




Timbral cues for learning to generalize musical instrument identity across pitch register

Stephen McAdams,^{1,a)}  Etienne Thoret,^{2,b)}  Grace Wang,³ and Marcel Montrey⁴ 

¹Schulich School of Music, McGill University, Montreal, Québec H3A 1E3, Canada

²Aix-Marseille University, Centre National de la Recherche Scientifique, Perception Representations Image Sound Music Laboratory, Unité Mixte de Recherche 7061, Laboratoire d'Informatique et Systèmes, Unité Mixte de Recherche 7020, 13009 Marseille, France

³Cognitive Science Program, McGill University, Montreal, Québec H3A 3R1, Canada

⁴Department of Psychology, McGill University, Montreal, Québec H3A 1G1, Canada

ABSTRACT:

Timbre provides an important cue to identify musical instruments. Many timbral attributes covary with other parameters like pitch. This study explores listeners' ability to construct categories of instrumental sound sources from sounds that vary in pitch. Nonmusicians identified 11 instruments from the woodwind, brass, percussion, and plucked and bowed string families. In experiment 1, they were trained to identify instruments playing a pitch of C4, and in experiments 2 and 3, they were trained with a five-tone sequence (F#3–F#4), exposing them to the way timbre varies with pitch. Participants were required to reach a threshold of 75% correct identification in training. In the testing phase, successful listeners heard single tones (experiments 1 and 2) or three-tone sequences from (A3–D#4) (experiment 3) across each instrument's full pitch range to test their ability to generalize identification from the learned sound(s). Identification generalization over pitch varies a great deal across instruments. No significant differences were found between single-pitch and multi-pitch training or testing conditions. Identification rates can be predicted moderately well by spectrograms or modulation spectra. These results suggest that listeners use the most relevant acoustical invariance to identify musical instrument sounds, also using previous experience with the tested instruments. © 2023 Acoustical Society of America. <https://doi.org/10.1121/10.0017100>

(Received 30 September 2022; revised 26 December 2022; accepted 12 January 2023; published online 2 February 2023)

[Editor: James F. Lynch]

Pages: 797–811

I. INTRODUCTION

When listening to music, we can often identify which section of the orchestra is playing, and how our favorite part in this concerto is the cello solo. So how do we perceive and learn to distinguish and identify these instrumental sound sources? In our sonorous world, timbre plays an important role in the perception of music. It is a multidimensional attribute of sound that accounts for many features unaccounted for by other sound attributes such as pitch, loudness, and duration (McAdams, 2019). Timbre, often referred to as “sound color,” has a multitudinous set of perceptual attributes that are often described with terms such as “brightness” or “richness” or “roughness” (Saitis and Weinzierl, 2019). It is one of the primary perceptual vehicles for recognition and identification of a sound source (McAdams, 1993, 2019). However, timbre is known to vary systematically with changes in pitch on a given instrument (Siedenburg *et al.*, 2021). For example, different names are given to the various registers of the clarinet: the dark, low chalumeau; the rich, middle clarion; and the bright, high altissimo registers. The timbral differences potentially complicate instrument identification across the whole pitch range. The question is whether

this poses a problem for instrument identification in practice, and if it does, whether the differences across pitch register can be learned.

Ecological psychologists propose that “knowledge acquisition involves the direct perception of an informational structure composed of systemic relationships; this informational structure is isomorphic to the actual invariant structure of whatever entity we are apprehending” [McCabe (1986), p. 30]. In this sense, systematic relations among different events produced by a sound source across variations in its mechanical properties (tube, string, or bar length, which vary with pitch) should be apparent in invariant properties of the acoustics of those events. As such, learning to identify an instrument at one pitch, should generalize to other pitches if the mechanical properties create an acoustical invariance over pitch, and listeners should be able to judge relations between instruments independently of pitch. Along these lines, Marozeau *et al.* (2003) found that timbre spaces derived from dissimilarity ratings for recorded musical instrument tones are similar at three different pitches (B3, C#4, and Bb4, where C4 is middle C), and that listeners were able to ignore pitch differences within an octave when they were asked to compare only the timbres of the tones. However, in a similar study, Marozeau and de Cheveigné (2007) varied the pitch over a range of 18 semitones (an octave and a half) for synthesized tones with different spectral centroids and asked listeners to rate the dissimilarities of

^{a)}Electronic mail: stephen.mcadams@mcgill.ca

^{b)}Also at: Institute of Language Communication and the Brain, 13100 Marseille, France.

pairs of tones. They found a dimension of pitch in the multi-dimensional space that was orthogonal to the timbre dimension and systematic distortion of the perceived spectral centroid relations due to the pitch changes. [Korsmit et al. \(2021\)](#) extended this study by employing the full pitch range of 11 instruments. They found a dimension related to pitch that was orthogonal to three dimensions related to timbre.

Studies with other approaches have focused on the way timbre-pitch covariation is characterized in our mental categories for musical instruments. [Handel and Erickson \(2001\)](#) examined how far timbre invariance could extend across pitches by investigating how well listeners could determine whether two instrumental notes at different pitches were played on identical or different instruments. They found that different kinds of errors in judgments occurred *above vs at or below* one octave pitch separation. At intervals greater than an octave (17–29 semitones), participants judged the instrument pairs to be from different instruments even when they were the same. They were able to distinguish same and different pairs at intervals of 5, 7, and 12 semitones. In a subsequent study, [Handel and Erickson \(2004\)](#) focused on whether timbre perception can independently affect listeners' ability to recognize one instrument at different pitches. In their second experiment, they investigated listeners' ability to identify an oddball instrument as a function of its pitch placement with respect to two other notes played in sequence by another instrument. An outlying pitch was most often chosen as the oddball, irrespective of its instrument, when the two woodwinds (clarinet and English horn) were paired. The task was performed correctly more often when a woodwind was paired with a brass instrument. The result shows that, despite timbre being the primary perceptual cue for identification, the listener still uses pitch differences secondarily to discriminate between instruments of similar timbre. Therefore, it is difficult to judge source timbre similarity independently of pitch, unless the timbres are strikingly different. However, [Steele and Williams \(2006\)](#) found that musician listeners can recognize sounds as coming from the same instrument at intervals of more than two octaves. Therefore, there do seem to be limits to timbral invariance across pitch that appear to depend on musical training.

The [Steele and Williams \(2006\)](#) result suggests that the timbre-pitch covariation can be learned. [Stilp et al. \(2010\)](#) have demonstrated that passive exposure to highly correlated acoustic properties results in a collapse of the two unitary dimensions (temporal envelope and spectral shape in their case) into a single perceptual dimension. They note that this is an important feature of perceptual learning given that natural sounds are complex and typically change along multiple acoustic dimensions that covary in accord with physical laws governing sound-producing sources. The adaptation to correlated attributes could be a mechanism for efficient coding of sound source properties ([Lewicki, 2002](#)), perhaps including those that lead to categorization and identification.

Various auditory features of sounds contribute to their identification ([McAdams, 1993](#)). Attack transients or the temporal envelope more globally, is one important feature

([Saldanha and Corso, 1964](#)). The of the resonator or sounding object is also crucial ([Giordano and McAdams, 2010](#)) as is, and perhaps more importantly, the action by which a sounding body is excited ([Lemaitre and Heller, 2012](#)).

This paper will contrast two hypotheses: (1) the acoustic invariance hypothesis states that the properties of a given instrument sound should be generalizable across pitches it produces; (2) the correlational learning hypothesis states that exposure to covariation of perceptual properties such as pitch and timbre in identification training should enhance performance, perhaps beyond the learned stimuli if the nature of the covariation can be extrapolated. We pose a number of questions.

- (1) Do we pick up invariant properties in the sound and then use those to categorize other sounds as coming from the same sound source? This model would predict a flat curve (to the extent to which they are completely invariant) of recognition across pitch, independently of the pitch at which an identification training stimulus is positioned.
- (2) Do we need to experience the way an important feature for identification varies with another feature in order to learn their correlation and use that to extrapolate to other instances? This model would predict better performance when correlated variation is provided in training, and the identification performance would be more constrained around a single training tone's pitch than around a wider range of training pitches, with perhaps increased performance at pitches outside the training set.
- (3) Do we need to experience all possible combinations of features (at least appropriately sampled across their ranges) in order to build a mental model of the sound source category? This model would predict bumps in the curves at the training pitches and lower performance beyond.

One aim of this paper, related to the first question, is to examine potential acoustic invariances and their relation to identification performance. From a biological perspective, many studies have used auditory models to assess timbre similarities and timbre perception ([Patil et al., 2012](#); [Thoret et al., 2021](#)). Historically, the modeling of sensory representations of sound has been based on waveforms and spectrograms. Incoming signals arrive at the cochlea and the mechanical waves excite the basilar member from base to apex. The selectivity of the basilar membrane excitation pattern is observed to be non-linear but involves a quasi-logarithmic scale due to the biomechanical properties of the membrane [for more details see [Thoret et al. \(2017\)](#)]. Therefore, the abstraction of the acoustic signal at the cochlear level can be interpreted as a log-frequency spectrogram, also known as the auditory spectrogram ([Chi et al., 2005](#)), although other studies have used a linear frequency scale [e.g., [Elliott et al. \(2013\)](#)]. More recently, studies have investigated the role of higher cortical networks such as the primary auditory cortex. These studies have revealed that

neurons of these areas seem to fire to specific acoustical patterns—spectral and temporal modulations—of an incoming acoustic signal (Shamma, 2001). Some studies have provided evidence for the prominent role of spectrotemporal modulations for timbre perception (Patil *et al.*, 2012; Elliott *et al.*, 2013; Thoret *et al.*, 2021) and sound source classification (Thoret *et al.*, 2016, 2017).

