Embody Listening and Timbre: Perceptual, Acoustical, and Neural Correlates

ZACHARY WALLMARK
Southern Methodist University, Meadows School of the Arts

MARCO IACOBONI
University of California, Los Angeles

CHOI DEBLIECK
University of Leuven, Leuven, Belgium

ROGER A. KENDALL
University of California, Los Angeles

Timbre plays an essential role in transmitting musical affect, and in recent years, our understanding of emotional expression in music has been enriched by contributions from the burgeoning field of embodied music cognition. However, little attention has been paid to timbre as a possible mediator between musical embodiment and affect. In three experiments, we investigated the embodied dimensions of timbre perception by focusing on timbral qualities considered "noisy" and aversive. In Experiment 1, participants rated brief isolated natural timbres scaled into ordinal levels of "noisiness." Experiment 2 employed the same design with a focus on polyphonic timbre, using brief (400 ms) excerpts from 6 popular music genres as stimuli. In Experiment 3, functional magnetic resonance imaging was used to explore neural activations associated with perception of stimuli from Experiment 1. Converging results from behavioral, acoustical, and fMRI data suggest a motor component to timbre processing, particularly timbral qualities considered "noisy," indicating a possible enactive mechanism in timbre processing. Activity in somatomotor areas, insula, and the limbic system increased the more participants disliked a timbre, and connectivity between the premotor cortex and insula relay decreased. Implications for recent theories of embodied music cognition, affect, and timbre semantics are discussed in conclusion.

Received: November 22, 2016, accepted June 27, 2017.

Key words: timbre perception, embodied cognition, fMRI, musical affect, noise
adaptation of the source-filter model, in which timbral properties are understood as the result of an interaction between physical actions (e.g., hitting, blowing, scraping) and material resonators (e.g., the vocal tract, the body of an instrument) (Handel, 1989, 1995), as well as the reverberant environment. Humans are capable of effortlessly discriminating between types of events and objects based on their timbre (for example, see studies on bouncing versus breaking events (W. H. Warren & Verbrugge, 1984); percussion mallet hardness (Freed, 1990); string bowing (Halmrast, Guettler, Bader, & Godoy, 2010); and trumpet mouthpiece depth (Poirson, Petiot, & Gilbert, 2005)).

The human ability to immediately grasp the identity of a sound source through the perceptual correlates of its mechanical and structural features has obvious adaptive value. However, within the context of music, it would seem that gross source identification is typically less relevant to listeners than the identification of expressive intentions through timbral cues. For example, we might hear a timpani strike as connoting anger or triumph, rather than merely hearing it as a timpani head struck quickly and forcefully with wooden mallets. Indeed, numerous studies have found that musicians are effortlessly discriminating between types of events and objects based on their timbre (for example, see studies on bouncing versus breaking events (W. H. Warren & Verbrugge, 1984); percussion mallet hardness (Freed, 1990); string bowing (Halmrast, Guettler, Bader, & Godoy, 2010); and trumpet mouthpiece depth (Poirson, Petiot, & Gilbert, 2005)).

By definition, timbre is a perception of the timbral characteristics of a sound that are perceived to be unique to a sound source. Timbre therefore is a perceptual construct of the static and dynamic aspects of a sound. In music, timbre is thought to be a prime factor in the perception of instrumental timbres and it is generally assumed that a timbral feature, such as timbral hardness, can influence the perceptual consonance or incongruity between a sound source and a context (Behrens & Green, 1993; Gabrielsson & Juslin, 1996; Verbrugge, 1984).

Timbre has been shown to play an important though largely unconscious role in affective response to music (Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005; Peretz, Gagnon, & Bouchard, 1998). Since it is difficult to isolate from other parameters of the musical experience, however, the exact nature of its contribution has long been overlooked. A small number of studies have systematically explored this question. Hailstone et al. (2009) found that perception of emotion conveyed by original melodic sequences was affected by the timbre of the instrument performing them. This result is consonant with Paraskeva and McAdams (1997), who demonstrated that timbral variation modulates perceived relaxation and tension in musical compositions. It is unclear in these studies, however, what mechanisms might help explain these effects. What aspects of timbre are modulating affective reactions, and how?

This is a beguiling question in part because timbre perception is largely excluded from conscious awareness. Fales (2002) calls the discrepancy between objective, physical measures of timbre and its corresponding perceptual frame of reference the “paradox of timbre.” The fact that timbre acoustics relates in a complex, multidimensional, and sometimes misleading way to perception means that the effects of timbre largely go unrecognized, or are misattributed to other more salient features of the musical signal or listening context. In short, though timbre can have an enormous influence on our affective responses to music, its effect is mainly implicit. This aloofness from conscious perception, moreover, may be an important component to its affective vehemence: studies on priming reveal that implicit primes exert a stronger affective sway on observers than explicit primes, since they are nonspecific and require little cognitive mediation (Murphy & Zajonc, 1993).

There is neurophysiological and behavioral evidence for timbre’s “subterranean impact” (Fales, 2002, p. 77). A study using EEG found that categorization of timbre occurs rapidly and preattentively, as reflected in mismatch negativity at the 100–200 ms latency range (Tervaniemi, Winkler, & Näätänen, 1997). Other studies have used this paradigm to explore affective categorization of timbre, with similar results (Goydke, Altenmüller, Möller, & Münte, 2004; Spreckelmeyer, Altenmüller, Colonius, & Münte, 2013). The processing speed of timbre is further evidenced in rapid music cognition studies, which demonstrate that reliable perceptual discriminations based on genre and other attributes of recorded music can occur in the 250–400 ms range, too short for the contribution of substantial melodic, harmonic, and rhythmic information (Filipic, Tillmann, & Bigand, 2010; Gjerdingen & Perrott, 2008; Krumhansl, 2010; Schellenberg, Iverson, & McKinnon, 1999).

Affective responses to timbre are not entirely of the “quick and dirty” variety, of course; they can also involve conscious deliberation. In a recent behavioral study, McAdams, Douglas, and Vempala (2017) found that listeners evaluate affective connotations of isolated instrumental timbres consistently, even across differing pitch registers. Eerola, Ferrer, and Alluri (2012) reported similar results with pitch controlled (D#4). Both of these studies revealed that only a small handful of acoustic variables drive most of the variance in affective responses; additionally, they converged on the key dimensions of valence, tension arousal, and energy arousal underlying the timbre appraisal process. Eerola
and colleagues theorized a role for mechanisms of embodied cognition in their results, inferring that affective properties of timbre are related to the “underlying physiological mechanisms that reflect the body-states related to affective experiences” (p. 63); however, this hypothesis was not explicitly tested.

Our study takes up where Eerola et al. (2012) leave off with an exploration of the behavioral, acoustical, and neural components corresponding to perhaps the strongest affective response one regularly has to timbre—the appraisal that a timbral quality is “noisy.” Noisy timbre is an intuitive place to begin mapping basic affective mechanisms, because the gestalt is reliably associated with both low valence and high energy arousal (Eerola et al., 2012). We designed three experiments—two using behavioral methods with isolated natural timbres and brief excerpts of popular music recordings as stimuli, and one using functional magnetic resonance imaging (fMRI)—in order to investigate the most fundamental affective axis of timbre, its valence, from a theoretical vantage point of embodied music cognition.

TIMBRE AND THE EMBODIED MIND

Embodied cognition is a contemporary research paradigm that emphasizes the primary role of the sensorimotor system in human thought processes (for an overview, see Shapiro, 2010). As defined by Thelen, Schöner, Scheier, and Smith (2001, p. 1):

To say that cognition is embodied means that it arises from bodily interactions with the world... Cognition depends on the kinds of experiences that come from having a body with particular perceptual and motor capacities that are inseparably linked and that together form the matrix within which memory, emotion, language and all other aspects of life are meshed.

By connecting the contingencies of the sensing body with the broader material and cultural ecosystem, embodied cognition offers an expansive framework for understanding the role of human perceptual systems in the process of producing, hearing, and making sense of musical sound. In recent years, perspectives from the embodied cognitive sciences have influenced music-theoretical discourse on a wide range of topics, including metaphor (Zbikowski, 2002), musical preference (Schafer & Sedlmeier, 2011; Sedlmeier, Weigelt, & Walther, 2011), organology (De Souza, 2017; Reybrouck, 2006), and musical expression (Cox, 2016; Godøy, 2003; Leman, 2007; Leman et al., 2009).

Researchers have also begun making inroads on the neuroscience of brain-body interactions in musical contexts. The human mirror neuron system (MNS) has been proposed as an important neural substrate for affective and motor responses to music (Leman, 2007; Molnar-Szakacs & Overy, 2006; Overy & Molnar-Szakacs, 2009). Mirror neurons cortically represent the physical actions of one individual in the sensorimotor regions of another, enabling the multi-modal coupling of perception and action (for a review, see Rizzolatti & Craighero, 2004). The shared representation of motor actions and intentions among multiple individuals, as reflected in co-activation of mirror neuron networks, has been labeled “motor resonance,” a term we will use throughout this article. Although cross-fertilization of mirror neuron research with music studies is still in its relatively early stages and remains as yet largely theoretical, motor resonance has been shown to play a role in a number of functions relevant to music, such as auditory action understanding (Aziz-Zadeh, Iacoboni, Zaidel, Wilson, & Mazziotta, 2004; Keysers et al., 2003; Kohler et al., 2002), mimetic learning (Buccino et al., 2004; Lahav, Saltzman, & Schlaug, 2007), phoneme processing (Wilson & Iacoboni, 2006), familiarity with instrumental sound (Margulis, Mlsna, Uppunda, Parrish, & Wong, 2009), affective response (J. E. Warren et al., 2006), and musical empathy (Wallmark, Deblieck, & Iacoboni, submitted).

Overy and Molnar-Szakacs (2006, 2009) hypothesized that motor resonance helps listeners to translate the affective motor patterns inherent in a musical signal to the expressive intentions of the performer through “shared affective motion experience” (SAME). They summarize: “The SAME model suggests that musical sound is perceived not only in terms of the auditory signal, but also in terms of the intentional, hierarchically organized sequences of expressive motor acts behind the signal... The expressive dynamics of heard sound gestures can be interpreted in terms of the expressive dynamics of personal vocal and physical gestures” (2009, p. 492). According to this model, musical input from the primary auditory cortex leads to motor resonance as the perceiver takes up the motional cues of the acoustic signal. Neural reenactment of sound-generating actions or other motoric implications of musical sound modulates affective response by way of a functional connection with the limbic system, with the insula serving as a conduit between the two regions (Carr, Iacoboni, Duboë, Mazziotta, & Lenzi, 2003; Menon & Levitin, 2005).

If timbral qualities are the result of “action-trajectories” (Godøy, 2003, p. 318) and we process sounds by
way of motor resonance with their implied actions, then we might expect to see confirmation of the SAME hypothesis at the neurophysiological level. Direct evidence for this, however, remains scant. Most studies have localized timbre processing in regions of the temporal lobe, including superior temporal gyrus, superior temporal sulcus, and Heschl’s gyrus (Alluri et al., 2012; Caclin et al., 2006; Halpern, Zatorre, Bouffard, & Johnson, 2004; Menon et al., 2002; Pantel, Roberts, Schulz, Engelen, & Ross, 2001; Platel et al., 1997; F. Samson, Zeffiro, Toussaint, & Belin, 2011; S.S Samson & Zatorre, 1994). Circumspect evidence for motor activity in timbre processing has been observed in a few studies. One fMRI study implicated the supplementary motor area (SMA) in a timbre imagery task, which was interpreted as evidence for subvocalization of gross timbre features (Halpern et al., 2004). Another fMRI study compared violinists’ and flutists’ perception of music on their own (Margulis et al., 2009). And a further imaging study revealed activation of the putamen (basal ganglia) when participants listened to “bright” qualities of timbre (Alluri et al., 2012). Additional studies using MEG (Pantev et al., 2001) and behavioral methods (Drost, Rieger, & Prinz, 2007) reported similar timbre-specific effects. It should be noted, however, that the relative dearth of motor evidence in the timbre neurophysiology literature is not itself evidence of absence: there is ample empirical support for motor resonance during music listening more generally (Zatorre, Chen, & Penhune, 2007), though without isolation of timbral elements it is impossible to attribute any specific effect to timbre. Indeed, all musical sounds are timbred in addition to being organized horizontally and vertically, so pulling apart how timbre alone contributes to embodied listening remains a challenging task.

