Perception and Processing of Pitch and Timbre in Human Cortex

A Dissertation

SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Emily Jean Allen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Andrew J. Oxenham, Ph.D.

April, 2018

Acknowledgements

I feel incredibly fortunate to have had the opportunity to conduct research in Dr. Andrew Oxenham's Auditory Perception and Cognition Lab (APC), as well as at the Center for Magnetic Resonance Research (CMRR) at the University of Minnesota, and in Dr. Elia Formisano's Auditory Perception and Cognition Lab at Maastricht University in The Netherlands. In each of these places I got to work with wonderful people doing fascinating, innovative research.

Thank you to Dr. Magda Wojtczak for taking me on as an undergraduate assistant, for helping me discover my love of research, and for her continued support over the years. A debt of gratitude is owed to Dr. Cheryl Olman for kindling my interest in fMRI as an undergraduate and her willingness to continue mentoring me with her limitless expertise in graduate school. Thank you to everyone that I interact with at the CMRR–conducting research at such a cutting-edge facility involves a lot of complicated troubleshooting and analyses. Fortunately, there are many wonderful, curious, hard-working people there all dedicated to helping each other do incredible work. In particular, thank you to Dr. Philip Burton for countless meetings, and always being there to help me climb the steep learning curve that is fMRI analysis. Thank you to Dr. Andrea Grant for her high-field and ultra-high-field training and for helping me resolve all manner of technical issues with MRI equipment. Thank you to Alex Bratch for all those hours spent in the ultra-high field with me getting comfortable working with the oldest (and fussiest) 7T in the world. Thank you also to Dr. Kendrick Kay and his Computational Visual Neuroscience lab for inviting me to be a part of lab meetings, neuroimaging journal clubs, coding sessions, and for the many educational and exciting research discussions and brainstorming sessions we have had. I look forward to future collaborations.

An immense thank you also to Dr. Michelle Moerel for guiding me through the complex world of computational modeling and for being an amazing colleague, to Dr. Elia Formisano for

taking me on as an exchange student and for being a great advisor to me for nine months in The Netherlands. Thank you also to Drs. Federico de Martino and Agustin Lage for their indispensable assistance with analyses. Thank you also to Ingrid Johnsrude for graciously inviting me to visit her in Canada in order to pick her brain (pun intended) about auditory cortex research.

Thank you to my friends and family for being present when I needed them, hugely supportive of my decision to obtain a second degree and apply to graduate school, and for being incredibly understanding when my social life drifted out to sea. Thank you to my dad, whose passion for psychology clearly had a profound impact on me and for demonstrating the value of a strong work ethic. Thank you to my kindred spirit, my mom, for her boundless kindness, wisdom, and encouragement. Thank you also to my sister, whose unparalleled determination to, not only pursue her passions, but also transcend all expectations in everything she does, has been an incredible motivator for me obtaining a PhD and continuing to push myself every day.

To my exceptional APC lab family over the years, including *almost doctor* Kelly Whiteford, Jordan Beim, Dr. Jackson Graves, Dr. Marion David, Dr. Anahita Mehta, Dr. Dorea Ruggles, Dr. Coral Dirks, Heather Kreft, Dr. Christophe Micheyl, Dr. Kyle Walsh, Dr. Eugene Brandewie, Dr. Ningyuan Wang, Andrew Byrne, Dr. Sebastien Santurette, Sachin Rai, Ryan Irey, Akshat Arneja, Shayeste Kia, Dr. Bess Borchert, Dr. Melanie Gregan, Dr. Evelyn Davies-Venn, Dr. Bonnie Lau, Adam Loper, Dr. Sam Mathias, Dr. Simon Christiansen, Dr. Sarah Madsen, Li Xiao, Hao Lu, and Erin O'Neill, Daniel Guest, Dr. Chhayakant Patro, my R.A., Zeeman Choo, and the countless delightful undergraduates we have had. Thank you all for your kindness and camaraderie. Thank you also to the Cognitive and Brain Sciences (CAB) crew and my cohort for their wonderful friendships and commiseration throughout graduate school. Thank you especially to Dr. Juraj Mesik for providing unfaltering support and companionship–for picking me back up at the lowest points, celebrating with me at the highest points, and everything in between.

Thank you to Dr. Neal Viemeister for teaching me the foundations of psychoacoustics, and to the rest of the incredible professors I have had throughout my graduate career: Drs.

Yuhong Jiang, Wilma Koutstaal, Steve Engel, Dan Kersten, Gordon Legge, Sheng He, Paul Schrater, and Niels Waller–I feel so fortunate to have gotten to know each of them and have learned so much through their instruction. Many thanks to Drs. Peggy Nelson, Bert Schlauch, Yang Zhang, Peter Watson, Benjamin Munson, and the Speech-Language-Hearing Sciences (SLHS) department for their excellent teaching and for fueling my interest in obtaining a minor in SLHS.

Thank you to my committee members, the aforementioned Drs. Cheryl Olman, Magda Wojtczak, and Peggy Nelson–three strong women in science who have each inspired me in so many ways. Finally, the ultimate thank you goes to my advisor, Dr. Andrew Oxenham, for believing in me, encouraging me, supporting me, and for being a truly outstanding mentor all these years.

This PhD dissertation research was supported by National Institutes of Health (NIH) Grant No. R01 DC005216, the University of Minnesota's College of Liberal Arts Brain Imaging Initiative, the Erasmus Mundus Student Exchange Network in Auditory Cognitive Neuroscience, the Netherlands Organization for Scientific Research (NWO; VENI grant 451-15-012, and VICI grant 453-12-002), the Dutch Province of Limburg, the Gloria J. Randahl Fellowship, the High Performance Connectome Upgrade for Human 3T MR Scanner 1S10OD017974-01, the Graduate Summer Research (GSR) fellowship, and the Graduate Research Partnership Program (GRPP).

Dedication

This dissertation is dedicated to my family, who raised me in an environment filled with music, thus nurturing my passion for auditory perception research.

Abstract

Pitch and timbre are integral components of auditory perception, yet our understanding of how they interact with one another and how they are processed cortically is enigmatic. Through a series of behavioral studies, neuroimaging, and computational modeling, we investigated these attributes. First, we looked at how variations in one dimension affect our perception of the other. Next, we explored how pitch and timbre are processed in the human cortex, in both a passive listening context and in the presence of attention, using univariate and multivariate analyses. Lastly, we used encoding models to predict cortical responses to timbre using natural orchestral sounds. We found that pitch and timbre interact with each other perceptually, and that musicians and non-musicians are similarly affected by these interactions. Our fMRI studies revealed that, in both passive and active listening conditions, pitch and timbre are processed in largely overlapping regions. However, their patterns of activation are separable, suggesting their underlying circuitry within these regions is unique. Finally, we found that a five-feature, subjectively derived encoding model could predict a significant portion of the variance in the cortical responses to timbre, suggesting our processing of timbral dimensions may align with our perceptual categorizations of them. Taken together, these findings help clarify aspects of both our perception and processing of pitch and timbre.

Table of Contents

List of Tables

Chapter 3

Chapter 4

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 1

Prologue

Pitch and timbre play central roles in both speech and music. Pitch allows us to hear intonation in a language, and notes in a melody. Timbre allows us to distinguish the vowels and consonants that make up words, and the unique sound qualities of different musical instruments. The combination of pitch and timbre enables us to identify a speaker's voice, as well as a particular piece of music.

Though cochlear implant technology has come a long way over the past several decades, pitch and timbre perception remain quite poor in cochlear implant users (e.g., Gfeller et al., 2002; Leal et al., 2003). Before we can perfect such auditory prostheses, we must first understand how sound is processed in a fully functioning auditory system, from the ear all the way up to the cortex. A good understanding of a healthy system will better enable us to aid those with various types of auditory disorders and hearing losses. David Poeppel and Tobias Overath, in the introduction of their book, *The Human Auditory Cortex* (2012), succinctly stated that, "…it is *cortical structures that lie at the basis of auditory perception and cognition*," (pp. 2). We cannot fully understand how sounds are perceived and brought into cognition without understanding how they are processed in the cortex.

> Unfortunately, we still have some way to go. As Plack et al. (2014), wrote in their review:
>
> Although we know a great deal about the psychophysics of pitch (for reviews see de Cheveigné, 2010; Plack & Oxenham, 2005) we still do not have a clear answer to some of the most basic questions regarding the underlying physiological mechanisms. First, we do not have a definitive account of how pitch is encoded in the auditory system. Second, we do not know how the pitch code is processed by the brain to produce a unified sensation. Finally, we do not know where in the auditory pathway this processing occurs, and which populations of neurons are involved.

It is fair to say that this statement applies to timbre as well. However, thanks to advances in neuroimaging methods, progress in these areas is becoming more likely.

This introductory chapter will provide a review of the literature on pitch and timbre perception and processing, starting with a general overview of these two psychoacoustic attributes, and will focus primarily on psychophysical and neuroimaging research.

Physical Correlates of Pitch and Timbre

Pitch and timbre are not new to scientific study. A notable controversy between Georg Simon Ohm and August Seebeck about precisely what aspects of a physical sound determine the pitch of a complex tone dates back to the mid-1800s (Turner, 2009). Also around this time Seebeck correctly postulated that the strengths of the upper harmonics of a complex sound influence its timbre. This relationship between timbre and spectrum was then popularized and universally attributed to Herman von Helmholtz (Turner, 2009).

The perceptual relationship between pitch and timbre, however, is still somewhat of a mystery. It is not fully understood how these two dimensions interact or influence one another. This is largely due to the challenges that lie in determining appropriate definitions for each (Houtsma, 1997). Pitch can be defined several different ways, and timbre is defined by what it is not. Subsequently, these attributes can be operationally defined and measured in a multitude of ways, making it difficult to pool the results of various experiments in order to develop coherent conclusions.

Pitch

American National Standard Acoustical Terminology previously defined pitch as a perceptual attribute of sound that can be ordered on a scale from low to high (ANSI, 1994). This definition

felt incomplete, however, as there are other attributes that can also be ordered on a scale from low to high–loudness and vertical location in space, for example. In recent years, this definition has been updated to, “That attribute of auditory sensation by which sounds are ordered on the scale used for melody in music,” (ANSI, 2013). This aligns with the more functional definition used in earlier studies that suggests that if a sequence of stimuli can carry a melody then the stimuli have a pitch (e.g., Burns & Viemeister, 1981), although even in this definition there lies some ambiguity, as it has been shown that manipulations in other perceptual dimensions, such as brightness and loudness, can be used by listeners to recognize well-known melodies (McDermott, Lehr, & Oxenham, 2008).