Incorporating spectrotemporal modulation analysis into timbre perception models may lead to greater understanding of both the mental representation of musical sounds and their storage in memory. Spectrotemporal modulations roughly correspond to the 2D Fourier transform of the spectrogram and are called the modulation power spectrum (MPS) (Singh and Theunissen, 2003). This representation reveals the regularities and periodicities of the spectrogram in the temporal and spectral dimensions. This neuromimetic mathematical formulation provides an efficient way to model perceptual dissimilarity judgments between instruments (Patil *et al.*, 2012; Thoret *et al.*, 2021) and musical instrument categorization (Thoret *et al.*, 2016, 2017), as well as providing a tool for automatic classification of different timbres (Patil *et al.*, 2012; Hemery and Aucouturier, 2015). Hence, the MPS provides a relevant tool to investigate which invariant acoustical structures might be relevant to memorize and identify musical instruments.

This paper investigates the relationship between timbre and pitch and addresses the question of whether we learn to identify instruments by the aspects of timbre that remain consistent across pitch (acoustically invariant properties) or by learning how the timbre varies with pitch (learned correlations), thereby extrapolating the timbre variation to identify instruments on untrained pitches. Furthermore, through an acoustic analysis, the paper also investigates whether the underlying generalization mechanism can be predicted from complete, unified acoustic representations, which would support the role of acoustic invariance across registers.

To test both aforementioned hypotheses, we focus on listeners' ability to generalize learning to identify musical instruments in a constrained pitch register to other registers under different training and testing stimulus conditions. We train participants to identify the selected instruments from either a single pitch (experiment 1) or a sequence of five pitches spanning an octave (experiments 2 and 3). We then test their ability to extrapolate their knowledge to identify the instrument from single tones (experiments 1 and 2) or three-pitch sequences spanning six semitones over the instrument's whole pitch range (experiment 3). This investigation is thus divided into three parts. Experiment 1 examines the ability to extrapolate identification from learning based on a single pitch in building mental models of instruments. Experiment 2 examines further the hypothesis by providing more stimulus samples to indicate how timbre varies with pitch in training to build an improved mental model of instruments, which is tested under the same conditions as experiment 1. Furthermore, we also hypothesize that by increasing the number of tones in each stimulus of the testing phase, we might improve the participants' success in applying their training and identifying the sounds correctly.

Therefore, in experiment 3, the provided information on pitch-timbre covariation remained the same as in experiment 2, but in the testing phase, the listeners are asked to identify instruments with a three-tone sequence (an augmented triad) in untrained registers. In order to better understand the timbral properties that underpin timbre-pitch covariation perception, we further analyze the results of the experiments in relation to the information in spectrograms and modulation power spectra, which may reveal relevant invariant acoustical structures involved in these recognition tasks.

II. METHODS

Methods common to all three experiments will be presented first, followed by specifics of each experiment.

A. General methods

1. Participants

All participants were nonmusicians, defined as a person having one year or less of musical training in elementary school and not having been involved in musical practice or study since then. All participants gave informed written consent and received compensation for their participation in the study. This study was certified for ethical compliance by the McGill University Research Ethics Board II.

2. Stimuli

The experimental sound stimuli were drawn from two collections: the *Vienna Symphonic Library* (2015) and the McGill University Master samples (Opolko and Wapnick, 2006). The sounds were produced by instruments playing at a mezzo forte level at different durations and were recorded using a sample rate of 44.1 kHz. To unify them, a 50-ms raised cosine fade-out amplitude envelope was used to create a constant duration of 500 ms. The initial attack portion was not modified as it contributes significantly to instrument identification (Saldanha and Corso, 1964). The levels of the sounds in the Sennheiser HD280 Pro headphones (Sennheiser Electronic GmbH, Wedemark, Germany) varied between 75.8 and 83.7 dB SPL as measured with a Brüel & Kjær type 2205 sound-level meter (A-weighting) and a Brüel and Kjær type 4153 artificial ear to which the headphones were coupled (Brüel and Kjær, Nærum, Denmark).

The traditional orchestral instruments used for the experiment were selected such that their playing ranges included the octave around middle C (C4, 261.6 Hz fundamental), which is the center pitch used to train participants to identify the instruments. We collected stimulus samples at intervals of three semitones starting from C4 to the lower and upper ends of each instrument's range, spanning from C1 (30.9 Hz) to F#7 (2960.0 Hz) for the instrument with the widest range, the harp.

Table I lists the instruments with their instrument families and playing range. Figure 1 further displays the playing ranges of each instrument in relation to each other: the red vertical bar highlights C4, the training pitch in experiment 1. The yellow

TABLE I. List of instruments sampled.

Family	Instrument	Pitch range
String	Cello (bowed)	C2–D#6
	Harp (plucked)	C1–F#7
	Acoustic guitar (plucked)	F#2–C6
Brass	Tenor trombone	F#2–F#4
	Tuba	C2–F#4
Woodwind	English horn	F#3–A5
	Clarinet	D#3–F#6
	Tenor saxophone	A2–D#5
Pitched percussion	Marimba	C2–C7
	Tubular bell	F#3–D#5
	Vibraphone	F#3–D#6

region highlights the octave that encircles C4, including pitches F#3, A3, C4, D#4, F#4 played in succession as an arpeggio for the training stimuli in experiments 2 and 3.

3. Procedure

After obtaining signed informed consent from the participants, they were seated in an audiometric booth and fitted with headphones. Prior to the start of the experimental study, participants were screened with a standardized pure-tone audiometric test separately in the left and right ears at octave-spaced frequencies from 125 Hz to 8000 Hz (International Organization for Standardization, 2004; Martin and Champlin, 2000). The participants were required to have threshold at or below 20 dB HL (relative to a standardized hearing threshold) to proceed to the main experiment.

The main experiment was divided into three phases: familiarization, training, and testing. During the familiarization stage, all the instrument names appeared on the screen

and a click on a name would produce the corresponding training stimulus for that instrument in each experiment. The participants took as long as needed for the familiarization stage. They were instructed to proceed to the training phase once they felt comfortable and familiar with the association between the instrument names and their corresponding C4 sounds.

In each trial of the training phase, the participant heard a training stimulus (C4 for experiment 1, arpeggio centered on C4 for experiments 2 and 3) and had to select the corresponding instrument name from the list of 11 instruments. They had a one-time replay button to hear the sound again. Feedback was provided. The name flashed green if the response was correct. If it was incorrect, the name flashed red and the correct name flashed green. The training phase was programmed in blocks of 11 trials corresponding to the 11 instruments, the order of which was randomized within each block. When the participant reached at least 75% accuracy (a score of nine or more correct out of 11 in each block) for four consecutive blocks, they moved from the training phase to the testing phase. However, if they did not reach this threshold within 20 blocks, the experimental program would terminate, and the participant would not move on to the testing phase.

Successful participants continued to the testing phase, where test stimuli across the entire range of each instrument were presented in randomized order (151 stimuli for experiment 1, 156 for experiment 2, 134 for experiment 3). Their task, similar to the previous phase, was to identify the sampled instrument for each test stimulus. A one-time replay button was provided, but no feedback was given in this phase. The trials were divided into three blocks, and participants had the option of taking breaks between blocks. Once they finished identifying the test stimuli, the experiment terminated.

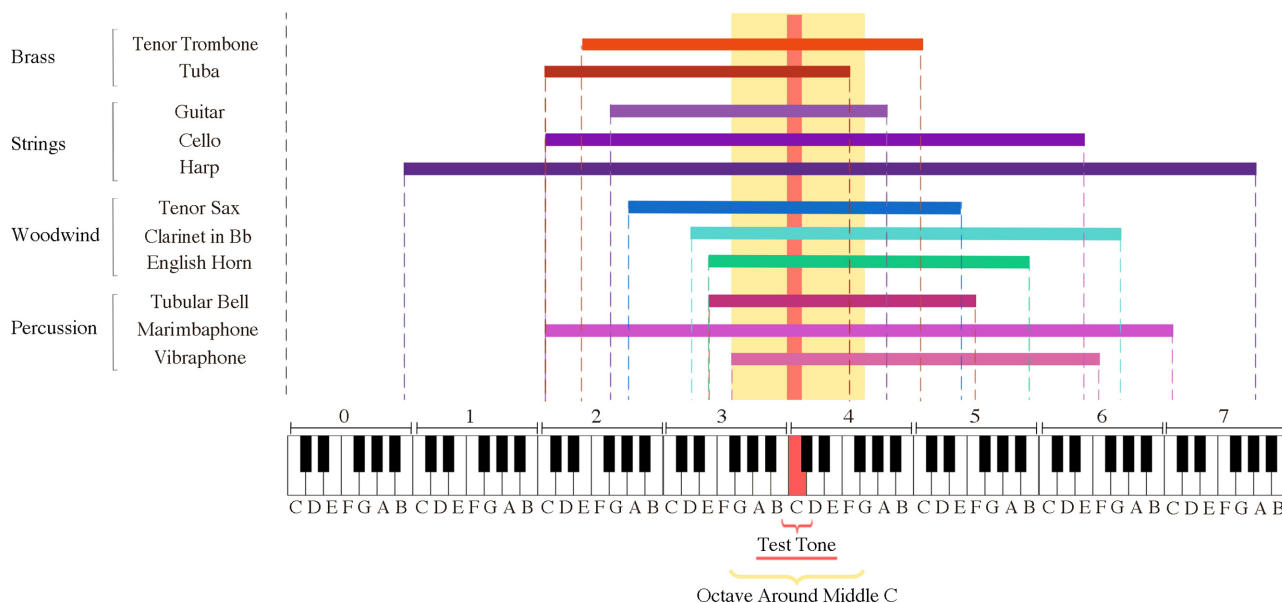


FIG. 1. (Color online) Instrument playing ranges and the range of stimuli used.

