ISSN 2043-0167

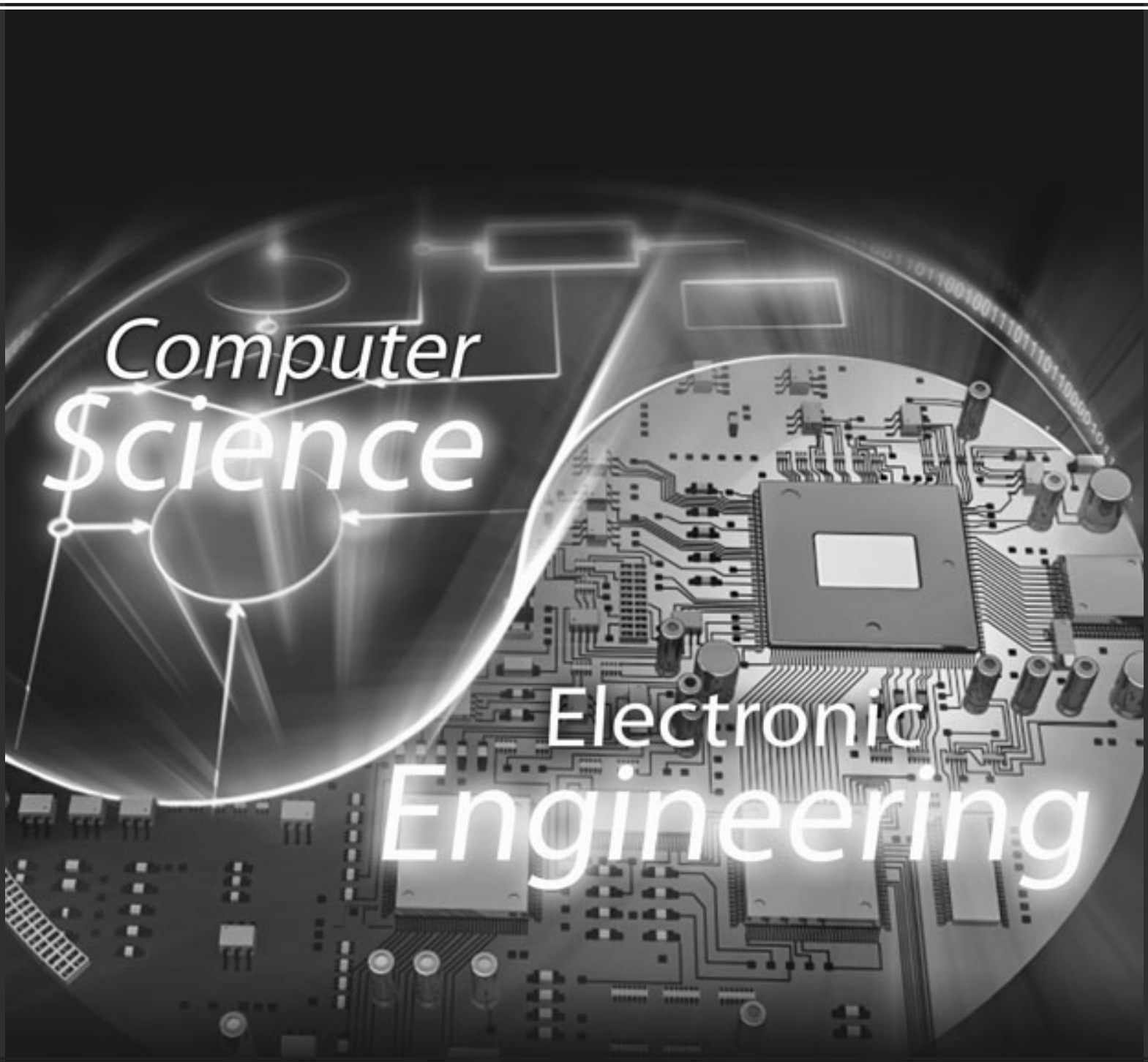
Recognising Cello Performers using Timbre Models

Magdalena Chudy and Simon Dixon

Queen Mary
University of London

EECSRR-12-01 February 2012

School of Electronic Engineering
and Computer Science



Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London

RECOGNISING CELLO PERFORMERS USING TIMBRE MODELS

Magdalena Chudy and Simon Dixon

{magdalena.chudy, simon.dixon}@eecs.qmul.ac.uk

Technical Report

February 2012

Abstract

In this work, we compare timbre features of various cello performers playing the same instrument in solo cello recordings. Using an automatic feature extraction framework, we investigate the differences in sound quality of the players. The motivation for this study comes from the fact that the performer's influence on acoustical characteristics is rarely considered when analysing audio recordings of various instruments. While even a trained musician cannot entirely change the way an instrument sounds, he is still able to modulate its sound properties obtaining a variety of individual sound colours according to his playing skills and musical expressiveness.

We explore the phenomenon, known amongst musicians as the "sound" of a player, which enables listeners to differentiate one player from another when they perform the same piece of music on the same instrument. We analyse sets of spectral features extracted from cello recordings of five players and model timbre characteristics of each performer. The proposed features include harmonic and noise (residual) spectra, Mel-frequency spectra and Mel-frequency cepstral coefficients. Classifiers such as k-Nearest Neighbours and Linear Discrimination Analysis trained on these models are able to distinguish the five performers with high accuracy.