The pitch percept is most closely associated with the repetition rate, or fundamental frequency (F0) of an acoustic waveform, and humans have been shown to be sensitive to pitch (as defined by the ability to recognize melodies and discriminate small differences in F0) in a range of periodicities from about 30 Hz to 5000 Hz (Attneave & Olson, 1971; Krumbholz, Patterson, & Pressnitzer, 2000; Oxenham, Micheyl, Keebler, Loper, & Santurette, 2011; Pressnitzer, Patterson, & Krumbholz, 2001). The pitch produced by sounds other than pure tones has taken on various monikers including “residue pitch,” “virtual pitch,” “low pitch,” “periodicity pitch,” “pitch-frequency,” “repetition pitch,” “synthetic pitch,” “musical pitch,” “the pitch of a complex tone,” and “the pitch of the missing fundamental” (e.g., Cariani & Delgutte, 1996). Further, pitch has been categorized into different dimensions, such as pitch chroma, and pitch height (e.g., Warren, Uppenkamp, Patterson, & Griffiths, 2003). Pitch height relates most closely to the repetition rate or fundamental frequency (F0) of a sound (e.g., a sound with a higher F0 tends to have a higher perceived pitch), whereas pitch chroma is a circular scale based upon a pitch’s location within an octave (e.g., the musical note “C”).

A pitch percept can be generated by a number of different stimulus types, including pure tones, wideband and narrowband harmonic complexes (e.g., Bendor & Wang, 2005; Micheyl,

Delhommeau, Perrot, & Oxenham, 2006), and even via certain manipulations of noise (e.g., Bilsen, 1966; Burns & Viemeister, 2014; Yost, 1996).

Timbre

Timbre, commonly referred to as the quality or color of a sound, was previously defined as everything by which a listener can distinguish between sounds with the same loudness and pitch (ANSI, 1994). In other words, it is anything that sets two sounds apart other than loudness or pitch and, arguably, other attributes often left out of definitions, such as duration, spatial location, and possibly even the reverberant qualities of an environment. Fortunately, ANSI's revised definition covers some of this: "That multidimensional attribute of auditory sensation which enables a listener to judge that two non-identical sounds, similarly presented and having the same loudness, pitch, spatial location, and duration, are dissimilar," (ANSI, 2013). Timbre has been eloquently referred to as a "waste-basket" category for anything that cannot be labeled pitch or loudness (McAdams & Bregman, 1979). Licklider (1951) aptly called timbre a "'multidimensional' dimension." Varying some of these dimensions can affect a sound's perceived "brightness", "clarity", "harshness", "fullness", and "noisiness", to name just a few (Stepanek, 2006). Unfortunately, as with most perceptual properties, timbral dimensions are difficult to quantify (Elliott, Hamilton, & Theunissen, 2013), much less label (Grey, 1977). With the vast array of dimensions that fall under the blanket definition of "timbre," it is no surprise that there are countless ways to manipulate and measure this attribute. As Grey states, "…a most important evaluation of any particular geometric mapping of similarities is its usefulness in interpreting the bases for the perceptual judgments." (pp. 2170). Put another way, the dimensions of timbre can be divided in many ways–the challenge lies in pairing these dimensions with unique, separable, percepts.

Multi-dimensional scaling

In order to measure and quantify timbral percepts, we must link these perceptual attributes to physical variables that can be manipulated when generating sounds. Given timbre's highly complex nature, or more specifically, its multidimensionality, there have been attempts to identify the perceptually salient aspects using multidimensional scaling (e.g., Grey, 1977). This approach utilizes subjective measures to determine how perceptually similar various timbral dimensions are, thus creating a geometric map that plots the subjective distances between a diverse set of stimuli as points in a space (Grey, 1977). Grey used digital additive synthesis (adding together partials controlled through time, amplitude, and frequency) to generate 16 realistic musical tones. He concluded that three dimensions best represented the perceptual relationships of his stimuli: one was related to the spectral energy distribution of the tone, and the other two were related to temporal patterns. One temporal pattern of importance was the presence of low-amplitude, high-frequency energy in the attack portion of a sound, while the other pattern was related to the synchronicity of higher harmonic transients and related spectral fluctuation through time (Grey, 1977).

Analysis of timbre by synthesis

Helmholtz, while correct about spectral content influencing timbre, believed it was specifically the steady-state portion of a sound that determined a tone's musical quality (Risset & Wessel, 1999). However, this can easily be disproven by taking sounds like piano notes, which have a sharp attack and long decay, and reversing them to have a long attack and sharp decay, significantly altering their timbres. In fact, as we now know, there is much more that contributes to the timbre of a sound than its spectrum. Although the three dimensions mentioned previously captured much of the variance between different instruments, the attempted recreation of natural

instrumental sounds via analysis through synthesis has led to the realization that more subtle aspects, such as the time course of individual harmonics, play an important role in our perception and recognition of different instruments. Conversely, some dimensions have been found to be "aurally irrelevant" and deemed unnecessary for a realistic synthesis, such as short-term amplitude fluctuations (Risset & Wessel, 1999).

Interactions

Given the previously mentioned concerns about different operational definitions for pitch and timbre, it is not surprising that the literature is mixed when it comes to how these perceptual dimensions interact. Although Marozeau et al. (2003) found pitch and timbre to be perceived independently, a later study by Marozeau and de Cheveigné (2007), acknowledging concerns about their previous study, revealed an influence of F0 on the perception of brightness (varied by altering spectral centroid). A general consensus seems to be that these two dimensions *can* influence each other (e.g., Krumhansl & Iverson, 1992; Russo & Thompson, 2005; Warrier & Zatorre, 2002). Silbert, Townsend, and Lentz (2009) explored a general framework for understanding interactions between perceptual dimensions based on signal detection theory (Green & Swets, 1966). They used concurrent changes in spectral centroid and F0 as an example of dimensional interactions and concluded that, for most of their seven listeners, the two dimensions were not processed independently. However, because they did not test identification performance for either dimension in isolation and only tested two values of each dimension, it is not clear how much interference each dimension produced on the other, or whether the effects were symmetric. It is also not clear what accounted for the relatively large individual differences observed in that study.

With some exceptions, such as the study by Silbert et al. (2009), the literature has tended to concentrate on how timbre influences pitch perception rather than the reverse. There are different hypotheses about how timbre influences pitch (e.g., Faulkner, 1985; Moore & Glasberg, 1990), but a dominating view is that changes in spectral timbre (on the dull-bright continuum) either produce a general distraction effect or are confused with changes in pitch height, based on F0 (e.g., Borchert et al., 2011; Moore & Glasberg, 1990; Singh & Hirsh, 1992; Warrier & Zatorre, 2002).

Congruence

Congruence of pitch and timbre is often related to the frequency content that is represented. If a complex tone has a high F0, and also has more energy devoted to its higher frequencies, making its timbre "brighter", "tinnier", or "sharper" (e.g., Fastl & Zwicker, 2007), for example, this would be considered a congruent pairing. If, however, a tone with a high F0 has more energy in the lower frequencies, leading to a "duller" or more "hollow" timbre, for example, this would be considered an incongruent pairing.

Melara and Marks (1990) reported the unexpected finding that subjects were significantly *slower* to respond in discrimination tasks, by about 15 ms, when the two dimensions were congruent (e.g, higher pitch was paired with a "twangy" timbre) than when they were incongruent (e.g., higher pitch was paired with a "hollow" timbre), and they found timbre judgments to be more strongly affected by pitch than the reverse. The delay in the congruent condition could indicate that the higher pitch and twangier timbre in this experiment were actually perceived as *less* congruent by subjects than higher pitch and hollower timbre. Moreover, the way in which timbre was manipulated was by varying the duty cycle, with lower duty cycles categorized as "hollow" timbres and higher duty cycles categorized as "twangy" timbres. It is possible that the duty cycles chosen for the experiment were not the best examples of "hollow" and "twangy", or

that this manipulation of timbre is not ideal for making congruence pairings with pitch. A final concern, which is shared by many studies (e.g., Beal, 1985; Pitt, 1994) is the limited number of stimuli used: a combination of two different duty cycles of square waves (0.1878 and 0.3128, labeled "twangy" and "hollow," respectively) were combined with two different F0s (900 Hz and 920 Hz), which, once again, limits the conclusions that can be drawn.

Equating for perceptual salience

Krumhansl and Iverson (1992) also found interactions between pitch and timbre for individual tones on speeded classification tasks, but used more musical sounds (notes F4 and C5 for the pitches, and a synthesized trumpet and piano for the timbres). They found that variation in the non-target dimension interfered with classification for both pitch and timbre, symmetrically. Again, however, a limitation of the study lies in the small number of stimuli used, and the fact that the differences in pitch and timbre were not equated for discriminability or perceptual salience. The importance of equating the dimensions of interest in terms of perceptual salience has been noted in both auditory and visual research by Melara and Mounts (1993, 1994).

Influence of Training and Musicianship

A study by Micheyl et al. (2006) found professional musicians to have much better difference limens (DLs) for pitch discrimination than non-musicians. However, it only took about four to eight hours of psychoacoustic training for the non-musicians to achieve performance comparable to that of the professional musicians. Musicians, however, showed little improvement with additional training, suggesting they were already at or near asymptotic performance prior to training. Little is known about differences between musicians and non-musicians in their ability to discriminate spectral shape (timbre), with or without the presence of F0 (pitch) changes. On

one hand, some benefit of musicianship in attending selectively to separate auditory dimensions beyond pitch might be expected; on the other hand, timbre discrimination may not be as highly trained in musicians as pitch discrimination because discriminating between very subtle spectral differences is not part of a typical ear-training program.

Musicians have also been found to have better performance in analytical listening in an informational masking context (Oxenham, Fligor, Mason, & Kidd, 2003). Attending to one dimension and ignoring another could be considered a form of analytic listening, so it may be that musicians are less susceptible to interference effects. In addition, reports of musicians understanding speech better in noise than non-musicians (Parbery-Clark, Skoe, & Kraus, 2009) suggest relatively generalized benefits in auditory perception, although more recent studies have failed to replicate these findings (Ruggles, Freyman, & Oxenham, 2014).

In a musical context, Beal (1985) found that musicians were better at recognizing when the same chord was played on two different instruments compared to non-musicians. However, this benefit of musicianship was only found when the chords were diatonic, suggesting that the successful referencing of familiar musical structures was the defining difference between musicians and non-musicians. In the absence of familiar musical structures and instruments, Borchert et al. (2011) found no significant benefit of musical training in a task that involved pitch discrimination between two sounds that varied widely in spectral shape.

Neural Correlates of Pitch and Timbre

Auditory Cortex

The auditory cortex is located in the superior temporal plane, which can be found within the Sylvian fissure of the temporal lobe. The three main parts of the auditory cortex are the core (i.e.,

primary auditory cortex), belt, and parabelt regions. Most projections from lower (brainstem and midbrain) auditory nuclei project to the core area, and processing seems to begin there before moving to the belt and parabelt regions. It is believed that the further out a region is from the core area, the more holistically sound is being processed (Plack, 2005, pp. 84), although our understanding of cortical auditory processing remains surprisingly limited. The primary auditory cortex (A1) is located on a convolution called the Heschl's gyrus, also known as the transverse temporal gyrus. Heschl's gyrus (HG) is named after Richard Ladislaus Heschl, an Austrian anatomist, who was the first person to describe this brain region. This area is difficult to study, partly due to the large variability in anatomical structure between subjects (e.g., Rademacher et al., 2001; Warrier et al., 2009). Humans can have between one and three HG within each hemisphere (Plack, Oxenham, & Fay, 2006, pp.153). Additionally, the location of A1 within the HG can be highly variable. Anterior to HG lies the planum polare (PP) while in the posterior direction lies the planum temporale (PT). There is some debate about which regions are considered core regions versus non-core regions in humans (Moerel, De Martino, & Formisano, 2014). Part of the issue rests on the fact that, for some people, A1 extends beyond HG, and is partially represented on PP and PT as well, while for others, non-primary areas can extend back onto HG (Clarke & Morosan, 2012).

