JASA ARTICLE

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 Montrey[4]
[1]Schulich School of Music, McGill University, Montreal, Québec H3A 1E3, Canada
[2]Aix-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
[3]Cognitive Science Program, McGill University, Montreal, Québec H3A 3R1, Canada
[4]Department 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 multidimensional 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

https://doi.org/10.1121/10.0017100

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 Bruel and Kjær type 4153 artificial ear to which the headphones were coupled (Bruel 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.

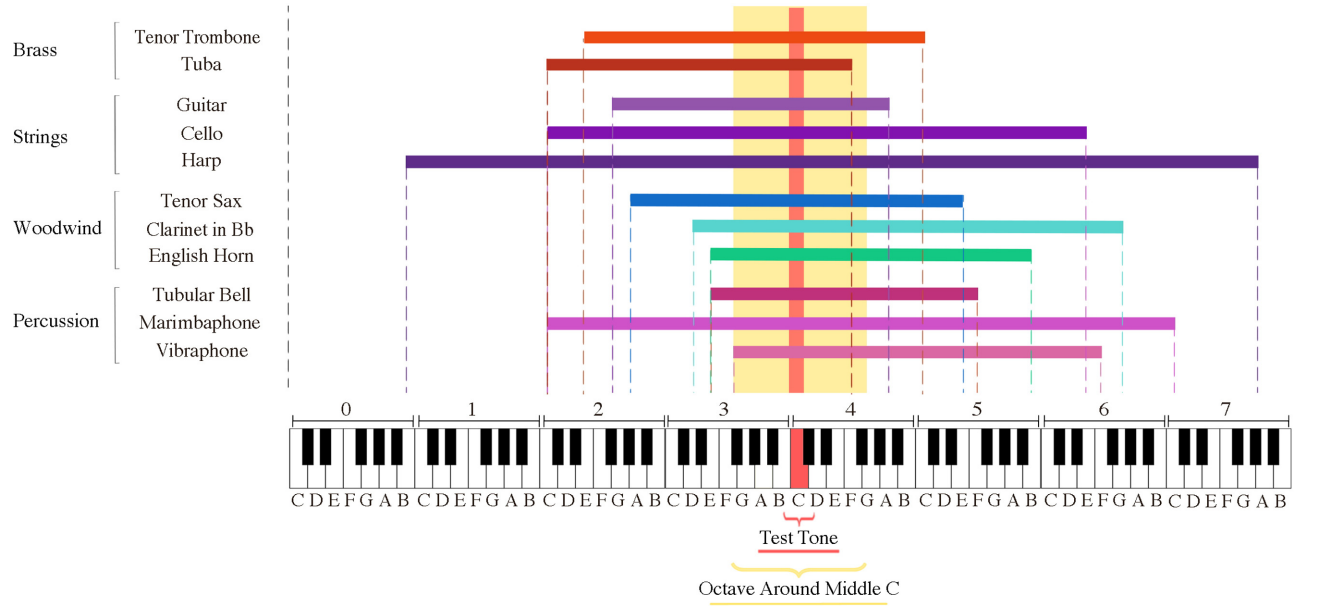

FIG. 1. (Color online) Instrument playing ranges and the range of stimuli used.

https://doi.org/10.1121/10.0017100

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 ¼ 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 ¼ 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 ¼ 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