At the end of the experiment, whether the participant moved on to the testing phase or not, they were asked to complete a questionnaire regarding their general demographics, music listening habits, and musical experience.

4. Apparatus

The participants completed both the screening and the main experiment seated in an IAC model 120act-3 double-walled audiometric booth (IAC Acoustics, Bronx, NY). Sounds stored on a Mac Pro 5 computer running OS 10.6.8 (Apple Computer, Inc., Cupertino, CA) were amplified through a Grace Design m904 monitor (Grace Digital Audio, San Diego, CA) and presented over Sennheiser HD280 Pro headphones (Sennheiser Electronic GmbH, Wedemark, Germany). The experimental session was programmed in the PsiExp computer environment (Smith, 1995). The levels of sounds were measured with a Brüel & Kjær type 2205 sound-level meter (A-weighting) with a Brüel & Kjær type 4153 artificial ear to which the headphones were coupled (Brüel & Kjær, Nærum, Denmark).

B. Experiment 1

1. Participants

Forty-one nonmusician participants were recruited through advertisement on Facebook and McGill Classified Marketplace. Participants' ages ranged from 20 to 50 years ($M = 20.7$, 26 females). One participant failed the audiometric screening test and 15 failed the training phase. Twenty-five participants completed the testing phase, indicating that they had learned the training stimuli and were therefore kept for the subsequent analyses of how listeners extrapolate from such learning.

2. Stimuli and procedure

The familiarization and training stimuli were single C4 tones produced by each instrument. The test stimuli were single tones drawn from the 156 stimulus tones across instruments and pitch registers (Fig. 1). Due to a programming error, five tones in the higher register (C5, D#5, F#5, A5, C6) of the guitar were not presented. There were thus 151 test stimuli. Comparative analyses across experiments will exclude these five stimuli.

C. Experiment 2

1. Participants

Twenty-seven nonmusician participants were recruited through advertisement on Facebook and McGill Classified Marketplace. None had participated in experiment 1. All listeners verbally confirmed that they had not previously participated in any other instrument identification study. Their ages ranged from 18 to 49 years ($M = 22.3$, 16 female). One participant failed the audiometric screening, and one failed the training phase. Twenty-five participants completed the testing phase.

2. Stimuli and procedure

The familiarization and training stimuli in experiment 2 consisted of an ascending arpeggio (F#3, A3, C4, Eb4, F#4) to provide information about covariation of pitch and timbre. The 156 test stimuli were single tones, identical to those of experiment 1, with the addition of the five missing guitar tones.

D. Experiment 3

1. Participants

Twenty-nine nonmusician participants were recruited through advertisement on Facebook and McGill Classified Marketplace. None had participated in experiments 1 and 2. All listeners verbally confirmed that they had not previously participated in any other instrument identification study. Their ages ranged from 18 to 28 years ($M = 21.8$, 18 female). All participants passed the audiometric screening, but four failed the training phase. Twenty-five participants completed the testing phase.

2. Stimuli and procedure

The familiarization and training stimuli in experiment 3 consisted of an ascending five-note arpeggio as in experiment 2. The test stimulus was an ascending three-note arpeggiated diminished triad; e.g., A3, C4, and D#4. Data points refer to the central pitch of the triad. Given that three-note arpeggios were presented instead of single notes, there were only 134 trials in this experiment, because the lowest and highest notes of each instrument could not be used.

III. RESULTS

The average number of training blocks to reach the threshold of at least 75% correct identification for four consecutive blocks was 10.6 blocks in experiment 1, 8.2 blocks in experiment 2, and 8.4 blocks in experiment 3. The difference between experiment 1 and experiments 2 and 3 combined was only marginally significant when corrected for unequal variances, $t(32.38) = 1.76$, $p = 0.087$. Furthermore, 15 participants failed the training phase in experiment 1 by not reaching threshold performance within 20 blocks, but there was only 1 failure in experiment 2 and 4 failures in experiment 3. A test of equal proportions between the failures in experiment 1 and the combined failures of experiments 2 and 3 was significant, $X^2(1) = 9.60$, $p = 0.002$. This result suggests that it is easier to learn the instrument identification task with more information about timbre-pitch covariation.

To investigate and compare the participants' results across pitch, we looked at identification performance for each instrument in all three experiments. Figures 2–5 represent the correct identification rate across pitch register for each instrument in the string, woodwind, brass, and percussion families, respectively, for all the experiments: Experiment 1 in solid blue, with single-note training and test, experiment 2 in long-dash red, with octave arpeggio

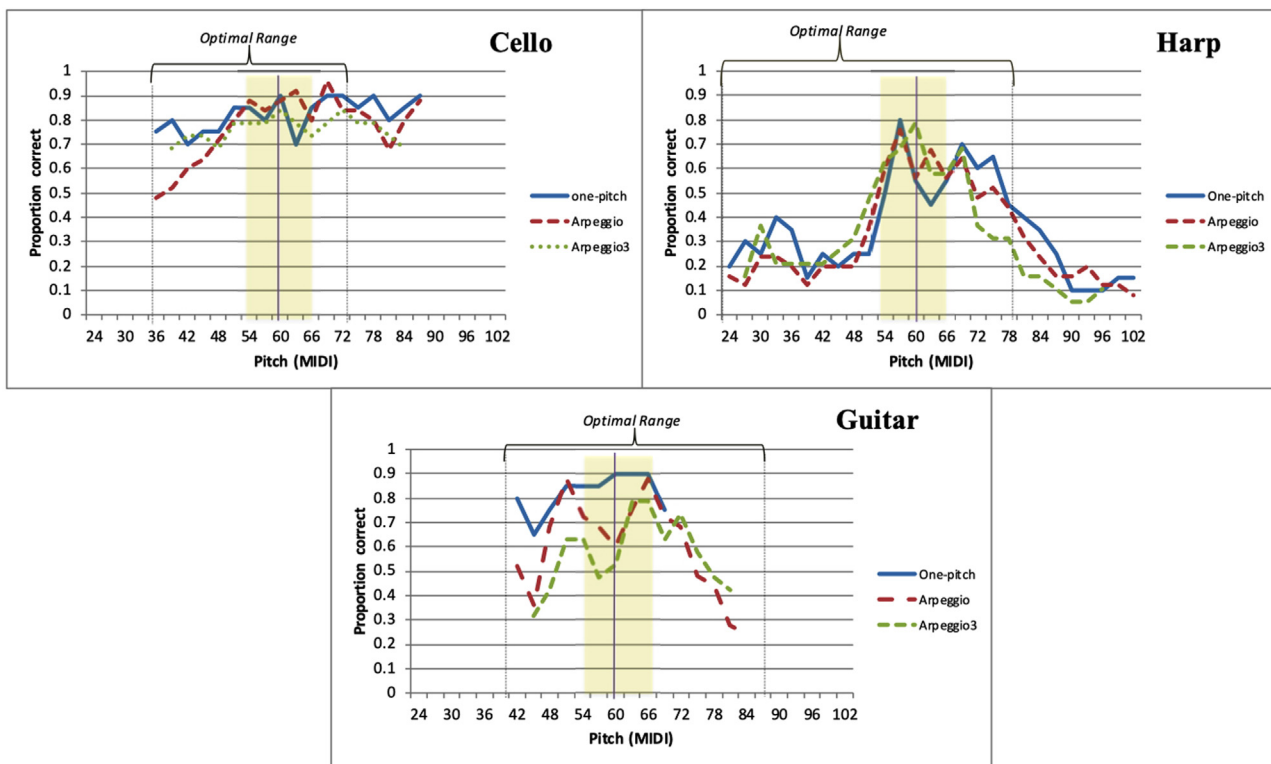


FIG. 2. (Color online) Proportion correct identification as a function of pitch for the string instruments.

training and single-note test, and experiment 3 in short-dash green, with octave arpeggio training and triad arpeggio test. Pitch is coded as MIDI number on the x axis with each integer representing a semitone. MIDI 60 is middle C or C4,

which is represented on the graph in the vertical line. The range of within-octave pitches, used in the training of experiments 2 and 3, is indicated in pale yellow. The optimal range used in orchestration is indicated within the graph