Contents

1 Introduction

Timbre, both as an auditory sensation and a physical property of a sound, although studied thoroughly for decades, still remains *terra incognita* in many aspects. Its complex nature is reflected in the fact that until now no precise definition of the phenomenon has been formulated, leaving space for numerous attempts at an exhaustive and comprehensive description.

The working definition provided by ANSI [2] defines timbre in terms of a sound perceptual attribute which enables distinguishing between two sounds having the same loudness, pitch and duration. In other words, timbre is what helps us to differentiate whether a musical tone is played on a piano or violin.

But the notion of timbre is far more capacious than this simple distinction. Called in psychoacoustics *tone quality* or *tone color*, timbre not only categorises the source of sound (e.g. musical instruments, human voices) but also captures the unique sound identity of instruments/voices belonging to the same family (when comparing two violins or two dramatic sopranos for example).

The focus of this research is the timbre, or sound of a player, a complex alloy of instrument acoustical characteristics and human individuality (see Fig. 1). What we perceive as a performer-specific sound quality is a combination of technical skills and perceptual abilities together with musical experience developed through years of practising and mastery in performance. Player timbre, seen as a specific skill, when applied to an instrument influences the physical process of sound production and therefore can be measured via acoustical properties of sound. It may act as an independent lower-level characteristic of a player. If timbre features are able to characterise a performer, then timbre dissimilarities can serve for performer discrimination.

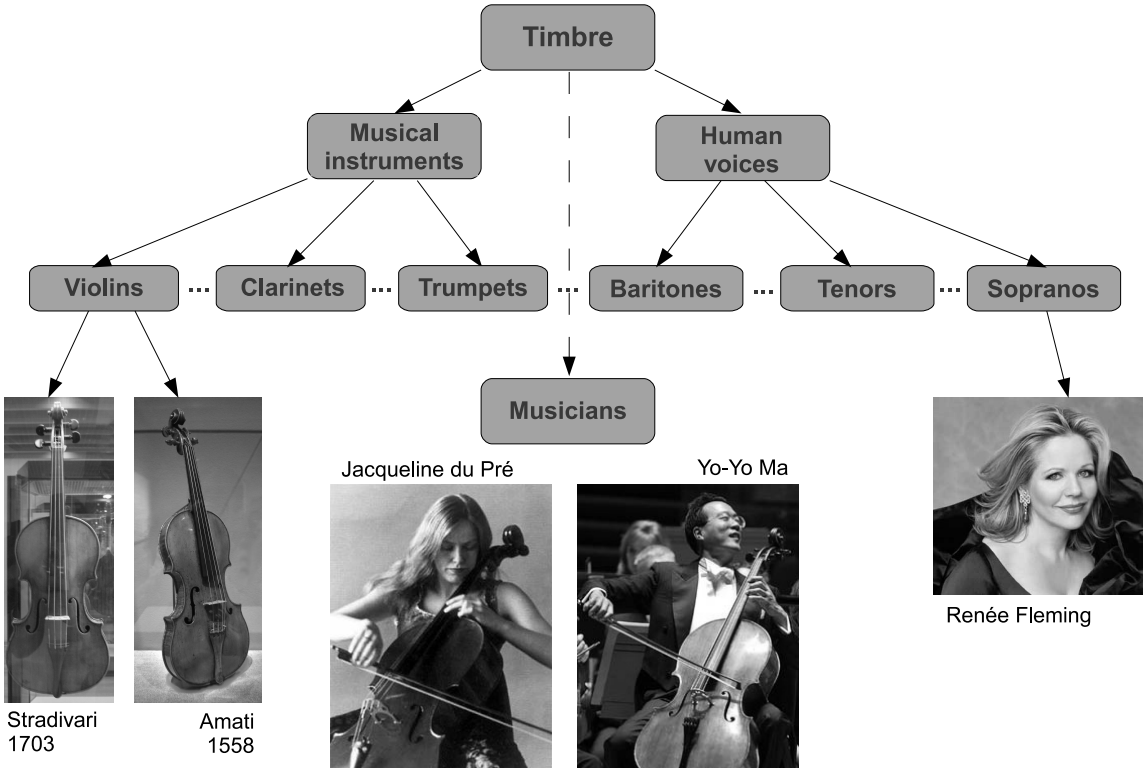


Figure 1: Factors determining timbre

2 Modelling timbre

A number of studies has been devoted to the question of which acoustical features are related to timbre and can serve as timbre descriptors. Schouten [9] introduced five major physical attributes of timbre: its "tonal/noiselike" character; the spectral envelope (a smooth curve over the amplitudes of the frequency components); the time (ADSR) envelope in terms of attack, decay, sustain and release of a sound plus transients; the fluctuations of spectral envelope and fundamental frequency; and the onset of a sound.

In order to find a general timbral profile of a performer, we considered a set of spectral features successfully used in music instrument recognition [5] and singer identification [8] applications. In the first instance, we turned our interest toward perceptually derived Mel filters as an important part of a feature extraction framework. The Mel scale was designed to mimic the entire sequence of pitches perceived by humans as equally spaced on the frequency axis. In reference to the original frequency range, it was found that we hear changes in pitch linearly up to 1 kHz and logarithmically above it [10]. A converting formula can be expressed as follows:

$$mel(f[Hz]) = 2595 \log_{10}\left(1 + \frac{f[Hz]}{700}\right) \tag{1}$$

Cepstrum transformation of the Mel scaled spectrum results in the Mel-frequency cepstrum whose coefficients (MFCCs) have become a popular feature for modelling various instrument timbres (e.g. [6, 7]) as well as for characterising singer voices [11].