Single-Unit Recordings

In awake, behaving macaque monkeys, single cortical neurons show a robust, systematic, spatial organization in A1 for characteristic frequencies (Recanzone, Guard, & Phan, 2000). Similar tonotopic organization can be found sub-cortically in the brainstem, midbrain, and thalamus (Saenz & Langers, 2014), reflecting the tonotopic organization established along the basilar membrane in the cochlea. Such robust tonotopic organization has not been found in belt or

parabelt areas, suggesting hierarchical processing and the extraction of features beyond the simple spectrum of the sound in these secondary cortical areas.

Bendor and Wang (2005) identified a cluster of pitch-selective neurons in the marmoset located in a region near the anterolateral border of A1 and the rostral field, and anterolateal and middle lateral nonprimary belt areas. The criteria they used to classify neurons as pitch-selective included significantly tuned responses to both pure tones and harmonic complex tones with a missing fundamental corresponding to the pure-tone frequency, but with components all outside the neuron's excitatory frequency response area. In this way, the pure tones and harmonic complexes were spectrally dissimilar (i.e., they had different timbres), but shared a common pitch. According to these findings, there exist neurons that respond to both individual frequencies and complex tones, suggesting that the processing of these sounds occurs in overlapping regions of the auditory cortices (Bendor & Wang, 2005). However, they also found neurons in this region that responded to narrowband or wideband complexes, but not to pure tones, suggesting some of these neurons are dedicated specifically to the integration of information from multiple components.

Taking a different approach, Bizley et al. (2009) used stimuli that varied in F0, spectral distribution and spatial location to identify neurons that were selective for one or more of those dimensions. Rather than finding neurons that were selective to only one of the features tested, they found instead a more distributed population code in the auditory cortices of ferrets. Over two-thirds of the units responded to at least two dimensions, most commonly pitch and timbre. In other words, there were more interactions for the two stimuli within the "what" domain than there were between the "what" and "where" domains. Additionally, azimuth ("where") sensitivity was found in deeper cortical layers, while pitch and timbre sensitivity were greater in more superficial layers. These findings suggest that pitch, location, and timbre sensitivity are interwoven and distributed across the core and belt areas of the auditory cortices (Bizley et al., 2009). A lack of

neurons responding selectively to single dimensions might predict the potential for more perceptual interference across these dimensions.

Although the auditory cortices of primates and other mammals are often researched in vivo, in hopes of drawing connections to human brains, many differences exist between them. One major difference is that, since Heschl's gyrus is relatively new, evolutionarily, monkeys such as macaques do not have them, and only a subset of chimpanzees do (Moerel et al., 2014). This, alone, is a strong argument for studying human brains, whenever possible, in order to understand how the human auditory cortex functions.

Neuroimaging

One obvious tool for attempting to identify the representations of pitch and timbre within the human auditory cortex, given its superior spatial resolution, is functional magnetic resonance imaging (fMRI). A great advantage of this tool is that it is non-invasive, and, thus, an ethically sound means for studying the human brain. fMRI has been successful for various perceptual studies, most notably in vision research, for identifying brain regions that seem to selectively process certain visual features and properties (e.g., Engel, Glover, & Wandell, 1997; Kanwisher, McDermott, & Chun, 1997). However, additional challenges arise when attempting to utilize this tool for auditory research. One of its greatest drawbacks is the sound generated by the scanner. MRI scanners are acoustically noisy devices, mainly due to the "ping" sound produced by large gradients switching rapidly during image readout (Blackman, Hall, & Kingdom, 2014), with much of the energy falling somewhere between 0.5 and 2 kHz. Additionally, the liquid helium pump, ventilation fan, and air-handling equipment all generate noise (Ravicz & Melcher, 2000).

Acoustic noise is an obvious concern when it comes to contamination of auditory stimuli. Therefore, cautionary steps must be taken, such as (1) providing the subjects with attenuating ear

buds, circumaural ear protectors, and/or noise cancelling headphones, (2) presenting stimuli at higher sound levels, (3) developing quieter pulse sequences (e.g., Idiyatullin, Corum, Park, & Garwood, 2006) and/or (4) adding silent gaps to pulse sequences, during which the auditory stimuli can be played with the least amount of acoustic interference from the scanner (Hall et al., 1999; Zaehle, Wüstenberg, Meyer, & Jäncke, 2004).

Tonotopy

Before delving into the more complex aspects of pitch and timbre, it is important to first address the more basic functional organization of human auditory cortex, starting at the level of frequency representation.

Given the relatively small size of A1, mapping its representation of frequency has been a challenge at the spatial resolution of standard neuroimaging techniques, in which a single voxel is measuring the response of hundreds of thousands of neurons (Saenz & Langers, 2014). As such, there have been debates about the precise orientation and number of the tonotopic gradients that exist in human auditory cortex. There is, however, a general agreement about a high-to-low-to-high frequency mapping spanning A1. This mirror-symmetric mapping aligns with non-human primate research and has been supported by ultra-high field strength (7T) imaging in humans (Formisano et al., 2003). This symmetric mapping also appears to be angulated or V-shaped in formation (e.g., Langers & van Dijk, 2012). However, in order to gauge the sharpness of this frequency tuning, pure tones and narrowband stimuli are most frequently used. A concern here is that studies limiting their stimuli to pure tones or narrowband stimuli cannot tell us whether it is pitch, or merely spectral content (possibly related to timbre) that drives the tonotopy observed in A1. In other words, is the tonotopy in A1 simply a reflection of the tonotopic organization found in the cochlea and auditory nerve, or have some features (such as pitch) already been extracted at

that level? With conflicting studies and inconclusive results, this remains an open question (Saenz & Langers, 2014).

"Pitch center"?

Though pitch processing has been researched more thoroughly than timbre processing, even the basic claim of there being a "pitch center" in the cortex is hotly debated (Bendor, 2012). There is growing evidence suggesting that this "center" may be located in the lateral portion of Heshl's gyrus (e.g., De Angelis et al., 2017; Norman-Haignere, Kanwisher, & McDermott, 2013; Patterson, Uppenkamp, Johnsrude, & Griffiths, 2002). One such fMRI study in search of this "center", looked for neural representations of pitch in the auditory cortex as a function of pitch salience, by manipulating the resolvability of complex tones (Penagos, Melcher, & Oxenham, 2004). Harmonics below about the 10^{th} are generally resolved, meaning they are individually represented along the basilar membrane. Conversely, those beyond the 10^{th} harmonic are unresolved, in that they are no longer individually represented, making the resulting pitch at the F0 less salient. Thus, Penagos et al. (2004) used lower spectral regions (340-1100 Hz) to produce more salient pitch percepts, and higher spectral regions (1200-2000 Hz) to produce weaker pitch percepts. The first two conditions, which contained a range of complex tones with low F0s (80-95 Hz), filtered into lower and higher spectral regions, respectively, were contrasted with conditions containing a higher F0 range (240-285 Hz). This was done so that resolved harmonics were present in both the lower and higher spectral regions, providing both conditions with relatively high pitch salience. Results revealed stronger activity during high-salience conditions relative to low-salience conditions, independent of spectral region or F0 range. Pitch salience corresponded to bilateral activation of the anterolateral end of HG, as well as some spread across the superior temporal gyrus (STG) of the left hemisphere, supporting the claim that anterior nonprimary auditory areas are important for pitch processing.

Imaging studies using iterated rippled noise

Studies have often used iterated rippled noise (IRN), due to the ease of manipulating the saliency of the pitch percept (e.g., Bilsen, 1966; Griffiths et al., 2010; Patterson et al., 2002). IRN is random noise (aperiodic) with a continuous spectrum. The noise is delayed and added back to itself in 'iterations.' This 'comb-filtered' ripple noise technique results in peaks that occur at F0 and corresponding harmonics, creating a pitch percept. If the iterations are attenuated, or the number of iterations is reduced, the pitch percept weakens (Shofner & Yost, 1997; Yost, 1996). Such manipulations provide well-controlled stimulus-versus-noise contrasts for neuroimaging purposes. However, it has been suggested that, since IRNs are frequently used as stimuli, perhaps the anatomical "center" being pinpointed in the brain is actually a region sensitive to qualities inherent to IRN itself, such as its varying spectro-temporal features, more so than pitch (Hall & Plack, 2009). In fact, Hall and Plack (2009) found a positive correlation between the strength of the spectro-temporal features of IRN and pitch saliency. Increasing the number of iterations increased both the strength of the spectro-temporal features and the pitch percept. This makes it is difficult to parse whether the lateral Heschl's gyrus is responding to pitch strength, spectro-temporal strength, or both. Barker, Plack, and Hall (2012), followed up on this, looking specifically at the spectro-temporal modulations found in IRN, by developing a "no-pitch IRN" stimulus. This stimulus contained only the modulations, and the temporal fine structure that provided the pitch-like percept was removed. As was suspected, the regions believed to be sensitive to pitch were also sensitive to these modulations, thus indicating that this putative "pitch" center may not be specific to pitch after all. Thus, the conclusions drawn from the many previous studies using IRN may warrant re-evaluation.

Multiple pitch regions?

Hall and Plack (2009) utilized a wide range of stimulus types to induce the percept of pitch, arguing, "…for a brain region to be confirmed as a general pitch center, it should respond to all pitch-evoking stimuli," (pp. 2). Their stimuli included a single frequency, a wideband complex, a resolved complex, an unresolved complex, a Huggins pitch, and iterated-ripple noise (IRN). IRN activated the lateral HG, as expected, but the other five stimulus types produced a wide range of activity throughout the auditory cortex. Though the lateral HG was not reliably activated by these stimuli, the PT was often active, suggesting that, if there is something similar to a general pitch center, it may occur later in the auditory processing stream than previously believed. Hall and Plack (2009) also point out that the multiple distinct regions found to be active for these stimuli may be regions involved in different levels of pitch processing. For example, it can be difficult to parse whether the regions being excited are responding to the frequency content of the sound, indicating lower-level processing, or pitch perception, which would be considered higher-level sound processing. What has been established is that wideband complexes, which contain a wide range of frequencies, are good for eliciting strong activation throughout a large portion of auditory cortex. Conversely, narrowband complexes and sweeps are suitable for finer brain mapping, such as tonotopy (Langers, Krumbholz, Bowtell, & Hall, 2014).

Lesion studies, PET, and fMRI

A lesion study was conducted by Samson and Zatorre (1994), in which 15 subjects had their left temporal lobes removed and another 15 had their right temporal lobes removed. In both cases the procedure was done in order to relieve medically intractable partial complex seizures. These 30 participants were compared to 15 normal controls. Two different timbre manipulations were used: varying the number of harmonics, a spectral manipulation of timbre, and varying the duration of the attack, a temporal manipulation. Based on their results, the right temporal lobe

was deemed necessary for both spectral and temporal timbre perception, as subjects with lesions in their right temporal lobes (RT group) performed worse on both tasks than the normal controls (NC group), as well as the subjects with left temporal lobe lesions (LT group). However, the authors did not report whether the LT group also performed significantly worse than the NC group. Their bar graph suggests that the control group performed best on both tasks, followed by the LT group, and finally the RT group. Based on this, it seems plausible that, in fact, both lobes may contribute to the processing of timbre.

