Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments

Etienne Thoret,[a] Philippe Depalle, and Stephen McAdams
Schulich School of Music, McGill University, Montreal, Quebec, Canada
etienne.thoret@mcgill.ca, philippe.depalle@mcgill.ca,
stephen.mcadams@mcgill.ca

Abstract: Modulation Power Spectra include dimensions of spectral and temporal modulation that contribute significantly to the perception of musical instrument timbres. Nevertheless, it remains unknown whether each instrument's identity is characterized by specific regions in this representation. A recognition task was applied to tuba, trombone, cello, saxophone, and clarinet sounds resynthesized with filtered spectrotemporal modulations. The most relevant parts of this representation for instrument identification were determined for each instrument. In addition, instruments that were confused with each other led to non-overlapping spectrotemporal modulation regions, suggesting that musical instrument timbres are characterized by specific spectrotemporal modulations.

[DMC]
Date Received: August 5, 2016 **Date Accepted:** November 9, 2016

1. Introduction

Acoustical correlates related to the perception of musical instrument timbre have been studied for some time (Miller and Carterette, 1975) and are mainly defined through dissimilarity experiments leading to a multidimensional representation. From this representation, timbre dimensions have been correlated with different acoustic descriptors such as attack time and spectral centroid (McAdams *et al.*, 1995). Recent studies have provided evidence for the prominent role of spectrotemporal modulations for timbre perception and sound source classification (Hemery and Aucouturier, 2015) and more specifically for musical instruments (Patil *et al.*, 2012; Elliott *et al.*, 2013). Moreover, these spectrotemporal modulations have been shown to be plausibly extracted through specific neural processes at the level of primary auditory cortex (Shamma, 2001). Nevertheless, it remains unknown which aspects of spectrotemporal modulations are relevant for the perception of musical instrument timbre and whether the same modulation properties are relevant across different musical instruments.

Here we tackled these questions for a small subset of musical instruments to determine which parts of the Modulation Power Spectrum (MPS) lead to the recognition of five different sustained instrument timbres (tuba, trombone, saxophone, clarinet, and cello). Based on a "molecular" approach (Gosselin and Schyns, 2001), a recognition task was set up in which listeners had to identify the instruments from a processed version of the original sounds composed from only a small part of their MPS. The aim was to determine which parts of the MPS are the most salient for the recognition of these five instruments.

2. The MPS of musical sounds

The MPS is defined here as the two-dimensional (2D) Fourier transform of the time-frequency representation of a sound signal (Elliott and Theunissen, 2009; Singh and Theunissen, 2003). More specifically, the time-frequency $X(t,f)$ representation itself is defined as the amplitude of the Fourier transform obtained with a Hamming window and is commonly known as the magnitude of the Short-Term Fourier Transform (STFT) or the Gabor Transform. Here, the length of the window equals 46.43 ms with a hop size of 11.61 ms and the sample rate is set to 44 100 Hz. These parameters were chosen to fulfill the time-frequency constraints for properly representing the harmonic structure of the stimuli (Rabiner and Schafer, 1978). The MPS is the amplitude of the successive Fourier transforms along the STFT temporal and frequency axes. This MPS representation is composed of two dimensions: the temporal modulations (in Hz) and the spectral modulations (in cycles/Hz). To summarize, the MPS is the 2D Fourier

[a] Author to whom correspondence should be addressed.

transform of the modulus of a linear time-frequency representation and can be understood as a representation of the spectral and temporal regularities of a spectrogram. A detailed description of how regularities of a spectrogram translate into the MPS is presented in Elliott and Theunissen (2009). Formally, MPS is defined by the following equation:

$$\mathrm{MPS}(s,r) = \int\int |X(t,f)| e^{-2\pi isf} e^{-2\pi irt} dfdt, \tag{1}$$

where s and r are the spectral and temporal modulations, respectively. The resolution of MPS (s,r) is constrained by the resolution of the time-frequency representation $X(t,f)$ mainly characterized by the sizes of the temporal Hamming windows and the overlap between two successive windows. They indeed define the upper and lower boundaries of the spectral and temporal modulations axes. Constrained by the uncertainty principle $\sigma_t \geq 1/4\pi\sigma_f$ where σ_t and σ_f correspond to the uncertainties along the temporal and spectral modulations dimensions, respectively, we here chose $\sigma_t = 11.61$ ms and $\sigma_f = 21.53$ Hz leading to upper boundaries of 43 Hz and 23.22 cycles/Hz, which correspond to values relevant for the auditory perception of sounds such as speech (Elliott and Theunissen, 2009).

