Exploring the perceived harshness of cello sounds by morphing and synthesis techniques

Jocelyn Rozé,[a] Mitsuko Aramaki, Richard Kronland-Martinet, and Sølvi Ystad
Aix Marseille Univ, CNRS, PRISM (Perception, Representation, Image, Sound, Music), 31 Chemin J. Aiguier, 13402 Marseille Cedex 20, France

(Received 7 June 2016; revised 23 February 2017; accepted 26 February 2017; published online 24 March 2017)

Cello bowing requires a very fine control of the musicians' gestures to ensure the quality of the perceived sound. When the interaction between the bow hair and the string is optimal, the sound is perceived as broad and round. On the other hand, when the gestural control becomes more approximate, the sound quality deteriorates and often becomes harsh, shrill, and quavering. In this study, such a timbre degradation, often described by French cellists as harshness (*décharnement*), is investigated from both signal and perceptual perspectives. Harsh sounds were obtained from experienced cellists subjected to a postural constraint. A signal approach based on Gabor masks enabled us to capture the main dissimilarities between round and harsh sounds. Two complementary methods perceptually validated these signal features: First, a predictive regression model of the perceived harshness was built from sound continua obtained by a morphing technique. Next, the signal structures identified by the model were validated within a perceptual timbre space, obtained by multidimensional scaling analysis on pairs of synthesized stimuli controlled in harshness. The results revealed that the perceived harshness was due to a combination between a more chaotic harmonic behavior, a formantic emergence, and a weaker attack slope. © *2017 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4978522]

[TRM] Pages: 2121–2136

I. INTRODUCTION

Musical expressivity has been the subject of a large number of studies for various instruments relating or not the musicians' body movements to the produced sounds. In the case of cello playing, the particular musician–instrument interaction, in close embrace type, emphasizes the implication of the corporeal aspects in the expressivity. We investigate this topic in the present study by focusing on a special acoustic feature of cello sounds, linked to a degradation of timbre quality, and frequently referred to as harshness (*décharnement* in French) by cellist teachers. A harsh sound (*son décharné*) is recognized among bowed string players as a sound in which timbre is impoverished, inducing a sensation of shrill, whistling, and loss in sound consistency.

Harsh notes are very frequent among young cellist students and often arise from an inappropriate bowing gesture (Guettler, 2002) and/or a bad postural coordination. Indeed, many studies have demonstrated that, in addition to instrumental gestures directly responsible for the sound production, musical interpretation is strongly influenced by accompanying (or ancillary) gestures, i.e., postural movements that are rarely learnt in music schools. Their musical significance was explored for the clarinet (Wanderley *et al.*, 2005; Desmet *et al.*, 2012), the piano (Thompson and Luck, 2012), the harp (Chadefaux *et al.*, 2013), and the violin (Van Zijl and Luck, 2013). For example, Wanderley *et al.* (2005) demonstrated that clarinetists might transcribe the phrasing and rhythmical elements of a musical score through arm flapping or waist and knee bending. In the same line of thinking, we were interested in exploring embodied music cognition of cellists, and how perceived musical features might result from their gestures and postural synergies. Leman (2008) introduced this notion of corporeal encoding as an interconnection between two learning mechanisms: a high-level reenacting mechanism, called action-perception loop, able to build a complex sequence of gestures (right bowing-arm positioning, left hand fingerings, etc.) in order to reach a local musical target, and a low-level mechanism, called sensorimotor loop, which continuously adjusts the musician's motor program (bow force, bow velocity, postural variables, etc.) to maintain the control of the sound production. In Rozé *et al.* (2016), we designed an experiment with these former studies in mind, by assessing the effects of different kinds of postural constraints on the cellist's musical expressivity along musical phrases. To ensure that expressive degradations were a consequence of postural constraints and not of technical weaknesses, the musicians that took part in the experiment were all professional or much experimented players. This experiment turned out to be particularly revealing, since in some difficult musical passages, the postural constraint affected the cellist's bowing gesture in a manner which might produce timbre degradations resulting in sounds perceived as harsh. The aim of the present study is to investigate this perceptual harshness in terms of signal properties.

Previous research related to the perception of harsh acoustic features for bowed-string instruments can be found in Stepanek and Otcenasek (2005), Stepanek (2006), and more recently in Fritz *et al.* (2012). These perceptual studies

[a]Electronic mail: roze@prism.cnrs.fr

aimed at correlating acoustic properties of violin sounds to a list of verbal attributes frequently used to characterize musical timbre. Multidimensional scaling (MDS) analysis was performed on the dissimilarity scores between pairs of stimuli in order to construct a perceptual space of verbal attributes, likely to explain the acoustic correlates along each one of its dimensions. For example, in the article by Fritz *et al.* (2012), the main dimension of the space assessing the "overall sound quality" corresponds to a spectral balance between adjectives of desirable sounds (warm, rich, mellow) and undesirable ones (metallic, cold, harsh). To correlate the amount of harshness perceived along this dimension with physical properties, the authors created synthetic violin sounds by modifying the energy levels in five-octave spectral bands, and performed dissimilarity tests on pairs of these sounds. They demonstrated that the undesirable perceptual effect of harshness corresponds to localized and excessive high-frequency energy, combining most of the time with poor low-frequency content. Thus, the harshness phenomenon primarily seems to be linked to a spectral energy transfer toward upper partials and indirectly to the notion of sound brightness (*brillance*). Rozé *et al.* (2016) characterized this spectral transfer more finely by designing a relative brightness descriptor related to the three frequency bands of the tristimulus criterion (Pollard and Jansson, 1982). Numerous studies (Grey and Gordon, 1978; McAdams, 1999; Caclin *et al.*, 2005) highlighted a strong correlation between brightness and spectral centroid, i.e., the center of gravity of the spectral energy distribution. According to Barthet *et al.* (2008) and Barthet *et al.* (2010), the brightness might also play an important role in the tension perceived within clarinet interpretation between a muffled tone judged as soft and a bright, more aggressive one. Analyses of cello sounds by Chudy *et al.* (2013) confirmed that boosted higher spectral harmonics would cause a tone to be more "tensed." As a consequence, sounds produced in our expressive experimental context and perceived as harsh might also comply with an increasingly aggressive and tense musical expression.

The aim of the present study is to better explain the harshness phenomenon from an acoustical point of view, based on certain assumptions linked to the way a postural constraint alters the bowing gesture, and hereby induces this sound degradation. The works of Demoucron (2008) or Schoonderwaldt (2009) provide some cues from investigations on the relationship between the brightness and the physical parameters of the bow–string interaction in the case of the violin. They computed the brightness determined by the spectral centroid values for different combinations of physical control parameters (bow force, bow velocity, bow-bridge distance), and projected them into Schelleng diagrams (Schelleng, 1973). Such diagrams constitute a bi-dimensional space defined by the bow force and the bow-bridge distance, and display the variations of acoustical parameters for a given bow velocity. Within Schelleng diagrams, spectral centroids of violin sounds increase with the bow force and decrease with the bow velocity or the bow-bridge distance. Guettler *et al.* (2003), Guettler (2003), and Schoonderwaldt *et al.* (2003) also showed that lowering the bow velocity or increasing the bow force resulted in higher frequency partials. Hereby, we may assume that harsh sounds, which are brighter than normal ones, result from potentially stronger or slower bowing gestures, or smaller bow-bridge distances ("sul ponticello"). Such sounds, therefore, might locate around the upper frontier of the Schelleng diagram within the "Raucous" area, above which the normal Helmholtz motion no longer exists.

In this paper, we are going to explore in depth the signal characteristics composing the notion of harshness for cello sounds, and validate it from a perceptual point of view by adapted listening tests. We expect that this phenomenon matches a complex multidimensional alteration affecting spectral, but maybe also spectro-temporal dimensions. Thus, we will first observe this sound quality degradation from a macroscopical point of view before inferring on the spectral and temporal descriptors of timbre that best characterize the harshness. Then, we will use two complementary methods to perceptually validate these descriptors. In the first method, the aim is to build a perceptual model of harshness composed of the most relevant signal descriptors. A morphing technique based on Gabor tools is presented to build continua between round and harsh sounds. Then the amount of perceived harshness is predicted by multilinear regression on all the descriptors to identify the most relevant descriptors in theory. In the second method, we assess the perceived sensation of harshness by controlling the envelopes of two relevant descriptors of the predictive model. The process uses cross-synthesis techniques and is perceptually validated by the construction of a perceptual timbre space issued from MDS analysis.

II. QUANTIFYING THE SOUND QUALITY DEGRADATION

In this part, we attempt to quantify the acoustic degradation of cello sounds, perceived as a harshness phenomenon, through a macroscopical comparative analysis and finer comparisons of suitably chosen acoustic descriptors. The presented sound corpus was used in Rozé *et al.* (2016) for a preliminary and quantitative exploration of the harsh degradation, based on a spectro-temporal duality between two descriptors. In the present paper, our aim is to refine this preliminary approach by proposing a full panel of signal descriptors that might account for the harshness phenomenon. An original approach based on Gabor masks is proposed in order to assess the descriptors that best characterize the timbre transformation between normal and harsh sounds.

A. Sound corpus

The sound corpus was created from data collected in the experiment conducted by Rozé *et al.* (2016). In the protocol of this experiment, we asked seven cellists to play a score in the most expressive way, while being subjected to four kinds of postural conditions. For the scope of this paper, we only deal with data collected in two opposite postural situations: a normal one, where the cellists could play naturally as in a performance context, and a physically fully constrained one, in which they had to play with the torso attached to the back of a chair by a safety race harness, and a neck collar limiting

their head movements. All the participants were professional or very experienced musicians to ensure that any loss of expressivity was due to the physical immobilization and no to potential weaknesses in their playing technique. They were asked to perform the experimental score with the same instrument at tempo ♩ = 45 bpm, in two articulation variants: detached (*détaché*) and *legato*. By using the same cello, we ensured that the physical characteristics of the instrument would not alter subsequent timbre analyses.

For each postural condition, the musicians' impressions were collected through a questionnaire assessing their feelings, specific difficulties that they may have encountered, as well as the perceived influence of the postural condition on their movements and expressive sound features. Most of them reported a degradation of their usual sound palette within the fully constrained postural situation, and especially the impression of producing "tighter, more tense sounds" ("*sons plus étriqués*, *tendus*" in French), "lacking of depth and natural resonances." Some of them also evoked more "harsh, shrill, and whistling notes" ("*notes plus décharnées, criardes et sifflantes*" in French), which were particularly common within a specific musical passage of the score (Fig. 1), containing more gestural and rhythmic difficulties. A thorough post-examination of the sound recordings revealed that in spite of the postural constraint, the cellists managed quite well to produce a proper Helmholtz motion, except in the specific musical passage described previously. According to the authors' subjective judgments, several notes located in the first bar of the passage were randomly affected by the harsh timbre degradations. The first note of the bar, a dotted sixteenth of pitch E3 (fundamental frequency 329.63 Hz) was most clearly and regularly perceived as harsh. As a consequence, we chose this note to create the sound corpus of the study and to perform harsh timbre comparisons. It should be mentioned that this note corresponds to a rhythmic difficulty requiring an excellent coordination between the right and left arms, since it occurs prior to a pushed bow stroke (*poussé d'archet*) combined with a fast-shift (*démanché rapide*) of the left arm. We may here suppose that the constraint altered this coordination between the two instrumental gestures, which resulted in a strong decrease in timbre quality for this note.