Several years later, another lesion study was conducted, looking at pitch perception (Johnsrude, Penhune, & Zatorre, 2000). Thirty-one subjects who had undergone surgical resectioning for relief of medically intractable seizures were compared to 14 controls. There were two pitch discrimination tasks: same-different (i.e., detection a change in frequency), and up-down (i.e., labeling the direction of the change). What they found was that for same-different tasks, the patient and control groups did not differ. However, when subjects had to determine whether a tone pair was rising or falling (a more challenging discrimination task) patients who had temporal lobe excisions that encroached upon the Heschl's gyrus in the right hemisphere performed significantly worse than controls. Interestingly, this was not the case for patients who had temporal lobe excisions that encroached upon the Heschl's gyrus in the left hemisphere, suggesting the right hemisphere is important for processing pitch direction. This also suggests that hemispheric differences may emerge as tasks become more difficult (i.e., more challenging analytical listening). However, only pure tones were used for this study, making it difficult to determine whether these same results would have occurred with the use of more complex or natural stimuli.

It is important to note that such lesion studies are imprecise and difficult to control. Moreover, lesioned brains are not "normal" brains, so it is difficult to draw conclusions about

how a fully-functioning auditory system works when studying patients with brain lesions. Therefore, we will now focus on neuroimaging studies of "normal" subjects.

A PET study by Zatorre and Belin (2001) in which variations in the temporal domain, in the form of fluctuating duty cycles, were found to be processed more in the left hemisphere. Conversely, fluctuations in the spectral domain, in the form of frequency density, were found to be more right hemisphere lateralized. Schönwiesner et al. (2005), using fMRI, found bilateral activation, but with an asymmetry in the nonprimary auditory cortex for spectral and temporal modulation. Specifically, the left superior temporal gyrus was found to be more sensitive to temporal modulation, and the right superior temporal gyrus was found to be more sensitive to spectral modulation. Santoro et al. (2014), however, found the differences between the spatial patterns for spectral and temporal processing, at a much higher spatial resolution, to be a bit more complex than this. The general consensus, however, is that there exist differences in how spectral and temporal information are processed at the cortical level.

An fMRI study by Warren et al. (2005) presented sequences of sound (7.5-8s in duration) that varied either the F0 (pitch) or spectral envelope (timbre) of harmonic complexes. This was contrasted with sequences varying the spectral envelope of noise. In order to maintain the subjects' attention, they were instructed to press a button at the end of each sequence. Both variation in F0 and spectral shape showed bilateral activation spanning superior temporal regions, including A1 in the medial Heschl's gyrus (HG), the lateral HG, and anterolateral planum temporale (PT). For alternating conditions, during which the spectro-temporal structure was constantly varying, Warren et al. (2005), argued that the computation of the spectral envelope requires the abstraction of spectral shape, which is a more abstract level of analysis. Such abstract levels of processing activated temporal lobe areas beyond those that were active for detailed spectro-temporal structure. These areas seem to be rightward-lateralized (from superior temporal plane lateral to PAC, inferiorly and anteriorly along STS). The findings suggest that, although

processing of both pitch and timbre is bilateral, there may be some differences between hemispheres at higher levels of processing. This evidence is consistent with the previously found left-hemisphere dominance for speech processing and right-hemisphere dominance for music processing (Zatorre, Belin, & Penhune, 2002). A potential weakness of the Warren et al. (2005) study is that the data were collected at a 1.5 T scanner, which has lower resolution than the currently standard 3 and 7 T scanners, making it more difficult to identify fine-grained differences in processing. Additionally, the stimuli varied in discrete ways (e.g., large variability in spectral shape), instead of varying along a continuum. Finally, the dimensions were not equated for perceptual salience.

Plasticity and Musicianship

Given that musicianship and training can improve one's pitch discrimination, a logical next step is to question how this ties in with brain plasticity, structure, and function. Differences between musicians and non-musicians have been identified, in terms of differences in gray matter volume in certain regions. Specifically, musicians have been shown to have 130% larger Heschl's gyri compared to non-musicians (Schneider et al., 2002). Additionally, Schneider et al. (2002) found the activity evoked by auditory stimuli, as measured by early-latency cortical response via MEG, to be 102% larger in musicians compared to non-musicians. What this study does not tell us, however, is whether musicians are born with these structural and functional differences, or if they develop as a consequence of musical training.

A study by Hyde et al., (2009) delves into this "nature versus nurture" debate in a longitudinal study. They looked at the brain structures of young children, around six years old, and compared those who had 15 months of musical training (weekly 30-minute private keyboard lessons) to controls (who still had a weekly 40-minute group music class). Even over this brief, relatively infrequent training period they found the private lessons to result in structural changes,

in the form of voxel expansion, in both motor and anatomical areas. These anatomical changes correlated with behavioral improvements in fine finger motor skills and auditory musical discrimination tests (Hyde et al., 2009). These anatomical and behavioral differences were not seen in the instrumental or control groups prior to the private lessons and no anatomical or behavioral changes were found in the controls at the end of the 15 months. Additionally, far-transfer measures (e.g., object assembly, block design, vocabulary subtests, and even phonemic awareness tests) showed no improvement after this training, suggesting that musical training may not always be generalizable to other skills, even if auditory.

A larger, longer-term longitudinal study supports this claim of non-generalizability. Yang et al. (2014) looked at 250 Chinese students, also around six years of age. The musician group received 3.5 hours of weekly musical training for around 43 months. Upon completion of this training, musicians were significantly better than non-musicians on musical achievement, as well as development of a second language. However, the academic benefits of the musical training ended there. The improvement in a second language is intriguing, however, as it suggests that, while Hyde et al. (2009) did not find a benefit of musical training on the phonemic awareness task, musical training may, in fact, be generalizable to other language-related auditory domains. Thus, the link between musicianship and language remains fuzzy.

Summary

Based on the research discussed, psychophysical studies of pitch and timbre, which will be explored further in the following chapter, suggest that these dimensions may not be perceptually independent. Exactly how pitch and timbre interact (e.g., in the form of distraction or confusion), and how similar these interference effects are across the two dimensions, has not been sufficiently addressed in the literature. Further, while there is some evidence suggesting that musical training

may lead to brain plasticity, revealed in the form of structural, functional, and/or behavioral changes, it is unclear whether musicianship reduces interference between pitch and timbre. Moreover, how these dimensions are processed at the level of the cortex is an even bigger mystery. The question of whether there is one "center" in the auditory cortex devoted to pitch processing continues to be explored. The goal of this dissertation work is to better understand how pitch and timbre interact, perceptually, as well as how they are processed, cortically. We will use a combination of psychophysics and fMRI techniques to address these questions.

Chapter 2

Symmetric interactions and interference between pitch and timbre

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. The Journal of the Acoustical Society of America, 135(3), 1371-1379.

Abstract

Variations in the spectral shape of harmonic tone complexes are perceived as timbre changes and can lead to poorer fundamental frequency (F0) or pitch discrimination. Less is known about the effects of F0 variations on spectral shape discrimination. The aims of the study were to determine whether the interactions between pitch and timbre are symmetric, and to test whether musical training affects listeners' ability to ignore variations in irrelevant perceptual dimensions. Difference limens (DLs) for F0 were measured with and without random, concurrent, variations in spectral centroid, and vice versa. Additionally, sensitivity was measured as the target parameter and the interfering parameter varied by the same amount, in terms of individual DLs. Results showed significant and similar interference between pitch (F0) and timbre (spectral centroid) dimensions, with upward spectral motion often confused for upward F0 motion, and *vice versa*. Musicians had better F0DLs than non-musicians on average, but similar spectral centroid DLs. Both groups showed similar interference effects, in terms of decreased sensitivity, in both dimensions. Results reveal symmetry in the interference effects between pitch and timbre, once differences in sensitivity between dimensions and subjects are controlled. Musical training does not reliably help to overcome these effects.

Introduction

The sounds we hear can be described in terms of multiple perceptual attributes, including pitch, timbre, and loudness. The present study focuses on pitch and timbre. Pitch has been defined as that perceptual attribute of sound that can be ordered on a scale from low to high (ANSI, 1994), although several researchers have suggested that it is multi-dimensional (e.g., Shepard, 1982), with the two most commonly cited dimensions being pitch height and pitch chroma, corresponding roughly to the physical attributes of fundamental frequency (F0) and position within an octave, respectively. The present study focuses on the dimension of pitch height. Timbre is associated with multiple acoustical and perceptual attributes (Grey, 1977). Its technical definition includes everything by which a listener can distinguish between sounds with the same loudness and pitch (ANSI, 1994), although duration (Plomp, 1970) and spatial location are also attributes that are not normally considered part of timbre. A primary determinant of timbre is the spectral centroid of a sound (Caclin, McAdams, Smith, & Winsberg, 2005). In general, a low-frequency emphasis in the spectral envelope leads to a "duller" sound, whereas more high-frequency emphasis leads to a "brighter", "tinnier", or "sharper" sound (e.g., Fastl and Zwicker, 2007).

Although some previous studies have shown pitch and timbre to be perceived independently (e.g., Marozeau et al., 2003) there are several examples of interference between them (Melara and Marks, 1990; Marozeau and de Cheveigné, 2007). Notably, variations in timbre are known to interfere with subjects' ability to discriminate small changes in pitch. There are different hypotheses regarding how this interference occurs (e.g., Faulkner, 1985; Moore and Glasberg, 1990), but a prevailing view is that changes in spectral timbre (on the dull-bright continuum) either produce a general distraction effect or are confused with changes in pitch

height, based on F0 (e.g., Borchert et al., 2011; Moore and Glasberg, 1990; Singh and Hirsh, 1992; Warrier and Zatorre, 2002).

Although many studies have examined the effect of spectral changes on F0 perception and discrimination, fewer have investigated the effects of F0 variation on spectral-shape discrimination. Beal (1985) conducted a study in which both the effect of timbre variation on discriminating between pitches and the effect of pitch variation on discriminating between timbres was observed. When listening to chord changes on different musical instruments, subjects found it challenging to ignore changes in timbre, i.e. switching between instruments, when attempting to focus exclusively on the pitches in musical chords. They had less difficulty ignoring chord changes when attempting to judge whether the two timbres were the same, suggesting an asymmetry between the dimensions of pitch and timbre. However, the salience or discriminability of the changes in the different dimensions was not controlled, and the timbres were limited to three distinctly different instruments (acoustic guitar, piano, and harpsichord). Beal (1985) also found differences in performance between musicians and non-musicians. Musicians were better at recognizing when the same chord was played on two different instruments, although the benefit of musicianship was only found when the chords were diatonic, suggesting that the successful referencing of familiar musical structures was the defining difference between musicians and non-musicians.