We also investigated discriminant properties of harmonic and residual spectra derived from the additive model of sound [1]. By decomposing an audio signal into a sum of sinusoids (harmonics) and a residual component (noise), this representation enables to track short time fluctuations of the amplitude of each harmonic and model the noise distribution. The definition of the sound $s(t)$ is given by

$$s(t) = \sum_{k=1}^{N} A_k(t)\ cos[\theta_k(t)]\ +\ e(t) \tag{2}$$

where $A_k(t)$ and $\theta_k(t)$ are the instantaneous amplitude and phase of the k^{th} sinusoid, N is the number of sinusoids, and $e(t)$ is the noise component at time t (in seconds).

Figure 2 illustrates consecutive stages of the feature extraction process. Each audio segment was analysed using the frame-based fast Fourier transform (FFT) with a Blackman-Harris window of 2048-sample length and 87.5% overlap which gave us 5.8 ms time resolution. The length of the FFT was set to 4096 points resulting in a 10.76 Hz frequency resolution. The minimum amplitude value was set at a level of -100 dB.

At the first stage, from each FFT frame, the harmonic and residual spectra were computed using the additive model. Then, all FFT frames, representing the full spectra at time points t, together with the residual counterparts, were sent to the Mel filter bank for calculating Mel-frequency spectra and residuals. Finally, MFCCs and residual MFCCs were obtained by logarithm and discrete cosine transform (DCT) operations on Mel-frequency spectra and Mel-frequency residual spectra respectively.

The spectral frames were subsequently averaged over time giving compact feature instances. Thus, the spectral content of each audio segment was captured by five variants of spectral characteristics: harmonic, Mel-frequency and Mel-frequency residual spectra, and MFCCs of the full and residual signals.

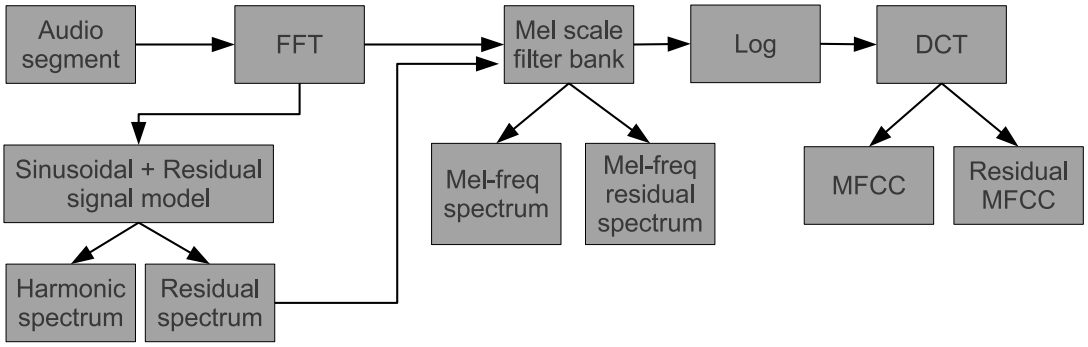


Figure 2: Feature extraction framework

3 Experiment Description

3.1 Sound Corpus

For the purpose of this study we exploited a set of dedicated solo cello recordings made by five musicians who performed a chosen repertoire on two different cellos[1]. The recorded material consists of two fragments of Bach's 1^{st}*Cello Suite*: *Prélude* (bars 1−22) and *Gigue* (bars 1−12). Each fragment was recorded twice by each player on each instrument, thus we collected 40 recordings in total. For further audio analysis the music signals were converted into mono channel .wav files with a sampling rate of 44.1 kHz and dynamic resolution of 16 bits per sample. To create a final dataset we divided each music fragment into 6 audio segments. The length of individual segments varied across performers giving approximately 11-12 s long excerpts from *Prélude* and 2-3 s long excerpts from *Gigue*. We intentionally differentiated the length of segments between the analysed music fragments. Our goal was to examine whether timbre characteristics extracted from shorter segments can be as representative for a performer as those extracted from the longer ones.

3.2 Feature Extraction

Having all 240 audio segments (24 segments per player performed on each cello) we used the feature extraction framework described in Sect. 2 to obtain sets of feature vectors. Each segment was then represented by a 50-point harmonic spectrum, 40-point Mel-freq spectrum and Mel-freq residual spectrum, 40 MFCCs and 40 MFCCs on the residual. Feature vectors calculated on the two repetitions of the same segment on the same cello were subsequently averaged to give 120 segment representatives in total. Figures 3(a)–3(d) show examples of feature representations.