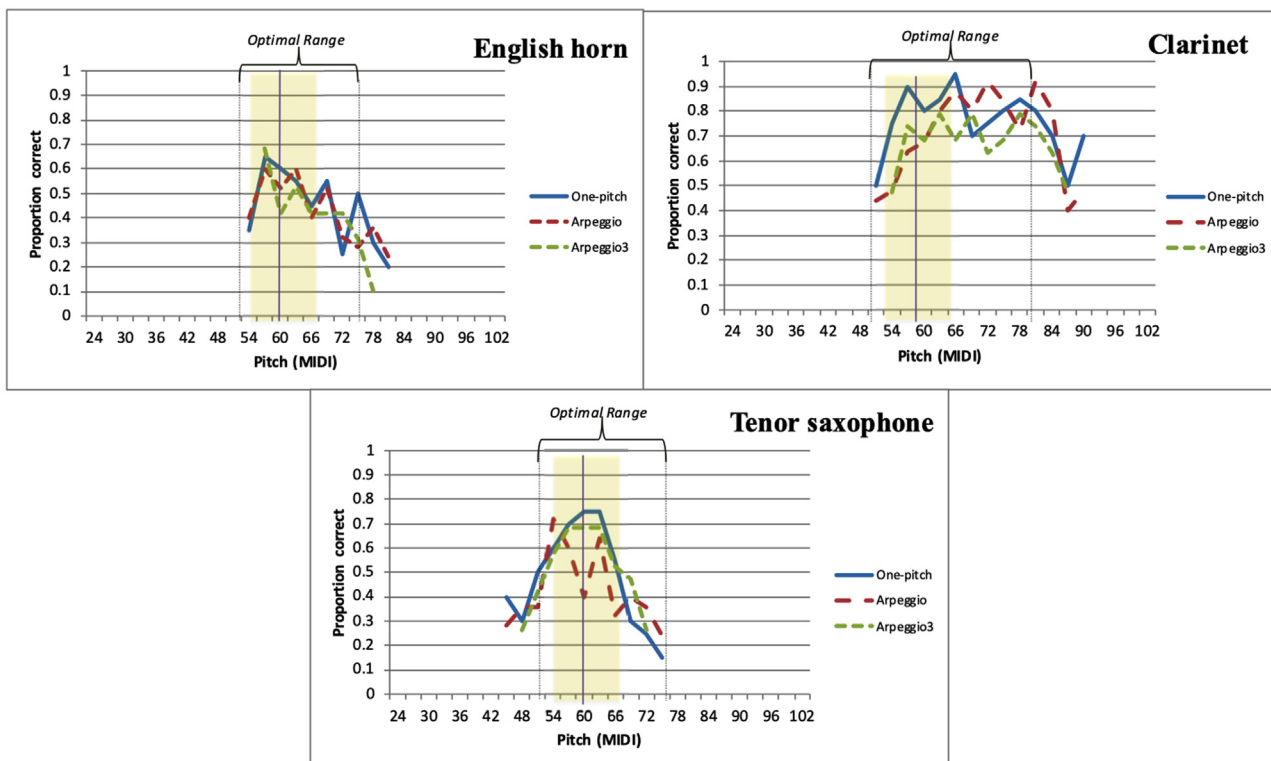


FIG. 3. (Color online) Proportion correct identification as a function of pitch for the woodwind instruments.

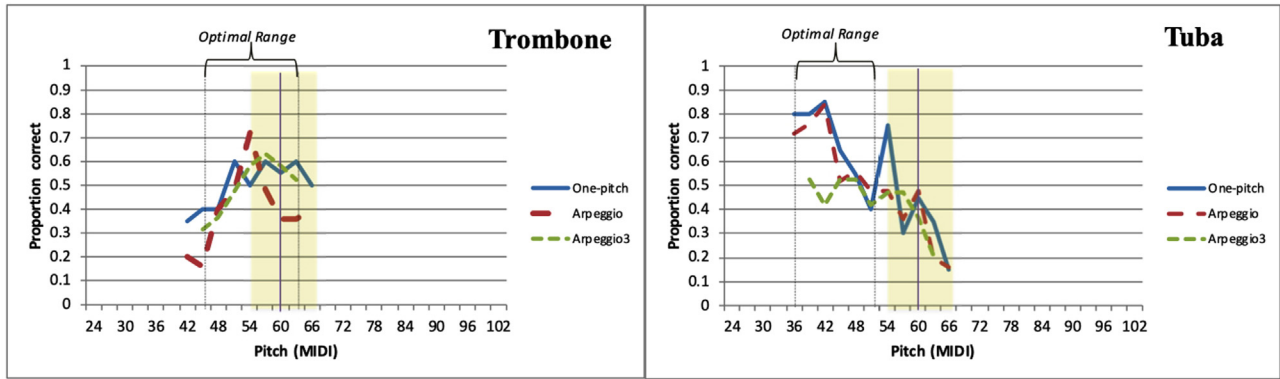


FIG. 4. (Color online) Proportion correct identification as a function of pitch for the brass instruments.

of the string, woodwind, and brass instruments (Adler, 2002). The optimal range is not indicated for percussion instruments as this concept is not meaningful in this case. There is indeed a separate mechanical system that produces each pitch which technically allows musicians to easily produce pitches over the whole range without any difficulties whereas it might be difficult for other instrumentalists for whom playability varies with the pitch to be produced.

The graphs present different patterns as a function of the instrument. Instruments that present a peak at the training pitches are harp and saxophone. Clarinet and guitar have broader regions beyond the training pitches, but performance falls off at more extreme pitches. Cello and tubular bells have relatively flat curves with no significant effect of

pitch, suggesting that some timbral invariant was extracted from the training set that allowed the other pitches to be identified easily. English horn and tuba have descending graphs. The English horn is better recognized in the training region, but the tuba is better recognized at pitches below the training region in its optimal register. The tuba curve extends over a surprisingly wide range of performance with especially low identification in the highest register of tuba, which is incidentally closest to the training tones.

The analysis of correct-response data were performed with generalized linear mixed effects modeling with fixed effects of pitch, experiment, and their interaction. Because pitch often has a quadratic relationship with measures of interest [e.g., McAdams *et al.* (2017)], we included both

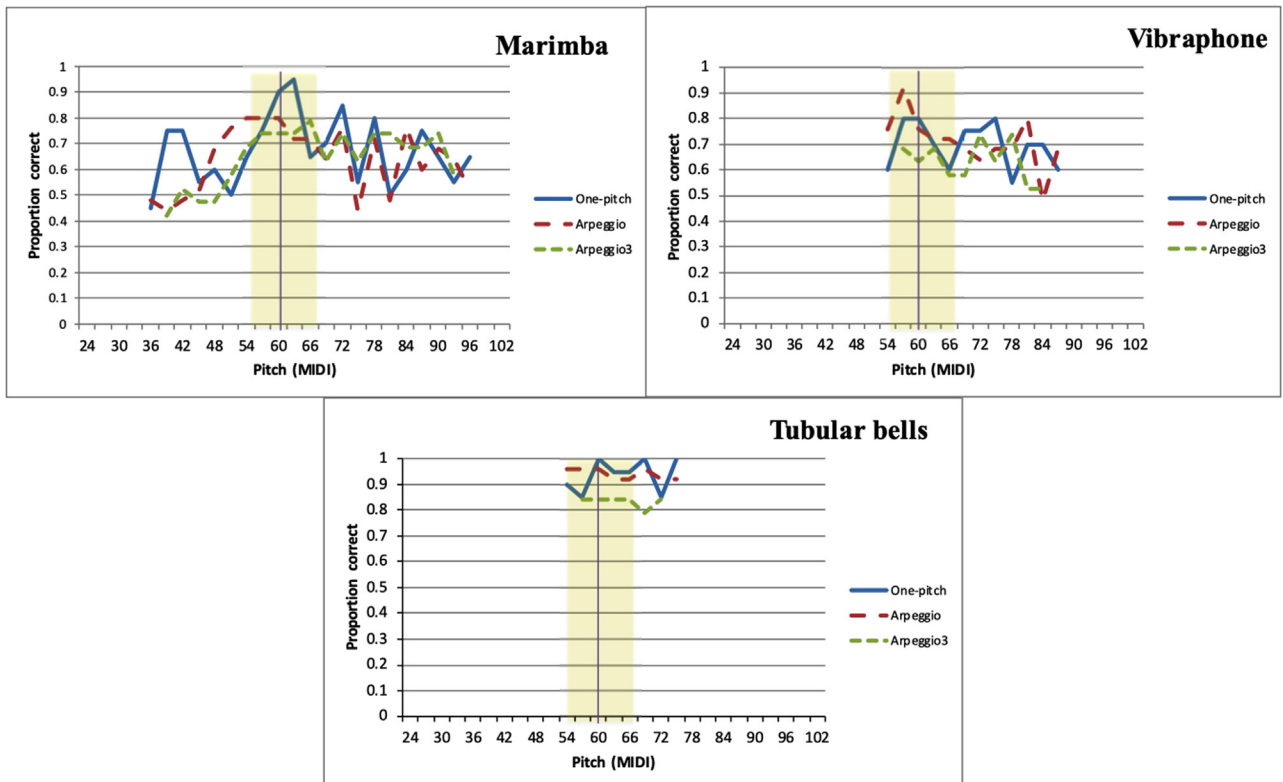


FIG. 5. (Color online) Proportion correct identification as a function of pitch for the percussion instruments.

linear and quadratic terms for this factor, minimizing the correlation between them using orthogonal polynomials. To fit an appropriate random effects structure, we followed the approach advocated by Barr *et al.* (2013), which involves finding the maximal random effects structure justified by the data (i.e., that does not result in a singular fit). This approach guards against the high type I error rates commonly found in intercept-only mixed effects models (Schielzeth and Forstmeier, 2009). We began by fitting a random intercept for participant and then attempted to add random slopes for both the quadratic and linear effects of pitch. No random effects for experiment were included because this was a between-subjects factor. If a model was singular, we first dropped the random slope for quadratic pitch, leaving in the linear random slope. If this wasn't enough, we then dropped the correlations between random slopes and random intercepts. In one case (Tuba), random slopes had to be dropped entirely. After fitting the maximal random effects structure, we then tested fixed effects. If the experiment by linear/quadratic pitch interaction was not significant, it was dropped from the model. Significant fixed effects are indicated in Table II.

The pitch factor was significant for most instruments except cello and tubular bells, indicating a lack of generalization from the training set to more distant pitches for nine of the instruments. Finally, the pitch × experiment interaction was only significant for cello, harp, and tuba. The experiment factor, which varies the amount of information provided on pitch-timbre covariation, was only marginally significant for trombone and marimba.

For clarinet, saxophone, and marimba, the identification curve is concave with respect to pitch. Identification is higher in a certain restricted pitch register for these instruments. For saxophone and marimba, there is a clear bump in the training region and performance falls rapidly at higher

and lower pitches for saxophone and more gradually for marimba, suggesting greater difficulty in generalizing identification across pitch for saxophone than for marimba. However, the clarinet's identification region is quite broad, indicating that some timbral characteristics generalize beyond the training region, but not across its entire range.