Pitt (1994) also compared musicians and non-musicians on pitch and timbre discrimination. In a categorization task, as subjects listened to different tones, they were to determine whether there was a pitch change, an instrument change (timbre change), both changed, or neither changed. Subjects were not required to report direction of change, however. Non-musicians were more strongly affected than musicians by variations in timbre when discriminating pitch, suggesting that non-musicians experienced greater difficulty processing the two dimensions independently. However, the number of stimuli used was again limited (two

different timbres: recordings of a trumpet and a piano, and two different pitches: 294 Hz and 417 Hz), and no attempt was made to equate the perceptual salience across the two dimensions, making direct comparisons difficult.

Melara and Marks (1990) found interactions between pitch and timbre for individual tones on speeded classification tasks. They attributed these interactions to failure in selective attention, or Garner interference. In one experiment, subjects were instructed to attend to either timbre changes or pitch changes, while both dimensions varied. Like Beal (1985) and Pitt (1994), however, a limited number of stimuli were used: a combination of two different duty cycles of square waves (0.1878 and 0.3128, labeled: "twangy" and "hollow", respectively) were combined with two different F0s (900 and 920 Hz). Krumhansl and Iverson (1992) also found interactions between pitch and timbre for individual tones on speeded classification tasks, but used more musical sounds (notes F4 and C5 for the pitches, and a synthesized trumpet and piano for the timbres). They found that variation in the non-target dimension interfered with classification for both pitch and timbre symmetrically. Again, however, a limitation of the study lies in the small number of stimuli used, and the fact that the differences in pitch and timbre were not equated for discriminability or perceptual salience. The importance of equating the dimensions of interest in terms of perceptual salience has been noted in both auditory and visual research by Melara and Mounts (1993, 1994).

More recently, Silbert et al. (2009) explored a general framework for understanding interactions between perceptual dimensions, based on signal detection theory (Green & Swets, 1966). They used concurrent changes in spectral centroid and F0 as an example of dimensional interactions and concluded that for most of their seven listeners the two dimensions were not processed independently. However, because they did not test identification performance for either dimension in isolation, and because they only tested two values of each dimension, it is not clear how much interference each dimension produced on the other, or whether the effects were

symmetric. It is also not clear what accounted for the relatively large individual differences observed in that study.

The present study explored the effects of spectral shape variation on F0 discrimination and *vice versa*. The two aims of the study were 1) to determine whether the interference and interactions between pitch and timbre were symmetric, and 2) to assess the effects of musical training on subjects' ability to ignore variations in irrelevant perceptual dimensions when performing a discrimination task. The first aim addresses the more general question of whether pitch has a privileged role in auditory perception. For instance, it is known that sensitivity to small changes in pitch is generally much greater than to changes in other dimensions (McDermott, Keebler, Micheyl, & Oxenham, 2010), and pitch has been cited as an exception to Miller's "seven plus-or-minus two" rule, in that musicians are able to perfectly identify more than just 9 pitch intervals (E.M. Burns, 1999). On the other hand, more recent work has suggested that some of the properties that were thought to make pitch "special" can also be found in other dimensions (such as timbre and loudness), when differences in basic sensitivity are equated (e.g., McDermott et al., 2008, 2010).

The second aim tackles the question of differences in basic perceptual skills between musicians and non-musicians. As mentioned above, Silbert et al. (2009) observed relatively large individual differences that were not accounted for. One factor may be the amount of prior musical training. There are some studies that have found better performance in musicians than non-musicians in tasks involving both pitch perception (e.g., (Micheyl et al., 2006) and analytic listening in an informational masking context (Oxenham et al., 2003). Attending to one dimension and ignoring another could be considered a form of analytic listening, so it may be that musicians are less susceptible to interference effects. In contrast to this expectation, Borchert et al. (2011) found no significant benefit of musical training in a task that involved pitch discrimination between two sounds that varied widely in spectral shape. Little is known about

differences between musicians and non-musicians in their ability to discriminate spectral shape, with or without the presence of F0 changes. On one hand, some benefit of musicianship in attending selectively to separate auditory dimensions beyond pitch might be expected; on the other hand, timbre discrimination may not be as highly trained in musicians as pitch discrimination, because discriminating between very subtle spectral differences is not part of a typical ear-training program.

Experiment 1 measured basic sensitivity to small changes in either F0 or spectral centroid, in the absence of variation in the non-target dimension. Experiment 2 used the individual difference limens (DLs) from Experiment 1 to examine the effects of random variations in either F0 or spectral centroid on listeners' ability to discriminate small changes in the other dimension. Finally, Experiment 3 provided a direct test of perceptual symmetry of the two dimensions by measuring performance in both dimensions using stimuli that varied by the same amount in terms of DLs measured in the individual subjects.

Experiment 1: Basic Pitch and Timbre Discrimination

Rationale

The goal of Experiment 1 was to find thresholds for each subject on basic pitch and timbre discrimination tasks. We did this by separately measuring DLs for F0 and spectral centroid of a bandpass-filtered harmonic tone complex. These DLs were then used in subsequent experiments to equate changes in F0 and spectral centroid in terms of basic sensitivity for each subject individually.

Methods

Stimuli. The stimuli were harmonic complex tones, 500 ms in duration with 20-ms raised-cosine onset and offset ramps and an overall level of 70 dB SPL. The components were added in sine phase. All harmonics of the complex tone up to 10,000 Hz were generated and then individually scaled to produce slopes of 24 dB/octave around the center frequency (CF), or spectral centroid, with no flat bandpass region. Thus, the 3-dB bandwidth of the filter was 1/8 octave. MATLAB (Mathworks, Natick, MA) was used to generate the stimuli and control the experimental procedures. All stimuli were generated via an L22 soundcard (LynxStudio, Costa Mesa, CA) with 24-bit resolution at a sampling rate of 44100 Hz, and were presented diotically through HD580 headphones (Sennheiser, Old Lyme, CT).

In the pitch-discrimination task, the CF of the filter was held constant at 1200 Hz. The nominal F0 value of 200 Hz was roved across trials by ±10% with uniform distribution. Each trial consisted of two presentation intervals, each containing a complex harmonic tone with the F0s differing by Δ_{F0}, expressed as a percentage of the F0 of the lower tone. The F0s of the two tones in each trial were geometrically centered around the nominal F0 value after roving.

In the timbre-discrimination task, the F0 of the complex tone was held constant at 200 Hz, and the nominal CF of the bandpass filter was roved between trials by ±10% around 1200 Hz, with uniform distribution. Within each trial, the CF of the filter differed across the two presentation intervals by Δ_{CF}, again expressed as a percentage of the lower CF, and the two CFs were geometrically centered around the nominal CF after roving. See Fig. 2.1 for a schematic diagram of changes in stimuli.

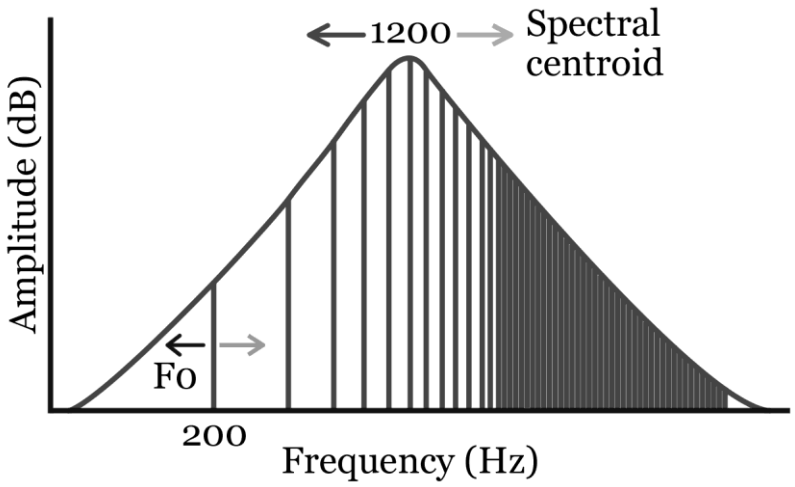

Fig. 1. Schematic diagram of the stimuli used in this study (plotted on log–log axes). Changing the F0 results in changes in the frequencies of the harmonics (represented by the vertical lines). Changing the center frequency of the filter results in changes in the spectral envelope of the sound and hence changes in the amplitudes (but not frequencies) of the harmonics.

Procedure

Prior to running the experiment, subjects were given basic definitions of pitch and timbre: pitch was related to notes on a musical scale, and timbre was related to sound quality differences between different musical instruments, using adjectives such as bright or dull. For comparison, they were told that a saxophone has a brighter timbre than a grand piano. Not surprisingly, subjects often had more difficulty grasping the concept of timbre, but were encouraged to use the practice runs and feedback to get a sense for what a brighter timbre sounded like, relative to a duller timbre. Subjects were tested individually in double-walled sound-attenuating chambers. The subjects' preliminary tasks were to compare tone pairs differing in either F0 or spectral centroid (i.e., "pitch" or "timbre"). In each trial, subjects were played two complex harmonic tones, separated by a silent interstimulus interval (ISI) of 300 ms. The task was to determine which of the two tones had the higher pitch or brighter timbre. The order of the tone presentations

was random, with the higher pitch (or timbre) being equally likely to be presented in the first or second interval. Two virtual boxes were displayed on a computer screen, which lit up with each corresponding tone. Subjects could select a box with the computer mouse or by pressing "1" or "2" on the keyboard, corresponding to the "1" and "2" displayed on the virtual boxes. Immediate feedback was provided after each trial, stating if the selection was "correct" or "wrong."

Each participant's DLs for F0 and spectral centroid were obtained using a standard two-alternative forced-choice procedure with a two-down one-up adaptive tracking rule that tracks the 70.7% correct point on the psychometric function (Levitt, 1971). The starting value of Δ_{F0} or Δ_{CF} was 200%. Initially, Δ_{F0} or Δ_{CF} was increased or decreased by a factor of 2. After the first reversal in the direction of the change in the tracking variable from "up" to "down", the factor was decreased to 1.26. After two more reversals, the factor was decreased to 1.12, which was the final step size. The run was terminated after six reversals at the final step size, and the DL in each run was the geometric mean of the value of Δ at those last six reversal points.

The first six runs performed by each subject in each condition were treated as practice. The next six runs in each condition were geometrically averaged to obtain the estimated DL for each subject. Each subject completed all testing in one dimension before proceeding to the other dimension, and the F0 and spectral centroid conditions were completed in counterbalanced order across subjects. Subjects were able to complete Experiment 1 in about 45 minutes on average, but the time varied for each participant, depending on the number and duration of breaks taken and the amount of time subjects took to make their responses.

Subjects

To avoid including subjects with severe F0 discrimination difficulties (Peretz, et al., 2009; Semal & Demany, 2006), only subjects whose F0DLs were 6% (about 1 semitone) or better

were included in the study. Since we have no estimate of an appropriate cutoff for "poor" spectral centroid discrimination, we did not exclude subjects based on exceeding a specific spectral centroid difference limen. After several subjects failed to reach the F0DL cutoff in the initial training phase, an additional training protocol was added, in which the between-trial roving of F0 or spectral centroid was eliminated. A total of 25 of the 57 subjects tested were given the non-roving practice trials. This appeared to make the task easier, and helped some subjects to subsequently improve their performance in the tasks with between-trial roving. Nevertheless a total of 12 subjects (7 of whom were given the non-roving practice) failed to achieve DLs of 6% or less. Eleven of the 12 disqualified subjects were non-musicians. The remaining 45 subjects (21 musicians and 24 non-musicians) took part in the experiment.