3.3 Performer Modelling

Comparing feature representatives between performers on various music segments and cellos, we bore in mind that every vector contains not only the mean spectral characteristics of the music segment (the notes played) but also spectral characteristics of the instrument, and then, on top of that, the spectral shaping due to the performer. In order to extract this "performer shape" we needed to suppress the

[1]The same audio database was used in the author's previous experiments [3, 4]

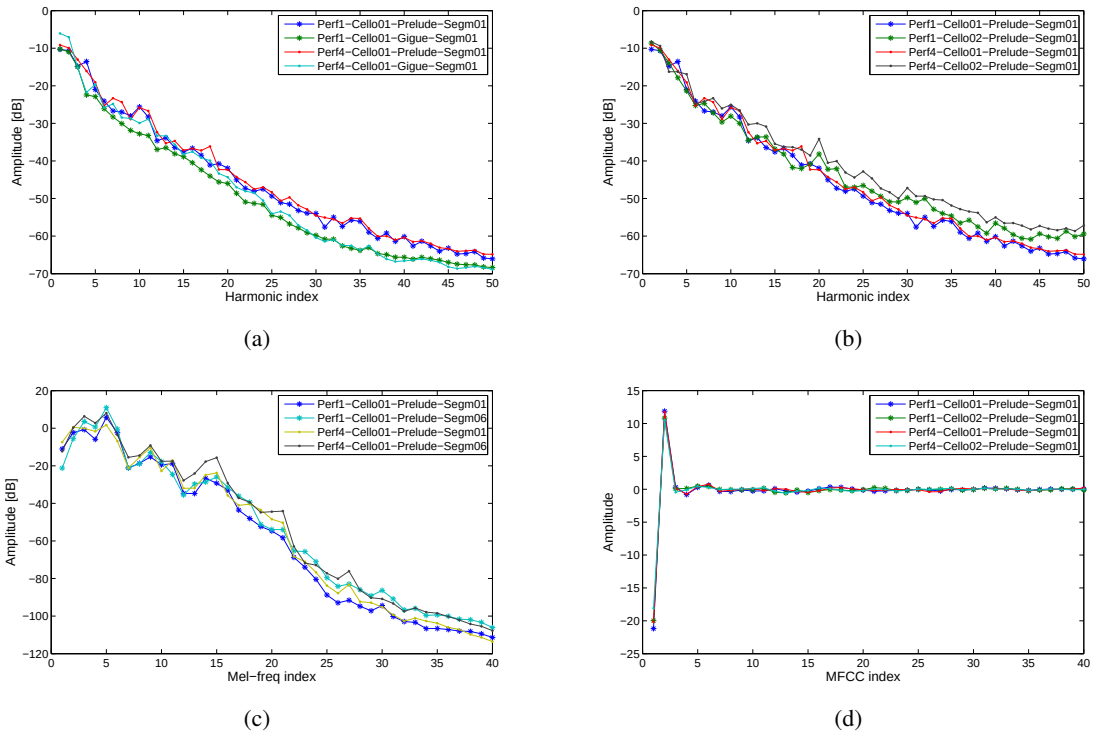


Figure 3: **(a)** Harmonic spectra of Perf1 and Perf4 playing Segment1 of *Prélude* and *Gigue* on Cello1, comparing the effect of player and piece; **(b)** Harmonic spectra of Perf1 and Perf4 playing Segment1 of *Prélude* on Cello1 and Cello2, comparing the effect of player and cello; **(c)** Mel-frequency spectra of Perf1 and Perf4 playing Segment1 and Segment6 of *Prélude* on Cello1, comparing the effect of player and segment; **(d)** MFCCs of Perf1 and Perf4 playing Segment1 of *Prélude* on Cello1 and Cello2, comparing the effect of player and cello

influence of both the music content and the instrument. The simplest way to do this was to calculate the mean feature vector across all five players on each audio segment and subtract it from individual feature vectors of the players (*centering* operation). Figure 4 illustrates the centered spectra of the players from the first segment of *Prélude* recorded on Cello1.

When one looks at the spectral shape (whether of a harmonic or Mel-frequency spectrum) it exhibits a natural descending tendency towards higher frequencies as they are always weaker in amplitude. The so called *spectral slope* is related to the nature of the sound source and can be expressed by a single coefficient (slope) of the line-of-best-fit. Treating a spectrum as data of any other kind, if a trend is observed it ought to be removed accordingly for data decorrelation. Therefore subtracting the mean vector removes this descending trend of the spectrum.

Moreover, the spectral slope is related to the spectral centroid (perceptual *brightness* of a sound in audio analysis) which indicates the proportion of the higher frequencies in the spectrum. Generally, the steeper the spectral slope, the lower is the spectral centroid and less “bright” is the sound.

We noticed that performers’ spectra have slightly different slopes, depending also on the cello and music segment. Expecting that it can improve differentiating capabilities of the features, we extended the centering procedure by removing individual trends first, and then subtracting the mean spectrum of a segment from the performers’ spectra (*detrending* operation). Figures 5–6 illustrate individual trends and the centered spectra of the players after detrending operation.