3. Filtering the MPS

In order to determine which parts of the MPS lead to the recognition of musical instruments, we employed a technique for filtering instrumental sounds in the spectro-temporal modulation domain. With this technique, a sound is processed by keeping only a part of its MPS, this filtered version is reconstructed, and whether the information that remains is relevant for the recognition of the initial instrument is then evaluated. This filtering process degrades the temporal and spectral regularities of the original sounds, which become more or less identifiable according to the remaining acoustical information. Hence, the MPS is first multiplied by a 2D Gaussian filter frequency response $G_{(\mu_s,\sigma_s),(\mu_r,\sigma_r)}(s,\, r)$ where μ_s, μ_r and σ_s, σ_r are the means and standard deviations (SDs) in the spectral and temporal modulation dimensions, respectively,

$$G_{(\mu_s,\sigma_s),(\mu_r,\sigma_r)}(s,\, r) = \exp\left(-\frac{1}{2}\left(\frac{s-\mu_s}{\sigma_s}\right)^2\right)\exp\left(-\frac{1}{2}\left(\frac{r-\mu_r}{\sigma_r}\right)^2\right). \tag{2}$$

It must be noted that the MPS and the filter G are composed of four quadrants with positive and negative spectral and temporal modulations. For the sake of simplicity and as the filter is perfectly symmetric in amplitude and anti-symmetric in phase in the spectral and temporal modulation dimensions, in the following, only values of positive spectral modulations and rates are considered. The MPS-filtered spectrogram $Y(t,f)$ can then be reconstructed by a 2D inverse Fourier transform of the processed MPS: $\mathrm{MPS}(s,r) \cdot G_{(\mu_s,\sigma_s),(\mu_r,\sigma_r)}(s,r)$. Note that $Y(t,f)$ is magnitude only, lacks the phase, and thus does not allow for perfect reconstruction of the waveform directly from standard reconstruction techniques such as the overlap add method (Rabiner and Schafer, 1978). We therefore used Griffin and Lim's (1984) algorithm in a MATLAB implementation provided by Slaney (1994) in order to iteratively build a signal, the STFT magnitude of which is as close as possible to $Y(t,f)$ in a quadratic sense. Twenty-five iterations lead to a correct reconstruction of the waveform for an acceptable computation time. Practically, the quality of the reconstruction is evaluated by computing the averaged relative log-error ratio ϵ in percent between the desired spectrogram $Y(t,f)$ and the STFT magnitude of the reconstructed waveform $Y_b(t,f)$,

$$\epsilon = 100\frac{1}{N_f N_t}\sum_{t_i=1}^{N_t}\sum_{f_i=1}^{N_f}\left|\frac{\log(Y(t_i,f_i)) - \log(Y_b(t_i,f_i))}{\log(Y(t_i,f_i))}\right|, \tag{3}$$

where N_f and N_t are the number of frequency and time bins, respectively. Note that $\log Z$ was floored to -100 when $\log Z$ is smaller.

Filtering the MPS implies different modifications of the original sound. For instance, when low spectral modulations and rates are retained after filtering, the resulting sound preserves the slow envelope variations and the coarse spectral structure of the original sound (e.g., formants and pitch). Conversely, retaining the high spectral modulations and rates keeps the fine temporal and spectral structure of the sound.

4. Materials and methods

4.1 Participants

Twenty-three participants (16 females) with ages between 18 and 30 ($M = 22.9$, SD = 3.2) took part in the experiment. All the participants were musicians having completed at least second-year university-level training. Participants provided informed consent, had normal hearing, and were compensated for their time.