Our study attempts to cross-fertilize theories of embodied cognition with timbre perception in order to test the hypothesis that processing aspects of timbre recruits sensorimotor and limbic regions of the brain. Motor involvement has been demonstrated for rhythm and beat processing (Grahn & Brett, 2007; Patel & Iversen, 2014; Popescu, Otsuka, & Ioannides, 2004), but not significantly, to our knowledge, for timbre. Similarly, while it is clear that limbic and paralimbic structures such as the amygdala, striatum, and parahippocampal gyrus play an important role in emotional responses to music in general (for a review, see Peretz, 2010), the contribution of the limbic system to timbre processing is not well understood.

NOISY TIMBRE

In order to explore which timbral percepts might be particularly potent as a locus of motor resonance, we chose to focus on musical timbres with prominent aperiodic, noise components, hereafter referred to as “noisy timbre.” Noise components in human vocalizations are generally a marker of high arousal and low valence: we tend to overdrive our voices—growling, screaming, and crying, for example—in response to intensely negative stimuli (Arnal, Flinker, Kleinschmidt, Giraud, & Poeppel, 2015; Johnstone & Scherer, 1999; Mende, Herzel, & Wermke, 1990). Many non-human animals exhibit similar action-perception associations with noisy timbre (Blumstein & Récapet, 2009; Fitch, Neubauer, & Herzel, 2002; Hauser, 2000; Scheumann, Hasting, Kotz, & Zimmermann, 2014); there are thus ethological reasons to suppose that the processing of noisy timbres may recruit phylogenetically ancient pathways, and may have developed under adaptive pressures (Morton, 1977). In their “superexpressive voices” theory, moreover, Justlin and Laukka (2003, p. 803) propose that instrumental timbres may convey certain affective connotations by way of acoustic similarities with the voice. Noisy timbre thus offers a promising test case for the application of theories of embodied music cognition to the realm of timbre. If timbre in isolation is capable of modulating motor resonance, it seems to us that this will be clearest to observe in auditory contexts that mimic, as closely as possible, the sound of the voice in high-exertion, ecologically urgent situations.

The use of noisy timbre is widespread in musical practices around the world, and has been explored in contexts as diverse as Beijing opera (Tsai, Wang, Wang, Shau, & Hsiao, 2010), rock and heavy metal (Berger & Fales, 2005), traditional Japanese music (Takemitsu, 2004), and African and Afro-diasporic musics (Avorgenbedor & Pyne, 1999; Fales & McAdams, 1994), among many other examples. Although the impression of a given quality of timbre as “noisy” is ubiquitous in everyday listening, the acoustic and psychological correlates of this commonplace appraisal are multifaceted and stubbornly elusive to define for empirical purposes: as Fales and McAdams (1994, p. 73) indicate, “both acoustically and perceptually, noise is intertwined with other elements of the music with tremendous complexity.” Furthermore, terminological ambiguities have made timbral noisiness a chimerical object of inquiry within both empirical and humanistic discourses. In this study we treat noisiness as a perceptual gestalt capable of modulating motor resonance response owing to the ecological link between acoustic noise components, exertion, valence, and affective response. Therefore,
while acknowledging the heterogeneity and cultural situatedness of any perception of sound as “noise,” for the sake of this study we subdivide noisy timbre into two psychoacoustic components, as both (1) a timbral quality of pitched sound that, in isolation, typically has negative valence, and (2) the acoustic determinants of this perception.

The musical acoustics literature has identified a broad range of spectral features as comprising the basis for the perception of noisiness in musical timbre (Elliott, Hamilton, & Theunissen, 2013), including inharmonicity, or spectral energy outside the harmonic series (Moore, Peters, & Peters, 1985); aperiodic components such as sidebands, broadband noise, and subharmonics (Gibiati & Castellengo, 2000); combination tones; and auditory roughness (Helmholtz, 1877/1954; Sethares, 1998). Such acoustic accounts have informed numerous case studies: for example, Fales and McAdams (1994) explored inharmonic components in the timbral makeup of African instruments and their perception as either “layered” or “fused” with the harmonic content of the tone. Vassilakis (2005) discussed the acoustics of roughness and its aesthetic uses in contexts such as the Middle Eastern mijwiz and Indian tambura drone. Similarly, Pressnitzer, McAdams, Winsberg, and Fineberg (2000) investigated the role of roughness in atonal orchestral music and its contribution to perceived dissonance. And Tsai et al. (2010), the closest precursor study to ours, studied biomechanical aspects of “growl-like timbre” through an embodied cognitive lens.

STUDY AIM

We designed a study to investigate perceptual, acoustic, and neural components of affective timbre listening. In order to explore the convergence between these three levels, we employed methods from cognitive psychology, acoustic analysis, and neuroimaging in three experiments. In Experiment 1, participants rated isolated natural monophonic timbres scaled into ordinal categories of noisiness according to a number of perceptual and affective dimensions. Behavioral data were then correlated with computationally extracted acoustic attributes. In order to compare results from Experiment 1 with perception of timbral noisiness in natural musical (i.e., polyphonic) settings, Experiment 2 employed a similar set of ratings and word selection tasks using very brief (400 ms) excerpts of popular music as stimuli. This experiment involved the same analytical procedures as the previous study, including low-level spectral feature extraction, correlations with behavioral results, and multidimensional scaling to assess perceptual similarity of stimuli in both experiments. Finally, in Experiment 3 we used functional magnetic resonance imaging (fMRI) to explore the neural dynamics of listening to isolated monophonic timbres scaled for differing levels of noisiness (same stimuli as Experiment 1). Behavioral data were included as covariates in our general linear model of blood-oxygen-level-dependent (BOLD) signal change with the aim of determining which regions of the brain are involved in the processing of noisy timbral stimuli.

Consistent with an embodied cognitive model of timbre perception, our study aimed to test the broad hypothesis that timbre processing—as exemplified here in the test case of noisy timbre—can involve multimodal, sensorimotor engagement. Acoustically, we wished to verify that parameters commonly labeled as “noisy” in timbre studies are actually perceived as such. We predicted that acoustically noisy qualities of monophonic and polyphonic timbre would be heard as reflecting higher exertion actions than non-noisy timbres, would have negative valence, and would be associated with low valence/high arousal primary emotions such as anger and fear. In Experiment 3, we hypothesized that noisy timbre would modulate activity in sensorimotor areas including the pre- and primary motor cortex and SMA, as well as somatosensory areas, indicating the possible involvement of an enactive mechanism in timbre processing. Furthermore, consistent with Overy and Molnar-Szakacs’s SAME hypothesis (2009), we postulated that motor processing might share functional connectivity with limbic circuitry by way of a relay in the insula.

Experiment 1

Method

PARTICIPANTS

Thirty-six participants (22 female, 14 males) were recruited from the UCLA community, age 18–31 \( (M = 20.3, SD = 2.07) \). Subjects self-reported their number of years of formal music training \( (M = 6.83 \text{ years}, SD = 5.93) \); they were all non-music majors with musical backgrounds ranging from no formal experience \( (7 \text{ subjects}) \) to 20 years \( (1 \text{ subject}) \). Participants were also involved in Experiment 2, and a random subset of this group also participated in Experiment 3. The order of the three experiments was randomized. To correct for non-independence of observations between the three experiments, relatively conservative statistical procedures and adjustments were employed throughout the entire study in order to place stringent limits on our inferential claims (e.g., Bonferroni post hoc testing and...
an imaging threshold of \( p < .01 \), to be discussed later). Subjects who participated in Experiments 1 and 2 only were compensated $20 for their time.

**STIMULI**

Stimuli consisted of twelve approximately 2 s (1.8–2.1 s) instrument and vocal timbres, 3 electric guitar, 3 tenor saxophone, 3 shakuhachi (an end-blown bamboo flute from Japan), and 3 female vocals, recorded by experienced musicians specifically for this experiment. Each set of three timbres was scaled with the performers’ assistance into ordinal levels of “noisiness”: (1) “normal,” (2) medium-noise condition, and (3) high-noise condition. Musicians used different tone-altering techniques in the medium- and high-noise conditions, including styles of electric guitar distortion, saxophone growling, shakuhachi noise techniques, and vocal “dist” tones (Borch, Sundberg, Lindestad, & Thalen, 2004). Following an ecological model of timbre perception (Giordano & McAdams, 2010; Handel, 1995), the four sound generators were selected to represent a range of different drivers, including electro-acoustic, reed, air-stream, and vocal production. Guitar, saxophone, and voice were familiar to participants, while the shakuhachi was largely unfamiliar. This is one of the few timbre studies we are aware of that uses the same sound generators in different timbral conditions as stimuli (Goydke et al., 2004; Spreckelmeyer et al., 2013). Stimuli recordings and additional details can be found in the Supplementary Materials section accompanying the online version of this paper.

In addition, two signals were included to demarcate the outer boundaries of acoustical noisiness: (1) a 2 s clip of pink noise generated with Audacity software, including 10 ms amplitude ramps, for the exemplar of “noise”; and (2) a 2 s sine wave, also including 10 ms amplitude ramps, for the exemplar of the opposite of noise. (These signals were excluded from the ANOVA detailed below due to asymmetry of study design.)

Natural signals were recorded using Apple Logic software, a Neumann TLM-103 microphone, and a MOTU 896 interface at a sampling rate of 44.1 kHz. Attack and cutoff were unaltered for maximum ecological validity. Fundamental frequency was the same for all signals (B♭3, ca. 233 Hz). Because perceived loudness can never really be equal for every listener (Hajda et al., 1997), we opted to equalize loudness of the stimuli manually through a process of independent inter-rater agreement between the authors (Z. W. and R. A. K.).

**PROCEDURE**

Participants were asked to rate perceptual and affective features of the randomly presented signals on a computer using Music Experiment Development System (MEDS) software (Kendall, 2002). Subjects listened to the mono signals through Bose SoundTrue headphones at a subjectively determined comfortable listening level in a quiet room. The duration of the full experiment was approximately 10 min.

The experiment consisted of five separate rating and word-selection tasks (5 x 12 signals = 60 total signals) presented in a predetermined order:

1. The perceived *bodily exertion* required to produce each timbre was measured using a numbered horizontal rating scale (0–100) with bipolar labels (low exertion–high exertion) consistent with the semantic differential paradigm (Osgood, Suci, & Tannenbaum, 1957). This scale was meant to capture the relative degree of physical effort implied in the production of a sound, either vocally or through other means, as a simple measure of embodied timbre perception. Noisy qualities of timbre in vocal and (non-synthetic) instrumental sound production typically index physical exertion, and are associated with low valence and high energy arousal (Juslin & Västfjäll, 2008; Scherer, 1995; Wallmark, 2014). We therefore hypothesized that noisy qualities of timbre would be associated with higher exertion than normal, non-noisy timbres.

2. *Valence* of each signal was assessed using a 0–100 horizontal rating scale with bipolar labels (dislike–like). The valence scale aimed to capture how strongly listeners like or dislike particular timbral qualities. Negative valence is an important indirect component to the perception of timbre as noisy (McDermott, 2012).

   We also wanted to explore certain basic semantic features of timbre as it relates to questions of noisiness and embodiment. Many adjectives and metaphors are used to describe qualities of sound (Wallmark & Kendall, in press), but for the present study we selected just two of the most ubiquitous, *bright* and *noisy*, employed as follows:

3. Perceived *brightness* of each signal, using a 0–100 Verbal Attribute Magnitude Estimation (VAME)-modified semantic differential scale (Kendall & Carterette, 1993a) (not bright–bright). The semantic dimension of *brightness* is a commonly acknowledged perceptual correlate of spectral centroid (Beauchamp, 1982; Schubert & Wolfe, 1991).
Perceived noisiness of each signal using a 0–100 VAME scale (not noisy–noisy), as our most direct semantic measure of timbral noisiness. The inclusion of this term is not merely tautological: despite its experiential ubiquity, little research has been done exploring the perception of noisiness in musical timbre and its acoustic correlates.

