A Flexible Bio-inspired Hierarchical Model for Analyzing Musical Timbre

Mohammad Adeli, Jean Rouat, *Senior Member, IEEE,* Sean Wood, Stéphane Molotchnikoff, and Eric Plourde, *Member, IEEE*

***Abstract*—A flexible and multipurpose bio-inspired hierarchical model for analyzing musical timbre is presented in this paper. Inspired by findings in the fields of neuroscience, computational neuroscience, and psychoacoustics, not only does the model extract spectral and temporal characteristics of a signal, but it also analyzes amplitude modulations on different timescales. It uses a cochlear filter bank to resolve the spectral components of a sound, lateral inhibition to enhance spectral resolution, and a modulation filter bank to extract the global temporal envelope and roughness of the sound from amplitude modulations. The model was evaluated in three applications. First, it was used to simulate subjective data from two roughness experiments. Second, it was used for musical instrument classification using the k-NN algorithm and a Bayesian network. Third, it was applied to find the features that characterize sounds whose timbres were labeled in an audiovisual experiment. The successful application of the proposed model in these diverse tasks revealed its potential in capturing timbral information.**

***Index Terms*—Timbre, cochlear filter bank, modulation filter bank, temporal envelope, time-averaged spectrum, instantaneous roughness, musical instrument classification, Bayesian network, multimodal timbre characterization.**

I. Introduction

A. Timbre

SOUNDS have three major characteristics: pitch, loudness and timbre. Pitch and loudness are perceptual measures related to the fundamental frequency and the intensity of a sound, respectively. Timbre is a quality that allows to distinguish between the sounds that have the same pitch, loudness and duration [1–4]. It is a multi-dimensional attribute and mainly depends on harmonic content, attack and decay, and vibrato [1, 5, 6]. Attack is the rapid increase of a sound's amplitude to its maximum value while decay is the gradual decrease of the amplitude. Vibrato is the periodic change in the pitch of a sound.

The attributes of timbre have been investigated in several studies [7–15]. In such studies, subjects are typically asked to rate timbral similarities (or dissimilarities) of all stimulus pairs for a stimulus set with equalized pitch, loudness, and duration. The similarity ratings are then analyzed using Multidimensional Scaling (MDS) to construct an N-dimensional timbre space in which the subjective similarities are represented by the distances among the stimuli [3, 16, 17]. A 3D space has been found in most studies, where spectral centroid and attack time are the first and the second dimensions. However, for the third dimension, different features such as spectral flux [9, 16], pitch strength, noisiness [14], and spectral deviation [11] have been proposed as there is no consensus on what it encodes [9, 16].

In general, timbre has many attributes, which can be categorized into three classes: temporal (e.g. attack time, decay time, temporal centroid [9–12, 14, 16]), spectral (e.g. spectral centroid, spectral deviation, noisiness, spectral skewness [9–12, 14, 16]), and spectro-temporal (e.g. spectral flux, roughness, fluctuation strength [9–11, 16]). Most of these features are formally defined in [18] and [19].

B. Applications and Models of Timbre

Timbre plays a key role in recognition and localization of sound sources [13, 20] as well as stream segregation in Auditory Scene Analysis (ASA) [21–24]. It has been widely used in applications such as instrument recognition in monophonic and polyphonic music [25, 26], music genre recognition [27], and music retrieval [27, 28].

Different models have been proposed for timbre. Chroma contours were used to describe music timbre in [29]. The contours were found by mapping the spectral content to a single octave (which was divided into 12 bins). Chroma-based features were also used in [28] where a variant of Mel-Frequency Cepstral Coefficients (MFCCs) called Pitch-Frequency Cepstral Coefficients (PFCCs) were estimated on a pitch scale (instead of the Mel scale) and mapped onto the 12 chroma bins. Leveau *et al.* [30] modeled sounds of an instrument as a linear combination of a set of harmonic atoms, where the atoms were defined as the sum of harmonic partials of that particular instrument. Timbre was encoded by the harmonic atoms. The timbre model proposed in [26] was based on the evolution of spectral envelope peaks derived from sinusoidal modeling, followed by principal component analysis (PCA). Patil *et al.* used the spectro-temporal receptive fields (STRFs) of neurons to model timbre [31]. The STRFs, which represent neurons' selectivity to spectral and temporal modulations, were modeled as 2D wavelets that were Gabor-shaped in frequency and exponential in time. Timbre was modeled as follows: a spectrogram was obtained by filtering a signal with a bank of constant-Q asymmetric bandpass filters. The spectrogram was then convolved with the 2D STRFs,

This research was supported by the grant FRQ-NT team program nb 172272 from the government of Québec.

M. Adeli, J. Rouat, S. Wood, S. Molotchnikoff, and E. Plourde are affiliated with the Neurocomputational and Intelligent Signal Processing Research Group (NECOTIS), Department of Electrical and Computer Engineering, Université de Sherbrooke, Sherbrooke, QC, J1K 2R1, Canada (e-mail: mohammad.adeli@usherbrooke.ca).

S. Molotchnikoff and J. Rouat are also affiliated with the Department of Biological Sciences, Université de Montréal, Montréal, QC, H3C 3J7, Canada.

resulting in a 4D representation, which was later collapsed over frequency and integrated over time. As a result, a 2D spectrotemporal modulation profile was obtained for timbre.

C. The Contributions of This Paper

The tendency to reduce the complexity of timbre representations and achieve high performance in specific tasks has favored models that only capture the distinctive timbral features of the sounds used in those specific tasks. However, a good model of timbre, regardless of its applications, should be able to capture not only the distinctive features such as spectral content, attack, and decay, but also other features such as amplitude modulations which are known to contribute to timbre richness.

In this paper, we propose a flexible and multipurpose bio-inspired hierarchical model for analyzing musical timbre. Inspired by findings in the fields of neuroscience, computational neuroscience, and psychoacoustics, not only does the model extract spectral and temporal characteristics of a signal, but it also analyzes amplitude modulations on different timescales to extract roughness and global temporal envelope. It uses a cochlear filter bank to resolve the spectral components of a sound, lateral inhibition to enhance spectral resolution, and a modulation filter bank to extract the global temporal envelope and roughness of the sound from amplitude modulations. As stated in section I-A, a timbre space is usually constructed from individual features such as spectral centroid and attack time, whereas in this paper, timbre is represented by three profiles (curves):

- a time-averaged spectrum, which contains important information about harmonics, pitch, and resonances (or formants).
- a global temporal envelope, which is estimated from slow amplitude modulations and encodes the global time evolution of the spectral components of a sound.
- an instantaneous roughness function, which is estimated from fast amplitude modulations and encodes the local fluctuations of the spectral components of a sound.

The model proposed in this paper provides a novel framework which extracts these profiles hierarchically, as opposed to other studies, where independent frameworks have been designed to extract spectral and temporal features. In addition, efficient integration of the existing roughness models [32–34] into timbre models seems very hard as this requires significant architectural modifications in both. Despite that, roughness estimation is an integral part of our timbre model. Moreover, unlike other roughness models, where only a single measure is estimated, we extract an instantaneous function for roughness as it varies with time.

In addition to the specific timbre representation described above, most of the spectral, temporal, and spectrotemporal features reviewed in section I-A can still be extracted from either the three profiles or other outputs at the different stages of the hierarchical model. Therefore, our model lends itself to a broad range of applications, which is one of its most important capabilities.

We evaluated the proposed model in three different applications to demonstrate its potential as a general framework for timbre representation: 1) evaluating roughness extraction and comparing results with subjective data from a previous psychoacoustic study [35] 2) musical instrument classification using k-NN and a Bayesian network and 3) selection of timbral features that best represent the sounds that had been labeled in an audio-visual experiment [36].

The rest of this paper is structured as follows: the model is described in section II and the applications and their results are presented in section III. Section IV includes the discussion of the model and the results while conclusions are presented in section V.

II. System Description

The timbre model comprises three main modules: a bank of cochlear filters, a lateral inhibition module and a roughness extraction module. These modules, as well as the estimation of the three profiles of timbre, i.e. the global temporal envelope, the time-averaged spectrum, and instantaneous roughness are described in detail in this section.

A. Cochlear Filter Bank

The cochlear filter bank is based on the ERBlet transform, which is provided with the large time/frequency analysis toolbox (LTFAT) [37, 38]. The ERBlet transform covers frequencies even as low as DC [37]. This is important as 1) the proposed model may be used in applications where very low-pitched notes such as E_0 (20.60 Hz) and F_0 (21.83 Hz) should be processed and 2) the envelopes of the cochlear filter bank outputs, which include very low frequency content, are analyzed by a modulation filter bank that is also based on the ERBlet transform (section II-C2).

In the ERBlet transform, filters are constructed from Gaussian windows in the frequency domain on the ERB (equivalent rectangular bandwidth [39]) scale, where the ERB of a given filter centered at f_c (in Hz) is given by:

$$ERB(f_c) = 24.7 + \frac{f_c}{9.265}\cdot \tag{1}$$

Furthermore, the number of ERBs below a given frequency f (in Hz) is given by:

$$ERB_{num}(f) = 9.265\ln(1 + \frac{f}{228.8455})\cdot \tag{2}$$

The spectrum of the input signal, which is obtained by the Fourier transform, is multiplied by the frequency response of each filter and transformed back to the time domain using the inverse Fourier transform. Thus, the ERBlet transform performs a type of multi-resolution analysis on a signal as the ERBlets have different bandwidths.