4.2 Stimuli

The stimuli were five arpeggios generated from samples of the Vienna Symphonic Library (2015). Five instruments (trombone, tuba, tenor saxophone, cello, and clarinet) and three pitches (F#3–185.0 Hz, C4–261.6 Hz, and F#4–370.0 Hz) were chosen. For each instrument, the three notes were equalized in loudness in a preliminary experiment. Their durations were all cut to 0.5 s with a raised cosine fade-out amplitude envelope of 50-ms duration to create a constant duration of 500 ms to avoid discrimination based on this criterion. The attack was preserved. Finally, arpeggios were generated by concatenating the three notes from the lowest to the highest with no silences between the notes. The peak level of the stimuli ranged from 58 to 71 dB Sound Pressure Level (*A-weighted*).

4.3 Apparatus

The experiment took place in an IAC Acoustics double-walled sound-isolation chamber (IAC Acoustics, Bronx, NY). Stimuli were sampled at 44 100 Hz and presented over Sennheiser HD280Pro headphones (Sennheiser Electronics GmbH, Wedemark, Germany) using a Macintosh computer (Apple Computer, Inc., Cupertino, CA) with digital-to-analog conversion on a Grace Design m904 monitor system (Grace Digital Audio, San Diego, CA). The experimental interface and data collection were programmed in the Max7 audio software environment (Cycling '74, San Francisco, CA) and the MATLAB software (The Mathworks, Inc., Natick, MA) interacting via the *udp* protocol, respectively.

4.4 Procedure

Participants first completed a standard pure-tone audiogram to ensure normal hearing with hearing thresholds of 20 dB Hearing Level or better at octave-spaced frequencies in the range of 250–8000 Hz. The task was 5-Alternative Forced Choice. In each trial, the participants were asked to recognize the instrument that played the arpeggios among the five instruments. They were asked to answer as quickly as possible after hearing the sounds. The triggering of the trials was controlled by the participant by clicking on a button to play the next stimulus. The experiment began with a training session of 15 trials (5 instruments × 3 repetitions) during which the participants performed the task with the original, unprocessed sounds. After having completed the training session, the participants began the main experiment, which was composed of 480 trials (5 instruments × 96 filters). For each instrument, the MPS was filtered with 96 Gaussian filters $G_{(\mu_s,\sigma_s),(\mu_r,\sigma_r)}$ with the following SDs: $\sigma_s = 4$ cycles/Hz and $\sigma_r = 5$ Hz overlapping by 25% in each dimension (12 rates and 8 spectral modulations, see Fig. 1). These SDs were determined by empirical tests in order to provide a good trade-off

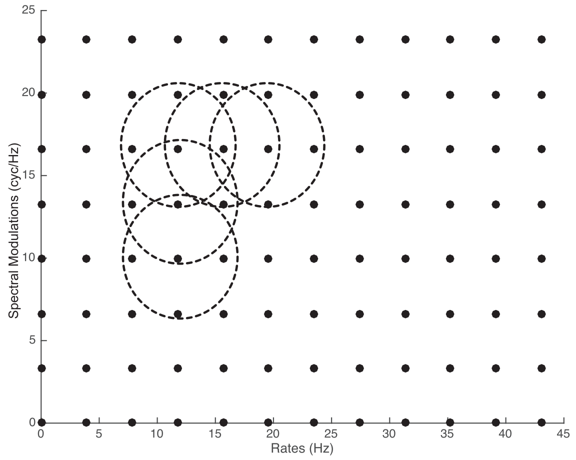

Fig. 1. Sampling of the MPS by 96 Gaussian MPS filters. The dots show the center value and the circles the SD of the 2D Gaussian distribution.

between accurate sampling and a reasonable number of filters for sampling the MPS. The relative log-error ratio [cf. Eq. (3)] for the 480 sounds equaled 10.25%. Hence in each trial, one among the five instrument arpeggios was filtered with one filter, and the participants had to recognize the original instrument. The order of presentation of the 480 trials was randomized for each participant.