The trombone graph is increasing concave with a narrow bump that peaks just below the training region in the middle of its optimal range. Instruments with decreasing concave curves include harp and guitar. The harp's bump falls in the training range, and it is better recognized at lower than at higher pitches. The guitar has a wider bump extending below the training range and is also better identified at lower pitches.

Cello and tubular bells have nearly perfect identification at some pitches, followed by marimba, clarinet, tuba, and vibraphone. Plucked strings (guitar and harp) are identified slightly less well in their best pitch range, followed by saxophone and trombone. The most poorly identified instrument overall is the English horn.

In subsequent analyses, given that there were no significant differences between the three experiments, we use the averaged data across all experiments to have a more general representation of identification performance. To examine the degree of generalization of identification across instruments, we determined the range of pitches identified between the maximum proportion correct and 75% of that maximum (allowing for some fluctuations around this point in marimba and vibraphone). This range and its expression as a percentage of the pitch range for each instrument in these experiments are shown in Table III. Cello, guitar, clarinet, and the percussion instruments have the greatest pitch range of above-threshold performance. Cello, clarinet, and percussion stand out as having the greatest degree of generalization across pitch in relation to each instrument's tessitura.

To further investigate trends in participants' performance in identifying the instruments, we analyzed identification confusions for each instrument across its pitch range on the average across experiments. We found that instruments are

TABLE II. Significant and marginal fixed effects in the generalized linear mixed models for each instrument.

Instrument	Fixed effect	X^{2a}	p (effect)	p (linear)	p (quad.)
Cello	Pitch × experiment	11.75	0.0193		
Harp	Pitch	53.63	<0.0001	0.0037	<0.0001
	Pitch × experiment	14.84	0.0050		
Guitar	Pitch	18.30	0.0001	0.0011	0.0340
English horn	Pitch	19.30	<0.0001	<0.0001	n.s.
Clarinet	Pitch	34.18	<0.0001	n.s.	<0.0001
Saxophone	Pitch	44.54	<0.0001	n.s.	<0.0001
Trombone	Pitch	17.68	0.0001	0.0257	0.0002
	<i>Experiment</i>	<i>5.37</i>	<i>0.0680</i>		
Tuba	Pitch	19.22	<0.0001	<0.0001	n.s.
	Pitch × experiment	14.72	0.0053		
Marimba	Pitch	13.38	0.0012	n.s.	0.0003
	<i>Experiment</i>	<i>5.55</i>	<i>0.0623</i>		
Vibraphone	Pitch	6.45	0.0400	0.0149	n.s.
Tubular bells	No significant effects	—	—		

^aWald X^2 is an omnibus test equivalent to ANOVA. Marginally significant effects are indicated in italics. Significant p -values for linear and quadratic terms for pitch are shown.

TABLE III. Pitch range of identification performance above 75% of maximum performance (Peak Pitch Range) and extent of that range relative to each instrument's tessitura (Relative Range).

Instrument	Peak pitch range	Relative range
Cello	A2–Eb6 (42 ST ^a)	82%
Harp	F#3–A4 (15 ST)	19%
Guitar	Eb3–C5 (21 ST)	50%
English horn	A3–Eb4 (6 ST)	22%
Clarinet	A3–C6 (27 ST)	69%
Saxophone	F#3–Eb4 (9 ST)	30%
Trombone	Eb3–C4 (9 ST)	38%
Tuba	C2–F#3 (18 ST)	60%
Marimba	Eb3–F#6 (39 ST)	65%
Vibraphone	F#3–A5 (27 ST)	82%
Tubular bells	F#3–Eb5 (21 ST)	100%

^aST = semitone.

usually confused within instrument family and within the type of excitation: sustained or impulsive. In the confusion graphs (Figs. 6 and 7), we include the plot of the correct identification of the instrument, which is always represented in solid blue, with other instruments that are often confused with it. Only instruments that were misidentified at least 10% of the time are shown. We also include the optimal ranges for each instrument from Adler (2002) at the top of the graphs.

Although there can be confusion between instruments within the subsections of the sustained (blown, bowed) and impulsive (struck, plucked) excitation groups, there is no confusion across the two groups, signifying that type of excitation provides unequivocal information about the sound events. The sustained sounds shown in Fig. 6 consist of cello, as the only bowed string, and all of the wind instruments. Some notable patterns include the fact that the listeners' identifications of instruments are better over the pitch range that orchestral instruments tend to occupy. Also, within excitation types, there are more confusions in extreme registers depending on the instrument, i.e., correct identifications decrease and incorrect identifications of other instruments increase.

Cello has a high correct identification rate, so the curve remained flat with rarely any confusion with other instruments. Within the woodwind family, tones in the middle to high registers are often confused with other instruments in the same family. For example, the English horn has an interesting pattern of increasing confusion with clarinet in higher register, and woodwind tones in the lower ranges are often mistaken for brass instruments. The brass instruments are also mainly confused within the family; trombone and tuba are often confused with each other in the appropriate registers, especially in the lower notes.

The impulsive sounds, shown in Fig. 7, consist of the plucked strings harp and guitar and the percussion instruments. The two plucked strings are often mistaken for each other throughout the registers. The harp in particular presents an interesting case, where it is more often identified as a guitar way below the guitar's range and as a mallet percussion at the top of its range. Vibraphone and marimba show a similar pattern of confusions to that of trombone and tuba, especially within the lower and higher ranges.



FIG. 6. (Color online) Confusion rates among instruments with sustained sounds. Proportion correct identification as a function of pitch, with the correct instrument represented with the solid line. Optimal ranges are drawn from the orchestration treatise of Adler (2002).

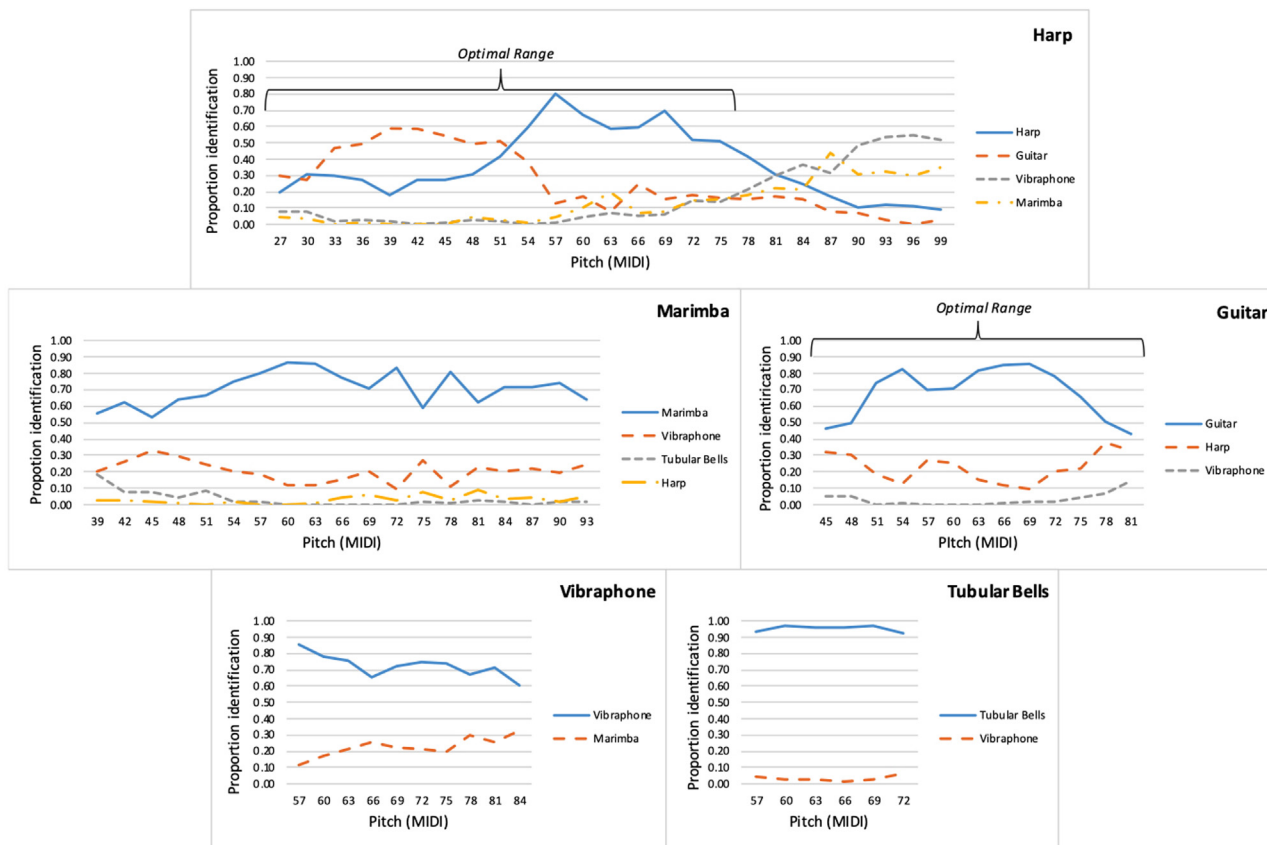


FIG. 7. (Color online) Confusion rates among instruments with impulsive sounds. Proportion correct identification as a function of pitch, with the correct instrument represented with the solid line.