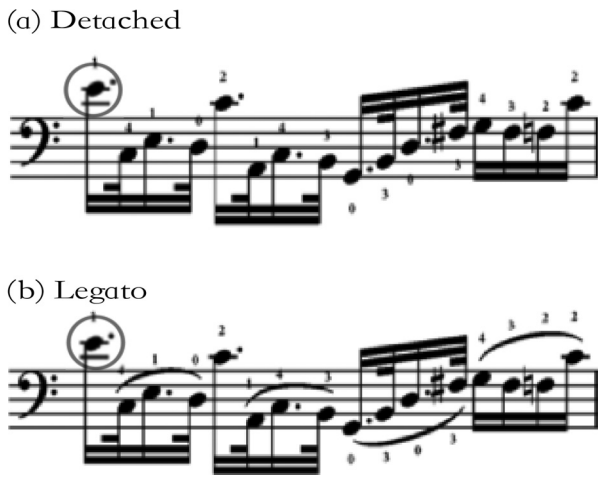

FIG. 1. Bar of the score selected to investigate the harshness. The analyzed note (E3) has been circled for each articulation variant: (a) *Detached* (Détaché) playing mode, (b) *Legato* playing mode.

The sound corpus was obtained by extracting 20 of these E3 notes from the recordings of all the cellists, whatever the articulations *détaché* or *legato*. The extraction process for the E3 notes relied on a pitch-tracking algorithm adapted from the MIR toolbox (Lartillot and Toiviainen, 2007). More precisely, these 20 notes were composed of 10 pairs of E3 notes, respectively, selected within the normal and constrained postural contexts for a given cellist and articulation. A pair of notes associates one of the round/beautiful sounds produced in the normal situation to one of the harsh/degraded sounds produced in the constrained situation (for the same cellist and articulation).

The selection of *round*/*harsh* pairs of notes was performed by the authors through an informal listening test, conducted on the audio recordings obtained after the experiment. The authors used high quality loudspeakers and headphones to detect significant harshness dissimilarities between sounds in the normal and the posturally constrained situation. Their musical expertise and knowledge of cello practice enabled them to select the sounds that best reflected the differences between the two postural situations. The relevance of the authors' judgment was further confirmed by the results of the first perceptual experiment presented in this paper [see Fig. 6(a), Sec. III E]. Sometimes, in the constrained situation, some cellists played the E3 without being able to prevent the bow from slipping on the D open string, which resulted in a sound that contained a second tone in the onset phase. We systematically removed this artifact as far as possible by means of a very selective notch filter centered on the D2 frequency.

The average length of the notes composing this corpus was 0.31 s, which is shorter than the theoretical duration of dotted sixteenths (0.5 s at tempo ♩ = 45 bpm). The standard deviation of their durations was 0.06 s. This suggests that the cellists consistently played the E3 notes more like sixteenths (theoretical duration 0.33 s at ♩ = 45 bpm) than like dotted sixteenths. Timbre discrimination of such short stimuli might not be obvious. However, their durations remained higher than the minimum duration required to recognize the timbre of a sound (around 100 ms in Schaeffer, 1966). The authors thus judged the task feasible, especially since the participants were cellists who are trained to constantly pay attention to the quality of their sounds.

B. Gabor tools

In the present study, Gabor transforms were used to reveal differences between pairs of round/harsh sounds composing the corpus. Gabor transforms, or short-time Fourier transforms (STFTs) with Gaussian analysis windows, are useful tools to get a macroscopic view of sound signal properties. They make it possible to represent the timbre of a sound as an image in the time-frequency space, defined on a suitably discretized time-frequency lattice. If $a,b > 0$ are time and frequency sampling constants, and L denotes the length of the analyzed signal, the time-frequency lattice of its Gabor transform is characterized by M, the number of

frequency channels, and N, the number of time steps, such that $L = Mb = Na$ (LTFAT toolbox of Søndergaard *et al.*, 2012). For a signal $x \in \mathbb{C}^L$, the transform can be written as coefficients $c_{m,n} \in \mathbb{C}^{M \times N}$ of the signal expansion in a family of Gabor atoms $g_{m,n} = e^{2i\pi mb(l-na)} g(l-na)$,

$$c_{m,n} = \langle x, g_{m,n} \rangle = \sum_{l=1}^{L-1} x(l) e^{-2i\pi mb(l-na)} \overline{g(l-na)}, \quad (1)$$

with $n \in [0, N-1], m \in [0, M-1]$. l is a discrete time variable along the length L of the signal, and a, $b > 0$ are the time and frequency sampling constants, respectively.

Differences between Gabor transforms for round and harsh sounds can further be observed by computing their time-frequency transfer function. This function is called a Gabor mask $\mathbf{m}$, and can be evaluated using a minimization process based on the least-square criterion (Depalle *et al.*, 2006)

$$\Psi(\mathbf{m}) = \| c_{m,n}(x_2) - \mathbb{M}_{\mathbf{m}} c_{m,n}(x_1) \|^2 + \lambda \| \mathbf{m} - 1 \|^2, \quad (2)$$

where $\lambda \in \mathbb{R}+$ is the Lagrange parameter introduced to control the norm of $\mathbf{m}$.

For the source and target signals, respectively, denoted $x_1, x_2 \in \mathbb{C}^L$, the mask $\mathbf{m}$ is estimated by considering that there is no difference between the signals when $\mathbf{m} = 1$,

$$\mathbf{m} = \frac{\| c_{m,n}(x_1) \| \| c_{m,n}(x_2) \| + \lambda}{\| c_{m,n}(x_1) \|^2 + \lambda} = \frac{C_1 C_2 + \lambda}{C_1^2 + \lambda}, \quad (3)$$

where C_1 and C_2 are the modulus of the Gabor transforms of the signals x_1 and x_2, respectively. As explained by Olivero *et al.* (2010) and Sciabica *et al.* (2012), λ plays the role of a regularization parameter, which enables us to reveal more or less detailed differences between signals.

C. Highlighting overall features of harshness

In this section we explain how to compute and tune a Gabor mask in order to capture the salient dissimilarities of timbre between the round and harsh versions of the note E3. One sound pair was chosen among the ten pairs of the corpus (Rozé, 2017). First of all, we matched the two signals of the pair in the time-frequency domain. In the frequency space, this operation was intrinsically achieved because the pitches to compare were identical. In the temporal space, however, we needed to synchronize the signal durations with an accurate time-scaling process. This implied changing the signal durations without affecting their pitch-contour and the time-evolution of the formantic structure. The procedure was achieved using *Adobe Audition* software (San José, CA) by reducing the duration of the longest signal so that it fitted with the shortest.

To capture global spectral tendencies within the Gabor mask, a sliding window size of $M = 2048$ bins was chosen. This corresponds to a lattice frequency resolution of $\sim$20 Hz. From Eq. (3), we controlled the Gabor mask estimation with different values of the regularization parameter λ to get the roughest ($\lambda \geq 1$) or finest ($\lambda \to 0$) differences between the signals. The time-frequency representation of the mask is displayed in Fig. 2 for $\lambda = 1$, and can be interpreted as an

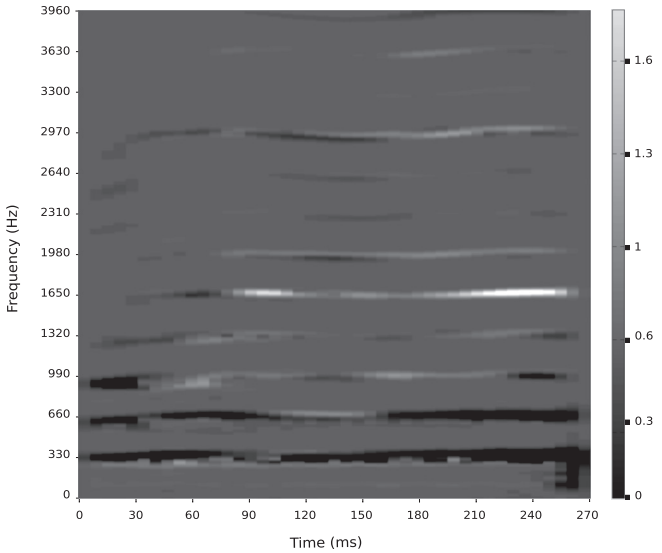

FIG. 2. Gabor mask computed between the round and harsh signals of a corpus pair for a regularization parameter $\lambda = 1$. Their macroscopic differences appear as deviations from $\mathbf{m} = 1$ (corresponding to no transformation). White frequency bands correspond to spectral energy reinforcements (mask deviations greater than one) of the harsh sound compared to the round sound, whereas black bands correspond to a decrease in spectral energy (mask deviations lower than one).

overview of the *harshnessing process* of a round sound. In Fig. 2, we can notice that black and white parts of the image mainly coincide with the frequency bands situated at multiple integers of the fundamental frequency of the note E3, which is logical since the analyzed sounds are instrumental and harmonic by nature. More interestingly, we observe as a whole that the harsh sound (relatively to the round one) presents strong energy reinforcements (white bands) in the middle-upper spectrum area, and strong energy decreases (black bands) in the lower spectrum area. This corresponds for the harsh sound to a transfer of spectral energy toward higher frequencies and the emergence of formants.

A detailed observation of the Gabor mask guides us toward three salient acoustic features that might explain the harshness phenomenon: an emergence of formantic areas, an energy loss in the attack part, and fluctuations in the temporal deployment of harmonic amplitudes. Regarding the formantic feature, the repartition of black and white bands clearly indicates a transfer of the spectral energy from lower to higher partials, with a noticeable reinforcement of the fifth component, the G♯5 ($330 \times 5 = 1650$ Hz). This may indicate that the harsh E3 note would globally sound brighter and in sympathetic resonance with its major third. It is important to note that the emergence of formants might be related to the physics resulting from an incorrect bowing gesture (Demoucron, 2008; Guettler, 2003), suggesting a slower bow speed for harsh sounds. Regarding the attack feature, we can notice consistent black portions located at the beginning of each harmonic band. This might correspond to a global delay in the birth of harmonics composing the harsh sound, globally implying a less marked attack than in the round sound. Finally, regarding the harmonic features, we observe amplitude fluctuations appearing in the temporal deployment of each partial, which suggests some differences

in the relative life cycle of partials between round and harsh sounds.

D. From Gabor masks to harshness descriptors

The three features revealed by the Gabor mask analysis suggest that the harshness is multidimensional by nature. Further investigations of perceived and observed timbre changes should be carried out by associating these observations with existing spectral, temporal, and spectro-temporal timbre descriptors.

1. Spectral descriptors

From a spectral viewpoint, the main phenomenon revealed by the Gabor mask sheds light on an energy increase in the high frequencies of the harsh signal, and more precisely an energy transfer from low to high frequencies with a formantic reinforcement around 1650 Hz. This transfer may correspond to an increase in the barycenter of the spectral energy distribution, commonly characterized as the spectral centroid (Peeters, 2004; Schoonderwaldt, 2009). Many studies refer to this descriptor as perceptually associated with the brightness (Grey and Gordon, 1978; McAdams, 1999; Caclin *et al.*, 2005; Merer *et al.*, 2007) and it has been validated as a reliable indicator of the timbre of bowed-string instruments like the violin (Stepanek and Otcenasek, 2005; Stepanek, 2006; Fritz *et al.*, 2012). From our observations of the differences between round and harsh sound signals, an increase in harshness might contribute to an increase in brightness.