Our final performer-adjusted datasets consisted of two variants of features: centered and detrended-

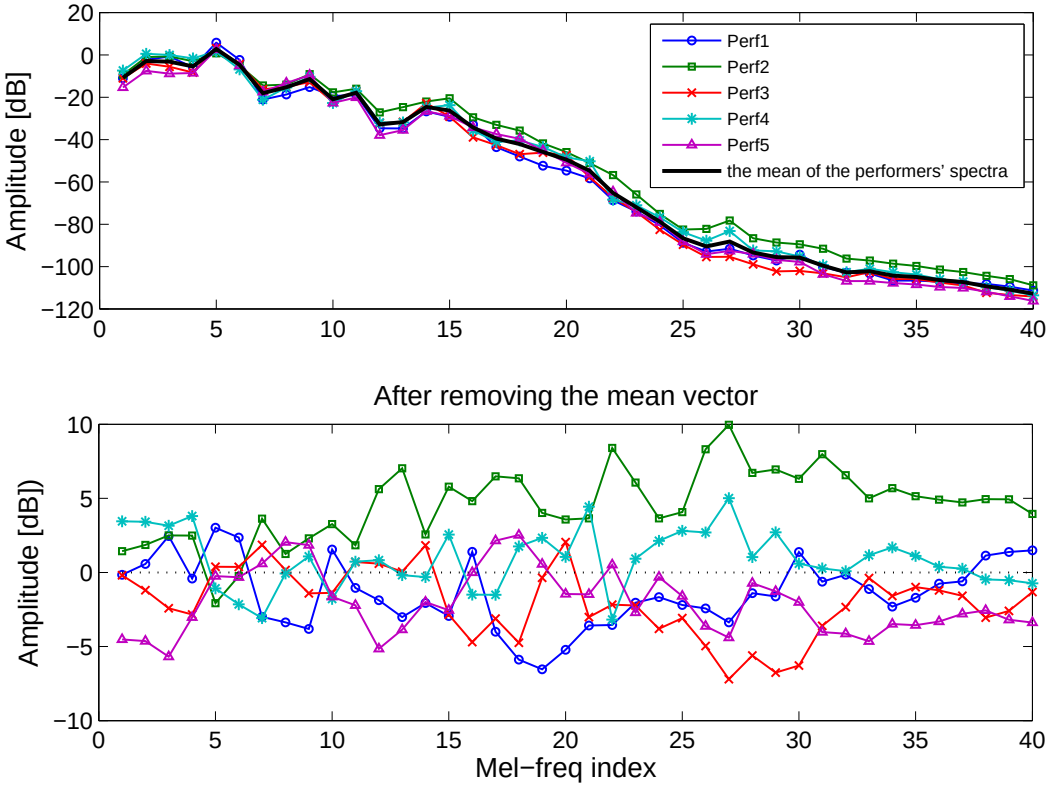


Figure 4: Mel-frequency spectra of five performers playing Segment1 of *Prélude* on Cello1, before and after centering

centered harmonic spectra, centered and detrended-centered Mel-frequency spectra and the residuals, centered MFCCs and the residuals.

3.4 Classification Methods

The next step was to test the obtained performer profiles with a range of classifiers, which also would be capable to reveal additional patterns within the data if such exist. We chose the k-nearest neighbour algorithm (k-NN) for its simplicity and robustness to noise in training data.

3.4.1 k-Nearest Neighbours

k-Nearest Neighbours is a supervised learning algorithm which maps inputs to desired outputs (labels) based on *supervised* training data. The general idea of this method is to calculate the distance from the input vector to the training samples to determine the k nearest neighbours. Majority voting on the collected neighbours assigns the unlabelled vector to the class represented by most of its k nearest neighbours. The main parameters of the classifier are the number of neighbours k and distance measure $dist$.

We ran a classification procedure using exhaustive search for finding the neighbours, with k set from 1 to 10 and $dist$ including the following measures: Chebychev, city block, correlation, cosine, Euclidean, Mahalanobis, Minkowski (with the exponent $p = 3, 4, 5$), standardised Euclidean, Spearman.

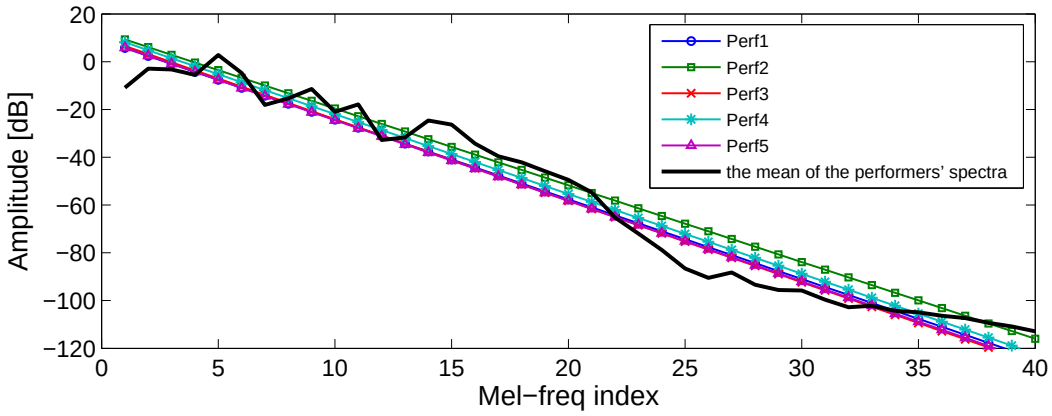


Figure 5: Individual trends of five performers playing Segment1 of *Prélude* on Cello1 derived from Mel-frequency spectra