Perceived emotion of each signal selected from a list of adjectives describing the emotion conveyed by the performer of each sound. The forced-choice categorization task consisted of the five primary emotions determined by Juslin (2001) to be generally communicable by performers, and used by Krumhansl (2010) in a similar study: (a) happiness, (b) sadness, (c) anger, (d) fear, and (e) tenderness. We included perceived primary emotions to confirm and explore categorical affective features of timbre, most notably the established link between acoustically noisy qualities of timbre and perceived anger and/or fear (Scherer & Oshinsky, 1977; Tsai et al., 2010). Perceived emotion selections were not meant to measure induced or “felt” emotions in response to different timbres, only the affective intentions of their performers.

**TABLE 1. Six Spectral Parameters of Noisy Timbre**

<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Spectral parameter</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF</td>
<td>High-frequency energy</td>
<td>Proportion of total energy above 1500 Hz (Juslin, 2000; Lartillot &amp; Toiviainen, 2007)</td>
</tr>
<tr>
<td>SC</td>
<td>Spectral centroid</td>
<td>Amplitude-weighted fulcrum point of energy distribution across the spectrum, as ratio between long-term average centroid and fundamental frequency through time (Kendall &amp; Carterette, 1996; Kendall, 2002)</td>
</tr>
<tr>
<td>IN</td>
<td>Inharmonicity</td>
<td>Frequency deviation of partials from ideal harmonic series (Lartillot &amp; Toiviainen, 2007)</td>
</tr>
<tr>
<td>SF</td>
<td>Spectral flatness</td>
<td>Wiener entropy of signal (ratio between geometric and arithmetic means of the spectrum)</td>
</tr>
<tr>
<td>ZC</td>
<td>Zero-cross rate</td>
<td>Number of signal changes per unit of time (Lartillot &amp; Toiviainen, 2007)</td>
</tr>
<tr>
<td>RO</td>
<td>Roughness</td>
<td>Sensory dissonance averaged through time (Sethares, 1998; Vassilakis, 2005)</td>
</tr>
</tbody>
</table>

**TABLE 2. Inter-acoustic Correlation Matrix for Isolated Timbres**

<table>
<thead>
<tr>
<th></th>
<th>HF</th>
<th>SC</th>
<th>IN</th>
<th>SF</th>
<th>ZC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF</td>
<td>.61*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SC</td>
<td>.57*</td>
<td>.70**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN</td>
<td>.37</td>
<td>.94**</td>
<td>.87**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF</td>
<td>.46</td>
<td>.93**</td>
<td>.37</td>
<td>.90**</td>
<td></td>
</tr>
<tr>
<td>ZC</td>
<td>.33</td>
<td>.89**</td>
<td>.68</td>
<td>.94**</td>
<td>.95**</td>
</tr>
</tbody>
</table>

Pearson’s r correlation coefficients between 14 signals; corrected for multiple comparisons (FDR). *p < .05, **p < .01.

**ACOUSTIC ANALYSIS OF STIMULI**

Spectral components of the signals were computationally extracted using MIRtoolbox1.4 (Lartillot & Toiviainen, 2007) in the MATLAB environment, and MEDS was used for spectral centroid calculations (Kendall, 2002). Measurements of all acoustic parameters were derived from a 12-order FFT with 50% overlap (50 ms per frame), producing approximately 40–45 frames for each signal at a sampling rate of 44.1 kHz. Overlap settings were selected to increase temporal resolution and spectral averaging at window edges, and the resulting calculations were based on an average across all frames. Similar FFT parameters were carried out in MEDS with the addition of a Bartlett window; transients were trimmed above a threshold value of 20, and the resulting calculations were a long-term average centroid across all remaining frames, as described by Kendall and Carterette (1996).

Elliott et al. (2013) found that perceived timbral noisiness was associated solely with spectral features of timbre; our analysis thus considered only spectral and not temporal or spectrotemporal timbral features. Six spectral parameters were assessed: high-frequency energy, spectral centroid, inharmonicity, spectral flatness, zero-cross rate, and auditory roughness, to be discussed in the next section (Table 1). As seen in Table 2, these parameters unsurprisingly exhibited strong inter-acoustic correlations: for example, spectral centroid was correlated with spectral flatness, r(12) = .94, and zero-cross rate, r = .93, both p < .001. (All p values for correlation coefficients throughout the rest of the article were...
adjusted for multiple comparisons using False Discovery Rate method (FDR); see Benjamini & Hochberg (1995).

Results
None of the subjects reported any difficulty in performing the experimental task, and acceptable inter-subject consistency in the ratings was observed ($M$ Cronbach’s $\alpha$ for all scales = .74). We first checked for normality of distribution of the behavioral and acoustic data using the Shapiro-Wilk test. To correct for violations of normality, 17 of the 48 perceptual variables and 1 of the 6 acoustic variables were transformed through an inverse-normal transformation process into normally distributed $z$-scores, as described by Templeton (2011), a process that roughly preserves the scale of the original data.

As summarized in Figure 1, a repeated-measures ANOVA (4 x 4 x 3) revealed significant effects for Perceptual, Instrument, and Timbre factors, with moderate to strong effect sizes: Perceptual variables, $F(3, 78) = 5.77, p < .01$, $\eta^2_p = .18$ (Greenhouse-Geisser corrected); Instrument, $F(3, 78) = 28.58, p < .0001$, $\eta^2_p = .52$; and Timbre, $F(2, 52) = 5.31, p < .01$, $\eta^2_p = .17$, in addition to all two- and three-way interactions, Perceptual/Instrument, $F(9, 234) = 11.81, p < .001$, $\eta^2_p = .31$; Perceptual/Timbre, $F(6, 156) = 23.93, p < .0001$, $\eta^2_p = .48$; Instrument/Timbre, $F(6, 156) = 9.98, p < .0001$, $\eta^2_p = .27$; and Perceptual/Instrument/Timbre, $F(18, 468) = 6.31, p < .0001$, $\eta^2_p = .20$. Post hoc testing (Bonferroni) of main effects and interactions confirmed the significance of differences between all Instrument means except guitar and saxophone (all $p < .05$), and between Timbres 2 and 3 ($p < .01$).

As shown in the first plot of Figure 1, means for the three timbre types clustered closely together on the exertion and brightness scales but diverged significantly for valence and noisiness ($p < .001$), with the two “noisy” conditions following a similar pattern. This might suggest that while the three timbral levels were conceived ordinally, their perception is more binary (i.e., “noisy” or not). (This dichotomous perceptual organization is confirmed at the neurophysiological level as well, as we will review later.) In the second plot, we can see that the interaction is primarily driven by the guitar and the female voice, which do not differ in exertion and noisiness means but cross substantially in the other perceptual dimensions. Surprisingly, the electric guitar was perceived as being significantly more pleasant than the voice ($p < .0001$), which was given the lowest valence ratings of all sound generators. In three-way interaction, voice and tenor saxophone followed the same pattern as the “noisy” timbres in plot 1, while guitar roughly reiterates the shape of the “normal” timbral category. The perceptual distance between guitar and voice—with saxophone and shakuhachi flute occupying the space between—drives this interaction, and was further confirmed by multidimensional scaling analysis (to be reported later).

Table 3 shows correlations between perceptual measures. We can see that bodily exertion correlated most
strongly with noisiness, $r(34) = .56$, $p = .13$, and with frequency of anger responses, $r = .68$, $p = .04$. Supporting the ecological view, the perception of exertion in timbre conditions 2 and 3 (the “noisy” conditions) was especially apparent among sound generators that actually require an elevated degree of effort for the production of distortion, such as the voice and saxophone. It is interesting and somewhat unexpected, however, that electric guitar signals generated a similar pattern: higher levels of exertion were heard in the distorted timbres, even though the physical means of producing guitar distortion do not require increased effort on behalf of the player. It could be argued that noisy acoustic qualities somehow “trick” the listener into hearing increased exertion, although no more physical work is required to produce distorted electric guitar than a clean tone. This result might also be mediated by familiarity with electric guitar playing technique, though such a measure was not assessed.

The valence scale yielded significant results as well, correlating with both semantic dimensions and two perceived emotions: brighter signals tended to elicit more positive values, while noisy signals were negatively correlated ($r = -.55$, $p = .13$). Not surprisingly, valence was moderately associated with frequency of happiness ($r = .56$, $p = .13$) selections, and low valence responses were strongly correlated with fear ($r = -.82$, $p < .01$). The positive relationship between valence and brightness was a novel finding: researchers have correlated impressions of brightness with high spectral centroid, and a large ratio of high-to-low frequency tends to have negative valence in both musical and non-musical contexts (Eerola et al., 2012; Kumar et al., 2012; McDermott, 2012). We therefore expected that brightness would be a negative attribute, but the opposite appeared to be the case.

Next we compared acoustic and behavioral data in order to explore how physical and perceptual frames of reference interrelate in the phenomenon of timbral noisiness (Table 4). All spectral parameters exhibited a strong correlation with at least one of the perceptual measures; for example, high-frequency energy was strongly associated with exertion ($r = .71$), noisiness ($r = .76$), and anger ($r = .68$), all $p < .05$. A multiple linear regression analysis was carried out to determine the strength of these six acoustic variables in predicting behavioral response. Results indicated that a combination of these collinear acoustic variables explained approximately 86% of variance in exertion ratings, $F(6, 7) = 7.25$, $p < .01$, $R^2 = .86$; 80% of brightness ratings, $F(6, 7) = 4.66$, $p < .05$, $R^2 = .80$; and 83% of noisiness ratings, $F(6, 7) = 5.76$, $p < .05$, $R^2 = .83$. Valence was not significantly explained by the acoustic variables, $F(6, 7) = 1.40$, $p = .33$, $R^2 = .55$.

To briefly summarize, Experiment 1 provided confirmation that noisy timbres—that is, those with larger high-frequency energy, spectral centroid, inharmonicity, spectral flatness, zero-cross, and roughness—seem to indicate a greater amount of bodily effort going into their production, and produce an affective response with lower valence than normal versions of the same sound generators. As expected, the perception of noisiness also

---

**TABLE 3. Inter-perceptual Correlation Matrix for Isolated Timbres**

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>V</th>
<th>B</th>
<th>N</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valence</td>
<td>-.17</td>
<td>.42</td>
<td>.55</td>
<td>.14</td>
<td>-.39</td>
<td>-.09</td>
<td>-.40</td>
<td>-.75</td>
</tr>
<tr>
<td>Brightness</td>
<td>.68</td>
<td>-.55</td>
<td>.03</td>
<td>.83</td>
<td>-.38</td>
<td>-.20</td>
<td>-.51</td>
<td>-.43</td>
</tr>
<tr>
<td>Noisiness</td>
<td>.14</td>
<td>.56</td>
<td>.83**</td>
<td>-.59</td>
<td>.09</td>
<td>.87**</td>
<td>.10</td>
<td>-.38</td>
</tr>
<tr>
<td>Happiness</td>
<td>-.39</td>
<td>0</td>
<td>-.58</td>
<td>.33</td>
<td>-.51</td>
<td>-.11</td>
<td>.16</td>
<td>-.74**</td>
</tr>
<tr>
<td>Anger</td>
<td>.68*</td>
<td>-.51</td>
<td>.20</td>
<td>.10</td>
<td>-.50</td>
<td>.16</td>
<td>.19</td>
<td></td>
</tr>
<tr>
<td>Fear</td>
<td>-.09</td>
<td>-.82**</td>
<td>-.58</td>
<td>.33</td>
<td>-.51</td>
<td>-.11</td>
<td>.16</td>
<td>.19</td>
</tr>
<tr>
<td>Tenderness</td>
<td>-.40</td>
<td>.41</td>
<td>-.40</td>
<td>-.75**</td>
<td>-.43</td>
<td>.38</td>
<td>-.11</td>
<td></td>
</tr>
</tbody>
</table>