All 45 subjects had normal hearing, defined as audiometric pure-tone thresholds of 20 dB HL or better at octave frequencies between 500 Hz and 8 kHz, and were recruited from the University of Minnesota community. Ages ranged from 19 to 59 years (mean: 25.3 years). Twenty-one subjects were categorized as musicians (12 females, nine males, age range: 19-59, mean: 26.3 years) with at least eight years of formal musical training, and 24 were categorized as non-musicians (13 females, 11 males, age range: 19-34, mean: 24.4 years), with two or less years of formal musical training. All protocols were approved by the University of Minnesota Institutional Review Board, and all subjects provided written informed consent.

Results

The results for musicians and non-musicians are shown in Figure 2. The average F0DL for musicians was 0.8%, whereas the non-musicians had an average F0DL of 1.9%. Musicians had an average spectral-centroid DL of 4.0%, while the non-musicians had an average DL of 5.0%. Mixed-model ANOVAs on the log-transformed DLs were used here and throughout this study,

with a Greenhouse-Geisser correction for lack of sphericity included where appropriate. A mixed-model ANOVA with a within-subject factor of dimension (F0 vs. spectral centroid) and a between-subject factor of musicianship showed a main effect of dimension [$F(1,43) = 226.72$, $p < 0.0001$, partial $\eta^2 = 0.84$], a main effect of musicianship [$F(1,43) = 10.91$, $p = 0.002$, partial $\eta^2 = 0.20$] and an interaction between dimension and musicianship [$F(1,43) = 0.87$, $p < 0.0001$, partial $\eta^2 = 0.26$].

A planned comparison revealed that musicians had significantly better F0DLs compared to non-musicians, [$t(43) = 4.05$, $p < 0.0001$, $r = 0.53$], but no significant difference was found between the groups' spectral centroid DLs [$t(33.7) = 1.36$, $p = 0.183$, $r = 0.23$]. Levene's test indicated unequal variances for the timbre condition [$F = 4.47$, $p = 0.04$], so degrees of freedom were adjusted from 43 to 33.7.

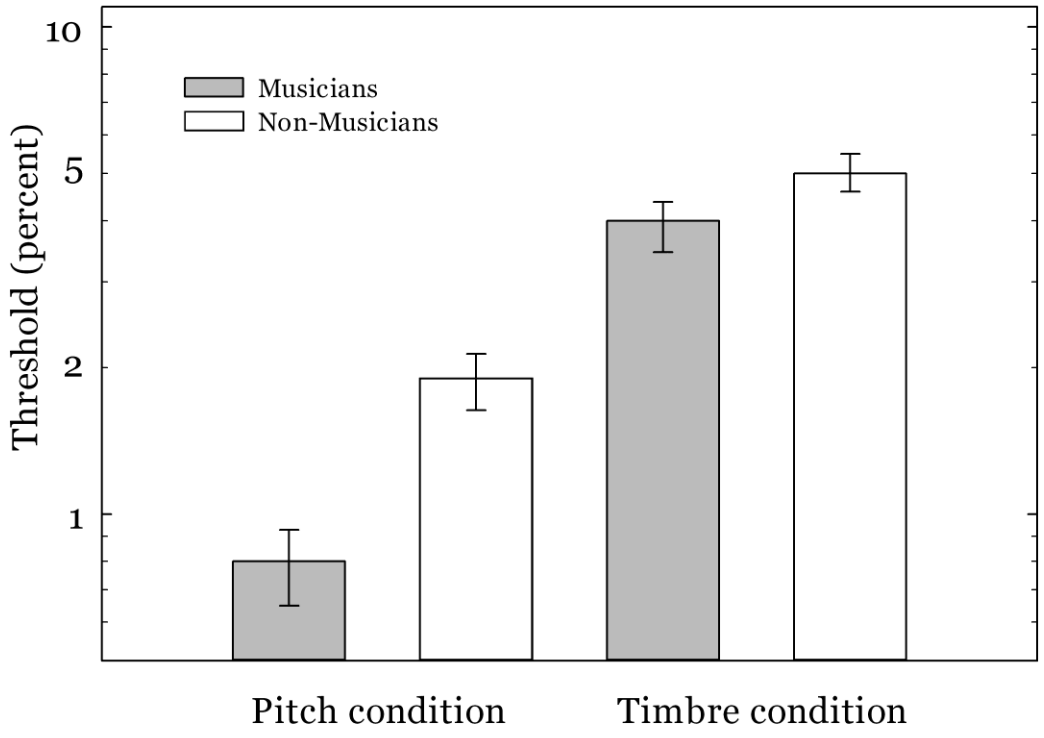

Fig. 2. Results from Experiment 1. Average DLs of musicians and non-musicians on basic pitch and timbre discrimination tasks. Error bars represent +/- one standard error of the mean.

Discussion

Musicians and non-musicians differed in their F0DLs, but had similar spectral centroid DLs. The differences in basic F0 discrimination with musical training are consistent with previous research that also used subjects with no extensive training (Micheyl et al., 2006). Based on earlier studies, however, we would expect the F0DLs from the non-musicians to converge with those of the musicians after more extensive practice. For instance, Micheyl et al. (2006) found that F0DLs from non-musicians reached the levels obtained by professional musicians after about 6 to 8 hours of practice, whereas our subjects typically had only around 20 minutes of practice before data were collected.

The lack of difference between musicians and non-musicians in sensitivity to spectral centroid is also consistent with previous research involving dissimilarity ratings (Caclin et al., 2005; McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995). The effect of musicianship on F0, but not spectral centroid, may be due to the fact that musicians regularly make fine judgments of pitch differences, for instance when tuning instruments, whereas fine timbre judgments tend to be less critical, since different musical instruments have rather distinct timbres. In addition, pitch changes define melodies, whereas the timbre of a particular instrument generally remains relatively constant. On the other hand, it could be argued that fine timbre discrimination is required when assessing the musical "color" of particular notes or a particular performance.

An alternative explanation as to why musicians did not have better spectral centroid DLs is that the stimuli in this experiment do not sound like musical instruments. These stimuli are

synthesized, and controlled exclusively by varying the location of the single spectral peak in the stimulus. Thus, it remains possible that musicians are more skilled at discriminating fine timbre differences in more natural musical sounds, perhaps even related to their own instrument. This idea is supported by previous research (Crummer, Walton, Wayman, Hantz, & Frisina, 1994; Pantev et al., 1998).

Finally, a potential limitation of excluding subjects with very poor F0 discrimination is that our population sample may be skewed towards better performance. Had we not excluded these subjects, based on the 6% F0DL cutoff, we would have likely seen a larger difference in F0DLs between the musician and non-musician groups, since 11 of the 12 subjects who were excluded were non-musicians.

Experiment 2: Thresholds as a Function of Amount of Interference

Rationale

The aim of Experiment 2 was to investigate the effects of variations in a non-target dimension on discrimination performance in the target dimension. This experiment involved similar stimuli and tasks to those used in Experiment 1, with the addition of random variation in the non-target dimension. Subjects were asked to attend to one dimension while ignoring the other. Shifts in F0 were paired with shifts in spectral centroid, in order to determine the effect of variations in one dimension on subjects' ability to discriminate changes in the other.

Methods

Stimuli. The stimuli were generated and presented in the same way as in Experiment 1. A standard adaptive two-alternative forced-choice procedure was again used. For this experiment, however, variations in the non-target dimension were introduced in each trial. The amount of

variation in the non-target dimension was based on multiples of the DL with no non-target variations, as measured in Experiment 1 for each subject individually (DL_0). Values tested were 0, 2, 5, 10, 25, 50, and 100DL_0, where zero indicates a lack of variation (i.e., a repeat of the conditions tested in Experiment 1). As in Experiment 1, the nominal F0 of the harmonic complex was 200 Hz, and the nominal CF (spectral centroid) was 1200 Hz. In each trial, both the nominal F0 and the nominal spectral centroid were roved independently by ±10%.

Procedure

In runs where the F0DL was adaptively tracked, the spectral centroid in each trial differed between the two intervals by a multiple of the centroid DL, as measured individually for each subject in Experiment 1, geometrically centered around the nominal centroid. The interval containing the higher centroid was selected randomly and independently from the F0 in each trial. In runs where the spectral centroid DL was adaptively tracked, the F0 between the two intervals also varied independently in multiples of the individual F0DL around the nominal F0 of 200 Hz, as described above for the spectral centroid variations. Thus the random variation in the non-target dimension was uninformative for the subjects' task.

The two parts of the experiment (the F0 task and the spectral centroid task) each contained seven conditions repeated three times, totaling 21 runs. The pitch and timbre tasks were performed in counterbalanced order across subjects, and all measurements of one dimension were completed before beginning measurements in the other dimension. No practice was given beyond the practice in basic discrimination received in Experiment 1. The basic discrimination tasks in Experiment 1 were performed just prior to starting Experiment 2. Completion of both experiments generally required two sessions, with the first session lasting two hours and the second session (which generally took place within a week of the first session) lasting between one and two hours. Participants were encouraged to take breaks when needed, to avoid fatigue effects.

Subjects

Thirty listeners took part in this experiment, all of whom had also participated in Experiment 1. Ages ranged from 19 to 59 years (mean: 27.95 years). Fifteen subjects were categorized as musicians (eight females, seven males, age range: 19-59, mean: 28.5 years) with at least eight years of formal musical training, and 15 were non-musicians (nine females, six males, age range: 19-34, mean: 24.3 years), with two or less years of formal musical training.

Results

The results of Experiment 2 are shown in Fig. 3. A mixed-model repeated-measures ANOVA on the log-transformed DLs was used to analyze the data. Within-subject factors were target dimension (F0 vs. spectral centroid) and amount of variation in the non-target dimension. The between-subjects factor was musicianship (musician vs. non-musician). Results showed a main effect of target dimension [$F(1,27) = 13.4$, $p = 0.001$, partial $\eta^2 = 0.33$], a main effect of variation in the non-target dimension [$F(6,22) = 18.5$, $p < 0.0001$, partial $\eta^2 = 0.39$], and a main effect of musicianship [$F(1,27) = 5.17$, $p = 0.031$, partial $\eta^2 = 0.16$]. The interaction between musicianship and dimension just failed to reach significance [$F(1,27) = 4.07$, $p = 0.054$, partial $\eta^2 = 0.13$], presumably reflecting the trend for musicians to perform better than non-musicians in the F0 dimension but not in the spectral centroid dimension. Indeed, separate ANOVAs revealed that musicians were significantly better than non-musicians on the F0 dimension [$F(1,27) = 6.41$, $p = 0.017$, partial $\eta^2 = 0.19$], while they were not significantly better than non-musicians on the spectral centroid dimension [$F(1,27) = 1.82$, $p = 0.188$, partial $\eta^2 = 0.06$]. No other interactions reached significance.