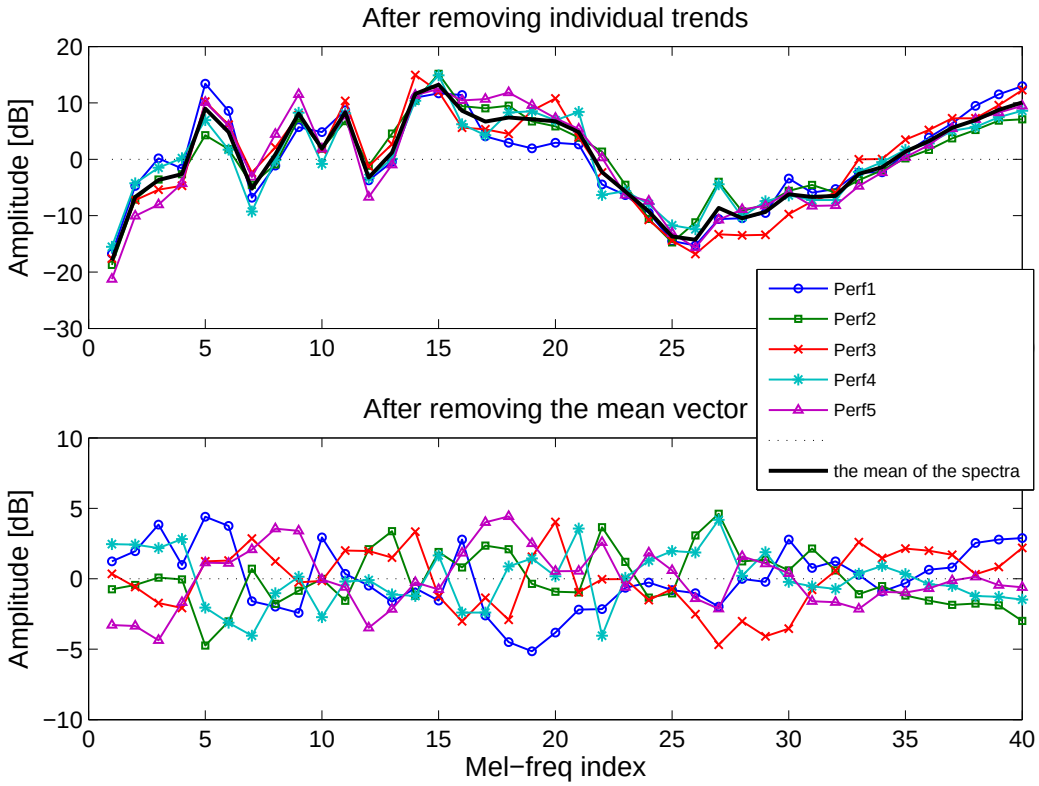


Figure 6: Mel-frequency spectra of five performers playing Segment1 of *Prélude* on Cello1, after detrending and centering

Classification performance can be biased if classes are not equally or proportionally represented in both training and testing sets. For each dataset, we ensured that each performer is represented by a set of 24 vectors calculated on 24 distinct audio segments (12 per each cello). To identify a performer p of a segment s, we used a leave-one-out procedure.

3.4.2 Linear Discriminant Analysis

Amongst statistical classifiers Discriminant Analysis (DA) is one of the methods that build a parametric model to fit training data and interpolate to classify new objects. It is also a supervised classifier as class labels are *a priori* defined in a training phase. Considering many classes of objects and multidimensional feature vectors characterising the classes, Linear Discriminant Analysis (LDA) finds a linear combination of features which separate them under a strong assumption that all groups have multivariate normal distribution and the same covariance matrix.

4 Results

In general, all classification methods we examined produced highly positive results reaching even 100% true positive rate (TP) in several settings, and showed a predominance of Mel-frequency based features in more accurate representation of the performers' timbres. The following sections provide more details.

4.1 k-Nearest Neighbours

We carried out k-NN based performer classification on all our datasets, i.e. harmonic spectra, Mel-frequency and Mel-frequency residual spectra, MFCCs and residual MFCCs, using both the centered and detrended-centered variants of feature vectors for comparison (with the exclusion of MFCC sets for which the detrending operation was not required). For all the variants we ran the identification experiments varying not only parameters k and $dist$ but also the feature vectors' length F for Mel-frequency spectra and MFCCs, where $F = \{10, 15, 20, 40\}$. This worked as a primitive feature selection method indicating the capability of particular Mel-bands to carry comprehensive spectral characteristics.

Table 1: k-NN results on harmonic spectra, vector length = 50

	Centered				Detr-centered			
length	**# k-NN**	**Distance**	**TP rate**	**FP rate**	**# k-NN**	**Distance**	**TP rate**	**FP rate**
50	9	corr	0.833	0.040	4	euc	0.867	0.032
	3	corr	0.825	0.041	6	seuc	0.858	0.034
	10	corr	0.825	0.042	6,7	cos,corr	0.850	0.036

Generally, detrended spectral features slightly outperform the centered ones in matching the performers' profiles (see Tables 1–3), attaining 100% identification recall for 20- and 40-point Mel-frequency spectra (vs 99.2 and 97.5% recall for centered spectra respectively). Surprisingly 20-point centered Mel- and residual spectra give higher TP rates than the 40-point (99.2 and 97.5% vs 97.5 and 96.7%), probably due to lower within-class variance, while the performance of detrended features decreases with decreasing vector length as expected.