Considering the harmonic nature of cello sounds, we decided to compute the harmonic spectral descriptors. For this purpose, a detection of the harmonic components and the extraction of their instantaneous amplitudes were performed beforehand. Knowing the fundamental frequency of the note analyzed (E3), this process could be achieved through decomposition in subbands centered on the signal harmonics (Barthet *et al.*, 2007). We could thus compute the harmonic spectral centroid (HSC) of the sound, which corresponds to the amplitude-weighted mean of its harmonic peaks in the spectrum

$$\mathrm{HSC} = f_0 \frac{\sum_{k=1}^{K} kA_k}{\sum_{k=1}^{K} A_k} \tag{4}$$

where k and A_k are, respectively, the bin index and the amplitude of the kth harmonic peak of the fundamental frequency f_0, and K is the total number of harmonics considered.

To characterize the spectral energy transfer occurring within the harshness phenomenon, we computed the harmonic tristimulus (Pollard and Jansson, 1982), which describes the spectral energy distribution in three frequency bands as the energy ratio between each band and the total number of harmonics. The first band contains the fundamental frequency, the second one the medium partials (two, three, four), and the last one the higher partials (five and more). Three coordinates were hereby obtained, corresponding to spectral centroid computations for each band

$$\mathrm{TR}_1 = \frac{A_1}{\sum_{k=1}^{K} A_k}; \quad \mathrm{TR}_2 = \frac{\sum_{k=2}^{4} A_k}{\sum_{k=1}^{K} A_k}; \quad \mathrm{TR}_3 = \frac{\sum_{k=5}^{K} A_k}{\sum_{k=1}^{K} A_k}, \tag{5}$$

where A_k is the amplitude of the kth harmonic peak and K is the total number of harmonics considered. From this classical definition of the tristimulus, we designed a new more compact ratio focusing on the spectral transfer related to the harshness phenomenon

$$\mathrm{TRIratio} = \frac{\mathrm{TR}_3}{\mathrm{TR}_1 + \mathrm{TR}_2}. \tag{6}$$

The values of this spectral transfer descriptor are expected to decrease for a round cello note and increase for a harsh one.

We characterized the emergence of formants revealed by the Gabor mask using MFCCs (mel-frequency cepstral coefficients; Davis and Mermelstein, 1980). This descriptor stands for a simplified view of the spectral envelope with just a few coefficients, and is therefore well suited to reveal particular spectral envelope variations such as appearances or shifts in formantic areas. It also takes into account the frequency selectivity of our auditory system by using frequency bands on a mel scale that are more selective for lower frequencies than for higher ones (Kim *et al.*, 2006; Bogert *et al.*, 1963). The MFCCs are linked to the notion of cepstrum, which has been intensively used to characterize the formants of the voice because of its acoustic behavior of source/filter type. A cello may also be modeled in this way, if we consider the bow friction on a string as the excitation source and the instrument-body as the filter. A harsh sound would result from an incorrect excitation of the string by the bow, which translates through the cello sound box into the shift of a formant (an energy transfer) from low- to high-frequency areas. MFCC coefficients c_i are computed through a DCT (discrete cosine transform) applied to the logarithmic spectral envelope as follows:

$$c_{\mathrm{i}} = \sum_{k=1}^{N_f} \log\left(E_k\right) \cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{N_f}\right], \quad 1 \le i \le N_c, \tag{7}$$

where E_k is the spectral energy measured in the critical band of the kth mel filter, N_f is the total number of mel filters, and $N_c \le N_f$ is the number of MFCC coefficients to compute.

Standard recommendations were followed for MFCC calculations (Kim *et al.*, 2006): a mel-filter bank built with a number of filters $N_f = 24$, and a bandwidth of $B_m = 2700$ mel (limiting the frequency range to 8000 Hz). With a typical number of MFCC coefficients $N_c = N_f/2 = 12$, we obtained the smooth spectral envelopes of a round and a harsh signal, presented in Fig. 3(a). The formantic structures of these two envelopes look quite different, but can clearly be explained by an examination of the first two MFCC coefficients. Indeed, the MFCC coefficients C_1 and C_2 [see Fig. 3(b)] are

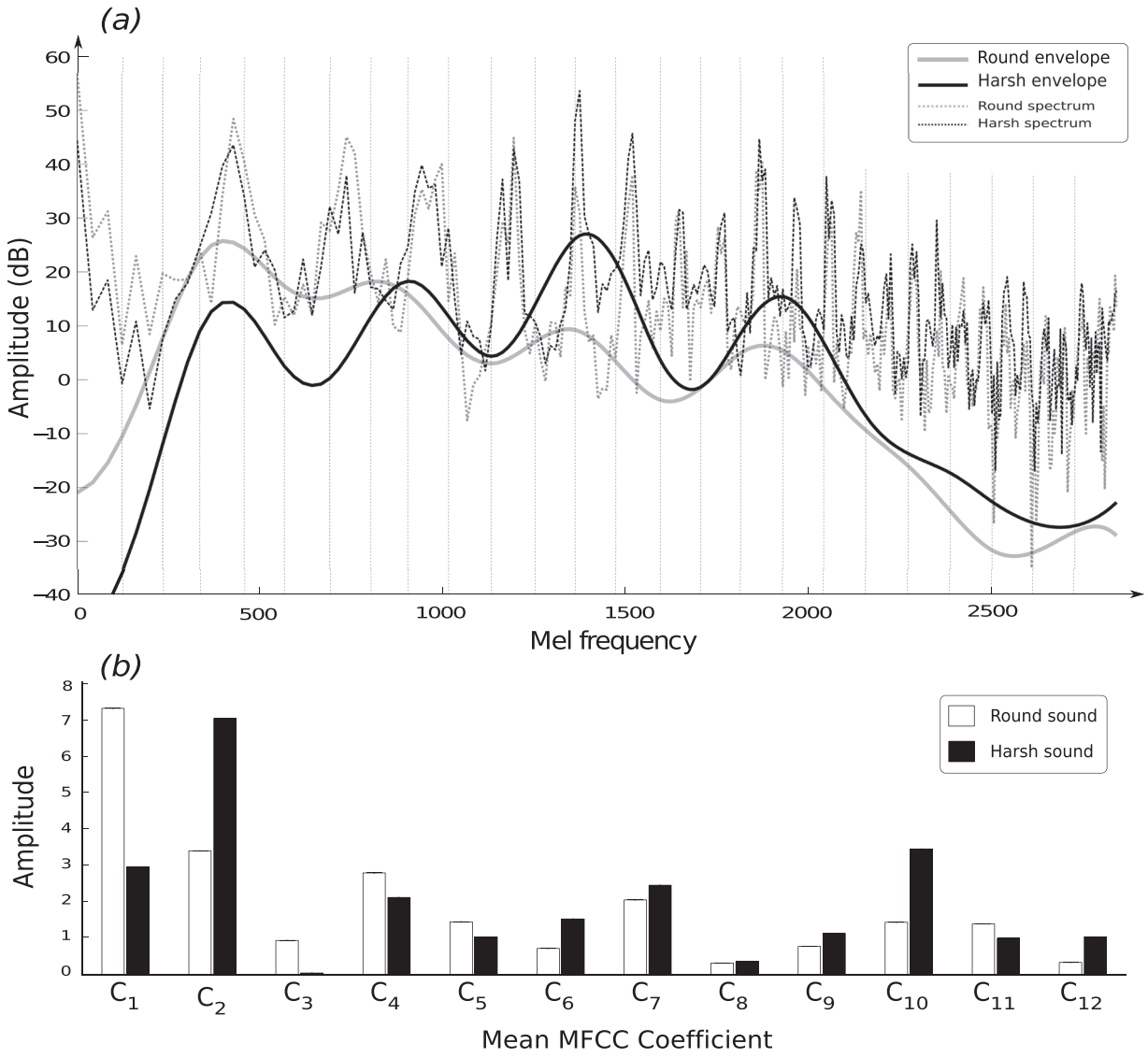

FIG. 3. (a) Formantic envelopes of the round and harsh signals (the same sound pair as in Fig. 2). The vertical lines correspond to the center frequencies of 24 critical bands in the mel-filter bank, from which spectral envelope energies are computed. (b) The 12 first mean MFCC coefficients associated to the previous round and harsh formantic envelopes. Absolute values of these coefficients are displayed to highlight the behavior difference between signals.

the founding parameters that describe the global spectral shape. C_1 induces a low-pass behavior of the spectrum, while C_2 induces a bandpass effect. Note that C_0 simply represents the mean log-energy output of the filterbank, which is not of interest in the present case when comparing spectral shapes. The MFCC analysis thus suggests that the shape of the spectral envelope evolves from a low-pass behavior for the round sound ($C_1 > C_2$), to a bandpass behavior for the harsh one ($C_1 < C_2$). This is coherent with the spectral transfer of harmonic energies accounted for by the tristimulus criterion [Eq. (5)].

In the same way as the TRIratio, we designed an MFCC ratio suitable for describing the harsh formantic behavior

$$\text{MFCCratio} = \frac{\|C_2\|}{\|C_1\|}. \tag{8}$$

Given the behavior of the MFCC coefficients, the values of this formantic descriptor should decrease for a round cello note and increase for a harsh one.

2. Temporal descriptors

From a purely temporal viewpoint, the Gabor masks shed light on a lack of energy at the beginning of the harsh signal. Perceptual studies of Grey (1977) and McAdams *et al.* (1995) have shown that the rise time of the energy of the signal during the attack phase plays a prominent role in instrument classification. From these statements, we deduced that a suitable descriptor of the attack phase might help to determine the timbre quality and thus contribute to the perception of the harshness. A common temporal descriptor is the log attack time (LAT) (Kim *et al.*, 2006):

$$\text{LAT} = \log_{10}(T_{\max} - T_{\text{threshold}}), \tag{9}$$

where $T_{\max}$ is the time needed for the signal envelope to reach its maximal value, and $T_{\text{threshold}}$ is the time needed for the envelope to exceed 2% of its maximal value.

Furthermore, to propose a descriptor that is less dependent on the energy differences between signals, we

computed an attack slope (ATS) or temporal increase (Peeters, 2004), defined as the temporal average slope of the energy during the attack phase

$$\mathrm{ATS} = \frac{\mathrm{AP}}{\mathrm{AT}}, \tag{10}$$

where AP is the peak value of the signal envelope reached at time $T_{\max}$, and $\mathrm{AT} = T_{\max} - T_{\text{threshold}}$ is the attack time. According to our observations, the values of this attack descriptor should increase for a round cello note and decrease for a harsh one.