Our cochlear filter bank consists of 84 filters from ERB_{num} 0.9981 to 42.4184 (2 filters per ERB), covering frequencies from 20 Hz to 22050 Hz. We designed these filters for a sampling frequency of 44100 Hz. Thanks to the ERBlet transform, the number of filters can easily be adapted for different applications to enhance the spectral resolution. In practice, one also faces a trade-off between the number of

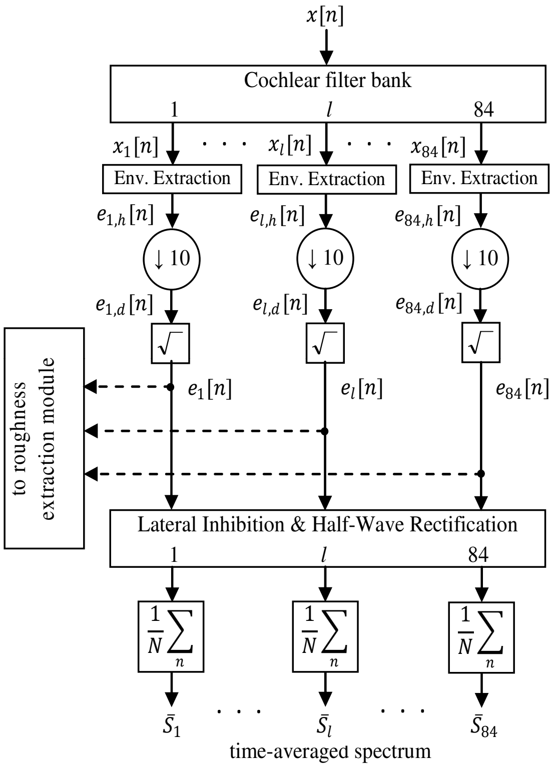

Fig. 1. Time-averaged spectrum computation: the cochlear filter bank decomposes the input into 84 components. The envelopes of these components are extracted, downsampled by a factor of 10, and then compressed. It is not necessary to use lowpass filters before downsampling for the signals used in section III. Lateral inhibition results in sharper peaks in the spectrum. The time-averaged spectrum extracted from the outputs of the lateral inhibition block is a characteristic of timbre. The downsampled and compressed envelopes $e_l[n]$ are used in Fig. 3 to estimate roughness and the global temporal envelope.

filters and the complexity of the system. In this research, 84 filters provided satisfactory results for the applications presented in section III. We also modified the design of the LTFAT's ERBlets to remove aliasing for the bandpass filters that were centered at very low or high frequencies. Moreover, the gains of the filters in the initial ERBlets were originally adjusted to allow for perfect reconstruction [37]. Since signal reconstruction was not necessary in the present study, the gains of the filters were equalized. This simple design choice does not affect the performance of the ERBlets, given that they are linear filters.

As Fig. 1 shows, the cochlear filter bank decomposes the digital signal $x[n]$ into its components $\{x_l[n]; l = 1, 2, ..., 84\}$, where l is the index of the cochlear filterbank channels and n denotes the discrete time. The envelope $e_{l,h}[n]$ of a given signal $x_l[n]$ is computed by the "Env. Extraction" block in Fig. 1 as the magnitude of the analytic signal $x_l^a[n] = x_l[n] + jx_l^h[n]$, where $x_l^h[n]$ is the Hilbert transform of $x_l[n]$. The envelopes of the filter bank outputs are of much lower frequencies than the input signal itself and are therefore downsampled by a factor of 10 (Fig. 1). It is not necessary to use lowpass filters before downsampling for the signals used in section III. The downsampled envelope $e_{l,d}[n]$ is computed from $e_{l,h}[n]$ by the following equation:

$$e_{l,d}[n] = e_{l,h}[10n];\;\; n = 0, 1, ..., \text{floor}(\frac{N_x - 1}{10}) \qquad (3)$$

where N_x is the length of the signal $x[n]$ and the function "floor" returns the integer part of a number. The envelope $e_{l,d}[n]$ is then compressed with the square root function. This is inspired by the compression applied by the outer hair cells in the auditory system [40]. The downsampled and compressed envelopes $e_l[n]$ are fed to the lateral inhibition and roughness extraction modules for further processing (Fig. 1).

B. Lateral Inhibition

Lateral inhibition, which exists in all sensory systems, is the capacity of an excited cell to reduce the activity of nearby cells to boost the contrast of their activities and sharpen their bandwidths [41, 42]. Lateral inhibition in the auditory system leads to sharper peaks in the spectrum and enhanced spectral resolution. In our model, it is implemented as follows:

$$e_l^{LI}[n] = e_l[n] - c_{l,l-1}e_{l-1}[n] - c_{l,l+1}e_{l+1}[n] \qquad (4)$$

where $e_{l-1}[n]$, $e_l[n]$, and $e_{l+1}[n]$ are the downsampled and compressed envelopes of filters $l-1$, l, and $l+1$, respectively, $e_l^{LI}[n]$ is the lateral inhibited envelope of channel l, and coefficient c for filters l and j is computed by:

$$\begin{aligned} c'_{l,j} &= <|F_l|, |F_j|>; j = l-1, l+1 && (5)\\ c_{l,j} &= \frac{c'_{l,j}}{c'_{l,l-1} + c'_{l,l+1}} && (6) \end{aligned}$$

where F_l and F_j are the frequency responses of filters l and j of the cochlear filter bank, and $< |F_l|, |F_j| >$ is the inner product of $|F_l|$ and $|F_j|$. In this implementation, only correlations with the two neighboring channels were used. As such, it is a local lateral inhibition model, which requires less computations and is common in the literature [41]. $c_{1,0}$ and $c_{84,85}$ were set to zero when lateral inhibition was applied to channels 1 and 84, respectively. Negative values of $e_l^{LI}[n]$ are removed by half-wave rectification:

$$e_l^{HW}[n] = \frac{e_l^{LI}[n] + |e_l^{LI}[n]|}{2} \qquad (7)$$

where $e_l^{HW}[n]$ is the half-wave rectified envelope of channel l.

As shown in Fig. 1, the time-averaged spectrum $\bar{S}_l$, which is one of the three main profiles extracted for timbre in this paper, is computed as follows:

$$\bar{S}_l = \frac{1}{N}\sum_{n=1}^{N} e_l^{HW}[n] \qquad (8)$$

where N is the length of the half-wave rectified envelope $e_l^{HW}[n]$. Fig. 2 presents examples of time-averaged spectrum with and without lateral inhibition for the note F_2 (87.31 Hz) of piano. As shown in this figure, lateral inhibition has enhanced the spectral resolution by separating the unresolved harmonics of this signal.

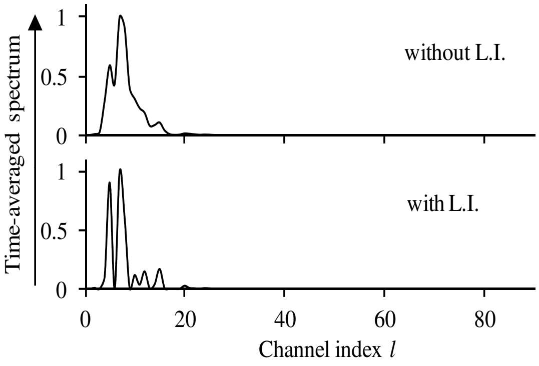

Fig. 2. Effect of lateral inhibition (L.I.) on the time-averaged spectrum: the time-averaged spectrum obtained for note F_2 (87.31 Hz) of piano without lateral inhibition (top) and with lateral inhibition (bottom). The unresolved harmonics in the top panel were resolved after lateral inhibition. Both graphs were normalized by their maximum values for display purposes.

C. Roughness Extraction Module

Roughness is a sensation caused by fast (30 Hz to 200 Hz) amplitude modulations of a sound [43, 44]. To estimate the roughness of a sound, we first estimate the roughness for all the channels of the cochlear filter bank separately and then add them together. In this subsection, we first present the factors influencing roughness perception, and then explain how roughness is computed for channel l and the whole signal from these factors.

1) Roughness determinants: The factors that influence the roughness perception are referred to as roughness determinants in this paper. Some of the known roughness determinants are introduced in this section. The perceived roughness of the pure AM tone

$$x[n] = \big(d + a\sin(2\pi f_m nT_s)\big)\sin(2\pi f_c nT_s)$$

depends on the carrier frequency f_c, the modulation frequency f_m, and the modulation depth $m = a/d$ [43, 44]. In general, when $f_c = 1000$ Hz and $f_m = 70$ Hz, any increase in m from 0 to 1.2 leads to the increase of roughness but after 1.2 it decreases. When $m = 1$ and $f_m = 70$ Hz, roughness is maximum at $f_c = 1000$ Hz and decreases for carrier frequencies lower or higher than 1000 Hz. When $m = 1$ and $f_c = 1000$ Hz, roughness is maximum at $f_m = 70$ Hz and decreases for modulation frequencies lower or higher than 70 Hz.

The above factors are not the only factors involved in roughness perception. In subjective experiments conducted in [35], it was observed that roughness perception depends on the shape of the amplitude fluctuations of a signal. Moreover, in [32], Duisters reported that although signals that have noise-like characteristics create large modulation depths, they are not perceived as rough. He thus proposed to decorrelate the filter bank outputs to overcome this problem. As explained below, the roughness module proposed in this paper accounts for all the aforementioned effects.

Though roughness typically depends on AMs from 30 Hz to 200 Hz, in this study, amplitude modulations from 30 Hz to 930 Hz are used to compute roughness because 1) AMs with frequencies up to 1000 Hz are still perceptible [45] and 2) for speech and sounds of some musical instruments, AM modulations are caused by pitch which is an important perceptual feature independent of timbre. Therefore, we decided to use a broader range of AM frequencies to be able to use high frequency AMs in future studies. In addition, the energy of the amplitude modulations from 10 Hz to 30 Hz is computed and used as a distinct feature. Also, amplitude modulations with frequencies less than 10 Hz are used to estimate the global temporal envelope of a sound. These specific choices of AM bands for the computation of the global temporal envelope, the energy of AMs from 10 Hz to 30 Hz, and roughness is due to the fact that amplitude modulations with $f_m < 10$ Hz are perceived as distinct events, AMs with frequencies between 10 Hz and 30 Hz are heard as acoustical flutter, and AMs with frequencies above 30 Hz fuse together and generate roughness [46]. The relevant details are presented below.

2) Roughness estimation in channel l: The process of roughness estimation for channel l of the cochlear filter bank is shown in Fig. 3. The downsampled and compressed envelope $e_l[n]$ is processed by a modulation filter bank that is also based on the ERBlet transform. The modulation filter bank comprises 15 filters which cover frequencies up to 930 Hz (1 filter per ERB from ERB_{num} 0 to 14). This is in agreement with other studies. For instance, in [32], 9 modulation filters were used. The low-pass filter (AM LPF in Fig. 3) gives the lowest frequency component of the input (less than 10 Hz), the second filter (AM BPF 1 in Fig. 3) covers frequencies from 10 to 30 Hz, and the rest of the filters (AM BPF 2 to AM BPF 14 in Fig. 3) give the higher frequency amplitude modulations (30 Hz to 930 Hz). The output of the second filter (AM BPF 1) is used to compute the energy of AMs in band [10 Hz, 30 Hz] and the outputs of the first filter (AM LPF) as well as those from 3 to 15 (AM BPF 2 to AM BPF 14 in Fig. 3) are used to compute the roughness.