What clearly emerges from the results is the choice of distance measures and their distribution between the two variants of features. Correlation and Spearman's rank correlation distances predominate

Table 2: k-NN results on Mel-freq spectra, vector length = 40, 20, 15, 10

	Centered				Detr-centered			
length	**# k-NN**	**Distance**	**TP rate**	**FP rate**	**# k-NN**	**Distance**	**TP rate**	**FP rate**
40	1-4	corr	0.975	0.006	1-4	seuc	1.000	0.000
	5	city	0.975	0.006	1,2,6	euc,cos,corr	1.000	0.000
	5	corr,euc	0.967	0.008	7-9	euc,cos,corr	1.000	0.000
20	1-10	corr	0.992	0.002	7,8	mink3	1.000	0.000
	7	spea	0.992	0.002	9,10	cos	1.000	0.000
	8-10	spea	0.983	0.004	3-8	cos,corr	0.992	0.002
15	5,6	corr	0.942	0.014	3,4	cos	0.975	0.006
10	3,4	corr	0.800	0.047	1,2	city	0.867	0.032

Table 3: k-NN results on Mel-freq residual spectra, vector length = 40, 20, 15, 10

	Centered				Detr-centered			
length	**# k-NN**	**Distance**	**TP rate**	**FP rate**	**# k-NN**	**Distance**	**TP rate**	**FP rate**
40	1,2	corr	0.967	0.008	1-3	cos,corr	0.992	0.002
20	3,4	spea	0.975	0.006	3	euc,seuc	0.983	0.004
15	3,4	spea	0.892	0.026	7	seuc	0.925	0.018
10	7	euc	0.775	0.053	6	euc	0.825	0.042

Table 4: k-NN results on MFCCs and residual MFCCs, vector length = 40, 20, 15, 10

	MFCCs				residual MFCCs			
length	**# k-NN**	**Distance**	**TP rate**	**FP rate**	**# k-NN**	**Distance**	**TP rate**	**FP rate**
40	1-4	seuc	1.000	0.000	3-10	spea	1.000	0.000
20	3	seuc	1.000	0.000	5-7	spea	0.992	0.002
15	5-8	maha	0.992	0.002	3-4	seuc	0.983	0.004
10	1-3	maha	0.950	0.012	5	seuc	0.908	0.022

within the centered spectra, while Euclidean, standardised Euclidean, cosine and correlation measures almost equally contribute to the best classification rates on detrended vectors. In regard to the role of parameter k, it seems that the optimal number of nearest neighbours varies with distance measure and the length of vectors but no specific tendency was observed.

It is worth noticing that the full spectrum features only slightly outperform the residuals (when comparing 100, 100, 97.5, 86.7% recall of Mel-frequency detrended spectra with 99.2, 98.3, 92.5, 82.5% recall of their residual counterparts for respective vector lengths = 40, 20, 15, 10). MFCCs and residual MFCCs (Tab. 4) in turn perform better than the spectra especially in classifying shorter feature vectors giving 100, 100, 99.2, 95% and 100, 99.2, 98.3, 90.8% TP rates respectively.

4.2 Linear Discriminant Analysis

For LDA-based experiments we used a standard stratified 10-fold cross validation procedure to obtain statistically significant estimation of the classifier performance. As previously, we exploited all five available datasets, also checking identification accuracy as a function of feature vector length.

For full length detrended-centered vectors of the harmonic, Mel-frequency and Mel-frequency residual spectra we were not able to obtain a positive definite covariance matrix. The negative eigenvalues related to the first two spectral variables (whether of the harmonic or Mel-frequency index) suggested that the detrending operation introduced a linear dependence into the data. In these cases, we carried out the classification discarding the two variables, bearing in mind that they might contain some important feature characteristics. Tables 5–8 illustrate the obtained results.

Table 5: LDA results on harmonic spectra, vector length = 50, 40, 30, 20

	Centered		Detr-centered	
length	**TP rate**	**FP rate**	**TP rate**	**FP rate**
50 (48)	0.900	0.024	(0.867)	(0.032)
40	0.858	0.034	0.875	0.030
30	0.842	0.038	0.833	0.040
20	0.758	0.056	0.842	0.038

Table 6: LDA results on Mel-freq spectra, vector length = 40, 20, 15, 10

	Centered		Detr-centered	
length	**TP rate**	**FP rate**	**TP rate**	**FP rate**
40 (38)	1.000	0.000	(1.000)	(0.000)
20	0.958	0.010	0.950	0.012
15	0.892	0.026	0.883	0.028
10	0.750	0.059	0.767	0.055

Table 7: LDA results on Mel-freq residual spectra, vector length = 40, 20, 15, 10

	Centered		Detr-centered	
length	**TP rate**	**FP rate**	**TP rate**	**FP rate**
40 (38)	1.000	0.000	(1.000)	(0.000)
20	0.933	0.016	0.933	0.016
15	0.900	0.024	0.850	0.036
10	0.792	0.049	0.767	0.055

Table 8: LDA results on MFCCs and residual MFCCs, vector length = 40, 20, 15, 10

	MFCCs		residual MFCCs	
length	**TP rate**	**FP rate**	**TP rate**	**FP rate**
40	0.992	0.002	1.000	0.000
20	0.992	0.002	0.983	0.004
15	0.992	0.002	0.983	0.004
10	0.917	0.020	0.900	0.024