IV. ACOUSTIC ANALYSIS

The previous results revealed that generalized learning of instrument identification with timbre-pitch variation is different across instruments. For example, whereas the identification rates appear to be relatively constant for the tubular bells, they clearly varied more for the tuba across pitches. In order to better understand these differences among the instruments and why some of them are better identified at different pitches, we conducted an analysis that tested whether acoustic representations in the form of spectrograms or modulation power spectra can explain the identification results.

Here, we aim to investigate the hypothesis that for a given instrument the variation in identification performance is correlated with the distance between such acoustic representations at the different pitches. Similarity in perceptually available features across pitches (i.e., some kind of acoustic invariance) would make generalization easier, whereas greater dissimilarity would hinder generalization. Timbre perception has historically been accounted for by acoustic representations that reveal the sounds spectrotemporal features and in particular spectrotemporal modulations (Patil *et al.*, 2012; Elliott *et al.*, 2013; Thoret *et al.*, 2016, 2017; Thoret *et al.*, 2021). We first describe the two different representations used here: spectrograms and modulation power spectra.

A. The spectrogram

Spectrograms represent the evolution of the sound's spectral content over time. They are mainly defined by their time/frequency resolution, i.e., the time window and the overlap between time windows at which the signal is framed to compute the spectra. The other aspect of spectrograms that is crucial is whether the frequency scale is linear or logarithmic. Although logarithmic scales are closer to the tonotopic mapping of frequency on the cochlear basilar membrane, linear scales remain largely used in the timbre literature because they are thought to provide equally accurate results. Here, we chose to compute both representations and to test their relevance in a methodologically agnostic fashion.

B. The modulation power spectrum

The MPS is a two-dimensional Fourier transform of the spectrogram (Elliott and Theunissen, 2009; Singh and Theunissen, 2003). The spectrotemporal modulation characterizes the spectral and temporal regularities of a spectrogram. It must be noted that two different kinds of MPS can be computed according to the choice of a linear or a logarithmic frequency scale in the spectrogram. In the following, both cases will be considered. Figure 8 illustrates the different steps in the computation of the MPS in the case of a linear frequency spectrogram. The short-time Fourier

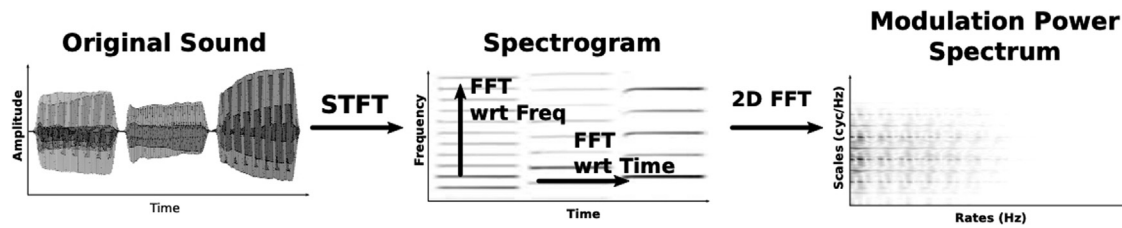


FIG. 8. Acoustic analysis process for the MPS in the case of a linear frequency scale. The STFT is first computed from the original sound excerpt to obtain the spectrogram, and then a two-dimensional fast Fourier transform (FFT) is applied to compute the MPS.

transform (STFT) is first computed. The MPS is then computed by applying two successive Fourier transforms (2D FFT) along the spectrogram’s temporal and frequency axes. The MPS is composed of two different dimensions: the temporal modulations (Hz)—referred to as “rate”—and the spectral modulations (cycle/Hz)—referred to as “scale” [for more detail, see Elliott and Theunissen (2009) and Thoret *et al.* (2017)]. Note that in the case of a logarithmic frequency spectrogram, the process remains the same with the difference that the spectrogram is computed with a constant-Q bandpass filterbank. The spectral modulations are then expressed in cycles/octave.

C. Distances between acoustic representations

We examine whether distances between acoustic representations can explain lower identification rates for some instruments. For each instrument, we computed the time-frequency representations for linear (TF_{lin}) and logarithmic (TF_{log}) frequency scales and then for each TF we computed the corresponding MPS (MPS_{lin} and MPS_{log}). We therefore tested two different factors: (1) the type of representation—time-frequency vs modulation power spectrum and (2) the type of frequency scale—linear vs logarithmic. Then, for each pair of pitches, we computed the Pearson correlation as a measure of distance between pairs of vectorized representations of sounds for each of the four acoustic representations.

For each representation and instrument, we compute a vector of distances of length $N - 1$, with N being the number of pitches for a given instrument. Each vector was composed of the distance between a reference note and all the other notes. The experimental reference note was middle C (MIDI 60), the pitch on which the participants were trained in experiment 1 or the center pitch of the training stimuli in experiments 2 and 3. These values, expressed as z-scores, are presented in Fig. 9 along with the z-transformed mean identification rate across all three experiments. What interests us is the similarity of shape, which is estimated as the mean squared error (MSE) between the two (Fig. 10). To get a global sense of the fit for the two main fixed effects, box plots of the variation of the four acoustic representations across each instrument and of the 11 instruments across each representation are shown in Fig. 11.

As seen in Fig. 9, the correspondence between the acoustic predictor (solid lines) and behavioral measure

(dashed dotted lines) is variable across instruments and input representations. The fit is quite good for some and not so good for others as measured by the MSE. The interaction between instrument and representation is very complex (Fig. 10). The instruments best predicted by MPS_{log} are marimba, trombone, and tubular bells. Those for TF_{log} are English horn, saxophone, trombone, and vibraphone. For MPS_{lin} , cello, clarinet, guitar, and harp are best, whereas for TF_{lin} , it is guitar, tuba, and vibraphone. The four representations provide roughly similar measures for some instruments (guitar, harp, marimba, and vibraphone, which are all impulsive sounds), but are quite different for others (notably English horn, saxophone, trombone, and tubular bells), making it difficult to select a representation that “best” accounts for the behavioral data (Fig. 11). As shown in the lower panel of Fig. 11, globally across instruments, none of the input representations stands out over the others in terms of predictive power. That being said, the acoustic representations do capture moderately well the variation in behavioral ratings (average MSE for $MPS_{lin} = 1.73\%$, $MPS_{log} = 1.48\%$, $TF_{lin} = 1.79\%$, $TF_{log} = 1.55\%$).

V. DISCUSSION

We studied how nonmusician listeners learn to generalize identification of a set of Western orchestral musical instruments from exposure to limited range of pitches across the full range of pitches these instruments can produce. We first provided a brief training session with feedback on a single pitch (C4—experiment 1) or on a pitch sequence covering one octave centered on that pitch (F#3-F#4—experiments 2 and 3). Many more listeners failed to reach criterion performance within the allotted number of trial blocks with single-pitch stimuli than with pitch sequences in which the timbral covariation with pitch was present. Listeners who successfully completed the training part also needed more blocks to reach criterion performance in experiment 1, than in experiments 2 and 3. Together these results suggest that exposure to timbre-pitch covariation enhanced the initial identification learning.

To test whether their learning could be generalized beyond the training pitch or pitch range, listeners attaining the criterion in the training session were subsequently asked to identify the instrument with either a single pitch (experiments 1 and 2) or an augmented triad sequence, centered on