From psychoacoustic experiments, roughness is known to be proportional to m^n, where m is the modulation depth and n varies from 1.4 to 2 depending on the experimental setup [43, 44]. The 2$^{\text{nd}}$ power of m [32, 33] is usually selected in roughness models and m is then multiplied by the roughness determinants mentioned in section II-C1. Therefore, the above proportionality for the instantaneous roughness $r_l[n]$ of channel l can be expressed as follows:

$$r_l[n] \propto \left(\frac{\gamma_{l-1,l} + \gamma_{l,l+1}}{2} \cdot H \cdot G_l \cdot Q_l \cdot W_l \cdot m_l[n]\right)^2 \quad (9)$$

where $m_l[n]$ is the instantaneous modulation depth in channel l , H and G_l are the weights of modulation and carrier frequencies respectively, Q_l is a factor for loudness normalization, W_l is the effect of waveform envelope shape, and $\gamma_{l-1,l}$ and $\gamma_{l,l+1}$ are the correlations of the output $f_l[n]$ (Fig. 3) of channel l with the respective outputs of channels $l-1$ and $l+1$, and the factor $(\gamma_{l-1,l} + \gamma_{l,l+1})/2$ [32, 33], which is explained in subsection II-C5, is used to decorrelate channel l with its neighbors.

There are two differences between our model and the previous ones: 1) we compute an instantaneous function rather than a single measure for roughness as it varies with time and

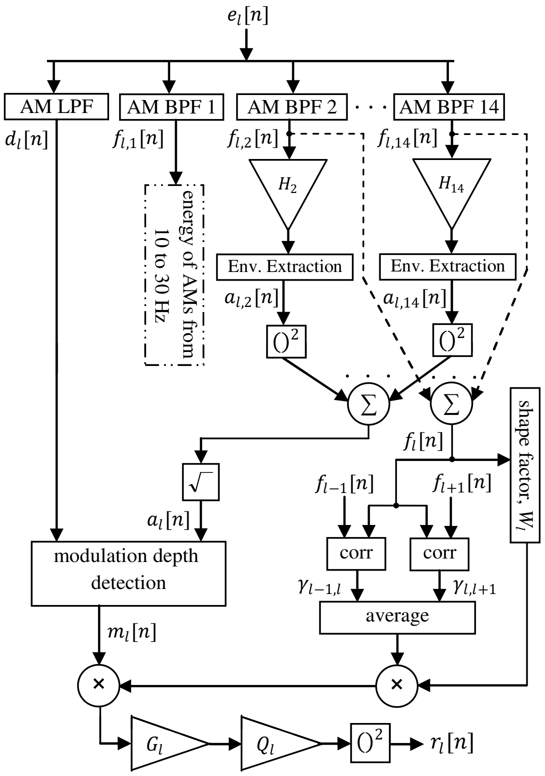

Fig. 3. Roughness estimation in channel l of the cochlear filter bank: the envelope $e_l[n]$ from Fig. 1 is analyzed by a modulation filter bank consisting of the filters AM LPF, AM BPF 1, AM BPF 2, ..., and AM BPF 14. The output of AM BPF 1 is used to compute the energy of AMs from 10 Hz to 30 Hz. Other outputs are involved in the estimation of modulation depth. The modulation depth is multiplied by the effects of modulation frequency (H_k), carrier frequency (G_l), envelope shape (W_l), loudness (Q_l), and the average of the correlations ($\gamma_{l-1,1}$ and $\gamma_{l,1+1}$) with neighboring channels. The components $d_l[n]$ are also used to estimate the global temporal envelope.

2) roughness estimation is integrated into a timbre model for the first time in this paper.

In the following subsections, the factors $m_l[n]$, W_l, Q_l, $\gamma_{l-1,l}$, and $\gamma_{l,l+1}$ are computed and the functions H and G are introduced.

3) Modulation depth estimation: For the sake of clarity of presentation, first we consider the case where the downsampled and compressed envelope of the channel l is $e_l[n] = d_l + a_l \sin(2\pi f_m n T_s^{'})$, with $T_s^{'}$ being the sampling period after decimation by 10. To estimate the modulation depth, parameters a_l and d_l first need to be estimated. The output of the first filter AM LPF in Fig. 3 gives d_l (the low frequency component). The AC component $a_l \mathrm{sin}(2\pi f_m n T_s^{'})$ of $e_l[n]$ is given by one of the other filters depending on the value of f_m. The envelope of the AC component a_l is computed by the "Env. Extraction" block in Fig. 3. Since only one of the filters from AM BPF 2 to AM BPF 14 is active, a_l is given by $a_l[n]$ in Fig. 3. The modulation depth m_l is simply the ratio of a_l to d_l.

In practice, the parameters a_l and d_l may change with time. Thus we now consider the following general form for $e_l[n]$:

$$e_l[n] = d_l[n] + F_l[n] \tag{10}$$

where $d_l[n]$ is assumed to have low frequency components (< 10 Hz) whereas $F_l[n]$ is assumed to have higher frequency components. $d_l[n]$ is still obtained by the low-pass channel of the modulation filter bank, but the AC component $F_l[n]$ is decomposed into its narrow-band components $f_{l,1}[n]$, $f_{l,2}[n]$, ... and $f_{l,14}[n]$, where $f_{l,k}[n] \approx a_{l,k}[n] \sin(2\pi f_k n T_s^{'})$, and k indicates the channels of the modulation filter bank. The amplitude $a_{l,k}[n]$ is computed by the "Env. Extraction" block in Fig. 3. An estimate for the envelope $a_l[n]$ of the AC component $F_l[n]$ can now be computed by:

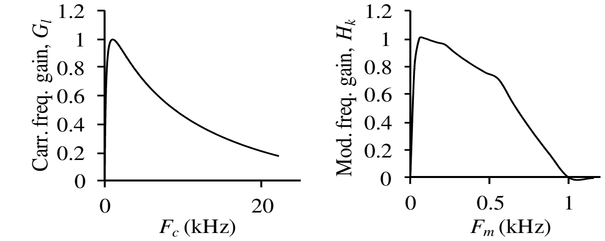

Fig. 4. Effects of carrier and modulation frequencies on roughness: G_l is the weight of the center (carrier) frequency F_c of channel l in the cochlear filter bank and H_k is the weight of the modulation frequency F_m (the center frequency of channel k in the modulation filter bank). G_l and H_k have their maxima at $F_c = 1000$ Hz and $F_m = 70$ Hz, respectively, as discussed in section II-C1. Similar functions have been used in other studies (see [32] for a summary).

$$a_l[n] \approx \sqrt{\sum_{k=2}^{14} (a_{l,k}[n])^2}. \tag{11}$$

The modulation depth, which is now a function of time, is found by $m_l[n] = a_l[n]/d_l[n]$. When $d_l[n]$ is zero or very small (when there is no or little activity in channel l), $m_l[n]$ can become very large which is not accurate. To circumvent this problem, the numerator and denominator of the fraction above are first divided by $d_{l,max}$, the maximum of $d_l[n]$, and then $m_l[n]$ is written as:

$$m_l[n] = \frac{a_l[n]/d_{l,max}}{d_l[n]/d_{l,max} + h(d_l[n]/d_{l,max})} \tag{12}$$

where the function $h(y)$ is chosen such that when y is large enough, $h(y) \approx 0$ and when y is zero or very small, $h(y)$ is large. In this research, the following function was chosen empirically:

$$h(y) = e^{-40y}. \tag{13}$$

4) Effects of modulation and carrier frequencies on roughness in channel l: Channel k of the modulation filter bank covers a specific range of amplitude modulation frequencies. H_k (Fig. 4) represents the weights of these modulation frequencies on roughness computations in channel l, as discussed in section II-C1. H_k is maximum at 70 Hz. Such weighting functions have been used in other studies and are reviewed in [32]. In (9), $m_l[n]$ needs to be computed separately and then multiplied by the other factors. However, in Fig. 3, the output $f_{l,k}[n]$ of the modulation filter bank is multiplied by the weighting function H_k before computing $m_l[n]$. In other words, it is no longer necessary to include H as a separate factor in (9) as the effect of the modulation frequency is implicitly applied before computing $m_l[n]$. The weighting function G_l, i.e. the effect of center (carrier) frequency of channel l on roughness is also depicted in Fig. 4. This effect is maximum at 1000 Hz as mentioned in section II-C1. The functions G_l and H_k simulate those used in other studies for the same purposes (see [32] for a summary).

5) Decorrelating channel l with its neighbors: As mentioned in subsection II-C1, signals with noise-like characteristics cause large modulation depths, but they are not perceived as rough [32, 33]. The neighboring channels of the cochlear filter bank have low correlations in such cases. Thus, correlation factors can be used to correct the high modulation depths. Correlation factors are defined as follows:

$$\gamma_{l-1,l} = \text{corr}(f_{l-1}[n], f_l[n]) \quad (14)$$

$$\gamma_{l,l+1} = \text{corr}(f_l[n], f_{l+1}[n]) \quad (15)$$

where the operator corr returns the Pearson correlation coefficient and $f_l[n] = \sum_{k=2}^{14} f_{l,k}[n]$ (Fig. 3). As shown in (9), the average of $\gamma_{l-1,l}$ and $\gamma_{l,l+1}$ is used as a roughness factor.

6) The effect of envelope shape on roughness in channel l: In [35], it was shown that roughness perception was affected by the shape of amplitude fluctuations of a signal. Sounds with reversed sawtooth envelopes were perceived as having greater roughness than those with sawtooth envelopes (Fig. 8). To account for this effect, a shape factor is introduced in the model. The shape factor is computed from the AC component of the envelope, $f_l[n]$. For that purpose, the discrete-time derivative $\dot{f}_l[n] = f_l[n+1] - f_l[n]$ is first computed. The shape factor is then defined by:

$$W_l = \mathrm{e}^{-\frac{1}{2}(\frac{T}{N}-0.5)} \quad (16)$$

where N is the length of $f_l[n]$ and T is the number of samples for which $\dot{f}_l[n] > 0$. When the overall rise time of envelope fluctuations is less than the overall decay time (reversed sawtooth), then $\frac{T}{N} > 0.5$ and the shape factor W_l has a value greater than 1. Otherwise, it is less than 1 (sawtooth).

7) Loudness normalization: To normalize the effect of loudness on roughness, the modulation depth is multiplied by factor Q_l, which was computed using the following equations:

$$\begin{aligned} e_{rms,l} &= \sqrt{\frac{1}{N}\sum_{n=1}^{N} e_l^2[n]} \\ Q_l &= \frac{e_{rms,l}}{\max[e_{rms,l}]} \end{aligned} \quad (17)$$

where $e_l[n]$ is the downsampled and compressed envelope of channel l (Fig. 1) and $e_{rms,l}$ is a measure for the loudness of channel l.