$N = 36$; corrected for multiple comparisons (FDR). *$p < .05$; **$p < .01$. Abbreviations: E = Exertion, V = Valence, B = Brightness, N = Noisiness, H = Happiness, S = Sadness, A = Anger, F = Fear, and T = Tenderness

**TABLE 4. Correlations Between Mean Perceptual Ratings and Spectral Parameters for Isolated Timbres**

<table>
<thead>
<tr>
<th></th>
<th>HF</th>
<th>SC</th>
<th>IN</th>
<th>SF</th>
<th>ZC</th>
<th>RO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exertion</td>
<td>.71*</td>
<td>.02</td>
<td>.30</td>
<td>-.18</td>
<td>-.09</td>
<td>-.15</td>
</tr>
<tr>
<td>Valence</td>
<td>-.33</td>
<td>-.50</td>
<td>-.48</td>
<td>-.38</td>
<td>-.41</td>
<td>-.32</td>
</tr>
<tr>
<td>Brightness</td>
<td>.09</td>
<td>-.53</td>
<td>-.50</td>
<td>-.68*</td>
<td>-.42</td>
<td>-.54</td>
</tr>
<tr>
<td>Noisiness</td>
<td>.76*</td>
<td>.67*</td>
<td>.49</td>
<td>.49</td>
<td>.69*</td>
<td>.58</td>
</tr>
<tr>
<td>Happiness</td>
<td>.03</td>
<td>-.25</td>
<td>-.37</td>
<td>-.31</td>
<td>-.07</td>
<td>-.16</td>
</tr>
<tr>
<td>Sadness</td>
<td>-.56</td>
<td>-.36</td>
<td>-.26</td>
<td>-.18</td>
<td>-.43</td>
<td>-.27</td>
</tr>
<tr>
<td>Anger</td>
<td>.68*</td>
<td>.38</td>
<td>.31</td>
<td>.14</td>
<td>.41</td>
<td>.30</td>
</tr>
<tr>
<td>Fear</td>
<td>.24</td>
<td>.56</td>
<td>.60</td>
<td>.49</td>
<td>.45</td>
<td>.33</td>
</tr>
<tr>
<td>Tenderness</td>
<td>-.37</td>
<td>-.17</td>
<td>-.90</td>
<td>-.03</td>
<td>-.27</td>
<td>-.11</td>
</tr>
</tbody>
</table>

$N = 36$; corrected for multiple comparisons (FDR). *$p < .05$; **$p < .01$. Abbreviations: HF = High Frequency, SC = Spectral Centroid, IN = Intensity, SF = Spectral Flatness, ZC = Zero-Cross, RO = Roughness.
appeared to increase linearly between the three timbre conditions, as shown in Figure 1.

**Experiment 2**

Polyphonic timbre has been defined as the “overall timbral mixture in a signal” as it progresses through time (Alluri & Toiviainen, 2010, p. 224), an aggregate of all sound generators and acoustical incidentals in a given musical context. In Experiment 2 we employed the same paradigm as Experiment 1 to compare appraisal of monophonic and polyphonic timbre perception in the form of brief excerpts (400 ms) of popular music songs. To isolate the timbral dimension from the set of brief stimuli, Experiment 2 combined approaches from both timbre perception and rapid music cognition research (or the study of music perception in very brief time intervals).

**Method**

**PARTICIPANTS AND PROCEDURE**

Subjects were the same as Experiment 1, with the exception of one subject omitted as an outlier due to a Euclidean distance of over 300 in hierarchical cluster analysis (single linkage). The procedure was also the same as Experiment 1, with a few minor alterations to account for the different stimuli.

**STIMULI**

Eighteen brief excerpts of popular music recordings were used as stimuli. Excerpts were selected to represent six genres of contemporary popular music: (1) Rock, (2) Electronic Dance Music, or EDM, (3) Hip-Hop/Rap, (4) Pop, (5) Heavy Metal, and (6) Country, with each genre represented by three excerpts (see Supplementary Materials for the list of excerpted songs). Stimuli were 400 ms, with 10 ms amplitude ramps on both ends of the signals. Loudness was equalized manually using the same method as the previous experiment.

Excerpts were chosen based on the criterion of genre representativeness, and were selected to include basic acoustic elements of the “timbral environments” (Ferrer, 2011) with which their representative genres are commonly associated. For example, rock and heavy metal excerpts included prominent distorted electric guitar and vocals, and country included common signifiers of “twang,” such as nasal vocals, fiddle, and slide guitar (Neal, in press). In order to isolate the timbral components of different singing styles without semantic confounds, excerpts with vocals were selected to exclude any identifiable words. Sources for the stimuli were recorded from the mid-1950s to the present. Rock excerpts, for example, were all from iconic songs from the “classic” era (mid 1960s to early 1970s) due to high genre identification levels achieved in a pilot test (N = 8), as opposed to contemporary rock, which exhibited ambiguity of identification in the pilot. Country excerpts originated in source recordings from the 1950s and 1960s for the same reason. Excerpts were meant to be “thin slices” brief enough that no melodic and harmonic elements, structural implications, or rhythmic patterns were discernable. A vague sense of tempo (e.g., fast or slow) could be observed in certain excerpts, but no other details were perceptible besides timbre, texture, and artifacts of the recording that might provide stylistic cues and time-period of release.

In other research, genre identification was found to be reasonably accurate at the 200–400 ms level, largely owing to timbral cues (Gjerdingen & Perrott, 2008; Krumhansl, 2010). To verify the genre representativeness criterion of our stimuli, we conducted another pilot genre identification experiment (N = 28) with a different group of participants (17 females, 11 males; age 18–22, M = 19.6, SD = 1.31; years music training, M = 4.20, SD = 4.99). Subjects were paid $10 for their participation. The pilot study yielded substantial consistency of genre identification: 82% of all categorizations were accurate. The lowest accuracy was rock, at 59%, and the highest was metal, with 95% accuracy (chance would have been 17%).

High-frequency energy, spectral centroid, spectral flatness, zero-cross rate, and roughness estimates were extracted from the signals using MIRtoolbox1.4, with similar parameters as used in Experiment 1 (12-order FFT with 50% overlap, 50 ms per frame, producing approximately 8–10 frames for each signal at a sampling rate of 44.1 kHz). Since the *mirinharmonicity* function in MIRtoolbox only handles monophonic signals, inharmonicity was omitted from this analysis. As in the previous experiment, there were significant inter-acoustic correlations among the descriptors, with the exception of roughness (Table 5).

**Results**

Moderate inter-subject consistency in the ratings was observed (M Cronbach’s α = .69). However, responses varied considerably from variable to variable: the greatest consistency was found in the noisiness scale (α = .91), indicating a fairly reliable understanding of how this semantic dimension maps onto the signals. Valence showed no consistency (α = .04), reflecting the diversity of musical likes and dislikes of the participant group.
correlated with valence, as well as the other positive qualities of brightness and tenderness, indicating the complexity of this ostensibly negative emotion in musical representation. This challenges the bipolar labeling of happy–sad employed in some paradigms of affect (Russell, 1980), since both may have positive valence in certain contexts.

These results resonate in interesting ways with the perceived emotion responses for each genre (Figure 2). It is clear, for example, that Heavy Metal was the outlier genre, with an overwhelming number of subjects (97%) hearing excerpts as conveying anger. Also of note, pop received the most happiness responses (73%), with little of any other single emotion represented, and also the highest frequency of brightness responses. Country, on the other hand, had the highest frequency of sadness and tenderness responses (32% and 50%, respectively) and the lowest exertion and noisiness ratings.

Following inter-acoustic and inter-perceptual correlations, a final correlation was performed to assess the relationship between acoustic parameters and behavioral data (Table 7). As before, generally speaking there was a consistency of direction between correlations across the acoustic parameters and perceptual categories. Thus, valence was negatively correlated with all spectral parameters investigated, as was brightness and sadness. Corroborating Experiment 1, moreover, it was clear that the perception of brightness in these excerpts differed markedly from psychoacoustic “brightness” as discussed in the timbre literature (Beauchamp, 1982; Schubert & Wolfe, 2006). This discrepancy is likely influenced by complex semantic, cultural, and even synesthetic associations. Metal, for example, was rated low in brightness despite high spectral centroid measurements, possibly because of cultural associations of metal music with darkness, both symbolically (lyrical content) and literally (album iconography, black clothes, etc.). Conversely, Pop—the brightest genre—might have been labeled as such for its emotional connotations.

### Table 5. Inter-acoustic Correlation Matrix for Brief Polyphonic Timbres

<table>
<thead>
<tr>
<th></th>
<th>HF</th>
<th>SC</th>
<th>SF</th>
<th>ZC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral centroid (SC)</td>
<td>.73**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral flatness (SF)</td>
<td>.53*</td>
<td>.75**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-cross (ZC)</td>
<td>.71**</td>
<td>.80**</td>
<td>.44</td>
<td></td>
</tr>
<tr>
<td>Roughness (RO)</td>
<td>-.22</td>
<td>-.10</td>
<td>-.22</td>
<td>.10</td>
</tr>
</tbody>
</table>

Pearson’s r correlation coefficients between 18 signals; corrected for multiple comparisons (FDR). *p < .05, **p < .01. Note: mirinharmonicity extraction tool in the MIRtoolbox is designed for monophonic signals with only one fundamental frequency. For this reason, inharmonicity was omitted from consideration in this experiment.

A Shapiro–Wilk test was performed to confirm normality of distribution of perceptual and acoustic data. Twenty-six of the 72 perceptual variables and 1 acoustic parameter were transformed according to the same procedure applied in Experiment 1 (Templeton, 2011).

Ratings for individual excerpts in each perceptual condition were averaged by subject to provide a single index of genre. We then performed a repeated-measures ANOVA to test the effect of Perceptual and Genre factors on ratings (4 x 6). The Perceptual factor main effect was significant, $F(3, 105) = 3.59, p = .02, \eta^2_p = .09$, as was Genre, $F(5, 175) = 20.35, p < .0001, \eta^2_p = .37$. Interaction effects between Perceptual/Gene factors were likewise highly significant, $F(15, 525) = 68.92, p < .0001, \eta^2_p = .66$. Bonferroni post hoc testing of the main effect for Perceptual dimension indicated that exertion was significantly different from valence, $p < .0001$, and valence from noisiness, $p = .02$. Among main effects of genre, Country showed the clearest differentiation from all others except (somewhat inexplicably) Hip-Hop.

Next, inter-behavioral correlations were generated, revealing similar patterns as Experiment 1 (Table 6). For example, exertion had a strongly correlated with valence, $r(33) = -.67$; noisiness, $r = .87$; anger, $r = .86$; and tenderness, $r = -.85$. Intriguingly, sadness was positively correlated with valence, as well as the other positive qualities of brightness and tenderness, indicating the complexity of this ostensibly negative emotion in musical representation. This challenges the bipolar labeling of happy–sad employed in some paradigms of affect (Russell, 1980), since both may have positive valence in certain contexts.

These results resonate in interesting ways with the perceived emotion responses for each genre (Figure 2). It is clear, for example, that Heavy Metal was the outlier genre, with an overwhelming number of subjects (97%) hearing excerpts as conveying anger. Also of note, pop received the most happiness responses (73%), with little of any other single emotion represented, and also the highest frequency of brightness responses. Country, on the other hand, had the highest frequency of sadness and tenderness responses (32% and 50%, respectively) and the lowest exertion and noisiness ratings.

### Table 6. Inter-perceptual Correlation Matrix for Brief Polyphonic Timbres

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>B</th>
<th>N</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valence</td>
<td>-.67**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brightness</td>
<td>-.61**</td>
<td>.71**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noisiness</td>
<td>.87**</td>
<td>-.63**</td>
<td>-.69**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happiness</td>
<td>-.36</td>
<td>.67**</td>
<td>.72**</td>
<td>-.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sadness</td>
<td>-.58*</td>
<td>.33</td>
<td>.45</td>
<td>-.76**</td>
<td>-.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anger</td>
<td>.86**</td>
<td>-.83**</td>
<td>-.88**</td>
<td>.87**</td>
<td>-.64**</td>
<td>-.58*</td>
<td></td>
</tr>
<tr>
<td>Fear</td>
<td>.38</td>
<td>-.02</td>
<td>-.16</td>
<td>.08</td>
<td>-.13</td>
<td>-.14</td>
<td>.17</td>
</tr>
<tr>
<td>Tenderness</td>
<td>-.85**</td>
<td>.36</td>
<td>-.36</td>
<td>-.74**</td>
<td>-.07</td>
<td>.69**</td>
<td>-.62**</td>
</tr>
</tbody>
</table>

$N = 35$; corrected for multiple comparisons (FDR). *p < .05, **p < .01.
(happiness), commercial appeal, cheerful, “bubblegum” quality, and so on. In short, there was no discernable relationship between brightness ratings and psychoacoustic “brightness” (as represented by high-frequency energy and spectral centroid), and in fact they were negatively correlated.