3. Spectro-temporal descriptor

The third feature of the Gabor mask suggested the presence of energy fluctuations or asynchrony in the temporal deployment of each harmonic amplitude. The instantaneous amplitudes of harmonic components were thus extracted for each round-harsh sound pair. In Fig. 4, we present this subband decomposition for the same pair used in Gabor mask analysis. Interestingly, it can be observed that the nature of harmonic interactions seems quite different between the two types of sounds. Indeed, the temporal harmonic structure deploys itself regularly in the round signal, whereas it is much more chaotic and disordered in the harsh one, with noticeable reinforcements in energy of the fifth component around 120 ms and 220 ms after onset. This spectro-temporal phenomenon might be efficiently captured by the HSV (harmonic spectral variation), a descriptor related to the spectral flux which focuses on harmonic components, and that has already been used by Chudy *et al.* (2013) to assess timbre changes in cello performances. The HSV is a descriptor related to the time-varying spectral content, describing the spectral variation of harmonic amplitudes between adjacent frames. At the frame level n, the HSV is defined as the complement to one of the normalized correlation between the amplitudes of harmonic peaks from two adjacent frames (Kim *et al.*, 2006)

$$\mathrm{HSV} = \frac{1}{N}\sum_{n=2}^{N}\left\{1 - \frac{\sum_{k=1}^{K}\left(A_{k,n-1}A_{k,n}\right)}{\sqrt{\sum_{k=1}^{K}A_{k,n-1}^2}\sqrt{\sum_{k=1}^{K}A_{k,n}^2}}\right\}, \tag{11}$$

where $A_{k,n}$ denotes the amplitude of the kth harmonic peak in the nth frame. K and N are, respectively, the total number of harmonics and frames considered. The values of this spectral fluctuation descriptor should decrease for a round cello note and increase for a harsh one.

E. Validating harshness descriptors

In order to assess the relevance of the six previously presented descriptors to characterize the harshness at the signal level, we performed statistical tests on the sound corpus, composed of round and harsh notes divided in two groups of ten samples. Mean values and standard deviation were then computed on the two groups for each descriptor. The results, presented in Fig. 5 and Table I, indicate that all the descriptors except the LAT are promising candidates for explaining the harshness. The relevance of each signal descriptor was evaluated by performing a simple paired two-tailed t-test, based on the null hypothesis that the means of the two groups are the same. We can also observe that, except for the LAT and ATS, the descriptor values are much more spread for harsh sounds than for round sounds. Regarding the spectral descriptors (HSC, TRIratio, and MFCCratio), this phenomenon suggests that harsh sounds would sound more or less bright, according to spectral energy increases in the upper or middle parts of the spectrum, respectively. By contrast, round sounds may be perceived as "deep" and "mellow," because their spectral energies are better localized on low-frequency components. In the same way, the spectro-temporal descriptor values (HSV) that reveal a stronger spectral spread within harsh sounds, suggests different levels of asynchrony in the temporal evolution of transients. By contrast, round sounds may be characterized overall by a more systematic synchronicity between the harmonic transients. In the two next parts, we will try to assess the perception of these signal tendencies through two complementary analysis methods.

III. BUILDING A PERCEPTUAL MODEL OF HARSHNESS

From the validation of harshness descriptors, we concluded on five acoustic descriptors that potentially could discriminate round and harsh sounds from a signal point of view. In this part, we propose an experimental protocol that aims at identifying the most prominent signal descriptors reflecting the variation of perceived harshness. This is based on a perceptual evaluation of intermediate sounds between round and harsh signals of a pair.

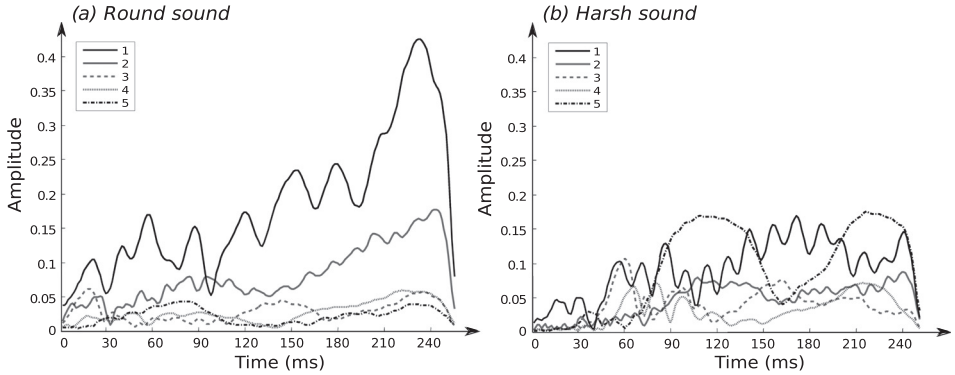

FIG. 4. Instantaneous amplitudes of the first five harmonics for the (a) round and (b) harsh sounds (sound pair of Fig. 2).

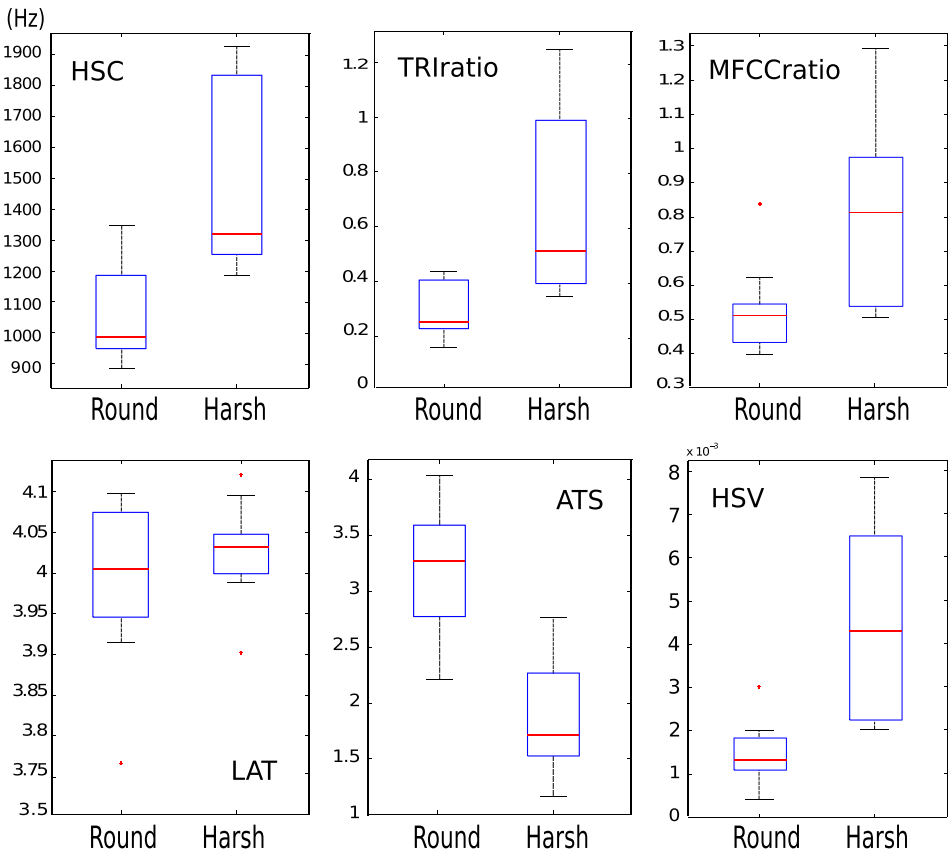

FIG. 5. (Color online) Comparison of spectral, temporal, and spectro-temporal features between the two groups of round and harsh sounds. The central marks are the medians, the edges of the boxes are the 25th and 75th percentiles.

A. Participants

Fifteen experienced cellists took part in the experiment. They were all volunteer teachers or students from musical schools or the conservatory of Marseille. All of them had extensive experience in cello playing, from 7 to 40 years of practice (the youngest had teacher recommendations). None of them had self-reported hearing problems.

B. Continua of stimuli

An intuitive way to assess an amount of perceived harshness consists in comparing a continuum of gradually harsher sounds with a reference sound judged to be of good quality (round; Aramaki *et al.*, 2009; Aramaki *et al.*, 2011). The ten sound pairs of the corpus were thus used as a base to build continua of stimuli for this experiment. For each sound pair, we carried out a morphing process to create a continuum of four sounds gradually deteriorating into harshness. This process first required a synchronization of the durations of the round/harsh signals for each pair. It was achieved with the *Adobe Audition* software, by using the same time alignment technique as for Gabor mask construction.

TABLE I. Results of paired t-tests on the six defined acoustical descriptors. The relevance of discriminating the round and harsh groups of ten sounds for each descriptor is given by the p-value p: $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$.

Descriptors	HSC	TRIratio	MFCCratio	LAT	ATS	HSV
$t(9)=$	5.69^{***}	4.1^{**}	4.13^{**}	1.21	-6.17^{***}	5.16^{***}

The morphing process between pairs of round and harsh sounds was possible thanks to the properties of the Gabor multipliers (Olivero *et al.*, 2010). Indeed, from Eq. (2), we define $\mathbb{M}_{\mathbf{m}}$, the Gabor multiplier of mask $\mathbf{m}$, which morphs into a target signal x_2 by a pointwise multiplication of its mask with the Gabor transform coefficients of the source signal x_1,

$$x_2 \simeq \mathbb{M}_{\mathbf{m}} x_1 = \sum_{m,n} \mathbf{m} c_{m,n}(x_1) h_{m,n}, \tag{12}$$

where $c_{m,n}(x_1)$ is the Gabor transform of the source signal x_1, and $h_{m,n}$ the dual synthesis window of $g_{m,n}$ [Eq. (1)].

To create the four sounds composing the continuum of a given pair, we provided the multiplier equation [Eq. (12)] with the round sound as the source signal x_1, and the harsh one as the target signal x_2. Intermediate sounds of these two signals were generated by computing several Gabor masks with gradually decreasing values of the regularization parameter λ [Eq. (3)]. The steps selected for λ have been chosen by the authors through an informal listening test to produce the most continuous evolution in harshness perception. For each sound pair, two Gabor masks corresponding to two gradual intermediary λ values were computed, and the two associated sounds resynthesized with the multiplier formula. The four sounds composing each continuum were finally equalized in loudness with the Loudness toolbox (Genesis, 2009). The whole procedure performed on each of the ten sound pairs resulted in a sound corpus composed of ten continua of four stimuli (cf. Rozé, 2017, for examples of continua).

Note that the morphing process that transforms the round sound into a harsh one is a monotonously decreasing mathematical transformation, which not necessarily reflects a linear perception of harshness. For this reason, in some continua, it was slightly more difficult to ensure the perceived continuity between the second synthetic stimulus and the harsh target sound x_2. The subject's scores observed in the following MUSHRA (multi stimulus test with hidden reference and anchor)-based listening test [Fig. 6(a)] confirmed this perceptual tendency.

C. MUSHRA-based procedure

The listening test of this experiment was designed with the MATLAB software, and installed on a laptop with Sennheiser HD-650 headphones (Wedemark, Germany). We performed it in the laboratory or in quiet places depending on the cellists' availabilities. Its conception was inspired from the MUSHRA comparison method (ITU-R Recommendation BS.1534-1, 2003), commonly used for audio codec quality assessment. The MUSHRA design relies on multiple comparisons of items with respect to a given reference. This reference is hidden among the set of items in order to potentially eliminate subjects who cannot recognize it.