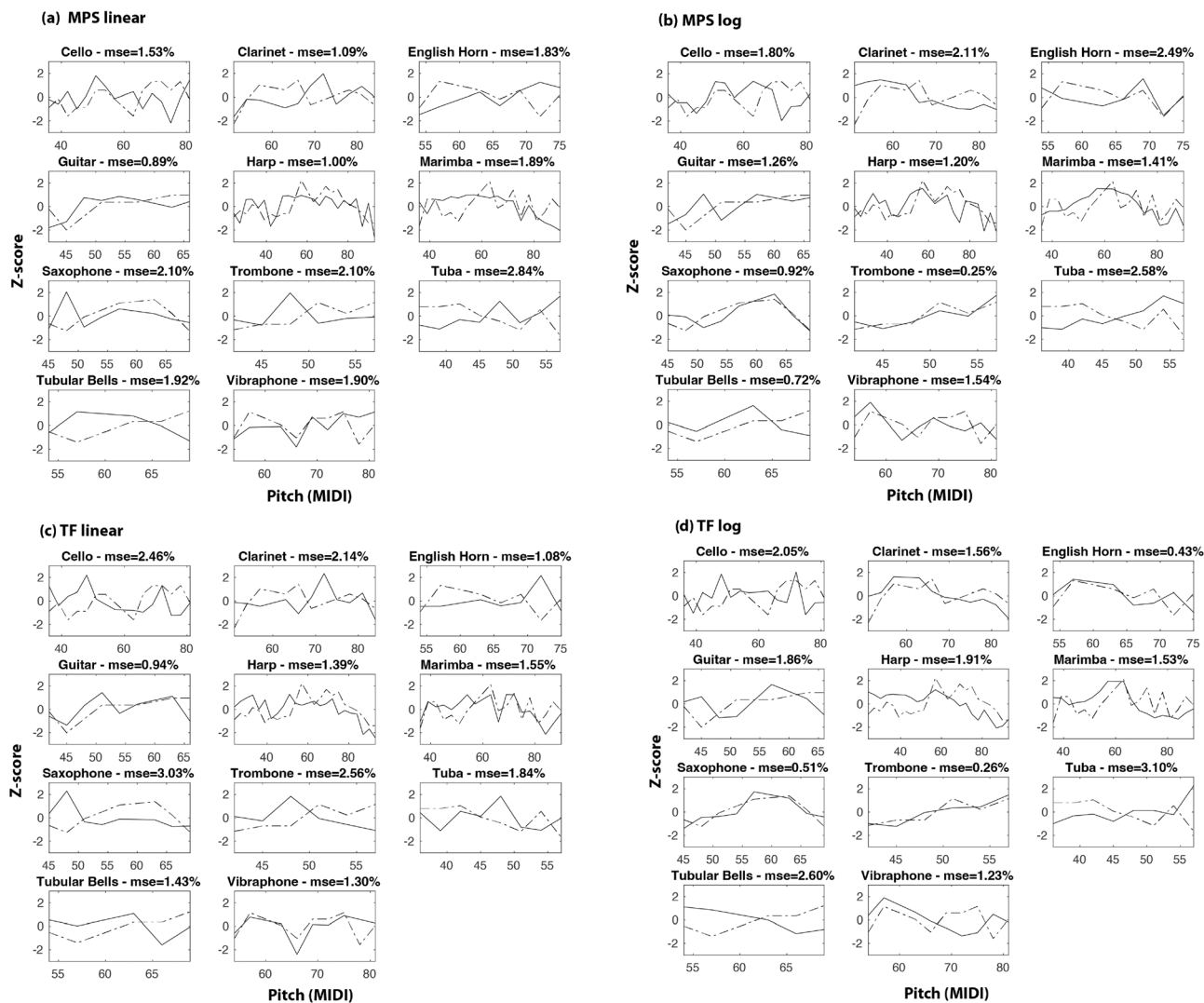


FIG. 9. Comparison of z-transformed mean identification performance across all three experiments at each pitch (dashed dotted) with z-transformed acoustic distances (solid) between sounds at each pitch and the reference pitch (C4, MIDI 60). Each instrument is analyzed with each representation (MPS or TF with linear or log frequency).

pitch ranges drawn from the full range of each instrument (experiment 3). We compared three hypotheses.

H1: Single sounds carry invariant acoustic information that specifies the instrument across its full pitch range.

H2: Exposure to the way timbre varies with pitch enhances learning of the training pitches and may allow listeners to extrapolate beyond the training range.

H3: Listeners need to learn all pitch-timbre combinations inherent to a given instrument in order to successfully identify that instrument across its full pitch range.

Different patterns were observed across instruments. Separate generalized linear mixed effects models of correct response data were estimated for each instrument with pitch (including both linear and quadratic terms), experiment, and their interaction as fixed effects. The main effect of experiment was significant for no instrument, indicating that timbre patterns of invariance across pitches do not enhance learning, arguing against H1 and H2.

The main effect of pitch was significant for all instruments except cello and tubular bells, indicating that learning did not generalize to all pitches for most of the instruments, again not supporting H1 or H2. Cello may have been identifiable across pitch as the only bowed string with bow noise and tubular bells as the only clearly inharmonic sound. Those distinguishing features could serve as invariants to enhance recognition over the whole range within the stimulus context of this experiment. However, this was not the case for the majority of instruments. The curves were linear decreasing over pitch for English horn, tuba, and vibraphone. The results for English horn and vibraphone would conform to H3 according to which learning is restricted to the encountered sounds: the training pitches are at the low end of their ranges, and higher pitches were not as easily identified. However, the tuba presents a surprising case given that performance was better at low notes and worse at higher notes that were in the training region. This may reflect previous experience (see Sec. VA of recognition

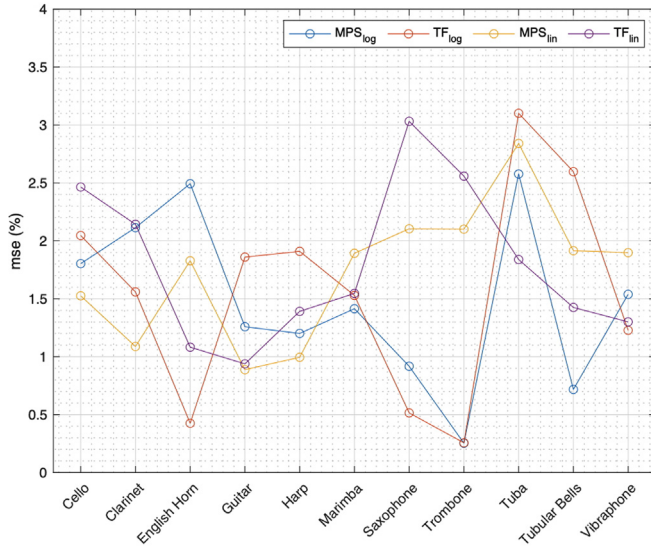


FIG. 10. (Color online) Mean squared error between identification performance and acoustic distance relative to the reference pitch for each instrument analyzed with each representation. Both variables are z-transformed.

below), because the tuba’s optimal (and likely most frequently encountered) range is below the training range. This instrument therefore conforms with none of the initial hypotheses. All of the other instruments had concave curves with highest identification performance in the vicinity of the training range, although the breadth of the peak was wider for some instruments such as the clarinet and harp. For these two instruments, certain acoustic invariances gleaned from the training set may still have been apparent in nearby adjacent regions.

The interaction between pitch and experiment was only significant for cello, harp, and tuba, surprisingly with lower performance at pitches below the training region in experiment 2 (cello), experiment 3 (tuba) or both (harp) compared to experiment 1. Performance was also lower at higher

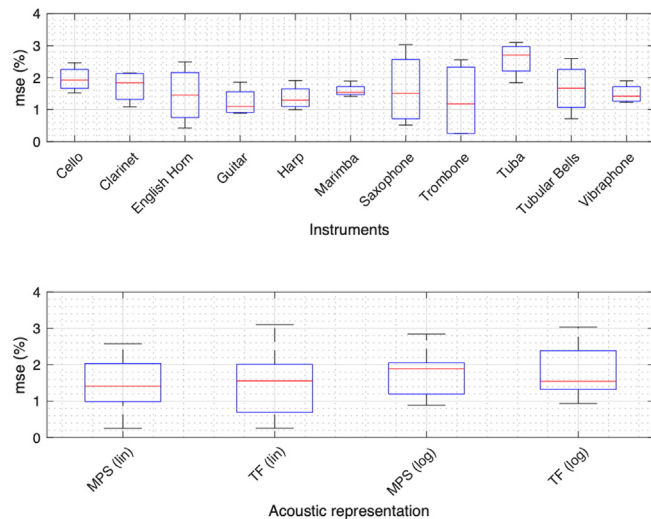


FIG. 11. (Color online) Boxplots of mean squared errors for each instrument across representations (upper panel) and each representation across instruments (lower panel).

pitches for harp in experiment 3. These results are counter-intuitive in that exposure to more information in training and text would be expected to increase performance at more distant pitches, not decrease it.

A. Bandwidth of recognition

To better understand the form and patterns of identification performance, we looked at the instruments’ optimal playing ranges in orchestration, i.e., the register that best represents the instrument, given that these ranges are likely encountered more often in music. We considered the approximate indication of optimal ranges for the wind and string instruments (Adler, 2002). In this discussion, we omit percussion instruments because the concept of optimal register is not relevant given that a separate mechanical system (bar and resonator) in percussion produces each pitch. We examined whether each instrument’s optimal range correlated with the range of identification performance exceeding 75% of maximum performance across experiments (Table III). The strings: cello, harp, and guitar have optimal ranges of C2–C5, Bb0–F5, and E2–E6, respectively. However, there doesn’t seem to be any clear bandwidth pattern in the strings’ identification performance. The optimal range of cello actually lies near its lower value of correct identification, and those of harp and guitar do not correlate to any pattern with the instrument identification.

The woodwinds—English horn, clarinet, and tenor saxophone—have optimal ranges of E3–D5, D3–B5, and Eb3–Eb5, respectively, which all correspond well with the range of correct identification in each instrument above 75% of maximum. The narrow bumps of English horn and saxophone have the highest proportion of correct identification at the ranges of A3–A4 and D#3–A4, respectively. The broad bump of clarinet has the highest proportion of correct identification in the range of D#4–C6, which also overlaps with its optimal range. Trombone and tuba have optimal ranges of A2–D4 and C2–E3, respectively, which also correspond with high identification performance in both instruments. The narrow bump of trombone has the highest identification rate around C3–A3. Tuba has a descending curve with its highest identification rate in its lower range around C2–C3, which is within the optimal playing range of C2–E3. The comparison of the tuba’s optimal range with identification rate provides a better understanding of its descending trend. Again, these results suggest a role of previous experience, as we are more often exposed to these optimal registers specific to each instrument; the timbre within those ranges becomes more familiar to listeners as a representation of that instrument, presumably prior to the experiment.

According to H2, one would expect that providing more covariation contexts in either the training or testing phase could expand the identified pitch range. However, the lack of a main effect of experiment does not support this hypothesis. Similarly to the work of Handel and Erickson (2001), listeners could only extrapolate the timbre of an instrument over a limited pitch range. In their case, the limit was one

octave (with a recognition rather than an identification task). However, in our data, this limit was found only for three instruments—English horn, tenor saxophone, and tenor trombone all had pitch ranges of less than an octave with at least 75% identification performance (Table III). English horn and trombone were also used in Handel and Erickson’s study. For the other instruments in our study, high performance ranged from 1.25 to 3.5 octaves, even for nonmusicians, similarly to the musicians’ performance in Steele and Williams (2006).

B. Confusions

We investigated the extent to which instruments are confused with one another. Confusions mostly occur within excitation categories (sustained and impulsive) and instrumental family. Although there are confusions within the sustained and impulsive groups, there were none between them, demonstrating the primacy both of excitation type as a feature for categorization of sounds based on temporal envelope (Lemaitre and Heller, 2012) and of onset characteristics for instrument identification (Siedenburg, 2019). Within-family confusions are apparent for woodwinds. English horn is at times confused with tenor saxophone in its lowest register and is more often misidentified as clarinet in its highest register. Trombone is labeled as tuba at its lowest pitch, and tuba is progressively labeled as trombone as its pitch increases, which fits with their optimal registers and thus perhaps with previous experience. Similarly, the two plucked string instruments, harp and guitar, are also often confused with each other. The confusions are strongest at lowest pitches for harp and more mildly for extreme pitches at both ends for guitar. Harp is most frequently misidentified as vibraphone or marimba in its extreme high register.