As in Experiment 1, we performed a multiple regression analysis to assess the relationship between acoustic variables and perceptual responses. The model produced erratic results, likely the result of multicollinearity of acoustic data.

Figure 3 summarizes mean ratings for the six genres in the four perceptual quality conditions. As expected, perceptually and acoustically noisier genres (Heavy Metal, Rock, and EDM) tended to be heard as more physically effortful to produce than the other three genres (Pop, Hip-Hop, Country). As in Experiment 1, signals with higher exertion scores tended to elicit lower valence ratings. Further, it is clear that the primary source of interactive variance is the opposite trending of Metal and Country (and to a lesser extent Hip-Hop). Also echoing Experiment 1, brightness revealed a good deal of ambiguity.

**Comparison with Experiment 1**

These results complement the major findings of Experiment 1, indicating that polyphonic and monophonic timbre share similar perceptual and acoustic properties among the basic perceptual and acoustic features assessed. For example, both experiments demonstrated that timbres considered noisy are consistently heard as indicating greater bodily exertion, and are strongly correlated with frequency of anger responses. Noisiness appears to be determined largely by these acoustic parameters, a finding that is consistent with vocal theories of timbral affect (Juslin & Laukka, 2003; Laukka, Juslin, & Bresin, 2005; Scherer & Oshinsky, 1977).

To visualize the degree of overlap in results, Figure 4 compares correlation coefficients between conditions in both experiments. To assess the significance of differences between correlations in the two experiments, we transformed Pearson’s $r$ values to $z$-scores using the Fisher $r$-to-$z$ transform (Fisher, 1970). As shown in the Supplementary Materials, perceptions of exertion produced the most statistically significant differences in correlations. This is understandable given the vast differences in implied action and affective agency between isolated sounds performed by a single musician (Experiment 1) and music performed by ensembles in “real world” contexts (Experiment 2). Responses to the valence scale were mostly similar between the two experiments, and differences in noisiness correlations were likewise largely insignificant, particularly with perceived emotion categories. Taken together, then, the first two experiments
produced remarkably consistent results given the pronounced differences in stimuli type and duration.

Finally, to further compare results from the two experiments and examine underlying perceptual similarities among the stimuli, we subjected ratings data to a multidimensional scaling analysis (MDS). The solution returned a very good fit in two dimensions for both sets of stimuli (together accounting for 99% of variance), as shown in Figure 5. For Experiment 1 stimuli (Figure 5A), sound generators clustered together in a circular distribution, despite the differences between the three timbral conditions. The first dimension seems to roughly correspond to negative valence, while the second dimension is more difficult to characterize. A possible label might be voice-like, with high loadings on the vocal signals and aerophones. Thus, saxophone 3 is negatively valenced and moderately voice-like, guitar 1 is positively valenced and not particularly voice-like, and the vocal signals span from positive to highly negative but are all (it goes without saying) quite voice-like.

Perceptual ratings from Experiment 2 (Figure 5B) also produced a good two-dimensional fit, with similarities clustering around genre. The most prominent dimension here corresponds to negative valence as well, with Heavy Metal excerpts clustering at one end of the spectrum and Pop, Country, and Hip-Hop at the other. (One

of the excerpts in the Pop cluster is actually an EDM excerpt that was ‘‘heard’’ as Pop, hence the asymmetry between these two genres in the MDS solution.) This result seems to confirm intuition in many ways, indicating a large perceptual difference between Heavy Metal and the other genres, especially Hip-Hop and Country. However, the close similarity between Country and Hip-Hop among the perceptual, affective, and semantic variables addressed here is somewhat mystifying. In contrast to Experiment 1, moreover, voice-like fails to capture the similarity described in Dimension 2. The upper-left quadrant of the MDS solution corresponds to low perceived exertion, but little else might seem to place these genres into close proximity.

In sum, Experiment 2 demonstrated that 400 ms is enough time to elicit consistent perceptual evaluations of selected popular music genres, suggesting the primacy of timbre in split-second appraisals of music (Filipic et al., 2010; Gjerdingen & Perrott, 2008; Krumhansl, 2010; Schellenberg et al., 1999). Furthermore, correlations between the behavioral and acoustic attributes of noisy polyphonic timbre appeared to be largely consistent with results found for isolated monophonic timbres, as explored in Experiment 1; that is, listeners tended to hear noisy timbral snapshots as physically effortful, disliked, and expressive of anger and, to a lesser extent, fear. The specific neural mechanisms responsible for this apparent underlying affective-motor structure were investigated in our third and final experiment.

Experiment 3

PARTICIPANTS
Fifteen participants (8 female, 7 male; 18–20 years old, \(M = 19.1, SD = 0.72\); years music training, \(M = 3.27, SD = 1.44\)) were recruited from the population involved in the first two experiments, representing a random sample of the other studies. Subjects had a range of musical backgrounds, from no formal training to over 10 years. Participants were ethnically diverse (6 white, 4 east Asian, 3 south Asian, 2 black), right-handed, had normal or corrected-to-normal vision, no history of neuropsychiatric disorder, and were not taking psychoactive medication. Written informed consent was obtained from all participants, and they received $50 compensation upon completion of the scan. The experiment was approved by the UCLA Institutional Review Board.

STIMULI
The identical set of twelve 1.8–2.1 s signals detailed in Experiment 1 were used in the imaging study.

MRI PROCEDURE
Subjects listened to the randomized stimuli while being scanned. Audio stimuli were timed and presented with Presentation software through noise-cancelling, magnet-compatible SereneSound headphones. In addition, subjects wore disposable foam earplugs to help minimize scanner noise. A sound check prior to the functional scan (conducted with the scanner running) allowed subjects to adjust the headphone volume to a subjectively determined comfortable listening level. Participants were then instructed to relax and keep their heads still while listening to the stimuli, and to keep their eyes open and their vision trained on a fixation.
cross presented through magnet-compatible LCD goggles.

We used a block design consisting of an alternation of 15–16 s baseline period of silence with randomized blocks of all 12 stimuli. Each signal was repeated 5 times in a row with 100 ms of silence between each onset (11–12 s total per signal in each block). The full block took approximately 135–140 s, and was repeated 3 times for a total duration of 405–420 s plus final baseline (around 7.25 m). See Figure 6 for a diagram of the design.

**DATA ACQUISITION, PREPROCESSING, AND STATISTICS**

Images were acquired on a Siemens 3T Trio MRI scanner. Functional runs employed a continuous scanning protocol comprising 231 T2-weighted echoplanar images (EPIs) [repetition time (TR) 2000 ms; echo time (TE) 28 ms; flip angle = 90°; 34 slices; slice thickness 4 mm; matrix 64 x 64; FOV 192 x 192 mm] sensitive to blood oxygenation-dependent (BOLD) contrast. To enable T1 equilibrium the first two volumes of each functional scan were automatically discarded before data collection commenced. Additionally, two sets of structural images were acquired for registration of functional data: a T2-weighted matched-bandwidth high-resolution scan with the same slice prescription as the EPI [repetition time (TR) 5000 ms; echo time (TE) 34 ms; flip angle = 90°; 34 slices; slice thickness 4 mm; matrix 128 x 128; FOV 192x192 mm]; and a T1 weighted magnetization prepared rapid-acquisition gradient echo image (MPRAGE) [TR, 1900 ms; TE 2.26 ms; flip angle = 9°; 176 sagittal slices; slice thickness 1 mm; matrix 256 x 256; FOV 250 x 250 mm].

Image preprocessing and data analysis were performed with FSL version 5.0.4. Images were realigned to the middle volume to compensate for any head motion using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002). Volumes were then examined manually for gross motion artifacts that cannot be corrected with simple realignment. When motion artifacts were detected, a nuisance regressor for each affected volume was included in the general linear model. Additionally, one run for one subject was excluded for excessive motion (more than 10% volumes exhibiting motion artifacts). Data were temporally filtered with a high-pass filter cutoff of 100 s and spatially smoothed with an 8 mm full width half maximum Gaussian kernel in three dimensions.

Statistical analyses were performed at the single subject level using a general linear model (GLM) with fMRI Expert Analysis Tool (FEAT, version 6.00). Separate regressors modeled the 12 individual stimuli. Task regressors were convolved with a canonical double-gamma hemodynamic response function. Contrasts included the following: (1) all timbres > baseline, which was used as an ROI for subsequent contrasts, (2) each of the 12 individual stimuli > baseline, (3) intra-instrument comparisons (e.g., Guitar 3 > Guitar 1), (4) inter-timbre comparisons (e.g., all condition 3 > all condition 1), and (5) each instrument > others (e.g., voice > others). First-level contrast estimates were
computed for each run and then registered to standard space (Montreal Neurological Institute, MNI) in three stages. The middle volume of each run of individual EPI data was registered first to the co-planar matched-bandwidth high-resolution T2-weighted image. Following this, the co-planar volume was registered to the T1-weighted MPRAGE. Both of these steps were carried out using FLIRT (affine transformations: EPI to co-planar, \( df = 6 \); co-planar to MPRAGE, \( df = 6 \)) (Jenkinson et al., 2002). The resulting transformation matrix was applied to the EPI volumes. Registration of the MPRAGE to MNI space (FSL’s MNI Avg152, T1 2x2x2mm) was carried out with FLIRT (affine transformation, \( df = 12 \)). Contrast estimates for each subject were then computed treating each of the three runs as a fixed effect. Finally, a group-level analysis was performed to calculate a group mean for each contrast treating each subject as a random effect using FSL FLAME (FMRIB’s local analysis of mixed effects) stage 1 and stage 2 (Beckmann, Jenkinson, & Smith, 2003). All images were thresholded at \( Z > 2.3, p < .01 \), corrected for multiple comparisons using cluster-based Gaussian random field theory controlling family-wise error across the whole brain at \( p < .05 \) (Forman et al., 1995).

**Results**

A region of interest (ROI) was determined from the task > baseline contrast, and subsequent contrasts were masked for this ROI. In addition to basic group-level contrasts, behavioral ratings were added as covariates to assess the neural correlates of *exertion, negative valence,* and *noisiness.* Due to strong loadings on *negative valence* found in the MDS, we opted to invert *valence* ratings from positive to negative in this analysis; also, *brightness* was not included due to incoherent results from the other experiments. In addition, gender, number of years of music training, and social-cognitive data (Wallmark et al., submitted) were factored into a GLM as covariates in order to determine if they significantly affected brain activity in the task > baseline ROI. Music training has been shown to modulate functional connectivity between auditory and motor areas (Palomar-Garcia, Zatorre, Ventura-Campos, Bueicheku, & Avila, 2016). An effect of gender was discovered in the brainstem, and for music training in the cerebellum; however, no significant effects for either variable were observed in the task > baseline ROI, including motor areas. Owing to scheduling complications we were only able to gather behavioral data from 10 of the subjects who were scanned (6 female; age \( M = 19.3, SD = 0.67 \); music training \( M = 3.5, SD = 1.58 \)).

In a repeated-measures ANOVA on behavioral data with Perceptual, Instrument, and Timbre factors among this smaller participant group (\( N = 10 \)), significant main effects were found for Instrument, \( F(3, 18) = 3.34, p < .05, \eta^2_p = .36 \); and Timbre, \( F(2, 12) = 7.51, p < .01, \eta^2_p = .56 \). The Perceptual variables of *exertion,* *negative valence,* and *noisiness* did not produce a significant main effect. Results of post hoc testing (Bonferroni) of Timbre means revealed statistically non-significant differences between the medium (2) and high-noise (3) conditions, \( t(9) = -0.8, p = .44 \). For this reason, we subsequently discarded the ordinality assumption of three levels of timbral noisiness in favor of a binary “normal” versus “noisy” organization of our group brain analyses.