4.5 Data analysis

For all participants and for all five instruments, a confusion matrix was computed from the responses to the MPS-filtered sounds and association scores were tested against chance level with a one-tailed Wilcoxon signed-rank test. The correct identification rates, i.e., the diagonal values of the confusion matrix, were also compared with those obtained from the identification of original sounds in the training session with a two-tailed Wilcoxon signed-rank test. The subsequent data analysis was inspired by a method proposed by Gosselin and Schyns (2001). In each trial, if the sound was properly associated with the instrument, the MPS-filter was added to a *CorrectMask* matrix. Across all trials, each MPS-filter was added to a *TotalMask* matrix. For each participant, a *ProportionMask* was derived by dividing *CorrectMask* by *TotalMask*. If no region had any special perceptual significance for recognition, *ProportionMask* would be homogeneous. On the contrary, if some regions were more important for recognition, they would have higher values than the other regions of the *ProportionMask*.

Note that our method differs from the so-called "bubbles" method proposed by Gosselin and Schyns (2001), which was initially used to determine the most salient parts of a face for gender and expressivity recognition. Although they used a self-calibrating method that adjusted the number of bubbles to converge on 75% correct recognition, here we only used single bubbles in order to determine their independent contribution to instrument recognition. Given that MPS-filters overlap each other, the resulting *ProportionMasks* represent the relative importance of each part of the MPS to the recognition of that instrument.

For each instrument and across participants, the statistical significance of these latter regions is determined with a one-tailed Wilcoxon signed-rank test between *ProportionMask* values and the averaged value of the *ProportionMask* ($\alpha = 0.01$). This average value is then used as a threshold for computing a Boolean matrix called *DiagnosticMask* out of *ProportionMask*.

5. Results

Table 1 presents the averaged confusion matrix across participants computed from responses to MPS-filtered sounds. All the instruments were recognized above chance ($z > 4$; $p < 0.001$). In addition, tuba, cello, and saxophone were significantly confused with trombone ($z = 2.99$; $p < 0.001$), saxophone ($z = 3.60$; $p < 0.05$), and cello ($z = 2.02$; $p < 0.001$), respectively. Moreover, all the MPS-filtered sounds were identified at significantly lower rates than were the original sounds (all above 85%, $z > 4$; $p < 0.001$) in the training session, except the tuba whose original sound was correctly identified 64% of the time ($z = 1.02$; $p = 0.15$), most likely due to the pitches being in a relatively high register for this instrument.

Figure 2 presents the *DiagnosticMask* of the five instruments. The clarinet is the instrument with the largest black area (39.3% of the MPS plane) leading to the best correct identification (63.4%), followed by the trombone (29.7% of the MPS plane, 61.6% correct identification) and the cello (15.8% of the MPS plane, 54.7% correct identification). For the tuba and the saxophone, only 5.8% and 7.6% of the MPS plane provide significantly high recognition but with relatively high correct identification rates (56.4% and 46.3%, respectively). More precisely, regions leading to the

Table 1. Confusion matrix in percent response averaged across participants computed from responses to the MPS-filtered sounds. Stimuli are presented in rows and responses in columns. Association rates significantly above chance are shown in bold. Note: ***$p < 0.001$; *$p < 0.05$.

	Trombone	Clarinet	Tuba	Cello	Saxophone
Trombone	**61.6*****	3.7	19.7	5.6	9.3
Clarinet	3.7	**63.4*****	8.5	12.8	11.7
Tuba	**30.3*****	3.4	**56.4*****	4.0	5.9
Cello	4.4	9.2	6.6	**54.7*****	**25.1***
Saxophone	5.9	7.7	5.6	**34.4*****	**46.3*****