A study by Giordano and McAdams (2010) proposed two possible reasons for the inability to discriminate instruments from the same family. They suggested that acoustical information is not present that differentiates between musical instruments from the same family in a reliable way across pitch variation and that even if acoustical information is present, it is less perceptually salient than pitch variations. In view of this, listeners are more likely to identify an instrument based on their knowledge of the instrument’s typical pitches, which seem to be at least partially responsible for the pattern of within-family confusions observed. We observed that certain timbral sound features might give rise to the confusions, even across instrument families. For example, listeners reported that the “metallic” characteristic caused confusions between harp and vibraphone, especially in the higher register. However, instruments that had little confusion even within families, such as tubular bells, demonstrated that perhaps large differences in mechanical properties of the sound source, and their inherent perceptual properties such as inharmonicity, are associated with significantly better identification ability.

C. Acoustic analyses

The acoustic predictor developed for comparison with identification performance was more or less successful depending on the sound representation used and the instrument being tested. These results echo those of Thoret *et al.* (2016, 2017) who related identification confusions to overlap in the MPS representation. In the current study, no single representation was systematically better at predicting the identification results than the others over all instruments. Furthermore, some instruments were better predicted than others. The spectrogram and modulation power spectrum representations were equivalent predictors across instruments, as were the use of linear and logarithmic frequency.

D. Limitations

One potential limitation of our study is that the training tones and sequences fall within different parts of each instrument’s range, as they vary from lower (English horn, clarinet, vibraphone, and tubular bells) to middle (cello, harp, guitar, saxophone, and marimba) or upper instrument registers (trombone, tuba). It would be interesting to study whether the variance of the training pitch range, according to each instrument’s register, affects identification. Although we specifically selected nonmusicians as participants, there may have been some exposure to certain instruments by their musical affinity or preconceived impressions about some of the instruments presented, as evidenced notably by the tuba results, where performance for untrained low pitches was better than for trained high ones.

VI. CONCLUSION

In this study, we examine the listener’s ability to identify musical instruments across timbres and pitch register to investigate how we learn and extrapolate the timbre-pitch covariance to correctly identify instruments. We observed that the combined results of the study showed that the generalization of pitch-timbre covariation is highly dependent on the specific instruments, the optimal range of each instrument, and the set of instruments selected for the experiment. Acoustic analyses provide a partial explanation of the behavioral data, but there was no apparent advantage of any of the different representations used—modulation power spectra and spectrograms, with linear or log frequency. Concerning our three competing hypotheses, acoustic invariance does not seem to be extractable from single pitches or even pitch sequences over an octave, providing no support for hypothesis 1. Also, listeners trained with pitch-timbre variation do no better than those trained with a single pitch and are largely unable to extrapolate from that acoustic behavior to untrained pitches. Providing limited pitch-timbre covariation at the testing phase does not improve performance either. Thus hypothesis 2 is not supported either. Further investigation is needed to confirm how much information concerning pitch-timbre co-varied information is needed to build an improved mental representation of the instruments, perhaps requiring

experience with the full pitch range of a to-be-identified instrument as proposed in hypothesis 3.

ACKNOWLEDGMENTS

We thank Bennett K. Smith for programming the experiments and Ayla Tse for running participants in experiment 1. This research was funded by grants from the Canadian Natural Sciences and Engineering Research Council (Grant Nos. RGPAS 478121-15, RGPIN-2020-04022), the Fonds de recherche du Québec—Société et culture (Grant No. SE-171434), and a Canada Research Chair (Grant No. 950-231872) awarded to S.M. E.T. was supported by Grant Nos. ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

- Adler, S. (2002). *Study of Orchestration*, 3rd ed. (W.W. Norton, New York).
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). “Random effects structure for confirmatory hypothesis testing: Keep it maximal,” *J. Mem. Lang.* **68**, 255–278.
- Chi, T., Ru, P., and Shamma, S. A. (2005). “Multiresolution spectrotemporal analysis of complex sounds,” *J. Acoust. Soc. Am.* **118**, 887–906.
- 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.
- Giordano, B. L., and McAdams, S. (2010). “Sound source mechanics and musical timbre perception: Evidence from previous studies,” *Mus. Percept.* **28**, 155–168.
- Handel, S., and Erickson, M. L. (2001). “A rule of thumb: The bandwidth for timbre invariance is one octave,” *Mus. Percept.* **19**, 121–126.
- Handel, S., and Erickson, M. L. (2004). “Sound source identification: The possible role of timbre transformations,” *Mus. Percept.* **21**, 587–610.
- 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), 80.
- International Organization for Standardization (2004). ISO 389–8, “Acoustics – Reference zero for the calibration of audiometric equipment—Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones” (International Organization for Standardization, Geneva).
- Korsmit, I., Adler, Y., Madahi, B., Smith, B. K., and McAdams, S. (2021). “Multidimensional scaling of timbral dissimilarities across pitch registers at different dynamics,” in *ICMPC/ESCOM 2021: Connectivity and Diversity in Music Cognition*, Sheffield, <https://drive.google.com/file/d/1aZcturif8IOTanfjEWRQpJ9JmiwgK9zW/view> (Last viewed 30 September 2022).
- Lemaitre, G., and Heller, L. M. (2012). “Auditory perception of material is fragile while action is strikingly robust,” *J. Acoust. Soc. Am.* **131**(2), 1337–1348.
- Lewicki, M. S. (2002). “Efficient coding of natural sounds,” *Nat. Neurosci.* **5**(4), 356–363.
- Marozeau, J., and de Cheveigné, A. (2007). “The effect of fundamental frequency on the brightness dimension of timbre,” *J. Acoust. Soc. Am.* **121**, 383–387.
- Marozeau, J., de Cheveigné, A., McAdams, S., and Winsberg, S. (2003). “The dependency of timbre on fundamental frequency,” *J. Acoust. Soc. Am.* **114**, 2946–2957.
- Martin, F. N., and Champlin, C. A. (2000). “Reconsidering the limits of normal hearing,” *J. Am. Acad. Audiol.* **11**(2), 64–66.
- McAdams, S. (1993). “Recognition of sound sources and events,” in *Thinking in Sound: The Cognitive Psychology of Human Audition*, edited by S. McAdams and E. Bigand (Oxford University Press, Oxford), pp. 146–198.
- McAdams, S. (2019). “The perceptual representation of timbre,” in *Timbre: Acoustics, Perception, and Cognition*, edited by K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay (Springer International Publishing, Cham), pp. 23–57.
- McAdams, S., Douglas, C., and Vempala, N. N. (2017). “Perception and modeling of affective qualities of musical instrument sounds across pitch registers,” *Front. Psychol.* **8**, 153.
- McCabe, V. (1986). “The direct perception of universals: A theory of knowledge acquisition,” in *Event Cognition: An Ecological Perspective*, edited by V. McCabe and G. J. Balzano (Lawrence Erlbaum, Mahwah, NJ), pp. 29–43.
- Opolko, F., and Wapnick, J. (2006). *McGill University Master Samples Collection on DVD* (McGill University, Montreal).
- 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.
- Saitis, C., and Weinzierl, S. (2019). “The semantics of timbre,” in *Timbre: Acoustics, Perception, and Cognition*, edited by K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay (Springer International Publishing, Cham), pp. 119–149.
- Saldanha, E. L., and Corso, J. F. (1964). “Timbre cues and the identification of musical instruments,” *J. Acoust. Soc. Am.* **36**, 2021–2026.
- Schielzeth, H., and Forstmeier, W. (2009). “Conclusions beyond support: Overconfident estimates in mixed models,” *Behav. Ecol.* **20**, 416–420.
- Shamma, S. (2001). “On the role of space and time in auditory processing,” *Trends Cog. Sci.* **5**, 340–348.
- Siedenburg, K. (2019). “Specifying the perceptual relevance of onset transients for musical instrument identification,” *J. Acoust. Soc. Am.* **145**, 1078–1087.
- Siedenburg, K., Jacobsen, S., and Reuter, C. (2021). “Spectral envelope position and shape in sustained musical instrument sounds,” *J. Acoust. Soc. Am.* **149**, 3715–3726.
- 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.
- Smith, B. K. (1995). “PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation,” in *Society for Music Perception and Cognition Conference '95*, University of California, Berkeley.
- Steele, K., and Williams, A. K. (2006). “Is the bandwidth for timbre invariance only one octave?,” *Music Percept.* **23**, 215–220.
- Stilp, C. E., Rogers, T. T., and Kluender, K. R. (2010). “Rapid efficient coding of correlated complex acoustic properties,” *Proc. Natl. Acad. Sci. U.S.A.* **107**(50), 21914–21919.
- Thoret, E., Caramiaux, B., Depalle, P., and McAdams, S. (2021). “Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre,” *Nat. Hum. Behav.* **5**(3), 369–377.
- Thoret, E., Depalle, P., and McAdams, S. (2016). “Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments,” *J. Acoust. Soc. Am.* **140**, EL478–EL483.
- Thoret, E., Depalle, P., and McAdams, S. (2017). “Perceptually salient regions of the modulation power spectrum for musical instrument identification,” *Front. Psychol.* **8**, 587.
- Vienna Symphonic Library (2015). <http://vsl.co.at/en> (Last viewed 30 September 2022).