8) Instantaneous and effective roughness estimation: Instantaneous roughness can now be estimated based on (9):

$$r_l[n] = \eta\left(\frac{\gamma_{l-1,l}+\gamma_{l,l+1}}{2}\cdot G_l \cdot Q_l \cdot W_l \cdot m_l[n]\right)^2 \quad (18)$$

$$R[n] = \sum_{l=1}^{84} r_l[n] \quad (19)$$

$$R_{eff} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} R^2[n]} \quad (20)$$

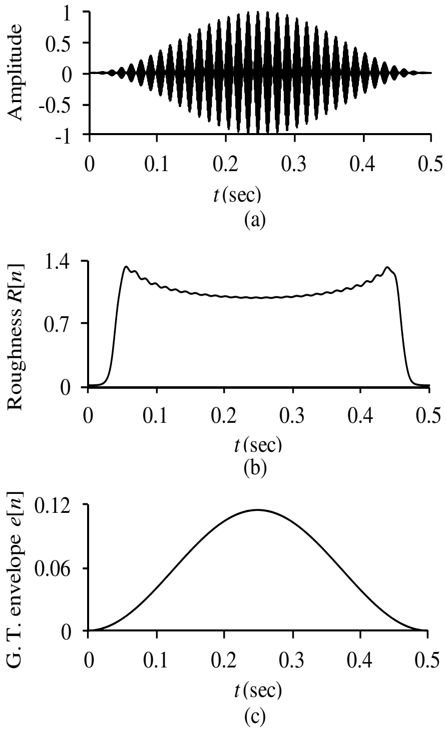

Fig. 5. Global temporal envelope and instantaneous roughness estimation: (a) Original signal. (b) The instantaneous roughness which encodes the local fluctuations of the signal envelope. (c) The estimated global envelope of the signal, which encodes the global evolution of the signal envelope. It has a lower scale than the exact envelope due to the compressions applied in Fig. 1 and removal of the AM modulations in Fig. 3.

where $r_l[n]$ is the instantaneous roughness of channel l, and $R[n]$ and R_{eff} are the instantaneous and effective roughness of the whole signal, respectively. $m_l[n]$ involves the effect of modulation frequency H_k, as stated in section II-C4. The constant η in (18) is computed as follows. Asper, the unit of roughness, is defined as the roughness of a pure AM tone with $f_c = 1000$ Hz, $f_m = 70$ Hz, $m = 1$ and a loudness of 60 dB SPL [43]. Therefore, we first computed the effective roughness R_{ref} of the above AM tone without considering the constant η in (18). The value $\eta = 1/R_{ref}$ guarantees that the effective roughness of the above AM tone becomes 1 Asper. η was found to be 1.38 for the proposed model. This value of η is used to compute the roughness of input sounds. An example of instantaneous roughness is shown in Fig. 5(b).

D. Global Temporal Envelope

It is necessary to estimate the global temporal envelope of a signal because an important set of timbral features depends on it. In this paper, the global temporal envelope $e[n]$ of the signal is defined as:

$$e[n] = \sqrt{\sum_{l=1}^{84} d_l^2[n]} \quad (21)$$

where the low frequency components $d_l[n]$ is estimated by the modulation filter bank in Fig. 3. In the estimation of $e[n]$, the temporal fine structure is removed by the cochlear filter bank, the envelopes of the filter outputs are compressed, and the AM modulations are removed by the modulation filter bank. Consequently, it has a lower magnitude than the exact envelope of the signal as shown in Fig. 5(c). Apart from being scaled down, $e[n]$ accurately estimates the shape of the global temporal envelope.

In summary, three major profiles are extracted for timbre using the proposed model: the Time-Averaged Spectrum found by (8), the Instantaneous Roughness found by (18), and the Global Temporal Envelope found by (21). In addition, most of the timbral features presented in section I-A can be computed from these profiles. In the next section, the proposed model is evaluated in three applications.

III. Applications

Since one of the goals of this study was to construct a flexible and multipurpose framework for timbre analysis, the proposed model was tested in three different applications: 1) comparison with subjective values of roughness, 2) musical instrument classification, and 3) feature selection for labeled timbres.

A. Comparison with Subjective Values of Roughness

In this section, the effective roughness evaluated by the proposed model is compared to subjective values of roughness. The subjective values that are used for that purpose were obtained from [35], in which Pressnitzer and McAdams conducted two subjective experiments to study the effects of carrier phase and waveform envelope shape on roughness perception. The same sounds were applied to our model and it was expected that the model would simulate the subjective roughness data from the above experiments. Both cases are presented below.

1) Effect of carrier phase on roughness perception: The synthetic sounds used in the first experiment of [35] were pseudo-AM (pAM) signals of the following form:

$$\begin{aligned} pAM[n] = {} & \frac{1}{2}\cos\big(2\pi(f_c - f_m)nT_s\big) \\ & + \cos(2\pi f_c n T_s + \phi) \\ & + \frac{1}{2}\cos\big(2\pi(f_c + f_m)nT_s\big) \end{aligned} \tag{22}$$

where the sampling period T_s was 1/44100 seconds. Seven sets of sounds were used in the experiment, each set consisting of 7 pAM tones which had the same carrier and modulation frequencies, but with 7 distinct values for ϕ: $-\pi/2$, $-\pi/3$, $-\pi/6$, 0, $\pi/6$, $\pi/3$, $\pi/2$. The pairs of carrier and modulation frequencies for these 7 sets were: (f_c,f_m) = (125, 30), (250, 40), (500, 50), (1000, 70), (2000, 70), (4000, 70) and (8000, 70), where all frequencies are in Hz. Some of these signals are shown in Fig. 6. The amplitude fluctuation of these sounds decreases with the absolute value of phase ϕ. Each sound was one second in duration and had raised-cosine onset and offset ramps of 50 ms. All sounds were presented to the subjects at 60 dB SPL.

Two groups of 15 subjects [35] took part in the experiment. They were first familiarized with the notion of roughness. To that end, tones with known carrier and modulation frequencies were presented to the participants. Participants could vary the modulation depth and observe roughness variations. After the training phase and a few practice trials, all sound pairs from a given stimulus set were presented to the subjects, and they were asked to compare the relative roughness of these sounds. Binary paired-comparison judgments were analyzed to find a psychophysical scale for roughness. The results showed that the absolute value of phase $|\phi|$ had a strong influence on roughness perception. In general, perceived roughness increased with the decrease of $|\phi|$. In other words, phases that caused larger amplitude modulations were perceived as having greater roughness. In addition, the sign of ϕ affected roughness perception, with positive phases causing higher roughness. For instance, a signal with phase $\pi/6$ was rougher than a signal with phase $-\pi/6$, assuming that they had the same carrier and modulation frequencies. This effect was observed for all values of f_c and ϕ except $f_c = 8$ kHz or $\phi = \pm\pi/2$.

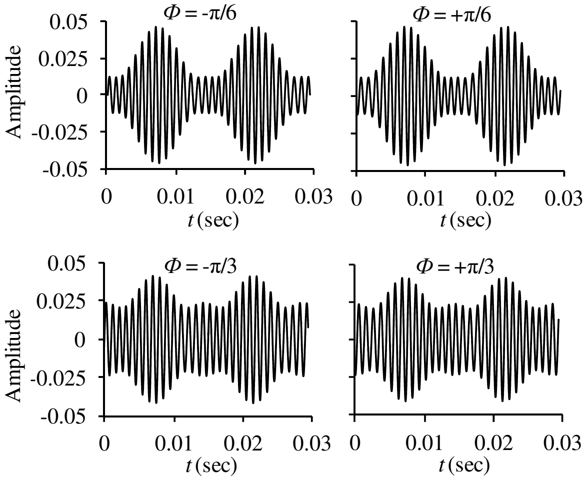

Fig. 6. Examples of the pseudo AM tones that were used to estimate the impact of carrier phase ϕ on roughness. The amplitude of envelope fluctuations decreases with $|\phi|$.

The same sounds were presented to the proposed model to estimate the effective roughness. Results are summarized in Fig. 7. As expected, no effect of phase sign was observed due to the fact that the model performs no carrier phase processing. In Fig. 7, estimated effective roughness curves for negative- and positive-phase signals are superimposed and look like a single curve. To quantitatively measure the degree of similarity between the estimated roughness curves in Fig. 7 and the equivalent subjective curves from [35], we computed the Pearson correlation coefficient for any given pair of (f_c, f_m). The correlation coefficients are presented in Table I. As seen in this table, the subjective roughness values and the values estimated by the model are highly correlated. Therefore, as in [35], roughness decreases with the absolute value of phase ϕ.

It is not possible to compare the perceived roughness of tones with different carrier and modulation frequencies (f_c, f_m), e.g. (250 Hz, 40 Hz) and (1000 Hz, 70 Hz), because the roughness scales obtained in [35] are relative and normalized.

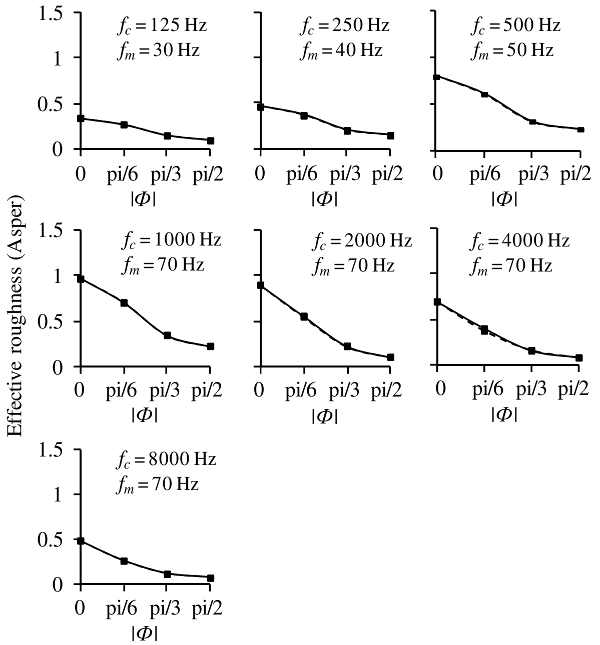

Fig. 7. The estimated effective roughness for the pseudo AM tones: dashed and solid lines show the effective roughness for negative and positive phases, respectively. They are superimposed and look like a single line in most figures because the model did not process phase information. In general, effective roughness decreased with the absolute value of phase ϕ.

TABLE I
CORRELATION COEFFICIENTS BETWEEN THE SUBJECTIVE ROUGHNESS FROM [35] AND THE OBJECTIVE ROUGHNESS FROM THE PROPOSED MODEL. THESE COEFFICIENTS WERE COMPUTED FOR THE pAM SIGNALS WITH CARRIER AND MODULATION FREQUENCIES (f_c, f_m) = 1: (125, 30), 2: (250, 40), 3: (500, 50), 4: (1000, 70), 5: (2000, 70), 6: (4000, 70) AND 7: (8000, 70). SUBJECTIVE AND ESTIMATED OBJECTIVE VALUES ARE HIGHLY CORRELATED.