Beginning with the task > baseline ROI (as seen in Figure 7), timbre listening generated activation across a large set of brain areas. Corroborating numerous previous studies, we found activity in areas of bilateral primary auditory cortex, including superior temporal gyrus, superior temporal sulcus, and Heschl’s gyrus (Alluri et al., 2012; Caclin et al., 2006; Halpern et al., 2004; Menon et al., 2002; Pantel et al., 2001; Platel et al., 1997; F. Samson et al., 2011; S. S. Samson & Zatorre, 1994). Less predictably, however, the timbre task revealed activity in a range of areas not typically associated with the processing of such low-level aural stimuli. For the baseline contrast, this included motor areas (supplementary motor area, premotor cortex), secondary somatosensory cortex (SII), limbic areas (bilateral amygdalae, hippocampus), and cerebellum. Motor activation is possibly attributable, at least in part, to the repetitive periodic onset of stimuli, which generated rhythmic regularity. The processing of a periodic beat has been found to recruit motor areas (Grahn, 2012). However, motor activation varied between the stimuli despite the same rhythmic regularity of stimuli onsets, indicating the role of timbre in sensorimotor activation. A complete list of brain regions active in this and other selected contrasts can be found in Table 8.

Vocal signals elicited a unique neural signature relative to the instruments (Belin, 2006). In addition to increased activation of auditory regions, vocal signals modulated activity in inferior parietal lobule and cerebellum to a greater degree than instrumental sounds (see Table 4a, voice > instruments). We next explored the effect of noisiness on the processing of vocal timbre, and found that “normal” vocal stimuli elicited higher levels of activation in bilateral temporal, limbic, and cingulate regions than the noisy versions, which produced no significant activations (Figure 8). Supporting the conclusions of J. E. Warren et al. (2006), who found
greater activity in sensorimotor mirror areas during perception of positively valenced non-verbal vocalizations compared to negative emotions, our results suggest heightened multisensorial engagement with the sound of the *pleasant* voice compared to negatively valenced acoustic qualities. Activity in bilateral visual areas was an intriguing finding as well, possibly reflecting enhanced imagery for pleasant vocal timbre.

Contrasts between passive listening tasks reveal similarities and differences in the processing of timbral stimuli, but not the neural dynamics underpinning how listeners *appraise* timbral stimuli. In order to investigate the brain regions involved in negative impressions of timbral qualities, *exertion*, *negative valence*, and *noisiness* ratings were added as covariates in the GLM for a second-level analysis, with one value entered per stimulus per subject. Since mean behavioral ratings were of comparative value only, for this analysis we opted instead to use the difference between ratings for the two timbre conditions as a measure of *intensity* of response. This allowed us to assess the neural correlates of the degree of affective polarity between two sets of timbres, rather than just their relative means. In Experiments 1 and 2, *negative valence* and *noisiness* were strongly correlated; however, among this smaller group the correlation was found to be moderate and non-significant, $r(8) = .52, p = .13$. We therefore carried out separate analyses on these two behavioral variables.

Activity in bilateral limbic regions, particularly the amygdalae and hippocampi, covaried with the intensity to which participants felt certain timbres to be “noisy” in comparison to normal signals (see Table 9). Though activation was found in both hemispheres, results were right-lateralized. This limbic result was found for all noisy timbres in aggregate (conditions 2 and 3), and for condition 2 alone. Somewhat unexpectedly, timbre condition 3 did not produce any patterns of covariance: although these two conditions were not significantly different according to behavioral ratings, timbre 2 (the “medium” noise condition) produced more robust differences of BOLD signal change than the “high” noise condition. A similar pattern was also observed in results of the *negative valence* covariate. The reason for this discrepancy in processing between the two conditions is not immediately apparent; it is likely the result of higher-level acoustic and perceptual attributes not addressed by the fairly simple features operationalized in the current study. Since the signals were not systematically manipulated to control for different acoustic attributes, perceptual ordinality of noisiness levels was a post hoc finding that held in Experiment 1 (see Figure 1) but not for subjects in this experiment. As discussed previously, the psychoacoustics of timbral noisiness is situational, multidimensional, and non-linear. The difference in brain activation corresponding to perceived *noisiness* and *negative valence* of these two conditions therefore appears to be one of kind rather than degree. Also of note, the perception of *exertion* did not appear to be significantly correlated with hemodynamic response.

Intensity of *negative valence* in response to noisy timbres was associated with activity in a broad range of brain areas (see Table 9). Figure 9 displays selected views of regions that modulated according to the degree to which participants disliked the noisy over the normal stimuli: the greater the self-reported *negative valence*, the more activity we observed in sensorimotor areas, visual areas, insula and putamen, bilateral amygdalae, and brain stem. The lower-right image in Figure 8 is a cut-away showing limbic areas involved in both appraised *negative valence* and *noisiness*, as well as where these covariates overlap. To confirm this result was not driven by outliers, we extracted parameter estimates (beta values) from peak voxels of selected areas of
Table 8. Results of Selected Contrasts

<table>
<thead>
<tr>
<th>Brain region</th>
<th>Cluster p value</th>
<th>Extent (# voxels)</th>
<th>MNI coordinates of peak voxels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td><strong>Task &gt; Baseline</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Left parietal and temporal</em></td>
<td>&lt;.00001</td>
<td>22,474</td>
<td>−68</td>
</tr>
<tr>
<td>Superior temporal cortex</td>
<td></td>
<td></td>
<td>−46</td>
</tr>
<tr>
<td>Premotor cortex/SMA</td>
<td></td>
<td></td>
<td>−62</td>
</tr>
<tr>
<td>Superior temporal gyrus</td>
<td></td>
<td></td>
<td>−58</td>
</tr>
<tr>
<td>Secondary somatosensory</td>
<td></td>
<td></td>
<td>−56</td>
</tr>
<tr>
<td><strong>Right parietal and temporal</strong></td>
<td>&lt;.00001</td>
<td>14,749</td>
<td>48</td>
</tr>
<tr>
<td>Planum polare</td>
<td></td>
<td></td>
<td>58</td>
</tr>
<tr>
<td>Secondary somatosensory</td>
<td></td>
<td></td>
<td>62</td>
</tr>
<tr>
<td>Insula</td>
<td></td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>Superior temporal gyrus</td>
<td></td>
<td></td>
<td>48</td>
</tr>
<tr>
<td><strong>Voice &gt; Instruments</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Left temporal</em></td>
<td>&lt;.00001</td>
<td>9,645</td>
<td>−66</td>
</tr>
<tr>
<td>Superior temporal gyrus</td>
<td></td>
<td></td>
<td>−58</td>
</tr>
<tr>
<td>Planum temporale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Right temporal</strong></td>
<td>&lt;.00001</td>
<td>7,578</td>
<td>66</td>
</tr>
<tr>
<td>Superior temporal gyrus</td>
<td></td>
<td></td>
<td>58</td>
</tr>
<tr>
<td>Inferior parietal lobule</td>
<td></td>
<td></td>
<td>52</td>
</tr>
<tr>
<td><strong>Left cerebellum</strong></td>
<td>&lt;.00001</td>
<td>2,123</td>
<td>−16</td>
</tr>
<tr>
<td>Crus I</td>
<td></td>
<td></td>
<td>−18</td>
</tr>
<tr>
<td><strong>Right cerebellum</strong></td>
<td>.0006</td>
<td>1,056</td>
<td>28</td>
</tr>
<tr>
<td>Crus I</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Crus II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td><strong>Normal voice (1) &gt; Noisy voice (2+3)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Right occipital</em></td>
<td>&lt;.00001</td>
<td>2,610</td>
<td>22</td>
</tr>
<tr>
<td>Occipital cortex</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Visual cortex V1</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Visual cortex V2</td>
<td></td>
<td></td>
<td>−2</td>
</tr>
<tr>
<td>Precuneous</td>
<td></td>
<td></td>
<td>22</td>
</tr>
<tr>
<td><strong>Left limbic and temporal</strong></td>
<td>&lt;.00001</td>
<td>1,789</td>
<td>−30</td>
</tr>
<tr>
<td>Hippocampus</td>
<td></td>
<td></td>
<td>−64</td>
</tr>
<tr>
<td>Superior temporal gyrus</td>
<td></td>
<td></td>
<td>−28</td>
</tr>
<tr>
<td>Amygdala</td>
<td></td>
<td></td>
<td>−50</td>
</tr>
<tr>
<td><strong>Left occipital</strong></td>
<td>.0002</td>
<td>1,398</td>
<td>−34</td>
</tr>
<tr>
<td>Visual cortex V2</td>
<td></td>
<td></td>
<td>−38</td>
</tr>
<tr>
<td>Lateral occipital cortex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bilateral cingulate</strong></td>
<td>.005</td>
<td>904</td>
<td>14</td>
</tr>
<tr>
<td>R cingulate gyrus</td>
<td></td>
<td></td>
<td>−10</td>
</tr>
<tr>
<td>L cingulate gyrus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Right parietal and occipital</strong></td>
<td>.03</td>
<td>664</td>
<td>54</td>
</tr>
<tr>
<td>Inferior parietal lobule</td>
<td></td>
<td></td>
<td>54</td>
</tr>
<tr>
<td>Superior parietal lobule</td>
<td></td>
<td></td>
<td>46</td>
</tr>
</tbody>
</table>

N = 15. Significant voxels were obtained at a threshold of Z > 2.3, p < .01 (cluster-corrected, p < .05). Brain region labels for all MNI coordinates based on the Juelich Histological Atlas (Mazziotta et al., 2001). Reverse contrasts (e.g., Instruments > Voice) did not reveal any significant activations.
interest and plotted them against behavioral ratings. Figure 10 displays correlations of activity in six selected regions with associated negative valence and noisiness values. No outliers were found based on the criterion of $\pm 3$ SDs from the mean.

Although cortical activation was bilateral for many regions, the right hemisphere was generally dominant in negative valence, which is consistent with studies on the lateralization of affective response (Alfano & Cimino, 2008). Recruitment of somatosensory areas has

**TABLE 9. Results of Selected Contrasts Correlated with Behavioral Data**

<table>
<thead>
<tr>
<th>Brain region and covariate</th>
<th>Cluster $p$ value</th>
<th>Extent (# voxels)</th>
<th>MNI coordinates of peak voxel</th>
<th>$Z$ max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy (2) &gt; Normal timbre (1) correlated with perceived noisiness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right limbic</td>
<td>.002</td>
<td>836</td>
<td>Amygdala</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Temporal fusiform cortex</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Inferior temporal gyrus</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Parahippocampal gyrus</td>
<td>24</td>
</tr>
<tr>
<td>Left limbic</td>
<td>.002</td>
<td>821</td>
<td>Inferior temporal gyrus</td>
<td>$-42$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hippocampus</td>
<td>$-24$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Temporal fusiform cortex</td>
<td>$-36$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Amygdala</td>
<td>$-32$</td>
</tr>
<tr>
<td>Noisy (2) &gt; Normal timbre (1) correlated with negative valence</td>
<td>&lt;.00001</td>
<td>14,901</td>
<td>Bilateral temporal, limbic, sensorimotor, and insular regions</td>
<td></td>
</tr>
<tr>
<td>L. amygdala</td>
<td>$-24$</td>
<td>$-6$</td>
<td>$-10$</td>
<td>4.03</td>
</tr>
<tr>
<td>R. insula</td>
<td>38</td>
<td>$-6$</td>
<td>6</td>
<td>3.96</td>
</tr>
<tr>
<td>L. planum temporale</td>
<td>$-54$</td>
<td>$-22$</td>
<td>2</td>
<td>3.94</td>
</tr>
<tr>
<td>SMA</td>
<td>2</td>
<td>$-14$</td>
<td>70</td>
<td>3.94</td>
</tr>
<tr>
<td>R. superior parietal lobule</td>
<td>18</td>
<td>$-40$</td>
<td>50</td>
<td>3.89</td>
</tr>
<tr>
<td>L. secondary somatosensory</td>
<td>$-42$</td>
<td>$-30$</td>
<td>22</td>
<td>3.86</td>
</tr>
<tr>
<td>Bilateral occipital</td>
<td>.0007</td>
<td>889</td>
<td>L. visual cortex V2</td>
<td>14</td>
</tr>
<tr>
<td>L. visual cortex V1</td>
<td>2</td>
<td>$-96$</td>
<td>4</td>
<td>3.44</td>
</tr>
<tr>
<td>R. visual cortex V2</td>
<td>$-6$</td>
<td>$-96$</td>
<td>22</td>
<td>3.43</td>
</tr>
<tr>
<td>R. visual cortex V1</td>
<td>$-6$</td>
<td>$-98$</td>
<td>12</td>
<td>3.42</td>
</tr>
</tbody>
</table>

$N = 10$. Significant voxels were obtained at a threshold of $Z > 2.3, p < .01$ (cluster-corrected, $p < .05$). Brain region labels for all MNI coordinates based on the Juelich Histological Atlas. Reverse contrasts (e.g., Normal > Noisy timbre) did not reveal any significant activations.