We designed an interface based on these principles (cf. Rozé, 2017), but suited to our perceptual comparisons of harshness. Each subject was randomly presented with ten successive trials corresponding to the ten continua of four stimuli to compare. Within each continuum, the presentation order of the four sounds was randomized. The round sound was chosen as the hidden reference of the MUSHRA test. Subjects were asked to rate the relative amount of perceived harshness between this reference and each of the four sounds of the continuum according to a continuous quality scale (CQS). This CQS consists of identical graphical scales divided into five equal intervals gradually composed from the bottom to the top scale of the adjectives "barely more harsh" (*a peine plus* in French), "a little more harsh" (*un peu plus*), "more harsh" (*plus*), "considerably more harsh" (*bien plus*), and "much more harsh" (*beaucoup plus*). Subjects

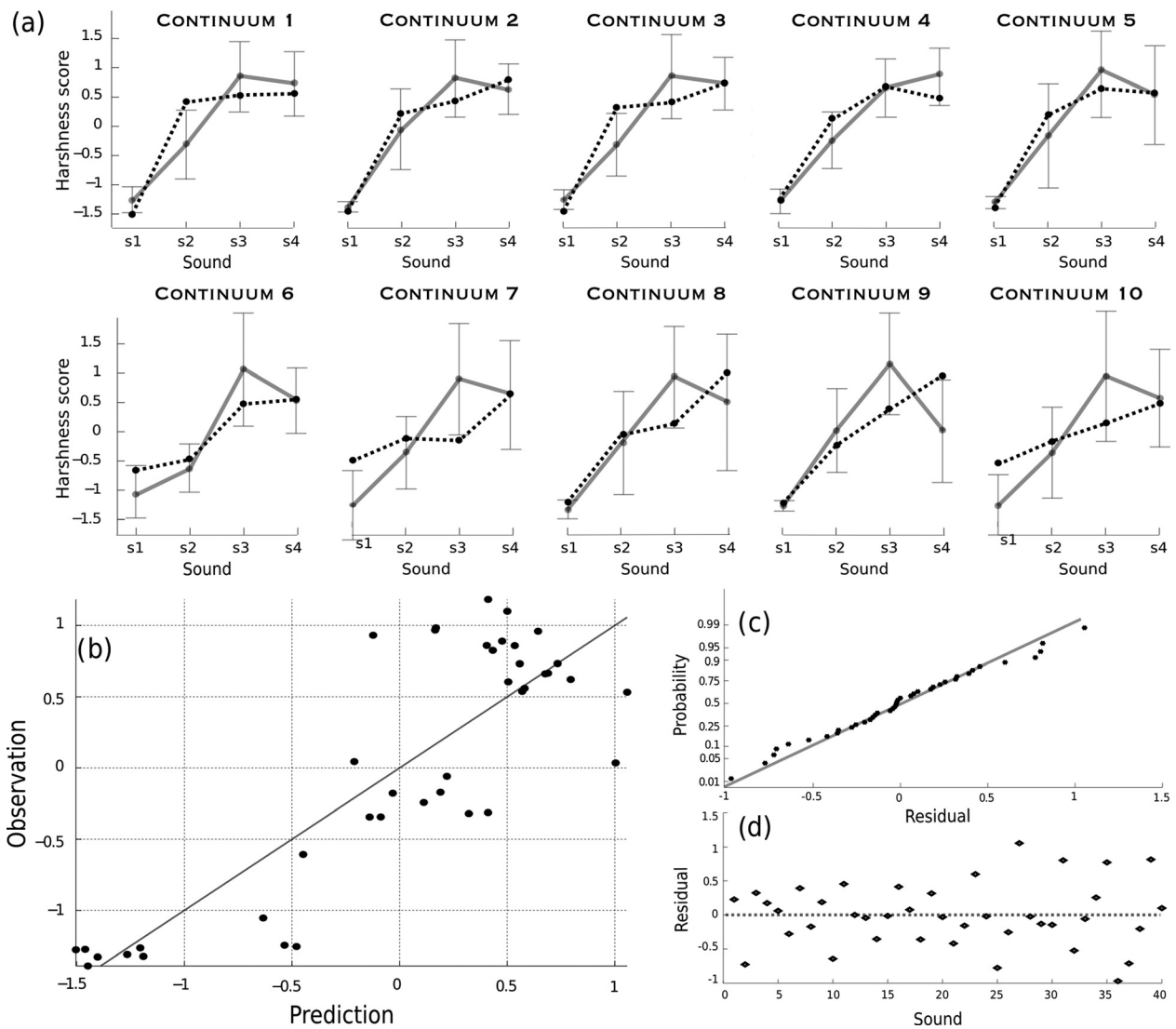

FIG. 6. Observed vs predicted scores of the least-square multiple linear regression model (a) by continuum (observed with the subjects' standard deviations in grey solid line, predicted in black dotted line) and (b) for all the continua ($R^2 = 0.724$, $n = 40$). (c) Normal probability plot of residuals. (d) Raw dispersion of residuals.

could rate 0 for a sound on this scale if they did not perceive a difference in harshness with respect to the reference, or even judge it identical. They could listen to the sounds as many times as desired, but it was not possible for them to return to previous continua once their choice was validated.

D. Data analysis

To investigate the relationships between the multidimensional set of acoustic features and the perceptual scores of the subjects, we carried out a multiple linear regression analysis. Two regression equations containing the most significant signal predictors were inferred. We here describe the different data preparation steps for the regression process.

1. Score checking

Subjects' ratings were most often zero for the sound corresponding to the hidden reference, although a small amount of harshness was sometimes attributed for this reference. We checked that the score of the round reference was always the lowest within each continuum. Consequently, we did not reject any continuum or subject.

2. Dataset building

For the 40 sounds composing the experimental plane, we computed the 5 acoustical descriptors from Sec. II E. In addition, the distribution of the four perceptual scores within each continuum was centered and normalized (mean 0, standard deviation 1) to compare the ten continua independently of their respective references. These perceptual scores were averaged on all the subjects for each sound.

3. Multiple linear regression

We performed a multiple linear regression analysis on the dataset in order to find a relevant model of signal descriptors which might explain the perceptual harshness scores (40 observations). This regression relies on typical least squares fitting, a process consisting in searching the fitting coefficients β_p, which minimize the mean square difference ϵ^2 between the model $\widehat{Y}$ (or prediction vector $X^p\beta_p$) and the response vector Y,

$$Y = \widehat{Y} + \epsilon = X^p\beta_p + \epsilon, \tag{13}$$

where $p = 5$ independent variables X^p of signal descriptors computed on the 40 sounds. To certify the robustness of this basic model, we checked the distribution of its residual vector ϵ, and validated the assumptions of linearity and homoscedasticity.

As a complementary approach, we also performed a coarser, stepwise regression to identify the most prominent descriptors of the basic predictive model. This stepwise regression has the advantage of automatically detecting multicollinearities among the predictors. It is here used as an additional approach to keep the less redundant descriptors that are likely to explain the perceived harshness scores.

E. Results

The results of the least-square multiple linear regression method are presented in Table II. They suggest that three out of the five regression coefficients are statistically different from zero (significance level 5%): The spectro-temporal descriptor HSV is predominant in the model ($p < 0.001$), the temporal descriptor ATS also seems to be relevant ($p < 0.01$), and the MFCCratio appears as the most relevant among the purely spectral descriptors ($p < 0.05$).

Figure 6 describes the fitting between the predictive model and the observed scores. In Fig. 6(a), we notice that the subjects actually perceived a global increase in amount of harshness within most sound continua. This perception was not fully linear, since they often rated the second synthetic stimuli (s3) as harsher than the real harsh sound (s4). This can be explained as a consequence of the morphing process that does not directly reflect human perception. In the same way, most of the predictions within each continuum follow an increasing tendency, which reveals that it was possible to find a linear combination of acoustic descriptors that fitted quite well the observed scores of harshness. This tendency is confirmed in Fig. 6(b) with a rather significant determination coefficient ($R^2 = 0.724$ and adjusted $R^2 = 0.683$). Similarly, the statistical distribution of the residuals [Fig. 6(c)] also tends to confirm that linearity and homoscedasticity assumptions are verified. Indeed, the linearity hypothesis appears in the normal probability plot of residuals, and homoscedasticity in the regular scattering of residuals on both sides of the axis $y = 0$. A Lilliefors test confirmed that the residual vector ϵ exhibits a normal distribution ($p > 0.5$).

From the process of least-square regression, we thus obtain a basic predictive model of perceived harshness. This can be written as an equation predicting the cellists' scores from the three best fitting descriptors

$$Y_{\text{basic}} = 0.29\text{MFCCratio} - 0.36\text{ATS} + 0.49\text{HSV}. \tag{14}$$

The stepwise regression approach provides a complementary model of lesser significance ($R^2 = 0.684$ and adjusted $R^2 = 0.667$), retaining only two prominent descriptors: the HSV ($p < 0.001$) and the ATS ($p < 0.001$),

$$Y_{\text{stepwise}} = 0.59\text{HSV} - 0.38\text{ATS} \tag{15}$$

F. Discussion

The global equation [Eq. (14)] suggests that the sensation of perceived harshness is multidimensional and characterized by a complex transformation of timbre involving spectral, temporal, and spectro-temporal aspects. Interestingly, we can

TABLE II. Results of the least-square multiple linear regression. The p-value p of each descriptor gives the relevance to add its coefficient β_p to the model: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Descriptors	HSC	TRIratio	MFCC ratio	ATS	HSV
Coefficients β_p	0.45	−0.58	0.29^{*}	-0.36^{**}	0.49^{***}

notice the negative coefficient of ATS, while positive coefficients are obtained for HSV and MFCCratio. This observation reveals an anti-correlation between temporal and spectral features, already observed in Sec. II D: An increase in perceived harshness induces a formantic emergence and more harmonic asynchrony, associated to a decrease in the temporal ATS. This is consistent with the Helmholtz motion, which builds more slowly or is more unstable. Furthermore, the stepwise regression eliminates the MFCCratio as a harshness predictor, which suggests a certain collinearity between the MFCCratio and the HSV descriptors. This dependence is confirmed by the correlation coefficient between the two descriptors ($r^2 = 0.65^{***}$), which can be explained from signal considerations, since the emergence of a formant corresponds to a sudden growth of upper partials within the harmonic's life cycle [see the evolution of the fifth component in Fig. 4(b) for an example]. This implies more crossings between the temporal evolutions of harmonic amplitudes, which should result in a stronger harmonic disorder as an indirect consequence.

Gabor multipliers turned out to be a quite efficient tool to create continua with increasingly harsher sounds. The degradations of timbre produced by this technique led to a model of perceived harshness. However, this approach presents some limitations since it proposes a statistical construction relying on the variations of a single feature (the regularization parameter λ) which has no significance in terms of acoustic descriptors. This is not sufficient to evaluate the individual perceptual relevance of each descriptor. Section IV aims at digging deeper into these aspects.

IV. TOWARD A CONTROL OF PERCEIVED HARSHNESS

In this section, we are going to assess to which extent each descriptor brought to light by the predictive model contribute to the perceived harshness. This procedure is essential as a perceptual validation of each descriptor. For this purpose, a synthesis process is used to manipulate/control the temporal and spectral signal envelopes. This control will further be perceptually validated by a paired comparison test, resulting in the elaboration of a perceptual harshness space.

A. Choice of control descriptors

The basic predictive model [Eq. (14)] of perceived harshness suggests that both spectral and temporal features are involved in this phenomenon. Among the three main descriptors (HSV, ATS, and MFCCratio) characterizing the harshness, the HSV was found to be predominant. However, this descriptor is not suitable for resynthesizing and controlling the harshness, since it encompasses too many details in the signal. Instead, we might take advantage of the spectro-temporal duality of the other descriptors unveiled in the previous experiment to attempt to control the harshness through only a few parameters: the coefficients composing the MFCCratio, which can be considered as a shaping parameter of the spectral sound envelope (and which is correlated with the HSV), and the ATS, defining the attack slope, likely to be considered as a shaping parameter of the temporal attack envelope.