	Carrier and modulation frequencies pairs						
	1	2	3	4	5	6	7
ϕ<0	0.91	0.96	0.87	0.95	1	0.95	0.92
ϕ>0	0.88	0.95	0.94	0.97	0.89	0.87	0.90

Therefore, the subjective roughness values cannot be compared with the values estimated by the proposed model at different (f_c, f_m). However, for the case where $\phi = 0$, the roughness values estimated by the model are completely in accordance with the facts presented in section II-C1. As seen in Fig. 6, when $\phi = 0$, roughness is maximum for (f_c, f_m) =(1000 Hz, 70 Hz) and gradually decreases for (2000 Hz, 70 Hz), (4000 Hz, 70 Hz), and (8000 Hz, 70 Hz). Similarly, it decreases for (500 Hz, 50 Hz), (250 Hz, 40 Hz), and (125 Hz, 30 Hz).

2) Effect of waveform envelope shape on roughness perception: In the second experiment, Pressnitzer and McAdams [35] investigated the impact of envelope shape on roughness perception. They used two classes of stimuli: sounds with sawtooth amplitude modulations (envelope fluctuations had slow rises but fast decays) and sounds with reversed sawtooth amplitude modulations (envelope fluctuations had fast rises but slow decays). Signals were generated as follows:

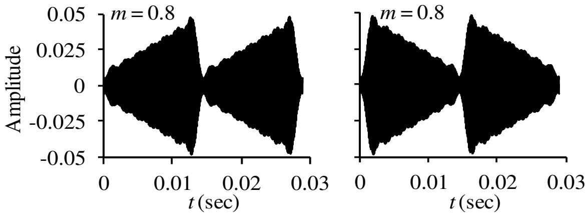

Fig. 8. Examples of sounds with sawtooth (left) and reversed sawtooth envelopes (right).

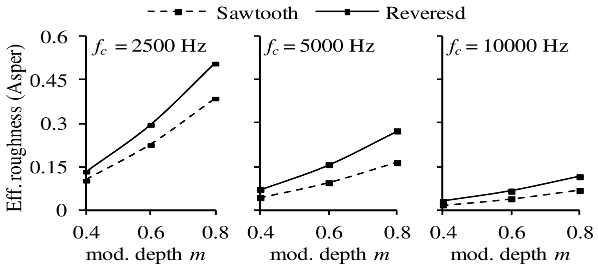

Fig. 9. Effect of waveform envelope shape on the roughness, estimated by the model: sounds with reversed sawtooth envelopes have greater roughness than those with sawtooth envelopes. Effective roughness of both types of sounds increased with the modulation depth. f_m is 70 Hz in all figures.

$$E[n] = \sum_{i=1}^{I} \frac{1}{i}\cos(2\pi i f_m n T_s + \phi) \quad (23)$$

$$x[n] = \left(1 + \frac{mE[n]}{\max\big[E[n]\big]}\right)\cos(2\pi f_c n T_s + \phi) \quad (24)$$

where I was chosen such that $If_m \leq \frac{1}{2}ERB(f_c)$. The function ERB is given by (1). The length N_x, the sampling period T_s, and the loudness of the sounds are the same as those used in experiment 1. For sawtooth and reversed-sawtooth signals, the phase ϕ was $-\pi/2$ and $\pi/2$, respectively. Two examples of these signals are plotted in Fig. 8. Modulation frequency f_m was 70 Hz for all the sounds and three different carrier frequencies f_c were used: 2500 Hz, 5000 Hz, and 10000 Hz. For each f_c, three modulation depths were considered: m = 0.4, 0.6, and 0.8. Ten subjects participated in the experiment and for each f_c they listened to all signal pairs and judged their relative roughness. Data analysis was similar to experiment 1. Results showed that roughness increased with modulation depth and sounds with reversed-sawtooth envelopes were rougher than those with sawtooth envelopes.

The same sounds were presented to our model to investigate the effect of envelope shape on roughness estimation. Results are presented in Fig. 9 and are consistent with the subjective results of the second experiment in [35]. Effective roughness increased with modulation depth at all carrier frequencies, with reversed sawtooth sounds having greater roughness. As in the previous subsection, we computed the Pearson correlation to measure the degree of similarity between the subjective roughness values and the roughness values estimated by the model for the sounds with sawtooth and reversed sawtooth

TABLE II

CORRELATION COEFFICIENTS BETWEEN THE SUBJECTIVE ROUGHNESS FROM [35] AND THE OBJECTIVE ROUGHNESS FROM THE PROPOSED MODEL FOR THE SAWTOOTH AND REVERSED SAWTOOTH TONES USED IN SECTION III-A2. THE MODULATION FREQUENCY OF ALL TONES IS $f_m = 70$ HZ. THE SUBJECTIVE AND OBJECTIVE VALUES ARE HIGHLY CORRELATED.

	Carrier frequency		
	2500 Hz	5000 Hz	10000 Hz
Sawtooth	1	0.99	1
Reversed sawtooth	0.91	0.94	0.92

envelopes. The correlation coefficients are presented in Table II. As seen in this table, the roughness values estimated by the model are highly correlated with the subjective roughness values in [35].

In summary, our model was able to simulate the subjective effects of the absolute value of the carrier phase and the envelope shape on roughness. The next subsection introduces its use for musical instrument classification.

B. Musical Instrument Classification

In this section, acoustic features and their use for classification are detailed and results are presented. Burred *et al.* [26] proposed an interesting model for timbre analysis where timbre was considered as the spectrotemporal envelopes estimated by sinusoidal modeling. Our bio-inspired hierarchical model extracts similar spectrotemporal characteristics and encodes them into time-averaged spectrum and global temporal envelope. We therefore would like to compare our model with the one proposed in [26].

1) Feature extraction and the sounds used in [26]: The timbral features that were used in [26] consisted of the ensemble of partials (spectral peaks) that were extracted using sinusoidal modeling and peak finding in overlapping windows. Principal component analysis (PCA) was applied to the extracted features for the sake of dimensionality reduction, where the 20 principal components that had the largest variances were selected. In the obtained twenty-dimensional space, each instrument was represented by a prototype conveying the evolution of its timbre. The prototypes were composed of $20R_{max}$ features, where R_{max}, the length of the longest partial trajectory, was not reported. A very conservative estimate that we computed for R_{max} is 50, taking into account the windowing strategy that was used in [26]. Therefore, the number of features might be at least 1000. Classification was performed by comparing the signal with the 5 prototypes in the obtained PCA space.

All the signals that were used in [26] were selected from the RWC musical instrument sound database [47]. They consisted of all the notes from C_4 to B_5 (covering two consecutive octaves) from 5 instruments: oboe, clarinet, trumpet, piano, and violin. Recordings from 2 to 3 examples of each instrument with 3 dynamics (*piano*, *mezzo-forte*, *forte*) and normal playing style were used which amounted to 1098 individual notes. The sampling frequency was 44100 Hz.

2) Feature extraction using the proposed model: Our model was tested in a musical classification task using the above signals. The features used were the time-averaged spectrum and the global temporal envelope presented in sections II-B and II-D. Roughness was not used as we wished to evaluate the system's capacity in encoding only the spectral and temporal characteristics of timbre in this task. The following dimensionality reduction strategies were applied to simplify signal representations:

- every two neighboring channels of the time-averaged spectrum were averaged, reducing the number of spectral features from 84 to 42. Forty-two spectral features were sufficient to encode the spectral characteristics of the signals used in this section. However, in other applications, all 84 spectral features may be used.
- 40 features were extracted from the onset of the global temporal envelope $e[n]$.
- 20 features were extracted from the offset of the global temporal envelope.

To compute the onset and offset features, the global temporal envelope was first divided into 60-ms non-overlapping windows. For any given window, a single temporal feature was then computed by averaging the global temporal envelope over the duration of that window. The first 40 temporal features (referred to as onset features) and the last 20 temporal features (referred to as offset features) of a sound were only used in this application, resulting in a total of 60 temporal features. Given the 60 temporal features and that each window was 60 ms, the minimum duration of the sound is 3.6 seconds (60 windows × 60 ms). However, some of the signals used in this study had shorter durations. To overcome this problem, the global temporal envelope was divided into three equal segments, roughly relating to attack and decay, sustain, and release segments of the envelope. Samples were added to the middle segment (sustain) by interpolating between the points using a cubic spline function. In other words, if the duration of a signal was 3.3 seconds (300 ms less than the minimum duration of 3.6 sec), the sustain segment of the global temporal envelope was augmented by interpolation to increase the length of the signal to 3.6 sec. In [26], the temporal trajectory of a given spectral peak was uniformly interpolated. This had an adverse effect on some of the signal representations, because the attack and decay, which are known to be important features of timbre, were stretched. As stated above, this adverse effect was avoided in this work by interpolating only the sustain segment, leaving the other 2 segments intact. In summary, a given signal is represented by 102 features: 42 spectral, 40 onset, and 20 offset features.

3) Classification methods used in this study: Classification methods that have previously been employed for instrument classification include hidden Markov models (HMM) [48], Gaussian mixture models (GMM) [49], support vector machines (SVM) [50], the k-nearest neighbors (k-NN) algorithm, [50, 51], Bayesian networks [50], and artificial neural networks (ANN) [52].

In this research, the k-NN algorithm and a Bayesian network were used. k-NN finds the k nearest neighbors of a given sound in the training set and assigns it to the class that includes the majority of these k neighbors. To that end, it requires a distance measure between sounds. An Euclidean distance was used where spectral, onset and offset features were weighted differently. The optimal weights and parameter k that gave rise to the minimum classification error were found as follows: the spectral features' weight, w_s, was fixed at 1, the onset and offset features' weights, w_{on} and w_{off}, were varied from 0 to 6 with steps of 0.1, and all odd numbers from 1 to 9 were tested for k. The optimal values $k = 1$, $w_{on} = 1$, and $w_{off} = 1.4$ were obtained by minimizing the average classification error for the signals introduced in section III-B1, after 100 runs of 10-fold random cross-validation. This minimization was performed for the 10 training folds altogether.