Similarly to the previous experiments, Mel-frequency spectra gave better TP rates then harmonic ones (100, 95.8, 89.2, 75% for the vector length = 40, 20, 15, 10 vs 90, 85.8, 84.2, 75.8% for the vector

length = 50, 40, 30, 20 respectively) comparing centered features. Again, MFCCs slightly outperform the rest of features in classifying shorter feature vectors (99.2, 99.2, 99.2, 91.7% recall for respective vector lengths = 40, 20, 15, 10). Detrended variants of spectra did not improve identification accuracy due to the classifier formulation and statistical dependencies occurring within the data. As previously, the residual Mel spectra (100, 93.3, 90, 79.2%) and residual MFCCs (100, 98.3, 98.3, 90%) produced worse TP rates than their counterparts, with the exclusion of the 100% recall for 40 residual MFCCs.

5 Discussion

The most important observation that comes out from the results is that multidimensional spectral characteristics of the music signal are mostly overcomplete and therefore can be reduced in dimension without losing their discriminative properties. For example, taking into account only the first twenty bands of the Mel spectrum or Mel coefficients, the identification recall is still very high, reaching even 100% depending on the feature variant and classifier.

This implied searching for more sophisticated methods of feature subspace selection and dimensionality reduction. We carried out additional LDA classification experiments on attributes selected by the greedy best-first search algorithm using centered and detrended Mel spectra. The results (see Tab. 9) considerably outperformed the previous scores (e.g. 98.3% and 97.5% recall for 13- and 10-point detrended vectors respectively), showing how sparse the spectral information is. What is interesting, from the Mel frequencies chosen by the selector, seven were identical for both feature variants indicating their importance and discriminative power.

Table 9: LDA results on Mel-freq spectra with selected Mel-freq subsets

	Centered		**Detr-centered**	
length	**TP rate**	**FP rate**	**TP rate**	**FP rate**
8 (10)	0.908	0.022	(0.975)	(0.006)
13	0.950	0.012	0.983	0.004

As it was already mentioned, Mel spectra and MFCCs revealed their predominant capability to model the players' spectral profiles confirmed by high identification rates. Moreover, simple linear transformation of feature vectors by removing instrument characteristics and music context increased their discriminative properties. Surprisingly, the residual counterparts appeared as informative as full spectra, and this revelation is worth highlighting. It means that the residual (noise) part of signal contains specific transients, due to bow-string interaction (on string instruments), which seem to be key components of a player spectral characteristics.

We performed classification experiments intentionally varying the segment length across the performers and the analysed music fragments. We aimed to find if this may influence the ability to recover performer identity. While calculating representative feature vectors of each performer-segment combination we averaged the spectral frames over time suppressing in this way the influence of music segment length. With the obtained results, we confirmed that for extracting a general spectral characteristics of a performer the length of music fragments is of little importance.

Although we achieved very good classification accuracy on proposed features and classifiers (up to 100%) we should also point out several drawbacks of the proposed approach: (i) working with dedicated recordings and experimenting on limited datasets (supervised data) makes the problem hard to generalise and non scalable; (ii) use of simplified parameter selection and data dimensionality reduction instead of other "smart" attribute selection methods such as PCA or factor analysis; (iii) the proposed timbre model of a player is not able to explain the nature of differences in sound quality between performers, but only confirms that they exist.

While obtaining quite satisfying representations ("timbral fingerprints") of each performer in the dataset, there is still a need for exploring temporal characteristics of sound production which can carry more information relating to physical actions of a player resulting in his/her unique tone quality.

References

[1] Amatriain X, Bonada J, Loscos A, Serra X (2002) Spectral Processing. In: Zölzer U (ed) DAFX: Digital Audio Effects. 2nd edn. Wiley, Chichester

[2] American Standard Acoustical Terminology (1960) Definition 12.9, Timbre. New York

[3] Chudy M (2008) Automatic identification of music performer using the linear prediction cepstral coefficients method. Archives of Acoustics 33(1):27–33

[4] Chudy M, Dixon S (2010) Towards music performer recognition using timbre features. In: Proc. of the 3^{rd} Int. Conf. of Students of Systematic Musicology:45–50

[5] Eronen A (2001) A comparison of features for musical instrument recognition. In: Proc. of the IEEE Workshop Applications of Signal Processing to Audio and Acoustics:753–756

[6] Eronen A (2003) Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In: Proc. of the 7^{th} Int. Symp. Signal Processing and its Applications (2):133–136

[7] Heittola T, Klapuri A, Virtanen T (2009) Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In: Proc. of the 10^{th} Int. Soc. for Music Information Retrieval Conf.:327–332

[8] Mesaros A, Virtanen T, Klapuri A (2007) Singer identification in polyphonic music using vocal separation and patterns recognition methods. In: Proc. of the 8^{th} Int. Soc. for Music Information Retrieval Conf.:375–378

[9] Schouten J F (1968) The perception of timbre. In: Proc. of the 6^{th} Int. Congr. Acoustics:35–44

[10] Stevens S S, Volkman J, Newman E B (1937) A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. 8(3):185-190.

[11] Tsai W-H, Wang H-M (2006) Automatic singer recognition of popular recordings via estimation and modeling of solo vocal signals. IEEE Trans. Audio, Speech and Language Processing 14(1):330–341