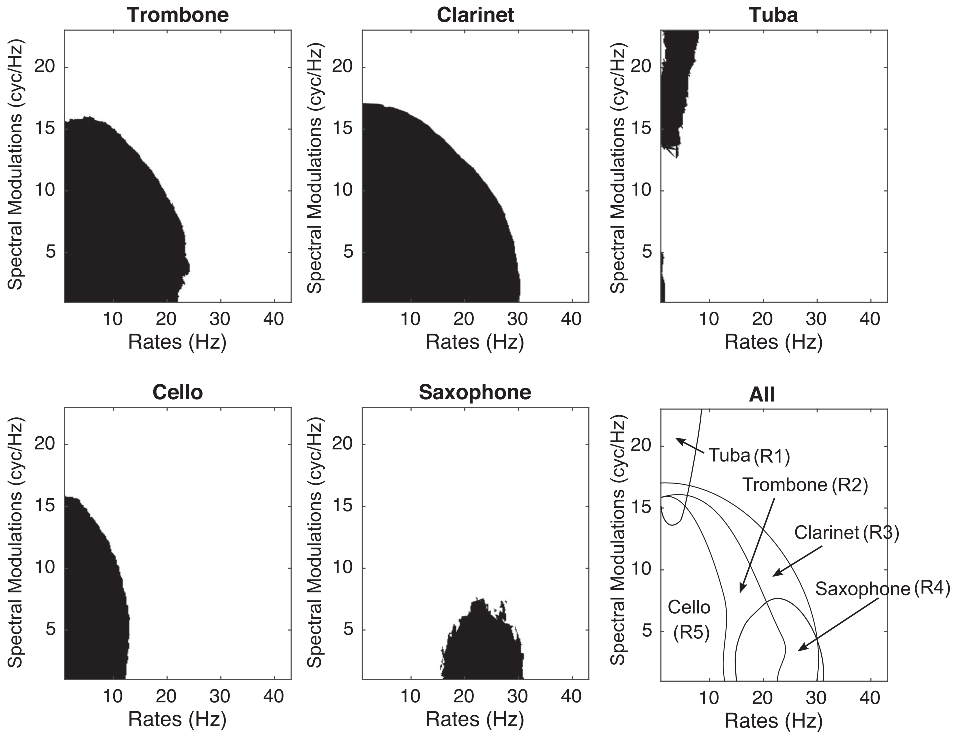

Fig. 2. DiagnosticMasks of the five instruments. The bottom-right plot represents the superposition of the contours of the five *DiagnosticMasks* smoothed in this representation.

recognition of trombone (R2), clarinet (R3), and cello (R5) are centered on low spectral modulations, from 0 to 15 cycles/Hz for the three instruments, and on low rates, from 0 to 20 Hz, 30 Hz, and 10 Hz, respectively. For tuba and saxophone, note that the relevant regions R1 and R4 are centered elsewhere in the MPS plane: from 13 to 23 cycles/Hz and 0 to 8 Hz for tuba and from 0 to 7.5 cycles/Hz and 0 to 30 Hz for saxophone. Note that 25.1% of regions leading to the recognition of tuba overlap with those of the trombone. More strikingly, the saxophone's relevant regions do not overlap with those of the cello. The fact that these two instruments were often confused with one another suggests that these regions (R1 and R4) are specific to the tuba's and saxophone's timbres with respect to the trombone's and cello's timbres, respectively.

6. Discussion and conclusion

In this study, we examined which parts of the MPS plane of a set of sustained instrument sounds are relevant for their recognition. We observed that tuba, trombone, cello, saxophone, and clarinet sounds present different relevant areas that allow them to be identified within this set of instrument samples. In particular, saxophone and tuba have relevant areas of very different shapes when compared to those of the three other instruments. Moreover, concerning the saxophone, its relevant area is nearly perfectly adjacent to that of the cello with which it was often confused. The interest of this result is that it provides new insight into the relevance of spectrotemporal modulations for musical instrument recognition. It is worth noting that this area not only enables the recognition of the saxophone but also stresses which specific acoustic cues of the MPS plane are characteristics of the saxophone's timbre that distinguish it from the cello's timbre. These confusions between the cello and the saxophone might be explained on mechanical grounds by the fact that reed woodwinds such as a saxophone exhibit the same self-oscillation process as the cello, i.e., Helmholtz motion (Ollivier *et al.*, 2004).