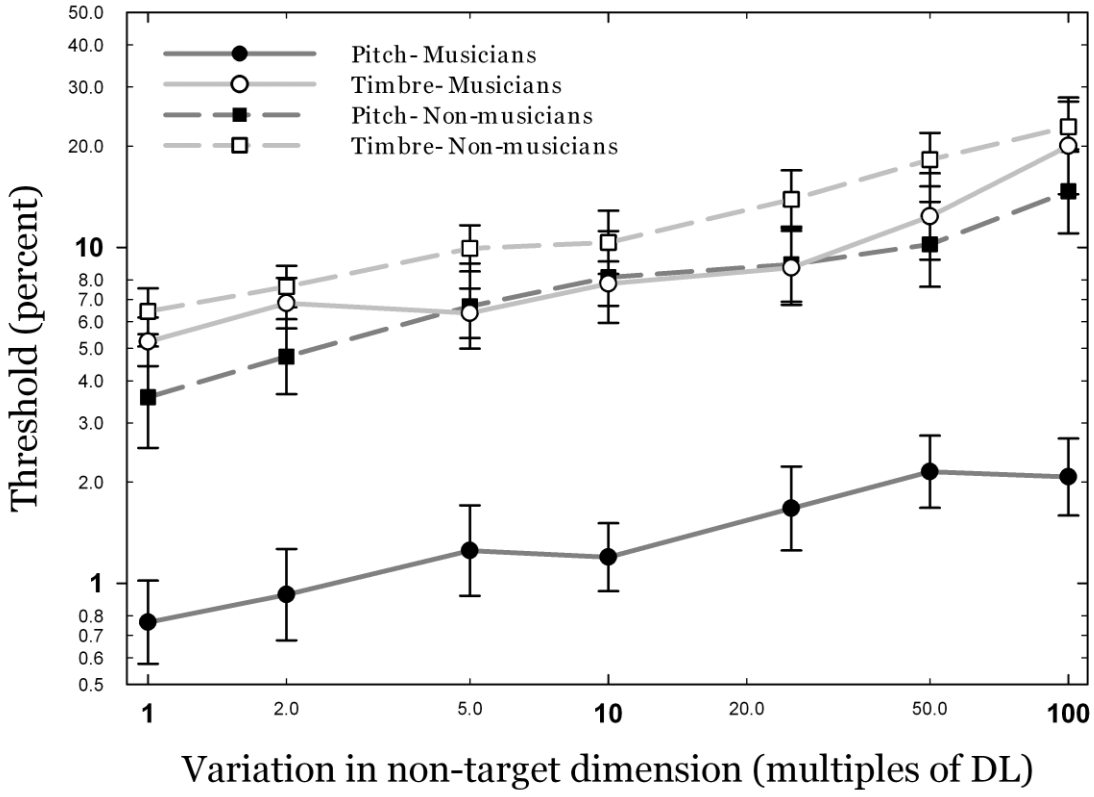

Fig. 3. Results from Experiment 2. Average DLs for musicians and non-musicians on pitch- and timbre- discrimination tasks are shown as a function of variation in the non-target dimension (in multiples of DL). Error bars represent +/- one standard error of the mean.

The amount of interference was assessed using the ratio of the DLs between the conditions with variation and the conditions with no variation in the non-target dimension; this measure is referred to as the "interference ratio." The interference ratio at the largest variation level ($100DL_0$) was 2.8 (i.e., 2.1% divided by 0.76%) and 4.1 (i.e., 14.7% divided by 3.6%) for the musicians and non-musicians, respectively, in the pitch target dimension. The same interference ratios in the timbre target dimension were 3.8 and 3.5 for the musicians and non-musicians, respectively. All four of these represented highly significant increases in DLs, based on paired-samples t-tests for F0 [$t(14) = 7.41$, $p < 0.0001$, $r = 0.89$] and spectral centroid [$t(14) =$

5.96, $p < 0.0001$, $r = 0.85$] for musicians, and for F0 [$t(14) = 5.04$, $p < 0.0001$, $r = 0.80$] and spectral centroid [$t(14) = 10.8$, $p < 0.0001$, $r = 0.95$] for non-musicians.

The fact that the original ANOVA found no significant interaction between musicianship and amount of variation in the non-target dimension suggests that the effect of interference was similar for both groups. This was confirmed in a new mixed-model ANOVA with the interference ratio as the dependent variable, target dimension and amount of non-target variation as the within-subject factors, and musicianship as the between-subjects factor. The results showed a significant main effect of non-target variation [$F(5,92.3) = 39.8$, $p < 0.0001$, partial $\eta^2 = 0.59$], but no main effect of the target dimension [$F(1,28) = 0.63$, $p = 0.434$, partial $\eta^2 = 0.02$], no main effect of musicianship [$F(1,28) = 1.24$, $p = 0.274$, partial $\eta^2 = 0.04$], and no significant interactions ($p > 0.24$ in all cases). This outcome confirms that the interference was similar for both pitch and timbre target dimensions, and that both musicians and non-musicians experienced similar amounts of interference in both dimensions.

Discussion

Variations in the non-target dimension led to increased (poorer) DLs in the target dimension for both F0 and spectral centroid, and for both musicians and non-musicians. The amount of interference (defined as the ratio between DLs with and without non-target variation) increased with increasing amount of variation, up to the maximum tested ($100DL_0$), although the greatest effect was observed between 0 and $10DL_0$.

Although musicians had generally lower F0DLs, their spectral-centroid DLs were similar to those of non-musicians, as found in Experiment 1. The effect of variations in both non-target dimensions was not significantly different for musicians and non-musicians, suggesting that musicians are as susceptible to interference due to random stimulus variations as non-musicians.

For both groups, when the variations were equated in terms of DL_0, the effects of F0 variation on spectral centroid discrimination and the effects of spectral centroid variation on F0 discrimination were symmetric – random variations in both dimensions produced substantial and similar interference. Thus, our results provide further support for the idea that pitch does not occupy a privileged position in auditory perception once differences in basic discrimination are equated (McDermott et al., 2010; McDermott & Oxenham, 2008).

Experiment 3: Congruent and Incongruent Interference

Rationale

In experiment 2, the direction of the variation in the non-target dimension was randomly selected on each trial and was independent of the direction of the change in the target dimension. Thresholds were determined using an adaptive procedure, and no attempt was made to separate trials with "congruent" motion (i.e., F0 and spectral centroid changed in the same direction) from trials with "incongruent" motion (i.e., F0 and spectral centroid changed in opposite directions). The interference produced by changes in the non-target dimension may reflect a "distraction" effect (Moore & Glasberg, 1990), produced by any task-irrelevant change, or it may reflect a partial inability on the part of subjects to distinguish between a change in timbre (i.e., higher brightness with increasing spectral centroid) from a change in pitch (i.e., higher pitch with increasing F0) (e.g., Russo and Thompson, 2005). It is also possible, in instances with large timbre variation, that an upward shift in spectral centroid induces an "octave error" (e.g., Robinson, 1993), causing subjects perceive the pitches an octave higher than the stimulus F0.

For this experiment, a method of constant stimuli was used. Congruent trials were randomly interleaved with incongruent trials, but the two categories were analyzed separately to determine whether changes in the non-target dimension produced systematic biases in responses

to the target dimension. Only relatively small variations in the dimensions were tested, making octave errors due to large spectral shifts less likely.

A second open question from Experiment 2 is whether multiples of DL_0 provide an appropriate scale along which to equate the perceptual salience of larger changes. If equal changes in terms of DL_0 result in equal salience, then presenting changes in both dimensions that are equal in terms of DL_0 should result in equal performance in both dimensions. The current experiment tested this hypothesis by presenting pairs of tones that varied in F0 and spectral centroid by the same amount, in terms of the individual DL_0s; the task varied (subjects were asked to judge either the pitch or timbre) but the stimuli were identical in the two conditions.

Methods

Stimuli and procedure. The method in which the stimuli were generated and presented was the same as that used in Experiments 1 and 2. However, this experiment used a method of constant stimuli, rather than an adaptive procedure. The subjects were presented with pairs of tones that varied in both F0 and spectral centroid by the same amount, in terms of the individual DL_0s, which had been determined in Experiment 1. The following five multiples of DL_0 were tested: 0.5, 1, 2, 3, and 5. Each trial had a pair of stimuli, as described in Experiment 2, in which both the F0 and the spectral centroid varied by one of the multiples of DL_0. In each block of 50 trials, half the trials had congruent pairings (F0 and spectral centroid changed in the same direction) and the other half had incongruent pairings (F0 and spectral centroid varied in opposite directions). Thus, each block included five repetitions of each condition and pairing type. The trials were evenly divided into separate blocks in which either pitch or timbre discrimination was measured. As in the previous two experiments, subjects were instructed to select which pitch was higher or which timbre was brighter in the tone pair, depending on which task they were

performing, and were instructed to ignore the other dimension. A total of ten blocks were run for each dimension, meaning the estimate of performance for each subject on each dimension was based on a total of 500 trials (100 trials per DL_0 multiple). Feedback was provided after each trial. Each subject completed all the measurements in one dimension before the other dimension was tested, and the order of presentation was counter-balanced across subjects. The experiment took around an hour to complete, but the time varied for each participant, depending on the number and duration of breaks taken and the amount of time subjects took to make their responses.

Subjects

A total of 20 subjects participated, all of whom also took part in Experiment 1. Five of these 20 subjects (four musicians, one non-musician) also participated in Experiment 2. The ages of the subjects ranged from 20 to 59 years (mean: 25.9 years). Ten subjects were categorized as musicians (6 females, 4 males, age range: 20-59, mean: 27.2 years) with at least eight years of formal musical training, and 10 were non-musicians (4 females, 6 males, age range: 21-34, mean: 24.6 years), with two or less years of formal musical training.

Results

The mean results in the different conditions for congruent and incongruent trials are shown in terms of proportion correct for musicians and non-musicians in the right and left panels of Fig. 4, respectively. Statistical analysis was performed on values of *d'* , converted from proportion correct by assuming unbiased responding to first and second intervals in each trial (Hacker, Ratcliff, Tables, & Ed, 1979). To avoid infinite values of *d'* when 100%-correct performance was achieved, a small correction factor was included, which effectively limited the maximum value of *d'* to 4.65, corresponding to a proportion correct of about 99.95%.