The validity of this assumption can be tested by carrying out a partial bivariate least-square linear regression on the purely spectral MFCCratio and purely temporal ATS predictor. This leads to a less relevant predictive model than Eq. (14) ($R^2 = 0.595$ and adjusted $R^2 = 0.573$), but with satisfying p-value predictors: MFCCratio ($p < 0.001$) and ATS ($p < 0.001$). We therefore decided to use these descriptors to control the perceived harshness.

B. Participants

The harshness, produced by this synthesis method, was assessed through a listening test based on dissimilarity ratings of sound pairs. The participants who took part in this listening test were the same 15 experienced cellists as in Sec. III A. Each of the two tests lasted for around 15–20 min for each subject. The tests were separated by a short pause. The total duration of these two listening tests was thus around 45 min per subject.

C. Stimuli

1. Cross-synthesis

A round/harsh representative sound pair was chosen by informal listening tests from the ten paired samples composing the sound corpus. For each signal of this pair, we used the technique of subband decomposition of the harmonics to design a synthetic calibrated sound composed of 25 harmonics. This number of harmonics was necessary to obtain a perceptually convincing cello sound.

From this root synthetic pair (a round and a harsh sound) called S1/S8, we created six intermediary synthetic stimuli using a cross-synthesis technique performed in the spectral and temporal domains. This technique requires the parameterization of spectral and temporal envelopes by means of a basis function expansion: $\text{env} = \sum_{k=1}^{K} C_k \phi_k$, where C_k are the coefficients of the expansion, and ϕ_k the set of basis functions to be combined linearly. This parameterization turns out to be a very flexible way to control and adjust the shape of the envelopes since it enables us to design envelopes, presenting intermediate spectral or temporal features of round and harsh sounds.

To apply intermediate envelopes to the source sound, a whitening process consisting in dividing the source signal by its own envelope (env_1) was first effectuated. The resulting (whitened) signal was then multiplied by the intermediate envelope (env_2). This cross-synthesis technique thus consisted in filtering the source sound by the transfer function $\text{env}_2/\text{env}_1$. To avoid divergence effects due to potential divisions by zero of the source envelope env_1, we turned the division of the transfer function into a difference of logarithmic envelopes (Zölzer, 2008): $\log(|\text{env}_2|/|\text{env}_1|) = \log|\text{env}_2| - \log|\text{env}_1|$. By applying the exponential operator to this logarithmic difference, the cross-synthesis process of envelopes could be achieved accordingly.

a. Spectral cross-synthesis. In the spectral domain, we parameterized a formantic envelope as an expansion of the mel-cepstrum coefficients C_k, with the discrete cosines

functions ϕ_k as a basis. Indeed, according to Eq. (7), we can rebuild a logarithmic envelope E_{spec} of the spectrum signal magnitude $|X(f)|$ from a vector of MFCC coefficients by means of an inverse DCT (Poli and Prandoni, 1997),

$$E_{\text{spec}}(\text{mel}(f)) = \sum_{k=1}^{N_c} C_k \cos\left(2\pi k \frac{\text{mel}}{B_m}\right), \quad 0 \le \text{mel} \le B_m \\ \simeq \log|X(f)|, \tag{16}$$

where N_c is the subset of coefficients chosen for resynthesizing the spectrum envelope. The bandwidth of the $N_f = 24$ mel-filter bank already used for computations of MFCCs is $B_m = 2700$ mel (Sec. II D). E_{spec} represents a smoothed version of the real spectral envelope, whose precision depends on the number of coefficients used for the reconstruction (we chose $N_c = N_f/2 = 12$). From the MFCC vector modeling the formantic envelope, we then obtained intermediate spectral envelopes between sounds S1 and S8 by tuning the coefficients C_1 and C_2 composing the MFCCratio.

b. Temporal cross-synthesis. In the time domain, we parameterized a temporal envelope as an expansion of B-spline basis functions ϕ_k weighted by coefficients C'_k. It approximated the temporal signal $e(t)$ by an envelope E_{temp} composed of the sum of a small set of scaled splines (Ramsay, 2006)

$$E_{\text{temp}}(t) = \sum_{k=1}^{m+L-1} C'_k B_k(t,\tau) \simeq e(t), \tag{17}$$

where $B_k(t,\tau)$ is the value at t of the B-spline basis function defined by an order m and a knot sequence τ of length $L-1$. The number of splines was defined by the order plus the number of interior knots $m+L-1$. A basis of five piecewise cubic splines (order $m=3$) was chosen to parameterize the temporal envelopes with quite a good fit. From there, we could design some intermediate temporal envelopes between sounds S1 and S8, by specifically tuning the coefficients C'_3 and C'_4 of their five-spline vectors. Indeed, this manipulation made it possible to move the peak attack amplitude forward or backward, which intrinsically corresponded to controlling the ATS.

2. Eight synthetic sounds

Figure 7 gives a synopsis of this stimuli design structured into two clusters, a round one and a harsh one. Each cluster is composed of the initial root sound (S1 or S8, respectively), and three synthetic sounds designed by cross-synthesis from the initial stimuli (cf. Rozé, 2017).

In the spectral domain, we created two sounds in each cluster. For the round cluster created from the round sound S1, the formantic envelope of the sound S3 was designed by decreasing C_1 and increasing C_2. Symmetrically, for the harsh cluster created from the harsh sound S8, the formantic envelope of the sound S6 was designed by increasing C_1 and decreasing C_2. The final values of the MFCC coefficients C_1 and C_2 hereby defined the intermediate spectral envelopes of

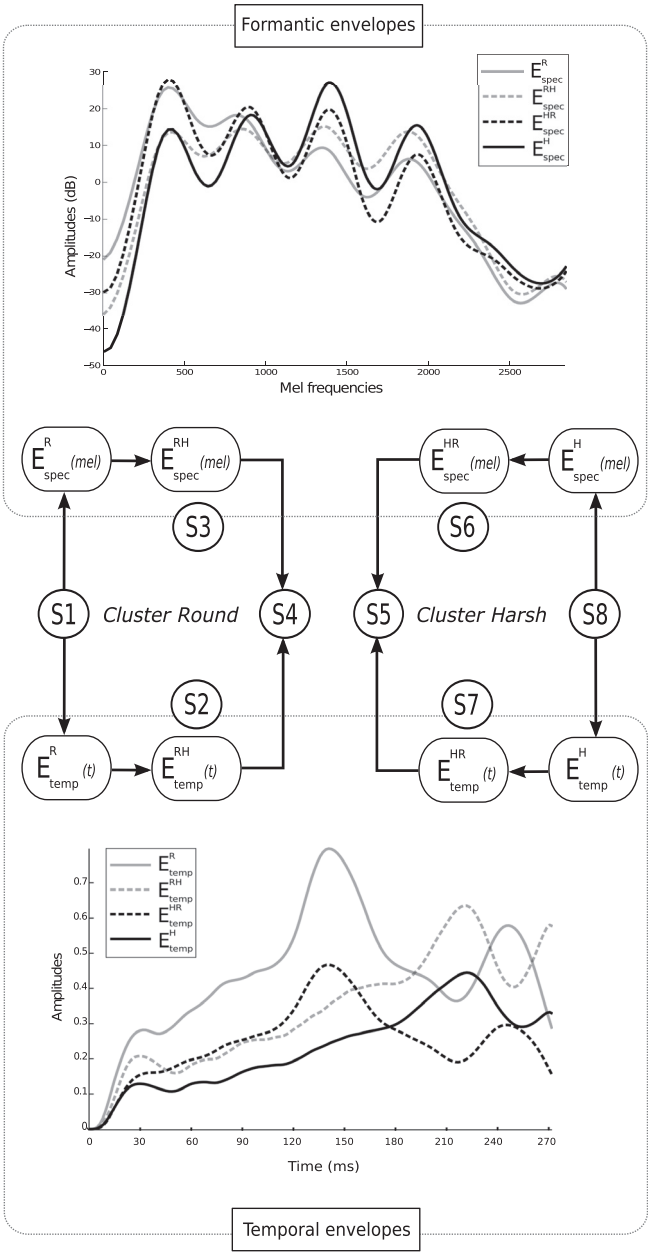

FIG. 7. Design by cross-synthesis of the eight synthetic stimuli. (Top) The two original spectral envelopes of S1 (E^R_{spec}) and S8 (E^H_{spec}) with their respective intermediaries (E^{RH}_{spec} and E^{HR}_{spec}). (Middle) Cluster of the four synthetic sounds built from the round sound S1 (mid-left) and from the harsh sound S8 (mid-right). (Bottom) The two original temporal envelopes of S1 (E^R_{temp}) and S8 (E^H_{temp}) with their respective intermediaries (E^{RH}_{temp} and E^{HR}_{temp}).

S1 and S8. The upper part of Fig. 7 presents the shapes of the formantic envelopes for these four sounds:

$$\widehat{S3} = \widehat{S1} \times \frac{E^{RH}_{\text{spec}}}{E^{R}_{\text{spec}}}; \quad \widehat{S6} = \widehat{S8} \times \frac{E^{HR}_{\text{spec}}}{E^{H}_{\text{spec}}},$$

where $\widehat{S_i}$ represents the spectrum of the sound S_i.

In the temporal domain, we also created two sounds in each cluster. For the cluster round created from the round sound S1, the sound S2 with a lower ATS was designed by decreasing C'_3 and increasing C'_4. Symmetrically, for the cluster harsh created from the harsh sound S8, the sound S7, with a higher ATS, was designed by increasing C'_3 and

decreasing C'_4. The final values of the spline coefficients C'_3 and C'_4 hereby defined the intermediate temporal envelopes of S1 and S8. The bottom part of Fig. 7 presents the shapes of the temporal envelopes for these four sounds:

$$\mathrm{S2} = \mathrm{S1} \times \frac{E^{RH}_{\mathrm{temp}}}{E^{R}_{\mathrm{temp}}}; \quad \mathrm{S7} = \mathrm{S8} \times \frac{E^{HR}_{\mathrm{temp}}}{E^{H}_{\mathrm{temp}}}.$$

The two last synthetic stimuli were created by hybrid crossings of the spectral and temporal envelopes in each cluster. For the cluster round, the sound S4 was designed by applying the temporal envelope of S2 to the sound S3. Symmetrically, for the cluster harsh, the sound S5 was designed by applying the temporal envelope of S7 to the sound S6:

$$\mathrm{S4} = \mathrm{S3} \times \frac{E^{RH}_{\mathrm{temp}}}{E_{\mathrm{temp}}(\mathrm{S3})}; \quad \mathrm{S5} = \mathrm{S6} \times \frac{E^{HR}_{\mathrm{temp}}}{E_{\mathrm{temp}}(\mathrm{S6})},$$

where $E_{\mathrm{temp}}(S_i)$ represents the temporal envelope of the sound S_i.

Finally, we equalized the loudness of the eight synthetic stimuli by the means of the Loudness toolbox (Genesis, 2009) to ensure timbre comparisons with the same sound level.