**FIGURE 8.** Normal voice > Noisy voice. Statistical parametric map representing fMRI signal change (Z score) in selected axial views. Images thresholded at $Z > 2.3, p < .01$. From left to right: SMA; Broca’s area, insula, and thalamus; limbic regions (amygdalae, hippocampus) and brain stem; and cerebellum and fusiform face area (FFA). (See color version of figure online).
be shown to correspond with perception of music-related action sounds (Keysers, Kaas, & Gazzola, 2010; Koelsch, Fritz, Cramon, Müller, & Friederici, 2006), though this association has not yet been established for features of timbre alone. In addition, putamen showed significant mean correlations. The role of putamen here resonates with the results of Alluri et al. (2012), who found a preferential involvement of this basal ganglia structure in processing timbral “brightness.” It is suggestive to note that putamen has been implicated in aversive auditory reactions (Zald & Pardo, 2002) as well as strongly negative social-cognitive responses (Zeki & Romaya, 2008).

**Psychophysiological Interaction Analysis (PPI)**

Finally, to test whether functional interactions between sensorimotor, insula, and limbic regions exist—a central tenant of the same hypothesis (Overy & Molnar-Szakacs, 2009)—and are correlated with negative valence ratings, we examined psychophysiological interactions (PPI) between noisy > normal data masked for task and negative valence ratings. To create the ROIs, time series were extracted from 10 mm diameter seeds created in standard (MNI 152) space centered on voxels based on maxima in regions where task > baseline and noisy > normal contrasts overlapped. Center voxel coordinates and noisy > normal contrast used for the PPI analyses are as follows: SMA ($x = 4, y = -14, z = 72$), right SII ($x = 50, y = -24, z = 20$), left SII ($x = -44, y = -28, z = 20$), left insula ($x = -34, y = -24, z = 12$), right amygdala ($x = 18, y = -8, z = 18$), and left amygdala ($x = -22, y = -8, z = -10$). For each analysis, we modeled activity using the following explanatory variables: Psychological, consisting of a boxcar function modeling the onset and duration of each task condition, convolved with a canonical double-gamma HRF; physiological, consisting of the ROIs time series (one per analysis); and a psychophysiological interaction (PPI) for each task condition, modeling the interaction between the expected BOLD response to each condition.
and the time series of interest. These separate PPI estimates were then contrasted at the group level. This allowed us to test for voxels that displayed significant changes in correlation with the time series of the ROI for any task contrast of interest, an approach known as generalized PPI, or gPPI (McLaren, Ries, Xu, & Johnson, 2012). gPPI analyses employed somatomotor regions (SMA, right SII, and left SII) and limbic-insula relay regions (left insula, left amygdala, and right amygdala). These sets were specifically chosen to test for functional connectivity between areas of positive and negative correlation with negative valence ratings.

First-level analyses were carried out using FEAT. A second-level analysis was carried out to examine correlations between PPI parameter estimates and negative valence ratings. Resultant images were cluster corrected at a Z-threshold of 2.3 (p value cutoff = .05), using FLAME 1.

PPI analysis of SMA, SII, and amygdalae did not yield significant results. However, left insula revealed
a reduction in connectivity with the right premotor cortex (−PPI), as shown in Figure 11 (cluster extent = 326 voxels, \( p = .04; Z \text{ max} = 2.98 \) at coordinates \( x = 28, y = -8, z = 60 \). The more intensely participants disliked noisy timbral qualities compared to normal versions of the same sound generators, that is, the more connectivity between these two regions diminished. If we assume that premotor areas support a simulation process that internally “mimics” the production of the sound, as we have been arguing, and that the insula is involved in affective response, reduced connectivity between these areas may represent an implicit control mechanism to attenuate the unpleasantness of the experience.

Discussion

In this study we demonstrated a motor component to timbre perception, specifically in relation to timbral qualities perceived as “noisy.” Behavioral, acoustic, and neuroimaging evidence support and extend the “shared affective motion experience” (SAME) hypothesis of Overy and Molnar-Szakacs (2009) at the basic level of noisy versus normal timbres. We verified certain physical correlates of perceptually noisy timbres in both isolated monophonic timbres (Experiment 1) and brief polyphonic timbres (Experiment 2), finding that high-frequency energy, spectral centroid, inharmonicity, spectral flatness, zero-cross rate, and auditory roughness are reliable indicators that a timbre will be perceived as the byproduct of high bodily exertion, will be disliked (low valence), considered “noisy” (high noisiness), and associated with high arousal/low valence primary emotions (anger and fear). This inventory of low-level spectral features is by no means exhaustive, but it does point to a set of fundamental acoustic parameters that betoken the perception of noisiness. Results largely confirm our hypothesis by indicating a set of acoustic parameters that consistently correlate with impressions of exertion, negative valence, noisiness, anger, and fear. For some sound generators, higher perceived exertion reflected actual increased effort involved in the production of noisier timbres (e.g., voice), though the pattern also held for generators that did not require additional exertion (e.g., guitar). This suggests that while ecological contingencies of sound production play a salient role in perception, acoustic characteristics alone, in certain instances, may be enough to signal intensity of physical exertion.

In Experiment 3, we found that isolated noisy timbres modulated activity in motor areas as an effect of negative valence, which we interpret as evidence of motor resonance to the physically effortful and affectively intense sound production of others. There are ethological reasons to postulate that noisy timbres would elicit a more vehement motor and limbic response than non-noisy timbres (Blumstein, Bryant, & Kaye, 2012). Noisy timbres possess enhanced salience owing in part to the physical arousal correlated with their production, typically in moments of heightened anger or alarm. According to findings in the neurophysiology of act-sound coupling, listening to sounds that are part of one’s motor repertory can involve co-representation of similar motor pathways; it would also seem that ecological relevance and urgency plays a role in modulating the degree to which we mirror the heard actions of others (Aglioti & Pazzaglia, 2010; Kohler et al., 2002). SMA has been implicated in internally generated movement, coordination of action sequences, and postural stability (Nguyen, Breakspear, & Cunnington, 2014), in addition to vividness of auditory imagery (Halpern et al., 2004; Lima et al., 2015), while the pre- and primary motor cortices are the main regions involved in action planning and execution.

EMBODIED TIMBRE PERCEPTION

Taken together, our results could reflect a neurophysiological propensity to link timbral qualities with their associated actions. This novel finding broadens the discourse of embodied music cognition by incorporating a parameter of musical sound that is not often interpreted through the framework of embodiment. As

FIGURE 11. Psychophysiological interaction (PPI) between left insula (−PPI variable (left insula seed) and intensity of negative valence in the noisy > normal contrast (cluster-corrected at Z threshold > 2.3, \( p < .01 \)). Juelpich Histological Atlas lists probability of this region being insula at 50% (in contrast to primary auditory cortex, at 20%) (Mazziotta et al., 2001). PPI revealed negative correlation of insula connectivity with right premotor cortex (local maxima: \( x = 28, y = -8, z = 60 \)). All coordinates reported in voxels in MNI space. (See color version of figure online)
suggested by Leman (2007), Godey and Leman (2010), Cox (2016), and others, musical expression is predicated on mimetic engagement with the gestural correlates of musical sounds, facilitated by the perceptual link between audition, action, and emotion. Research into embodied auditory-motor interactions has thus typically focused on the most fundamental of music’s gestural components: rhythm and beat (Grahn & Brett, 2007; Patel & Iversen, 2014; Popescu et al., 2004). Given the importance of periodic tactus to the temporal organization of music, such scholarly attention is more than justified. However, within the discourse of embodied music cognition there has been the tendency to equate embodiment with aspects of rhythm and pulse—indeed, the term “embodiment” is often used synonymously with coordinated or unconscious movement to a beat. Consequently, timbre has been unwittingly marginalized from many discussions of the motor dimensions of music listening. In demonstrating a role for the human sensorimotor system in timbre perception and appraisal, this is the first study to explicitly connect timbre to motor resonance.

Timbre discrimination has been localized to temporal regions of the brain involved in auditory processing (Alluri et al., 2012; Caclin et al., 2006; Halpern et al., 2004; Menon et al., 2002; Pantev et al., 2001; Patel et al., 1997; F. Samson et al., 2011; S. S. Samson & Zatorre, 1994). Imagery for timbre, moreover, has been considered a matter of the “inner ear” not the “inner voice”; that is, primarily a sensory and not a motor phenomenon (Crowder, 1989; Kalakoski, 2001). These conclusions are consonant with the information processing account of timbre (McAdams, 1993), and with the broader belief that timbre represents what musical sounds are, not what they do (Fales, 2002). Our results augment the current consensus by suggesting that timbre, in addition to its obvious auditory dimensions, makes audible the motions (and emotions) that contribute to its production, which are in turn mirrored by the listener in perception. Following this ecological, embodied interpretation, people may not just passively listen to different timbres—in some instances, they may enact and tacitly rehearse some of the underlying physical determinants of sound production, whether through subvocalization (Halpern et al., 2004; Smith, Wilson, & Reisberg, 1995), autobiographical—specific act-sound associations (Margulis et al., 2009), or other mechanisms of audio-motor coupling, with related limbic contributions.

The voice appears to be categorically distinct as a stimulus for embodied enaction, though in a different manner to other timbral stimuli. Vocal signals were found to cluster together separate from instruments in our MDS analysis (Figure 5A), and the label voice-like comprised one dimension of the solution. Furthermore, while we found motor, somatosensory, and limbic involvement as an effect of negative valence in response to noisy timbres in aggregate, we observed activation of temporal, limbic, and occipital areas in the processing of pleasant vocal timbre only, with non-significant results for the noisy voice (Figure 7). Affective valence of vocal timbre is processed earlier than vocal identification; an ERP study found that we preattentively classify voices by valence in under 200 ms (Spreckelmeyer, Kutas, Urbach, Altenmüller, & Münte, 2009). J. E. Warren et al. (2006) demonstrated preferential engagement of motor areas in the perception of positively valenced nonverbal vocal sound. Additionally, Koelsch et al. (2006) observed activation in the Rolandic operculum, an important larynx control region, while subjects listened to “pleasant” music, suggesting that listeners are more likely to subvocalize to music they like over music they dislike. Although our normal > noisy voice contrast did not reveal activity in the operculum, increased activity in the SMA, which is an important region for movement onset, may indicate motor resonance in response to the pleasant vocal sounds of others (Berlucchi & Aglioti, 1997; Mukamel, Ekstrom, Kaplan, Iacoboni, & Fried, 2010), though not necessarily of vocal control regions specifically. Our result thus appears to confirm that we are more neurophysiologically sensitive to the sound of voices (in general) than non-voices, and to positively valenced voices more specifically. Why instruments would be more selective to negative valence, however, requires further study.