The Bayesian network that was used as the second classifier is illustrated in Fig. 10. Each circle represents a single feature which is considered a random variable. The sets $S = \{S_i,\ i = 1, 2, ..., 42\}$, $A = \{A_i,\ i = 1, 2, ..., 40\}$, and $D = \{D_i,\ i = 1, 2, ..., 20\}$ are the spectral, onset, and offset features, respectively. The three sets of features are assumed to be conditionally independent of each other for a given instrument but there are first order Markovian dependencies between features within any given set. For instance, for spectral features, S_{42} depends on S_{41}, S_{41} depends on S_{40} and so on. The variable C represents the class (or the instrument) of a signal. The joint distribution of all the variables can be factored as follows:

$$\begin{aligned} P(S, A, D, C) = P(C) \cdot P(S_1|C) \cdot \prod_{i=2}^{42} P(S_i|S_{i-1}, C) \cdot \\ P(A_1|C) \cdot \prod_{i=2}^{40} P(A_i|A_{i-1}, C) \cdot \\ P(D_1|C) \cdot \prod_{i=2}^{20} P(D_i|D_{i-1}, C). \end{aligned} \tag{25}$$

Individual features (spectral, offset, and onset) were modeled as Gaussian random variables and C was assumed to have a polynomial distribution. These choices were motivated by the empirical histograms of the variables. The unknown parameters of the distributions were estimated using maximum likelihood (ML) optimization. The classification of a test signal with feature sets A, D and S was done by maximizing the posterior probability:

$$P(C|S, A, D) \propto P(S, A, D, C) \tag{26}$$

where $P(S, A, D)$ is constant for all classes and its reciprocal is the coefficient of the above proportionality.

Both classifiers (the k-NN and the Bayesian network) were trained and tested with distinct feature sets (extracted from the signals presented in section III-B1) and their combinations using 10-fold random cross-validation.

4) Classification results: Classification results for the 5 instruments are presented in Table III. These are the average classification results after 100 runs of 10-fold cross-validation. The results presented in this subsection are of the form $\mu \pm \sigma$ where μ is the average classification rate for a given instrument and σ is its standard deviation for 100 runs of 10-fold cross-validation. The average classification accuracy for piano was $100\% \pm 0$ using both the k-NN and the Bayesian network (95.81% in [26]). This was expected because piano has both a fast onset and offset with no sustain, and therefore, both methods have classified the piano notes correctly. The average classification accuracy for clarinet was $96.30\% \pm 0.81$ with the k-NN and $92.92\% \pm 0.51$ with the Bayesian network (92.52% in [26]). On average, oboe was correctly classified $88.20\% \pm 1.13$ and $91.56\% \pm 0.54$ of the time using the k-NN and the Bayesian network, respectively (95.10% in [26]). The average classification accuracies of the k-NN and the Bayesian network for violin were $96.76\% \pm 0.71$ and $98.88\% \pm 0.4$, respectively (95.45% in [26]). The average classification accuracies of the k-NN and the Bayesian network for trumpet were $88.88\% \pm 1.03$ and $96.33\% \pm 0.51$, respectively (96.53% in [26]). Oboe and trumpet, which had similar timbres, were confused by the k-NN. The Bayesian network misclassified clarinet as oboe and violin 4.66% and 2.32% of the time, respectively. It also misclassified oboe as violin 6.69% of the time. The overall average classification accuracy was $94.03\% \pm 4.69$ for the k-NN and $95.94\% \pm 3.65$ for the Bayesian network (95.08% in [26]). The k-NN and the Bayesian network used only 102 features whereas at least 1000 features were used in [26].

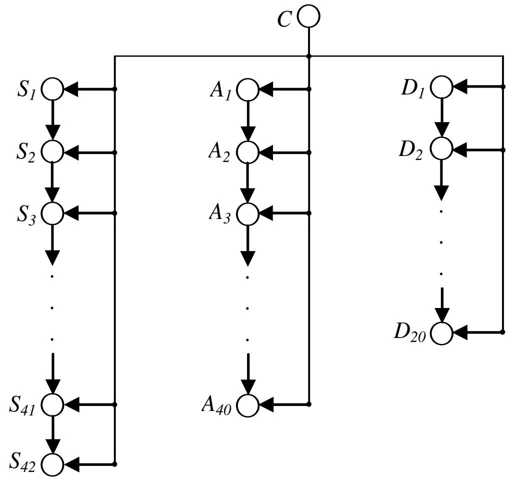

Fig. 10. The Bayesian network used as the second classifier: each circle represents a feature which was modeled as a random variable. Variable C represents the class or instrument. Sets $S = \{S_i,\ i = 1, 2, ..., 42\}$, $A = \{A_i,\ i = 1, 2, ..., 40\}$ and $D = \{D_i,\ i = 1, 2, ..., 20\}$ include the spectral, onset and offset features, respectively.

To test the importance of different feature sets and their impact on the overall classification accuracy, both classifiers were trained on distinct feature sets (spectral, onset and offset) and their various combinations. The results are presented in Table IV. The overall average classification accuracy of both classifiers for spectral and onset features were higher than offset features. However, when k-NN was trained on all features, offset features contributed a great deal to the overall performance (approximately 4%). For the Bayesian network, offset features had no significant impact on overall performance when all features were used.

All the connections between features of every feature set were removed in the Bayesian network to test the effect of dependencies between features on the classification performance. For instance, in Fig. 10, there were no longer arrows from S_1

TABLE III
CONFUSION MATRIX FOR THE k-NN, THE BAYESIAN NETWORK (BN), AND THE MODEL PROPOSED IN [26]. ALL NUMBERS ARE PERCENTAGES. THE OVERALL AVERAGE CLASSIFICATION ACCURACY IS 94.03% FOR THE k-NN, 95.94% FOR THE BN, AND 95.08% FOR [26].

Presented	**method**	**Detected**				
		piano	clarinet	oboe	violin	trumpet
piano	k-NN	**100**	0.00	0.00	0.00	0.00
	BN	**100**	0.00	0.00	0.00	0.00
	[26]	**95.81**	1.40	0.47	0.00	2.33
clarinet	k-NN	0.00	**96.30**	0.46	1.85	1.39
	BN	0.02	**92.92**	4.66	2.32	0.08
	[26]	1.40	**92.52**	5.14	0.93	0.00
oboe	k-NN	0.00	6.25	**88.20**	0.69	4.86
	BN	0.00	0.73	**91.56**	6.69	1.02
	[26]	0.00	2.10	**95.10**	2.10	0.70
violin	k-NN	0.46	0.93	0.46	**96.76**	1.39
	BN	0.00	0.43	0.69	**98.88**	0.00
	[26]	1.07	0.53	0.00	**95.45**	2.94
trumpet	k-NN	0.00	3.78	5.56	1.78	**88.88**
	BN	0.00	1.51	0.72	1.44	**96.33**
	[26]	0.00	0.00	0.00	3.47	**96.53**

TABLE IV
OVERALL AVERAGE CLASSIFICATION ACCURACY OF THE k-NN AND THE BAYESIAN NETWORK (BN) FOR DISTINCT FEATURE SETS AND THEIR COMBINATIONS: 1) OFFSET ONLY, 2) ONSET ONLY, 3) SPECTRAL ONLY, 4) SPECTRAL & OFFSET, 5) ONSET & OFFSET, 6) SPECTRAL & ONSET, 7) ALL FEATURES.

	Feature sets						
	1	2	3	4	5	6	7
k-NN	71.6	83.19	84.35	88.52	89.72	90.05	94.03
BN	78.37	89.59	85.74	92.76	90.13	95.48	95.94

to S_2 or from S_2 to S_3, etc. The overall average classification accuracy decreased from 95.94% to 90.60%. Therefore, this classifier was more successful in encoding the signal characteristics when the first order Markovian dependencies were present. This was expected as the time evolution of frequency components of acoustic signals is structured and not random.

C. Timbral Feature Selection

The model's potential for timbre representation was tested in a third application. The idea was to use the model to find the best features that could distinguish between three timbral classes consisting of 23 sounds. We used these sounds in an earlier audio-visual experiment, where 119 subjects were asked to select a visual shape out of a set of three for each [36]. Results showed that there existed a strong correspondence between timbre and visual shapes. Accordingly, sounds can be classified into three groups (timbral classes) based on the selected visual shapes that represent the perceptual similarities/dissimilarities of the timbres of these sounds. In this study, we hypothesized that sounds associated with the same visual shape should share timbral features. The goal was to find a space of features in which sounds associated with the same visual shape were nearby. In other words, we wished to find out which features invoked the selection of a particular shape for a given sound.

All the 23 sounds used in [36] were 1 second in duration and were sampled at 44100 Hz. They included notes from piano, cello, guitar, marimba and saxophone all of which had been multiplied by a Hanning window with a duration of 1 second to equalize the onset and offset such that only the spectral features of timbre were preserved. Therefore, it was more difficult for the subjects to distinguish between the timbres of these instruments as their sounds had similar onsets and offsets. For each instrument, four notes were selected: G_2 (98 Hz), D_3 (146.83 Hz), G_3 (196 Hz), and B_3 (246.94 Hz). The other 3 sounds were notes from gong, crash cymbals and triangle.

Sounds associated with the visual shapes labeled as S_1, S_2, and, S_3 in [36] are referred to as class 1, class 2, and class 3, respectively. Sounds derived from crash cymbals, gong and triangle belong to class 1. Class 2 consists of the notes G_3 (196 Hz) and B_3 (246.94 Hz) from cello and guitar, and all the four notes from saxophone. Class 3 contains the notes G_2 (98 Hz) and D_3 (146.83 Hz) derived from cello and guitar and all the four notes from piano and marimba.

To obtain the features that were involved in the selection of visual shapes for the above 23 sounds, 15 known timbral features were extracted from the sounds: spectral and temporal centroids as well as standard deviations, Kurtosises, skewnesses and spreads, spectral flatness and flux, log attack time, effective roughness, and energy of AMs from 10 Hz to 30 Hz. Spectral and temporal features were computed from the time-averaged spectrum (8) and the global temporal envelope (21), respectively. All features were normalized to have zero means and unit standard deviations.

The following procedure was performed for all possible combinations of the 15 features in N-dimensional spaces, where N (the number of features) varied from 1 to 5. There were a total of 4943 feature spaces to examine: 15 1D spaces, 105 2D spaces, 455 3D spaces, 1365 4D spaces, and 3003 5D spaces. We used the k-means algorithm to find three clusters of sounds in a given feature space. The 3 clusters were then compared to the 3 timbral classes from the audiovisual experiment. In general, there are 6 combinations to map the 3 clusters to the 3 known timbral classes. The combination with the minimum classification error was kept for that feature space. Classification error was defined as the percentage of the 23 sounds that were not correctly assigned to their known timbral classes. In the end, we searched for the space with the minimum classification error by comparing all the 4943 feature spaces.