D. Paired-comparison procedure

The listening test of this experiment was performed subsequent to the previous one (Sec. III C) with the same equipment and in the same conditions. We designed an interface (cf. Rozé, 2017) to evaluate the amount of perceived harshness between pairs of the eight synthetic stimuli. The combinations of pairs were presented in both ways (direct AB and reverse BA), and in a global random way for each subject to avoid any effect of presentation order. Since identical pairs were not presented, a total of 56 pairs (8×7) was evaluated.

For each sound pair presented, the subject was asked to process two tasks: first, choose the sound of the pair A/B perceived as the most harsh. It was possible to check a third item entitled "none" if the two sounds were considered equal in terms of harshness. Then, assess the difference of perceived harshness of the checked sound relatively to the other one, according to a CQS, corresponding to a graphic scale divided into three equal intervals, i.e., "a little bit more harsh," "more harsh," and "much more harsh." The subjects were informed that they could rate zero on this scale, if they judged the two sounds of a pair identical in terms of harshness, even if they were structurally different. In this case, they had to check the "none" item. Conversely, the harshness rating was only available when a non-zero rating was chosen on the scale. They could listen to the sounds of each pair as many times as they desired, but it was not possible for them to return to previous pairs once their choice was validated.

E. Data analysis

To explore the repartition of the subjects' ratings, we performed a metric MDS analysis. We provided the MDS with a dissimilarity matrix composed of the harshness proportion rated by the subjects between pairs of synthetic stimuli. This dissimilarity matrix was designed as a full non-symmetric 8×8 table of 56 harshness ratings, including the 2 ways of presenting the stimuli (direct AB in the upper-triangular and reverse BA in the lower-triangular). Since only symmetrical matrices can be used in MDS analyses, we averaged the dissimilarity ratings of each symmetric element and stored this average in an upper-triangular matrix of 28 (56/2) distances. This process was repeated for each participant, and finally we obtained a global dissimilarity matrix by averaging the individual matrices of all subjects. The quality of the obtained MDS configuration was evaluated by taking into account two criteria: a scree-elbow criterion of the Kruskal stress evolution (for the choice of the right number of perceptual dimensions in the new space), and an analysis of the Shepard diagram (for the preservation of the distances in the new space relative to initial dissimilarities). Finally, each dimension of the MDS space was correlated with the salient signal descriptors of harshness previously obtained.

F. Results and discussion

The goodness-of-fit criterion, corresponding to a minimization of the Kruskal stress, indicated that three dimensions were sufficient to get an optimal MDS configuration (stress ≈ 0.093). Figure 8 presents the tridimensional space of the eight synthetic stimuli obtained by the MDS, with its two main bi-dimensional projections. This system depicts a perceptual harshness space. Table III reports the correlations between each MDS dimension and the five acoustic descriptors highlighted in Sec. II E.

On the first spatial dimension, a very strong correlation can be observed with the HSVs descriptor, and to a lesser extent with the temporal ATS. The high HSV score suggests that the perceived harshness essentially increases with a loss of synchronism between the harmonics of the signal. The ATS score, weaker and negatively correlated to the first dimension, reveals a concurrent loss of ATS in the temporal signal envelope. The predominance of HSV in these results is a bit surprising, given that it was not directly used to create the stimuli. However, these variations in harmonic chaotic behavior actually emerged as an indirect consequence of synthesis transformations on the formantic envelopes.

Interestingly, the first dimension of Fig. 8(a) splits the three-dimensional (3D) space into two groups matching the clustering design of the synthetic stimuli: At the left, we obtain the round cluster (S1, S2, S3, S4) and at the right the harsh one (S8, S7, S6, S5). Besides, the pairs of stimuli that only differ by their ATS (S1/S2, S3/S4, S5/S6, and S7/S8) rate quite similarly on this axis. Therefore, in the round cluster, the intermediary pair of stimuli S3/S4 was perceived as harsher than the S1/S2 pair because of a stronger harmonic asynchrony coupled with a smallest temporal ATS. Symmetrically, in the harsh cluster, the intermediary pair of stimuli S5/S6 was perceived as rounder than the S7/S8 pair because of less harmonic asynchrony coupled with a greatest temporal ATS. These results validate the physical relevance of the HSV and ATS descriptors in the predictive harshness

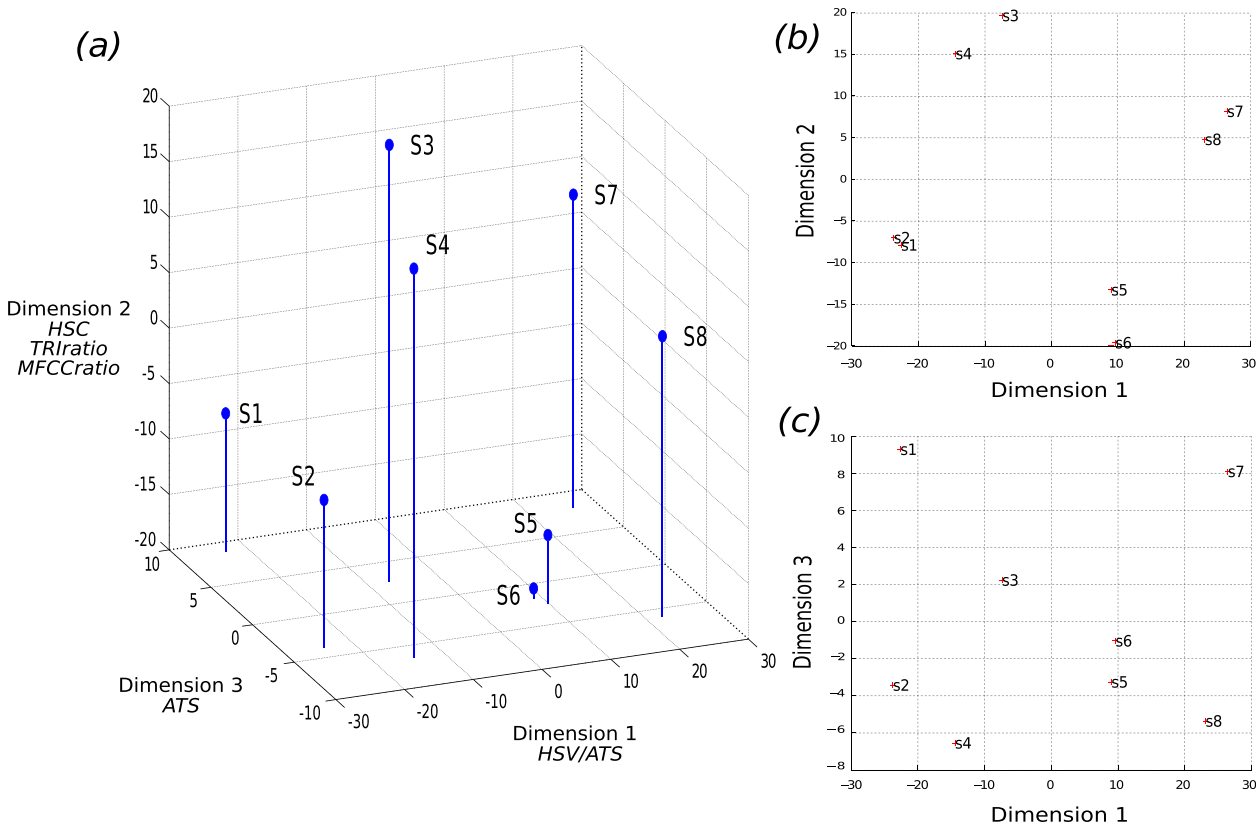

FIG. 8. (Color online) (a) Perceptual space resulting from the metric MDS achieved in three dimensions (Kruskal stress = 0.093). The spatial configuration is projected on the two main pairs of dimensions: (b) Dim1 vs Dim2, (c) Dim1 vs Dim3.

model [Eq. (14)]. It is also consistent with the stepwise complementary approach of the previous experiment [Eq. (15)], and confirms the central role of the simultaneous opposite variations of HSV/ATS in harshness perception.

Referring back to Table III, the second spatial dimension correlates quite well with all the purely spectral descriptors HSC, TRIratio, and MFCCratio. The strong correlation with the MFCCratio validates the physical coherence of this descriptor in the predictive model of harshness [Eq. (14)]. Besides, the HSC and TRIratio descriptors arise as corollaries from modifications of the spectral shape, since the emergence of a formant necessarily implies a transfer of spectral energy. In Fig. 8(b), the projection of the MDS configuration on the first two dimensions shows that the stimuli do not follow the cluster design any longer. Certain sounds synthesized from intermediate formantic envelopes were actually rated outside the bounds of the two root sounds (S1 and S8). On the second dimension, both stimuli S3/S4, formantically harsher than S1, seem hereafter to be perceived with more degraded quality than the root harsh sound S8. And symmetrically, both stimuli S5/S6, formantically less harsh than S8, now seem to be perceived with better timbral quality than the round root sound S1. This result turns out to comply with the evolution of the MFCCratio computed from intermediate values of the first two Mel cepstrum coefficients ($C1$ and $C2$). Therefore, the second dimension should be interpreted with respect to the spectral energy distribution, and tends to characterize the harshness phenomenon in terms of perceived brightness.

TABLE III. Correlations between the salient signal descriptors of harshness and the coordinates of the eight stimuli along the three dimensions of the MDS space, p-values: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Significant values appear in boldface.

Acoustic correlates	Dim 1	Dim 2	Dim 3
HSC	0.52	**0.84****	0
TRIratio	0.58	**0.79***	−0.01
MFCCratio	0.04	**0.84****	0.1
ATS	**−0.79***	−0.37	0.32
HSV	**0.98*****	0.02	−0.02

Finally, Table III does not reveal at first sight any acoustic correlates of the third spatial dimension. However, the projection of the MDS configuration on the dimensions one and three of the perceptual space [Fig. 8(c)] reveals a surprising organization of the stimuli. Indeed, they are distributed on either side of the third axis, according to pairs only differing by the ATS of their temporal envelope. With the exception of the pair S5/S6 perceived a little bit confusedly, the sounds composed of a harsh temporal envelope are negatively rated (S2, S4, S8), whereas those composed of a round temporal envelope are positively rated (S1, S3, S7). This third dimension thus reflects that subjects only detected changes in temporal ATS, without significant consequences on their perception of harshness.

V. CONCLUSION

In this paper, we explored a deterioration of the timbre quality for cello sounds, often characterized as sound

harshness among professional cellists. By using analysis-synthesis techniques and suitable listening tests, this perceptually undesirable feature was modeled into a combination of some relevant signal descriptors. First, a signal-based approach using Gabor masks (Sec. II) revealed that a harsh sound may essentially be characterized by three facets: an energy transfer or formantic shift toward higher order partials of its spectrum, a weakened ATS of its temporal envelope, and more fluctuations in the harmonics' life cycle. Then, two complementary experiments enabled us to perceptually assess the effect of these signal features. On the one hand, a least-square linear regression (Sec. III) performed on the subjective ratings of continuously harsher sounds, allowed us to build a predictive model of the harshness. This model shed light on the spectro-temporal duality potentially responsible for the harsh perception. On the other hand, a MDS was carried out (Sec. IV), in order to perceptually validate this signal duality, and better assess the harshness contribution of each identified descriptor. To this aim, we attempted to create rounder and harsher synthetic stimuli simply by controlling their spectral and temporal envelope shapes. The MDS, performed on the perceived dissimilarities of pairs of these synthetic stimuli, were conclusive. As a matter of fact, the method gave rise to a perceptual harshness space, which two main dimensions were strongly correlated to the acoustic descriptors identified by the predictive model. The perceived harshness was thus partly due to the combination of a more chaotic harmonic behavior and a weaker ATS, which corresponds to a more unstable Helmholtz motion, building up more slowly. The shift in formant and energy transfer toward the upper partials of the spectrum appeared as corollaries of the loss in synchronicity between signal transients.