This experiment reveals an interesting localization of the perceptually relevant spectrotemporal modulation representation for the recognition of musical instruments. These results can be put into perspective with those of Suied *et al.* (2012) and Isnard *et al.* (2016), which highlighted the point that severely impoverished sounds, from either their spectrotemporal modulation or their spectrotemporal representations, remain recognizable and still convey relevant information for their recognition. Here we complement these studies by precisely defining, and specifying the relevant spectrotemporal modulations for the recognition of a small subset of musical sounds. Other experiments are needed in order to fully understand the reduction of this kind of representation for a larger subset of musical sounds, e.g., impulsive sounds. Moreover, experiments with different subsets of sounds are needed to evaluate the effect of context. For example, one might expect that by removing the cello from the subset, the relevant spectrotemporal modulations of the saxophone would be different as it was most often confused with the cello.

Finally, this experiment validates the relevance of this molecular method to determine perceptually sparse spectrotemporal modulation representations of sounds. Nevertheless, supplementary experiments are also necessary here, in particular, to determine the influence on the recognition of processed sounds of the size of the filters along each dimension, i.e., the Gaussian SDs.

Acknowledgments

The authors are thankful to Grace Wang for help running participants. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada awarded to S.M. (Grant Nos. RGPIN-2015-05208 and RGPAS-478121-15) and to P.D. (Grant No. RGPIN-262808-2012), as well as a Canada Research Chair awarded to S.M.

References and links

Elliott, T. M., Hamilton, L. S., and Theunissen, F. E. (**2013**). “Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones,” J. Acoust. Soc. Am. **133**, 389–404.

Elliott, T. M., and Theunissen, F. E. (**2009**). “The modulation transfer function for speech intelligibility,” PLoS Comput. Biol. **5**, e1000302.

Gosselin, F., and Schyns, P. G. (**2001**). “Bubbles: A technique to reveal the use of information in recognition tasks,” Vision Res. **41**, 2261–2271.

Griffin, D. W., and Lim, J. S. (**1984**). “Signal estimation from modified short-time Fourier transform,” IEEE Trans. Acoust. Speech Signal Process. **32**, 236–243.

Hemery, E., and Aucouturier, J.-J. (**2015**). “One hundred ways to process time, frequency, rate and scale in the central auditory system: A pattern-recognition meta-analysis,” Front. Comput. Neurosci. **9**, 80.

Isnard, V., Taffou, M., Viaud-Delmon, I., and Suied, C. (**2016**). “Auditory sketches: Very sparse representations of sounds are still recognizable,” PloS One **11**, e0150313.

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.

Miller, J. R., and Carterette, E. C. (**1975**). “Perceptual space for musical structures,” J. Acoust. Soc. Am. **58**, 711–720.

Ollivier, S., Dalmont, J. P., and Kergomard, J. (**2004**). “Idealized models of reed woodwinds. Part I: Analogy with the bowed string,” Acta Acust. Acust. **90**(6), 1192–1203.

Patil, K., Pressnitzer, D., Shamma, S., and Elhilali, M. (**2012**). “Music in our ears: The biological bases of musical timbre perception,” PLoS Comput. Biol. **8**, e1002759.

Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice Hall, Englewood Cliffs, NJ).

Shamma, S. (**2001**). “On the role of space and time in auditory processing,” Trends Cogn. Sci. **5**, 340–348.

Slaney, M. (**1994**). “An introduction to auditory model inversion,” Interval Technical Report IRC1994. https://engineering.purdue.edu/%7emalcolm/interval/1994-014/.

Singh, N. C., and Theunissen, F. E. (**2003**). “Modulation spectra of natural sounds and ethological theories of auditory processing,” J. Acoust. Soc. Am. **114**, 3394–3411.

Suied, C., Drémeau, A., Pressnitzer, D., and Daudet, L. (**2012**). “Auditory sketches: Sparse representations of sounds based on perceptual models,” in *From Sounds to Music and Emotions*, edited by M. Aramaki, M. Barthet, R. Kronland-Martinet, and S. Ystad (Springer, Berlin, Heidelberg).

Vienna Symphonic Library (**2015**). http://vsl.co.at/en (Last viewed November 23, 2016).