The minimum classification error of 4.35% was achieved for a 2D space whose dimensions were log attack time and spectral centroid, two important features that had also been found by other studies [9, 15]. Attack time was computed as the time required for the global temporal envelope $e[n]$ to increase from 0.1max$\big[e[n]\big]$ to its maximum max$\big[e[n]\big]$. Spectral centroid was defined by:

$$SC = \frac{\sum_{l=1}^{L} \bar{S}_l . ERB_{num}(f_{c,l})}{\sum_{l=1}^{L} \bar{S}_l} \tag{27}$$

where L is 84, $\bar{S}$ is the time-averaged spectrum from (8), $f_{c,l}$ is the center frequency of the l^{th} filter, and the function ERB_{num} is defined in (2). This definition of spectral centroid is different from the conventional one that uses a linear frequency scale

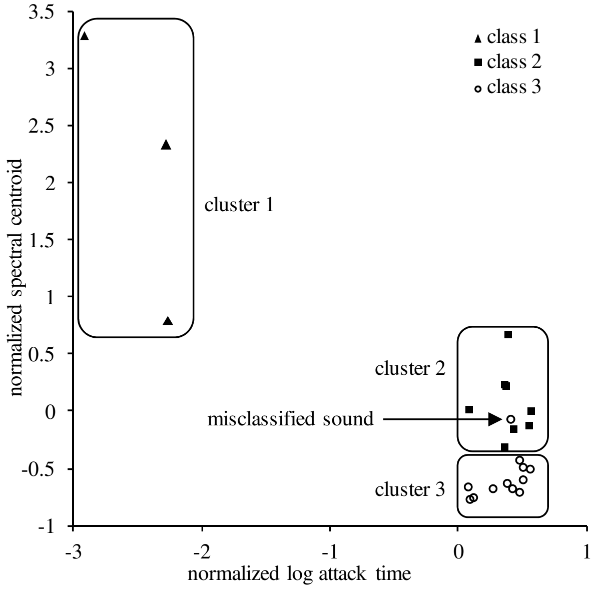

Fig. 11. The timbre space obtained for the labeled sounds used in the feature selection task. The log attack time and spectral centroid were found to be the best descriptors for the three timbral classes. The 3 clusters obtained by the k-means algorithms for the 3 timbral classes of [36] are shown. The relative positioning of the three timbral classes in this space is in accordance with the perceptual associations obtained in [36].

instead of the warped scale $ERB_{num}(f_{c,l})$, and can be considered as a subjective measure for frequency. The representation of the sounds in the obtained 2D space is shown in Fig. 11. In this figure, triangles, squares, and circles represent the sounds in classes 1, 2, and 3, respectively. The sounds in class 1 had much faster attack time (because they were not weighted with a Hanning window) and higher spectral centroid. The sounds in classes 2 and 3 had similar attack time. This was expected as they had been multiplied by a Hanning window which had removed onset differences. Despite the attack time, spectral centroid distinguishes between the sounds in class 2 and those in class 3.

We now compare the 2D space in Fig 11 to the results presented for the grayscale shapes (represented by CMG confidence measure) in Table 1 of [36]. This comparison is also valid for the colored shapes (represented by CMC confidence measure) in the same table. In [36], on average, for the sounds in class 1, 89% of subjects have selected the shape S_1, whereas only 11% have selected the shape S_2 (the shape of class 2) or S_3 (the shape of class 3) altogether. This shows that almost all subjects have agreed that the timbres of class 1 are perceptually very distant from those of class 2 and 3. This is clearly seen in the obtained 2D space in Fig. 11 where there is a large distance between the sounds of class 1 and those of classes 2 and 3. On the other hand, for the sounds in class 2, on average 60% of the subjects have selected the shape S_2, whereas 24% have selected the shape S_3 (the shape of class 3). Similarly, for the sounds in class 3, on average 66% of the subjects have selected the shape S_3 whereas 30% have selected the shape S_2 (the shape of class 2). Thus, there is a considerable percentage of the subjects who have selected the shape S_2 for class 3 and the shape S_3 for class 2. This is because of the similarity between some timbres of class 2 and some timbres of class 3. For instance, as shown in Table 1 of [36], the notes G_2 (98 Hz) and D_3 (146.83 Hz) derived from cello and guitar have been associated with the shape S_3 (class 3) whereas the notes G_3 (196 Hz) and B_3 (246.94 Hz) derived from the same instruments have been associated with the shape S_2 (class 2). However, notes of an instrument, regardless of the pitch difference, have similar timbres. This is the reason why, even though the k-means algorithm have separated the classes 2 and 3 in Fig. 11, some sounds in class 2 are close to some sounds in class 3. Therefore, the 2D timbre space in Fig. 11 is completely in accordance with the perceptual timbre-shape associations found in [36].

IV. Discussion and Future Work

We designed a novel bio-inspired hierarchical timbre model that 1) is able to extract all the important timbral features hierarchically (as biological systems do) rather than extract them using separate parallel subsystems and 2) can be used in various applications rather than a specific application.

We used the proposed model in three different applications to verify its potential to capture different aspects of timbre. In the first application, the proposed model was able to simulate the subjective values of roughness. Though roughness is usually considered to be a single quantity, one of the novelties of the proposed model is that it extracts an instantaneous roughness function because amplitude modulations (and consequently roughness) vary with time. The objective roughness values estimated by the proposed model were highly correlated with the subjective values in [35]. Therefore, the proposed model effectively encoded amplitude modulations and successfully simulated their perceptual characteristics.

Roughness is an important feature that contributes to the richness of timbre, however there are other important features that convey very important information about timbres of instruments: spectral features such as harmonic structure and resonances, and temporal features such as attack and decay. In the second application, the proposed model proved highly efficient in capturing and encoding spectral and temporal features for the purpose of instrument classification. Though our classification results were comparable to those in [26], our model provided much lower dimensional representations for timbre. The presented results were achieved using only 102 features while in [26] at least 1000 features were used. The performance of the system is comparable to the state-of-the-art methods of instrument classification. However, other systems may achieve higher rates under different circumstances. For instance, the bio-inspired model in [31] achieved a classification rate of 98.7% though at a higher cost (and using different classifiers and a different dataset). In [31], the system included 30976 filters (128 × 22 × 11) and provided 242 dimensional feature spaces for signals whereas the proposed model included only 1260 filters (84 × 15) and extracted 102 features for the signals used in section III.

In the third application, the goal was to find a timbre space that best characterizes three timbral classes. Though it would have been possible to compute features using linear methods such as the Fourier transform, the features that we used were extracted by the proposed model, which simulates nonlinearities of the auditory system e.g. nonuniform spectral resolution and compression. Attack time and spectral centroid proved to be the best features for this application. The representation of the three timbral classes in the obtained timbre space (constructed by these 2 features) is in complete agreement with the perceptual timbral qualities obtained in [36]. This is also consistent with other studies, where spectral centroid and attack time were shown to be the two main timbre dimensions, though using different settings.

The proposed model currently does not account for the effect of the sign of carrier phase on roughness. More evidence on this effect is required to incorporate a phase processing module into the system. Another important characteristic of timbre that is not encoded by the system is vibrato. Though, in general, the leakage of energy between the adjacent channels can be used to estimate vibrato, further investigation is required to appropriately quantify vibrato. In addition, since we have only modeled timbre in this work, it would be of interest to integrate mechanisms for pitch and loudness perception into the proposed system to fully represent acoustic signals.

In future studies, the model will be evaluated in other contexts, such as speech and music separation, music genre classification, and speech recognition using larger and more varied databases. In order to use the model in the context of auditory scene analysis (e.g. cocktail party problem and instrument recognition in polyphonic music), more modules such as sequential or parallel grouping based on pitch, and onset and offset times, should be developed and integrated into the model. It is also of interest to compare the performance of the proposed profiles of timbre with the well-known Mel-scale Cepstral Coefficients (MFCCs) in different applications.

V. Conclusion

We presented a multipurpose bio-inspired hierarchical model that extracts three profiles for timbre: time-averaged spectrum, global temporal envelope, and instantaneous roughness. The model was tested in three applications. First, it successfully simulated the subjective roughness data obtained in [35]. Second, it was used to classify musical instruments, where the k-NN algorithm and a Bayesian network achieved classification rates of 94.03 % and 95.94%, respectively, using only 102 features. Finally, it successfully obtained a timbre space for three classes of sounds with labeled timbres. Spectral centroid and log attack time were obtained as the features that best described the perceptual qualities of the labeled sounds. Regarding the results of these diverse tests, we have shown that the proposed model has high potential for encoding spectral, temporal and spectrotemporal characteristics of timbre and is applicable in various contexts.

Authors' contributions

This study was conceived by J.R. and M.A. M.A. designed the model, conducted the three tests on the model, analyzed the results and wrote the manuscript. J.R. supervised the research and also contributed to the model design, tests, analysis of results, and manuscript writing. S.W. contributed to the musical instrument classification test and manuscript writing. S.M. and E.P. contributed to the analysis of results and manuscript writing.

Acknowledgment

The LTFAT toolbox was very beneficial in this work. The authors would like to thank all those who have contributed to the development of this toolbox. The authors would also like to thank the anonymous reviewers whose insightful comments and suggestions led to the improvement of both the evaluation of the proposed system and the clarity of the paper.

References

[1] M. Sonn, "American national psychoacoustical terminology (ANSI S3.20)," American National Standards Institute (ANSI), p. 67, 1973.

[2] J. C. Risset and D. L. Wessel, "Exploration of timbre by analysis and synthesis," in *The Psychology of Music*, 2nd ed., D. Deutsch, Ed. San Diego, CA, USA: Elsevier, 1999, vol. 28, pp. 113–169.

[3] S. McAdams, "Perspectives on the contribution of timbre to musical structure," *Computer Music Journal*, vol. 23, no. 3, pp. 85–102, Sep. 1999.

[4] S. Mcadams and B. L. Giordano, "The perception of musical timbre," in *Oxford Handbook of Music Psychology*, S. Hallam, I. Cross, and M. Thaut, Eds. New York: Oxford University Press, 2009, pp. 72–80.

[5] K. W. Berger, "Some factors in the recognition of timbre," *J Acoust Soc Am*, vol. 36, no. 10, pp. 1888–1891, 1964.

[6] A. J. M. Houtsma, "Pitch and timbre: Definition, meaning and use," *Journal of New Music Research*, vol. 26, no. 2, pp. 104–115, Jun. 1997.

[7] G. von Bismarck, "Sharpness as an attribute of the timbre of steady sounds," *Acta Acust united Ac*, vol. 30, no. 3, pp. 159–172, 1974.

[8] ——, "Timbre of steady sounds: A factorial investigation of its verbal attributes," *Acta Acust united Ac*, vol. 30, no. 3, pp. 146–159, 1974.

[9] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Amer.*, vol. 61, no. 5, pp. 1270–1277, 1977.

[10] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *J Acoust Soc Am*, vol. 63, no. 5, pp. 1493–1500, 1978.