Limbic and Insula Response to Noisy Timbre

Finally, completing the hypothesized SAME music-affective network, we observed activation in amygdala and hippocampus when subjects listened to the “noisy” timbres compared to normal versions. This confirmed our prediction of limbic recruitment, perhaps indicating greater intensity, arousal, and ecological urgency signaled by timbral qualities perceived to be “noisy” (Blumstein et al., 2012). Numerous studies have implicated the amygdala in emotion processing (Phelps & LeDoux, 2005)—particularly negative, high arousal emotions such as fear (Noesselt, Driver, Heinz, & Dolan, 2005) and aggression (Beaver, Lawrence, Passamonti, & Calder, 2008)—and the amygdala has been found to be preferentially involved in aversive responses to sound (Blood, Zatorre, Bermudez, & Evans, 1999; Kumar et al., 2012; Scott et al., 1997; Zald & Pardo, 2002). Although the...
precise role of amygdala in music emotion processing remains unclear, it is generally agreed that it acts as a “relevance detector” for ecologically meaningful percepts (Sander, Grafman, & Zalla, 2003).

Our results suggest the amygdala might play a role in the processing of “relevant” low-level acoustical features of noisy timbre, even when removed from musical context. The finding that noisy timbre drives limbic activity more so than normal timbres produced by the same sound generators may have implications for our understanding of the role of timbre in affective response to music more generally (Eerola et al., 2012; McAdams et al., 2017). We theorize that part of the pleasure derived from music with prominent noisy timbres may be a case of “limbic reversal” as theorized by Huron (2006). Noisy timbres, due to the mechanisms outlined above, might in some instances trigger a negative limbic reaction; however, upon appraisal, listeners are likely to deem these timbral qualities harmless, with limbic activity reinterpreted post hoc in a positive light. Following this interpretation, it would be interesting in further research to explore how musical preference and enculturation might modulate timbre-driven limbic activity—for example, when listeners hear acoustically noisy but well-liked timbres (e.g., fans of heavy metal).

The current study also revealed a role for the insula in negative affective response to timbre. As theorized by Overy and Molnar-Szakacs (2009) and confirmed by psychophysiological interaction analysis, insula and premotor cortex were functionally connected in the processing of noisy timbre. The importance of insula as a relay between sensorimotor and limbic areas has been discussed in many studies (Carr et al., 2003; Di Cesare et al., 2014; Phan, Wager, Taylor, & Liberzon, 2002; Wicker et al., 2003). Extending these studies, our result provides additional evidence for the connectivity of the insula with premotor areas, and offers tentative support for the role of insula in musical affect. Though Overy and Molnar-Szakacs hypothesized insula involvement irrespective of affective type, we found that insula activation was specific to negative valence, although PPI analysis demonstrated that connectivity with the premotor cortex was reduced in response to unpleasant timbres. Functional dissociation between insula and premotor cortex when participants heard noisy, disliked timbres suggests to us a possible implicit control mechanism for the down-regulation of negative affect.

In addition to activity in motor areas and the limbic-insula relay, we observed increased activation in secondary somatosensory cortex (SII) to negatively valenced timbral qualities. Activation of tactile areas has been documented in the perception of music-related action sounds (Keysers et al., 2010; Koelsch et al., 2006), though never, to our knowledge, for timbre alone. One could argue that the recruitment of the secondary somatosensory cortex in the perception of negatively valenced musical timbres may indicate the somatic mirroring of what it feels like to produce a given quality of sound (e.g., the sensation of the throat vibrating, the string pressed under finger). The degree of somatomotor activity is sensitive to the autobiography of the hearer, though it is likely that even individuals lacking experience with an instrument may construe instrumental timbre by way of analogy with the voice (Juslin & Laukka, 2003).

**TIMBRE, TACTILITY, AND METAPHOR**

What does noisy timbre have to do with the sense of touch? Multisensory interaction between hearing and touching is not unique to timbre: neurophysiological slipperiness between these sense modalities has been explained by the broad functional overlap of auditory and tactile sensory regions (Foxe et al., 2002; Schurmann, Caetano, Hlushchuk, Jousmäki, & Hari, 2006). Eitan and Rothschild (2010) found that metaphors of tactile engagement are ubiquitous in listeners’ conceptualization of pitch height, loudness, and timbre. Of these, timbre—particularly noisy timbre—is a uniquely transparent locus of tactile mapping, as reflected in common metaphoric vocabulary for qualities of sound (e.g., “harsh,” “rough,” “coarse”) (Wallmark, 2014). In addition to cross-modal interactions between hearing and tactility, moreover, metaphors involving texture (e.g., “she had a rough day” versus “she had a bad day”) have been found to activate parietal operculum (OP1 and OP3) areas of SII, regions also associated with texture perception (Lacey, Stilla, & Sathian, 2012). Could the processing of timbral qualities commonly described with tactile metaphors such as “rough” activate areas of SII also involved in feeling rough things and comprehending the metaphor “rough,” even without verbal priming?

No imaging research to our knowledge has been carried out directly investigating involvement of tactile processing areas in timbre perception, and without a localizer task, interpretations of our results must necessarily remain provisional. Nevertheless, it is suggestive that noisy timbres in our study modulated activation in the precise regions of SII implicated by Lacey et al. (2012) (see Figure 9, plot D). The involvement of SII during the perception of specifically noisy qualities of timbre suggests to us that perhaps “rough” qualities of musical sound are indeed processed with
the contribution of texture areas of the brain. If this is the case, we would posit two plausible overlapping explanations. First, the metaphorical mapping of negatively valenced tactile sensation onto noisy qualities of timbre is possibly so ingrained among native English speakers (and many other languages) as to influence perception (Zacharakis, Pastiadiis, & Reiss, 2014). We may hear certain timbres as “rough” even when not verbally prompted to do so. Second, according to the neural theory of metaphor and simulation semantics (Feldman, 2006; Fernandino et al., 2015; Gallese & Lakoff, 2005), processing metaphors grounded in a somatomotor source domain may involve covert somatomotor simulation. Applied to timbre, it could be argued that even without direct verbal mediation, tactile metaphors for timbre reflect the simultaneous processing of certain noisy timbral qualities in both auditory and somatosensory regions. This conclusion would support cognitive linguistic work on the bodily basis of conceptual metaphor (Lakoff & Johnson, 1999).

We theorize, moreover, that ecological contingencies of sound production and perception may ground timbre processing via certain shared, image schematic gestalt structures (Johnson, 1987). Sounds are perceived according to the physical dynamics underlying them (Giordano & McAdams, 2010; Handel, 1995): “rough” qualities of timbre, which are generally associated with high bodily exertion and negative valence, as demonstrated in this study, are mapped with metaphors sharing similar affective and experiential features. In short, producing “rough,” “harsh” sounds often involves rougher, harsher physical actions, which are in turn conceptualized and conventionalized through related metaphors. Although sensation precedes lexical activation, common tactile metaphors for noisy timbre may reflect the “somatic marking” of auditory perception according to the bodily mechanics and affective states betokened by a given quality of timbre (Damasio, 2000). A follow-up with explicit control of the verbal parameters and a texture localizer is of course needed to investigate this hypothesis in greater depth.

LIMITATIONS

There are a number of limitations to the present study. First, the possible influence of gender, music training, and familiarity was not accounted for in our design. There is no a priori hypothesis supporting a gender difference in timbre perception; however, the effect of music training remains unclear, and could have influenced response. Filipic et al. (2010), for example, found no difference between musicians and non-musicians in the perception of brief timbral stimuli. On the other hand, McAdams et al. (2017) reported a three-way interaction between training, pitch register, and instrument family in perceptual evaluations of timbre. Although our stimuli were controlled for pitch, it is possible that familiarity affected behavioral evaluations and neural activity (Margulis et al., 2009). Furthermore, musicians have been shown to possess greater connectivity in sensorimotor regions during music listening than non-musicians (Alluri et al., 2017).

In addition to the relatively small N and stimuli set and the lack of experimental control over acoustic parameters, moreover, behavioral ratings scales were arguably a coarse way to quantify perceived exertion, valence, noisiness, and emotion perceived. In the future, it might be profitable to measure these attributes via indirect procedures that are not as heavily mediated by language. Psychophysiological methods may also be useful in order to assess the arousal dimension of timbre without cognitive input. It could also be illuminating to incorporate temporal and spectro-temporal components into an account of timbral noisiness: although steady-state spectrum appears to be of primary importance in evaluations of noisiness and in timbre perception more generally (Elliott et al., 2013; Hajda et al., 1997), time-variant features not addressed here certainly mediate the psychodynamics of listening to noisy timbre as well.

Regarding Experiment 3, it remains unclear what patterns of area-specific motor activity are revealed by the gross motor resonance we observed. In follow-up studies, it will be valuable to assess the somatotopic mapping of this activity in order to explore motor representations in relation to participants’ backgrounds (Margulis et al., 2009); for example, do guitarists process guitar timbre with greater resonance in finger areas of the pre-/primary motor cortex than non-guitarists, whose motor resonance is reflected primarily in larynx control centers (Buccino et al., 2004)? This remains a question for further study. Finally, we must acknowledge the possible limitation of scanner noise, as well as the flaws of a block design necessitating repetition of the stimuli. In future research, we believe that sparse-sampling may allow for cleaner data acquisition of brief timbral stimuli (Perrachione & Ghosh, 2013).

Conclusion

This study used behavioral, acoustic, and neuroimaging techniques to explore some perceptual, affective features of timbre processing. By implicating a broad range of
motor and emotional circuitry in the perception of timbre, this study contributes to our understanding of musical processing as a thoroughly embodied, multimodal process involving the contribution of a range of motor, somatosensory, and limbic regions. Motor regions have been implicated in rhythm perception and the processing of music in general (Zatorre et al., 2007), but have never conclusively been shown to play a role in timbre perception. We show here that motor resonance is involved in the processing of timbre, specifically “noisy” timbral qualities that reflect higher exertion, low valence body-affective states.

The study has a number of implications for our understanding of the neuroscience of musical affect. Somatomotor resonance was involved in processing timbres that generated strongly negative appraisals—the more intensely a listener disliked a given timbral quality, the more involvement we observed in the somatomotor-insula-limbic pathway partially hypoth-

ized in the SAME model, though connectivity was attenuated between insula and motor areas. This might jostle against expectations: motor resonance has been shown to play a role in musical preference, not aversion (Schäfer & Sedlmeier, 2011; Sedlmeier et al., 2011; J. E. Warren et al., 2006). Moreover, the role of SII in aversive response to timbre is an intriguing finding that requires additional study: somatosensory regions responsible for tactile processing appear to play a role in the perception of noisy timbre, indicating a possible cross-modal role for touch areas in the affective processing of musical sound qualities often described by way of tactile metaphors. Indeed, we are inclined to believe that certain common descriptive words for timbre (e.g., rough, smooth), which appear in many languages (Wallmark & Kendall, in press), may have a basis in embodied listening. It is a novel finding that such impoverished musical stimuli (isolated 2 s timbres) would elicit a consistent motor and limbic signature: if even brief, ecologically invalid stimuli produce a motor resonance response within the restrictive confines of the MRI magnet, then it stands to reason that naturalistic timbres in musical and social context could be far more efficacious as loci for embodied music cognition.

**Author Note**

For generous support the authors wish to thank the Brain Mapping Medical Research Organization, Brain Mapping Support Foundation, Pierson-Lovelace Foundation, The Ahmanson Foundation, William M. and Linda R. Dietel Philanthropic Fund at the Northern Piedmont Community Foundation, Tamkin Foundation, Jennifer Jones-Simon Foundation, Capital Group Companies Charitable Foundation, Robson Family, and Northstar Fund. We thank our action editor Mari Terraniemi and three anonymous reviewers for their invaluable feedback on our manuscript. Additionally, we would like to acknowledge Kevin Blankenship, Katy Cross, Nina Eidheim, Robert Fink, Mark Johnson, Stephen McAdams, Marita Meyer, and Mark Samples for input and assistance at various stages of this research. Experiment 3 was sponsored by a UCLA Transdisciplinary Seed Grant.

Correspondence concerning this article should be addressed to Zachary Wallmark, SMU Meadows School of the Arts Division of Music, PO Box 750356, Dallas, Texas 75275, USA. E-mail: zwallmark@smu.edu

**References**