It should be noted that the perceptual space created by investigating the harshness phenomenon presents some interesting analogies with the recognized timbre spaces in the litterature (Grey, 1977; McAdams *et al.*, 1995; Barthet *et al.*, 2010). As a matter of fact, we can roughly consider by analogy with the terms of McAdams *et al.* (1995), our first dimension as the "spectral fluctuations," characterizing the variations of synchronicity in the transients of harmonics. Our second dimension may match with the "spectral energy distribution," and the third one may be interpreted in terms of "temporal patterns" (even though this dimension is not itself directly linked to harshness perception). We expect similar trends for notes that differ from the one investigated in this analysis, at least with respect to the signal descriptors. Nevertheless, further investigations are needed to verify the robustness of the perceived harshness phenomenon on other parts of the score.

The results of this study naturally encourage the pursuit of additional investigations of the gestures responsible for this sound degradation. In particular, it seems important to assess the coherency with physics-based synthesis models (Demoucron, 2008), regarding the relationships between the acoustical features composing the harshness and the control gestures of the musician. Particularly, the transfer of spectral energy responsible for the increase in brightness is likely to coincide with a stronger bow force or a slower bow velocity (Guettler *et al.*, 2003; Guettler, 2003). Similarly, the weak ATS is believed to match with an improper Helmholtz motion, caused by a lack of firmness in the onset of the note produced by the string–bow interaction. In addition to the changes directly related to the physical parameters of the bow, it could be very interesting to investigate how the postural constraint impacted the gestural coordination and consequently the timbre of the sound. For example, we may ask whether the cellists compensated their postural constraint by consistently adjusting the joint movements of their right arm, and if these specific adjustments resulted in the harsh sounds that we collected for this study. If this is the case, it might indicate that postural movements of the chest and the head play an important role in motor coordination patterns of the right arm's joints, and indirectly in the maintenance of a good (round) timbre quality.

A potential application of these results consists in designing new technological tools that enable real-time measures of harshness. This technology-enhanced learning would be of great interest for young cello students that are not always aware of how their gestures can be optimized in order to improve their sound quality. Inappropriate gestures might hereby be corrected by the feedback of such devices.

ACKNOWLEDGMENTS

This work is partly supported by the French National Research Agency and is part of the "Sonimove" project (Grant No. ANR-14-CE24-0018). We would like to thank the reviewers for their constructive comments related to the manuscript.

Aramaki, M., Besson, M., Kronland-Martinet, R., and Ystad, S. (**2011**). "Controlling the perceived material in an impact sound synthesizer," IEEE Trans. Audio, Speech, Lang. Process. **19**(2), 301–314.

Aramaki, M., Gondre, C., Kronland-Martinet, R., Voinier, T., and Ystad, S. (**2009**). "Thinking the sounds: An intuitive control of an impact sound synthesizer," in *International Conference on Auditory Display (ICAD'09)*, pp. 119–124.

Barthet, M., Guillemain, P., Kronland-Martinet, R., and Ystad, S. (**2010**). "From clarinet control to timbre perception," Acta Acust. Acust. **96**(4), 678–689.

Barthet, M., Kronland-Martinet, R., and Ystad, S. (**2008**). "Improving musical expressiveness by time-varying brightness shaping," in *Computer Music Modeling and Retrieval. Sense of Sounds* (Springer-Verlag Berlin-Heidelberg), pp. 313–336.

Barthet, M., Kronland-Martinet, R., Ystad, S., and Depalle, P. (**2007**). "The effect of timbre in clarinet interpretation," in *International Computer Music Conference (ICMC)* (Copenhagen, Denmark), pp. 59–62.

Bogert, B. P., Healy, M. J., and Tukey, J. W. (**1963**). "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, Vol. 15, pp. 209–243.

Caclin, A., McAdams, S., Smith, B. K., and Winsberg, S. (**2005**). "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," J. Acoust. Soc. Am. **118**(1), 471–482.

Chadefaux, D., Le Carrou, J.-L., Wanderley, M. M., Fabre, B., and Daudet, L. (**2013**). "Gestural strategies in the harp performance," Acta Acust. Acust. **99**(6), 986–996.

Chudy, M., Pérez Carrillo, A., and Dixon, S. (**2013**). "On the relation between gesture, tone production and perception in classical cello performance," in *Proceedings of Meetings on Acoustics ICA2013*, Vol. 19, No. 1, p. 035017.

Davis, S., and Mermelstein, P. (**1980**). "Experiments in syllable-based recognition of continuous speech," IEEE Trans. Acoust., Speech Signal Process. **28**, 357–366.

Demoucron, M. (**2008**). "On the control of virtual violins—Physical modelling and control of bowed string instruments," Ph.D. thesis, Université Pierre et Marie Curie-Paris VI; Royal Institute of Technology, Stockholm.
Depalle, P., Kronland-Martinet, R., and Torrésani, B. (**2006**). "Time-frequency multipliers for sound synthesis," in *Proc. the Wavelet XII Conference* (SPIE Annual Symposium, San Diego, CA), pp. 221–224.
Desmet, F., Nijs, L., Demey, M., Lesaffre, M., Martens, J.-P., and Leman, M. (**2012**). "Assessing a clarinet player's performer gestures in relation to locally intended musical targets," J. New Music Res. **41**(1), 31–48.
Fritz, C., Blackwell, A. F., Cross, I., Woodhouse, J., and Moore, B. C. (**2012**). "Exploring violin sound quality: Investigating English timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties," J. Acoust. Soc. Am. **131**(1), 783–794.
Genesis (**2009**). "Loudness toolbox," available at http://genesis-acoustics.com/en/loudness_online-32.html (Last viewed January 3, 2017).
Grey, J. M. (**1977**). "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. **61**(5), 1270–1277.
Grey, J. M., and Gordon, J. W. (**1978**). "Perceptual effects of spectral modifications on musical timbres," J. Acoust. Soc. Am. **63**(5), 1493–1500.
Guettler, K. (**2002**). "On the creation of the Helmholtz motion in bowed strings," Acta Acust. Acust. **88**(6), 970–985.
Guettler, K. (**2003**). "A closer look at the string player's bowing gestures," CASJ (Series II) **4**(7), 12–16.
Guettler, K., Schoonderwaldt, E., and Askenfelt, A. (**2003**). "Bow speed or bowing position—Which one influence spectrum the most?," in *Proceedings of the Stockholm Music Acoustic Conference (SMAC)*.
ITU-R Recommendation, B. S. 1534-1 (**2003**). "Method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA)," Technical Report, International Telecommunication Union, Geneva, Switzerland.
Kim, H.-G., Moreau, N., and Sikora, T. (**2006**). *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval* (John Wiley & Sons, Hoboken, NJ).
Lartillot, O., and Toiviainen, P. (**2007**). "A MATLAB toolbox for musical feature extraction from audio," in *International Conference on Digital Audio Effects*, pp. 237–244.
Leman, M. (**2008**). *Embodied Music Cognition and Mediation Technology* (MIT Press, Cambridge, MA).
McAdams, S. (**1999**). "Perspectives on the contribution of timbre to musical structure," Comput. Music J. **23**(3), 85–102.
McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (**1995**). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," Psychol. Res. **58**(3), 177–192.
Merer, A., Ystad, S., Kronland-Martinet, R., and Aramaki, M. (**2007**). "Semiotics of sounds evoking motions: Categorization and acoustic features," in *International Symposium on Computer Music Modeling and Retrieval* (Springer, Berlin Heidelberg), pp. 139–158.
Olivero, A., Torrésani, B., and Kronland-Martinet, R. (**2010**). "A new method for Gabor multipliers estimation: Application to sound morphing," in *European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, pp. 507–511.
Peeters, G. (**2004**). "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Technical Report, IRCAM.
Poli, G. D., and Prandoni, P. (**1997**). "Sonological models for timbre characterization," J. New Music Res. **26**(2), 170–197.
Pollard, H. F., and Jansson, E. V. (**1982**). "A tristimulus method for the specification of musical timbre," Acta Acust. Acust. **51**(3), 162–171.
Ramsay, J. O. (**2006**). *Functional Data Analysis* (John Wiley & Sons, Inc., Hoboken, NJ).
Rozé, J. (**2017**). "Exploring the perceived harshness of cello sounds by morphing and synthesis techniques," available at http://www.lma.cnrs-mrs.fr/~kronland/Cello_Harshness (Last viewed January 3, 2017).
Rozé, J., Aramaki, M., Kronland-Martinet, R., Voinier, T., Bourdin, C., Chadefaux, D., Dufrenne, M., and Ystad, S. (**2016**). "Assessing the influence of constraints on cellists' postural displacements and musical expressivity," in *Music, Mind and Embodiment—Post-Proceedings of CMMR* (Springer, New York), Vol. 9617 of LNCS, pp. 22–41.
Schaeffer, P. (**1966**). *Traité des Objets Musicaux* (*Treaty of Musical Objects*) (Editions du Seuil, Paris).
Schelleng, J. C. (**1973**). "The bowed string and the player," J. Acoust. Soc. Am. **53**(1), 26–41.
Schoonderwaldt, E. (**2009**). "The violinist's sound palette: Spectral centroid, pitch flattening and anomalous low frequencies," Acta Acust. Acust. **95**(5), 901–914.
Schoonderwaldt, E., Guettler, K., and Askenfelt, A. (**2003**). "Effect of the width of the bow hair on the violin string spectrum," in *Proceedings of the Stockholm Music Acoustics Conference (SMAC)*.
Sciabica, J.-F., Olivero, A., Roussarie, V., Ystad, S., and Kronland-Martinet, R. (**2012**). "Dissimilarity test modelling by time-frequency representation applied to engine sound," in *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*.
Søndergaard, P. L., Torrésani, B., and Balazs, P. (**2012**). "The linear time frequency analysis toolbox," Int. J. Wavelets, Multiresol. Inform. Process. **10**(04), 1250032.
Stepanek, J. (**2006**). "Musical sound timbre: Verbal description and dimensions," in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pp. 121–126.
Stepanek, J., and Otcenasek, Z. (**2005**). "Acoustical correlates of the main features of violin timbre perception," in *Proceedings of the Conference on Interdisciplinary Musicology*, pp. 1–9.
Thompson, M. R., and Luck, G. (**2012**). "Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music," Musicae Sci. **16**(1), 19–40.
Van Zijl, A. G., and Luck, G. (**2013**). "Moved through music: The effect of experienced emotions on performers' movement characteristics," Psychol. Music **41**(2), 175–197.
Wanderley, M. M., Vines, B. W., Middleton, N., McKay, C., and Hatch, W. (**2005**). "The musical significance of clarinetists' ancillary gestures: An exploration of the field," J. New Music Res. **34**(1), 97–113.
Zölzer, U. (**2008**). *Digital Audio Signal Processing* (John Wiley & Sons, Hoboken, NJ).