[11] C. L. Krumhansl, "Why is musical timbre so hard to understand," in *Structure and Perception of Electroacoustic Sound and Music*, S. Nielzen and O. Olsson, Eds. Excerpta Medica, 1989, pp. 43–54.

[12] P. Iverson and C. L. Krumhansl, "Isolating the dynamic attributes of musical timbre," *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 2595–2603, 1993.

[13] S. Handel, "Timbre perception and auditory object identification," in *Hearing (Handbook of Perception and Cognition)*, 2nd ed., B. C. J. Moore, Ed. San Diego, California, USA: Academic Press, 1995, pp. 425–461.

[14] S. Lakatos, "A common perceptual space for harmonic and percussive timbres," *Perception Psychophysics*, vol. 62, no. 7, pp. 1426–1439, 2000.

[15] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," *J Acoust Soc Am*, vol. 118, no. 1, p. 471, 2005.

[16] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychol. Res.*, vol. 58, no. 3, pp. 177–192, 1995.

[17] J. M. Hajda, R. A. Kendall, E. C. Carterette, and M. L. Harshberger, "Methodological issues in timbre research," in *Perception and Cognition of Music*, I. Deliège and J. Sloboda, Eds. East Sussex, UK: Psychology Press, 1997, pp. 253–307.

[18] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *J Acoust Soc Am*, vol. 130, no. 5, pp. 2902–2916, 2011.

[19] X. Zhang and Z. W. Ras, "Analysis of sound features for music timbre recognition," in *Int. Conf. Multimedia and Ubiquitous Engineering (MUE)*. Seoul, Korea: IEEE, Apr. 2007, pp. 3–8.

[20] S. McAdams, "Recognition of sound sources and events," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, S. McAdams and E. Bigand, Eds. Oxford Univ. Press, Apr. 1993, pp. 146–198.

[21] G. L. Dannenbring and A. S. Bregman, "Stream segregation and the illusion of overlap," *Journal of Experimental Psychology Human Perception and Performance*, vol. 2, no. 4, p. 544, 1976.

[22] B. Roberts and A. S. Bregman, "Effects of the pattern of spectral spacing on the perceptual fusion of harmonics," *J Acoust Soc Am*, vol. 90, no. 6, pp. 3050–3060, 1991.

[23] P. Iverson, "Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes," *Journal of Experimental Psychology Human Perception and Performance*, vol. 21, no. 4, pp. 751–763, 1995.

[24] B. C. Moore and H. Gockel, "Factors influencing sequential stream segregation," *Acta Acust united Ac*, vol. 88, no. 3, pp. 320–333, 2002.

[25] I. Fujinaga, "Machine recognition of timbre using steady-state tone of acoustic musical instruments," in *Proceedings of the International Computer Music Conference*. Ann Arbor, MI, USA: Citeseer, 1998, pp. 207–210.

[26] J. J. Burred, A. Robel, and T. Sikora, "Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds," *IEEE Trans Audio Speech Lang Process*, vol. 18, no. 3, pp. 663–674, Mar. 2010.

[27] J.-J. Aucouturier, F. Pachet, and M. Sandler, ""The way it Sounds": timbre models for analysis and retrieval of music signals," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1028–1035, Dec. 2005.

[28] M. Muller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Trans Audio Speech Lang Process*, vol. 18, no. 3, pp. 649–662, Mar. 2010.

[29] H. Ezzaidi, M. Bahoura, and G. E. Hall, "Towards a characterization of musical timbre based on chroma contours," in *Communications in Computer and Information Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 322, ch. 17, pp. 162–171.

[30] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans Audio Speech Lang Process*, vol. 16, no. 1, pp. 116–128, Jan. 2008.

[31] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: the biological bases of musical timbre perception," *PLoS Comput. Biol.*, vol. 8, no. 11, p. e1002759, Nov. 2012.

[32] R. Duisters, "The modeling of auditory roughness for signals with temporally asymmetric envelopes," Master's thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 2005.

[33] A. Kohlrausch, D. Hermes, and R. Duisters, "Modeling roughness perception for sounds with ramped and damped temporal envelopes," in *Proceedings of Forum Acusticum*, Budapest, Hungary, 2005, pp. 1719–1724.

[34] V. Vencovsky, "Modeling roughness perception for complex stimuli using a model of cochlear hydrodynamics," in *International Symposium on Musical Acoustics (ISMA 2014)*, Le Mans, France, 2014, pp. 484–488.

[35] D. Pressnitzer and S. McAdams, "Two phase effects in roughness perception," *J Acoust Soc Am*, vol. 105, no. 5, pp. 2773–2782, 1999.

[36] M. Adeli, J. Rouat, and S. Molotchnikoff, "Audiovisual correspondence between musical timbre and visual shapes," *Front. Hum. Neurosci.*, vol. 8, 2014.

[37] T. Necciari, P. Balazs, N. Holighaus, and P. Sondergaard, "The ERBlet transform: An auditory-based time-frequency representation with perfect reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. Vancouver, BC, Canada: IEEE, May. 2013, pp. 498–502.

[38] P. L. Søndergaard, B. Torrésani, and P. Balazs, "The linear time frequency analysis toolbox," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 10, no. 04, pp. 1–26, 2012.

[39] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1, pp. 103–138, 1990.

[40] L. Robles and M. A. Ruggero, "Mechanics of the mammalian cochlea," *Physiol. Rev.*, vol. 81, no. 3, pp. 1305–1352, 2001.

[41] S. A. Shamma, "Speech processing in the auditory system ii: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," *J Acoust Soc Am*, vol. 78, no. 5, pp. 1622–1632, 1985.

[42] C. Blakemore and E. A. Tobin, "Lateral inhibition between orientation detectors in the cat's visual cortex," *Exp. Brain Res.*, vol. 15, no. 4, pp. 439–440, Sep. 1972.

[43] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 22.

[44] W. De Baene, A. Vandierendonck, M. Leman, A. Widmann, and M. Tervaniemi, "Roughness perception in sounds: behavioral and ERP evidence," *Biological Psychology*, vol. 67, no. 3, pp. 319–330, Nov. 2004.

[45] P. X. Joris, C. E. Schreiner, and A. Rees, "Neural processing of amplitude-modulated sounds," *Physiological Reviews*, vol. 84, no. 2, pp. 541–577, Apr. 2004.

[46] T. Miyazaki, J. Thompson, T. Fujioka, and B. Ross, "Sound envelope encoding in the auditory cortex revealed by neuromagnetic responses in the theta to gamma frequency bands," *Brain Research*, vol. 1506, pp. 64–75, Apr. 2013.

[47] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database." in *International Society for Music Information Retrieval Conference,(ISMIR 2003)*, vol. 3, Baltimore, MD, USA, 2003, pp. 229–230.

[48] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Classification of musical patterns using variable duration hidden markov models," *IEEE Trans Audio Speech Lang Process*, vol. 14, no. 5, pp. 1795–1807, Sep. 2006.

[49] J. Marques and P. J. Moreno, "A study of musical instrument classification using gaussian mixture models and support vector machines," *Cambridge Research Laboratory Technical Report Series CRL*, vol. 4, 1999.

[50] P. J. Donnelly and J. W. Sheppard, "Classification of musical timbre using bayesian networks," *Computer Music Journal*, vol. 37, no. 4, pp. 70–86, Dec. 2013.

[51] I. Kaminskyj and T. Czaszejko, "Automatic recognition of isolated monophonic musical instrument sounds using kNNC," *J. INTELL. INF. SYST.*, vol. 24, no. 2/3, pp. 199–221, Mar. 2005.

[52] Q. Ding and N. Zhang, "Classification of recorded musical instruments sounds based on neural networks," in *Proc. IEEE Symp. Computational Intelligence in Image and Signal Process., (CIISP).* Honolulu, HI, USA: IEEE, Apr. 2007, pp. 157–162.

Mohammad Adeli received his bachelor's degree in electrical engineering from Isfahan University of Technology, Isfahan, Iran in 2001. In 2004, he received his master's degree in biomedical engineering from Sharif University of Technology, Tehran, Iran. He also received a PhD degree in electrical engineering from Université de Sherbrooke, QC, Canada in 2015.

His research interests include signal processing, audio processing using bio-inspired methods, roughness and timbre modeling, machine learning, computational neuroscience, perceptual audiovisual correspondences, and bio-inspired auditory to visual substitution systems.

Jean Rouat holds a M.Sc. degree in Physics from Univ. de Bretagne, France (1981), an E. & E. M.Sc.A. degree in speech coding and speech recognition from Univ. Sherbrooke (1984) and an E. & E. Ph.D. in cognitive and statistical speech recognition jointly with Univ. de Sherbrooke and McGill Univ. (1988). His post-doc has been in psychoacoustics with the MRC, App. Psych. Unit, Cambridge, UK and in electrophysiology with the Institute of physiology, Lausanne, Switzerland. He is now with Univ. de Sherbrooke where he founded the Computational Neuroscience and Intelligent Signal Processing Research group (NECOTIS). He is also adjunct prof. in the biological sc. dep., Montreal Univ. His translational research links neuroscience and engineering for the creation of new technologies and a better understanding of learning multimodal representations. Development of hardware low power consumption Neural Processing Units for a sustainable development, interactions with artists for multimedia and musical creations are examples of transfers that he leads based on the knowledge he gains from neuroscience. He is leading funded projects to develop sensory substitution and intelligent systems. He is also leading an interdisciplinary CHIST-ERA european consortium (IGLU - Interactive Grounded Language Understanding) for the development of an intelligent agent that learns through interaction.

Stéphane Molotchnikoff received his Ph.D. degree in neuroscience from State University of New York (Buffalo) in 1973. He joined Université de Montréal the same year. Full professor at the Département de Sciences Biologiques, he is teaching physiology. The research interests of his laboratory are neuronal plasticity following adaptation using model of orientation selectivity in visual cortex of mammals and neural assembly or connectomes selectivity in response to sensory stimuli.

Eric Plourde received both the B.Ing. degree in electrical engineering and the M.Sc.A. degree in biomedical engineering from the Ecole Polytechnique de Montréal, QC, Canada in 2003. He completed a Ph.D. degree in electrical engineering from McGill University, Montreal, QC, Canada in 2009. From 2009 to 2011, he was a Postdoctoral Fellow in the Neuroscience Statistics Research Laboratory with joint appointments at the Massachusetts General Hospital, Harvard Medical School and the Massachusetts Institute of Technology. He joined the Université de Sherbrooke in 2011, where he is an Assistant Professor within the Department of Electrical and Computer Engineering. His research interests include neural signal processing as well as speech processing with emphasis on speech enhancement and perceptually/biologically inspired processing.