, and Andy D. Hunt. Timbral description of musical instruments. In *9th International Conference of Music Perception and Cognition*, 2006.

[11] Charalampos Saitis, Bruno L. Giordano, Claudia Fritz, and Gary P. Scavone. Perceptual evaluation of violins: A quantitative analysis of preference judgments by experienced players. *J Acoust Soc Am.*, 2012.

[12] Charalampos Saitis, Claudia Fritz, Catherine Guastavino, Bruno L. Giordano, and Gary P. Scavone. Investigating consistency in verbal descriptions of violin preference by experienced players. In *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennal Conference of the European Society for the Cognitive Sciences of Music*, 2012.

[13] Jan Štěpánek. Evaluation of timbre of violin tones according to selected verbal attributes. In *32nd International Acoustical Conference*, 2002.

[14] Massimiliano Zanoni, Daniele Ciminieri, Augusto Sarti, and Stefano Tubaro. Searching for dominant high-level features for music information retrieval. In *20th European Signal Processing Conference (EUSIPCO 2012)*, 2012.

[15] C. Fritz, A. F. Blackwell, I. Cross, B. C. J. Moore, , and J. Woodhouse. Investigating english violin timbre descriptors. In *In Proceedings of the 10th International Conference on Music Perception & Cognition (ICMPC 10)*, 2008.

[16] Jan Štěpánek. Musical sound timbre: Verbal description and dimensions. In *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06), Montreal, Canada*, 2006.

[17] Rie Hirai, Kajiro Watanabe, Kazuyuki Kobayashi, and Yosuke Kurihara. Measurement and Evaluation of Violin Tone Quality. In *SICE Annual Conference (SICE)*, pages 2790–2793, 2011.

[18] E. Schmidt, D. Turnbull, and Y. E. Kim. Feature selection for content-based , time-varying musical emotion regression. In *In proceedings of the International Conference on Multimedia Information Retrieval*, 2010.

[19] YH Yang, YC Lin, and YF Su. A Regression Approach to Music Emotion Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):448–457, 2008.

[20] S Rho and Byeong-jun Han. Svr-based music mood classification and context-based music recommendation. In *MM '09 Proceedings of the 17th ACM international conference on Multimedia*, pages 713–716, 2009.

[21] A. Sen and M. Srivastava. *Regression analysis theory methods and applications*. Springer, New York, 1990.

[22] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. *Springer: Statistical Computing Journal,*, 14(3):199222, August 2004.

[23] D.P. Solomatine and D.L. Shrestha. Adaboost.rt: a boosting algorithm for regression problems. *Proceedings of the IEEE International Joint Conference on Neural Networks*, pages 1163–1168, 2004.

[24] Richard S. Zemel and Toniann Pitassi. A gradient-based boosting algorithm for regression problems. In *In Advances in Neural Information Processing Systems*, pages 696–702, 2001.

[25] O. Lartillot and P. Toiviainen. Mir in matlab (ii): A toolbox for musical feature extraction from audio. In *2007 International Society for Music Information Retrieval conference (ISMIR)*, 2007.

[26] W. A. Sethares. *Tuning, Timbre, Spectrum, Scale*. London: Springer-Verlag., 1998.

[27] Lie Lu, Dan Liu, and H.J. Zhang. Automatic mood detection and tracking of music audio signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):5–18, 2006.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2011.