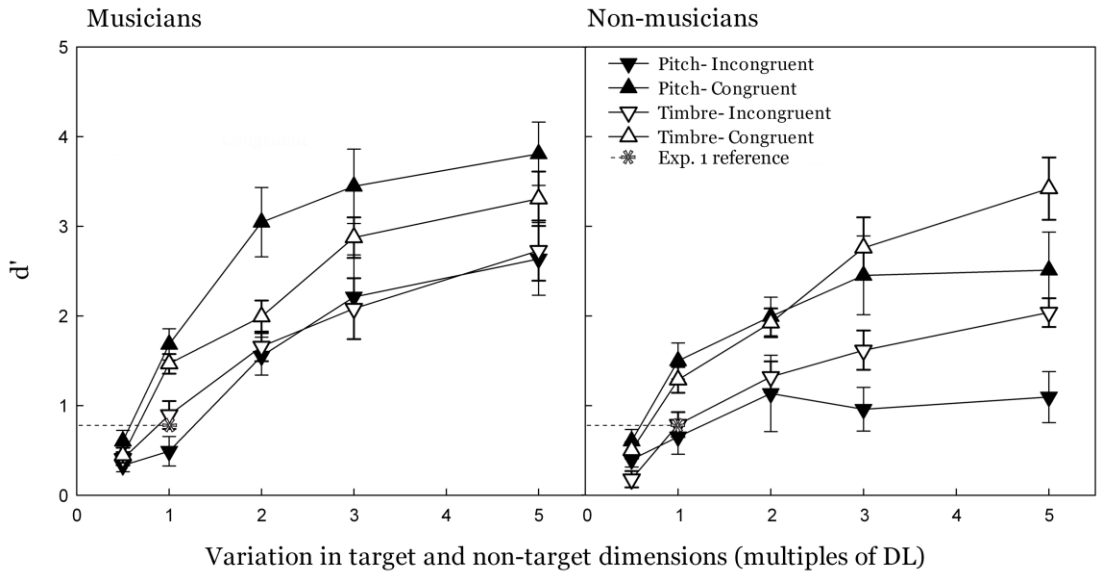

Fig. 4. Experiment 3: Values of *d'* are shown for congruent and incongruent stimuli pairings for pitch and timbre tasks, as a function of amount of variation in target and non-target dimensions (in multiples of DL). Musicians' scores are shown in the left panel, and non-musicians' scores are shown in the right panel. The asterisk in each panel is shown at the point corresponding to the DL in Experiment 1. Error bars represent +/- one standard error of the mean.

A mixed-model ANOVA was performed on the *d'* values with within-subject factors of target dimension (F0 or spectral centroid), congruence (congruent or incongruent changes between intervals), and amount of variation (0.5 through $5DL_0$), and a between-subjects factor of musicianship. A significant main effect of congruence was found [$F(1,18) = 66.9$, $p < 0.0001$, partial $\eta^2 = 0.79$, reflecting the observation that performance was generally better in congruent than in incongruent trials. The main effect of amount of variation was also significant [$F(2.23,40.2) = 108$, $p < 0.0001$, partial $\eta^2 = 0.86$, $\varepsilon = 0.56$], reflecting the observation that performance improved as the size of the F0 or spectral-centroid difference increased. Finally, the

main effect of target dimension (F0 or spectral centroid) was not significant [$F(1,18) = 0.04$, $p = 0.847$, partial $\eta^2 = 0.002$], showing that overall levels of performance were similar in the two dimensions.

A significant interaction between the amount of variation and congruence was also found [$F(2.35,42.3) = 7.78$, $p < 0.0001$, partial $\eta^2 = 0.302$, $\varepsilon = 0.59$], possibly reflecting the widening gap between the congruent and incongruent performance with increasing amount of variation. Additionally, a significant interaction was found between dimension and congruence [$F(1, 18) = 6.77$, $p = 0.018$, partial $\eta^2 = 0.273$], indicating that congruence differentially affected F0 and spectral centroid, with the congruence effect being larger when the target dimension was pitch than when it was timbre. However, performance in congruent trials was significantly higher than performance in incongruent trials for both F0 [$F(1,18) = 43.7$, $p < 0.0001$, partial $\eta^2 = 0.71$], and spectral centroid [$F(1,18) = 49.8$, $p < 0.0001$, partial $\eta^2 = 0.73$].

There was a significant effect of musicianship [$F(1,18) = 8.03$, $p = 0.011$, partial $\eta^2 = 0.309$], and a significant interaction between amount of variation and musicianship [$F(4,72)=4.44$, $p = 0.003$, partial $\eta^2 = 0.198$]. These effects seem to reflect the somewhat worse performance of non-musicians, particularly at larger levels of variation. No significant interaction was found between dimension and musicianship [$F(1,18) = 2.28$, $p = 0.148$, partial $\eta^2 = 0.112$], indicating that the two groups performed similarly across the F0 and spectral centroid conditions. Additionally, no significant interaction was found between congruence and musicianship [$F(1,18) = 0.30$, $p = 0.591$, partial $\eta^2 = 0.016$], suggesting that these groups were similarly affected by whether the dimensions were congruent or incongruent. There was one significant 3-way interaction for dimension, variation, and musicianship [$F(1,18) = 0.30$, $p = 0.024$, partial $\eta^2 = 0.143$], suggesting that the groups may be differentially affected by the amount of variation across dimensions. However, the 3-way interaction for congruence, dimension, and musicianship was

not significant (p = 0.167), suggesting the two groups were similarly affected by congruence across the dimensions. There was no significant 4-way interaction.

An asterisk in each panel of Fig. 4 is shown at the point corresponding to the DL in Experiment 1. By definition, based on our tracking procedure, the DL was 70.7%, which in a 2-interval 2-alternative forced-choice task corresponds to a d' of about 0.77 (Hacker et al., 1979). The asterisks fall closer to the downward than to the upward triangles, suggesting at face value that performance was enhanced in the congruent trials, but not degraded in the incongruent trials, relative to no variation. However, this outcome may be related to improvements with practice, as all the subjects through necessity participated in Experiment 1 (asterisks) before embarking on Experiment 3. Thus, without this potential confound, it may be expected that congruence would lead to improved performance, whereas incongruence would lead to poorer performance, relative to no irrelevant changes.

Discussion

The first important finding from this experiment is that performance in the congruent trials (where the variation in the non-target dimension was in the same direction as that in the target dimension) was better than performance in incongruent trials. This outcome suggests that variations in the non-target dimension did not just provide a distraction, but were confused to some extent with changes in the target dimension. This confusion could be of at least two types: the first possibility is that the two dimensions are not perceptually separable, and that a change in spectral centroid may induce a change in the pitch percept (and vice versa). This seems unlikely, as pitch-matching experiments using harmonic stimuli with widely different spectral content have not shown large or systematic biases in pitch away from the underlying F0 (Oxenham et al., 2011; Walliser, 1969). The second, and more plausible, possibility is that changes in F0 and spectral

centroid elicit changes in pitch and timbre, respectively, but that subjects sometimes confuse the two, and therefore respond to the inappropriate dimension. When the dimensions change in a congruent manner, an inappropriate response will still be correct, thereby leading to higher performance in the congruent than in the incongruent trials. This would suggest the confusion is more post-sensory, which aligns with the conclusions of Silbert *et al.* (2009). Nevertheless, as variations in both F0 and spectral centroid elicit changes along the tonotopic dimension in the auditory periphery, there remains a possible basis for sensory confusion.

The second important finding is that overall performance in the F0- and spectral-centroid-discrimination tasks (averaged across congruent and incongruent conditions) was similar when variations in the two dimensions were equated in terms of multiples of DL_0 for each dimension separately. This finding suggests that salience (and coding accuracy) in the two dimensions may be equated using basic discrimination thresholds, at least for differences up to multiples of $5DL_0$. However, performance was not identical, as indicated by the significant interaction of dimension and congruence, suggesting that equivalence only holds when both congruent and incongruent trials are employed in roughly equal measure. In addition, we cannot rule out the possibility that more differences might be revealed through the use of other measures, such as reaction time.

The third important finding is that musicians and non-musicians showed similarities in terms of overall performance on the pitch and timbre tasks, as well as similarities in how they were affected by congruence. The main effect of musicianship and the interaction with amount of variation reflect some differences between the groups, but the general pattern of results was quite similar. Taken together with the results from Experiment 2, where no significant effect of musicianship was found on the amount of interference, the outcome suggests that musicians' superior analytic listening ability, as demonstrated in an informational masking task that involves attending to one frequency while ignoring others (Oxenham et al., 2003), does not extend to attending to one perceptual dimension while ignoring another.

Finally, it is worth noting that any differences observed between groups may depend to some extent on how the groups are defined. Although many studies have compared the performance of musicians and non-musicians, there are no uniform criteria that are used to distinguish between the two groups. We defined musicians as those with at least 8 years of formal musical training; however, no ear-training test was used to verify musical ability (e.g., Oxenham et al., 2003), no requirement was made that they were currently active musicians, and there was no maximum age allowed by which musical training should have commenced. Similarly, although non-musicians were defined as those with 2 years or less of formal training, it is possible that at least some members of this had informal experience with listening or performing music. Thus, as with any study comparing these two groups, the conclusions are qualified by the specific definitions of musical training used here.

Conclusions

Difference limens for F0 and spectral centroid (perceptually, pitch and timbre) were measured in groups of listeners with and without musical training in a two-alternative forced-choice paradigm. The following results were obtained:

1) In line with earlier studies, F0DLs were better in musicians than in untrained listeners without musical training. However, DLs for spectral centroid were not significantly different between the two groups.
2) Discrimination thresholds in either F0 or spectral centroid were impaired by random variations in the non-target dimension. The amount of interference was similar for the two dimensions, and was similar for both musicians and non-musicians.

3) Performance was better when the interference varied coherently with the target (i.e., both F0 and spectral centroid increased from the first to the second interval) than when the varied in opposite dimensions. This outcome suggests that listeners sometimes confused changes across the two dimensions. Musicians were no less susceptible to this "confusion" than non-musicians.

Overall the results provide evidence that judgments in pitch and timbre (in terms of F0 and spectral centroid, respectively) are similarly affected by random variations in the other dimension, suggesting relatively symmetric processes. In addition, musical training does not appear to provide strong immunity from interference effects in either dimension.

Acknowledgements

The work was supported by National Institutes of Health Grant No. R01 DC005216.

Chapter 3

Representations of Pitch and Timbre Variation in Human Auditory Cortex

Allen, E. J., Burton, P. C., Olman, C. A., & Oxenham A. J. (2017). Representations of pitch and timbre variation in human auditory cortex. Journal of Neuroscience, 1284-1293.

Abstract

Pitch and timbre are two primary dimensions of auditory perception, but how they are represented in the human brain remains a matter of contention. Some animal studies of auditory cortical processing have suggested modular processing, with different brain regions preferentially coding for pitch or timbre, whereas other studies have suggested a distributed code for different attributes across the same population of neurons. This study tested whether variations in pitch and timbre elicit activity in distinct regions of the human temporal lobes. Listeners were presented with sequences of sounds that varied in either fundamental frequency (eliciting changes in pitch) or spectral centroid (eliciting changes in brightness, an important attribute of timbre), with the degree of pitch or timbre variation in each sequence parametrically manipulated. The BOLD responses from auditory cortex increased with increasing sequence variance along each perceptual dimension. The spatial extent, region, and laterality of the cortical regions most responsive to variations in pitch or timbre at the univariate level of analysis were largely overlapping. However, patterns of activation in response to pitch or timbre variations were discriminable in most subjects at an individual level using multi-voxel pattern analysis, suggesting a distributed coding of the two dimensions bilaterally in human auditory cortex.

Key words: auditory cortex; fMRI; Heschl's gyrus; perception; pitch; timbre

Significance Statement

Pitch and timbre are two crucial aspects of auditory perception. Pitch governs our perception of musical melodies and harmonies, and conveys both prosodic and (in tone languages) lexical information in speech. Brightness – an aspect of timbre or sound quality – allows us to distinguish different musical instruments and speech sounds. Frequency-mapping studies have revealed tonotopic organization in primary auditory cortex, but the use of pure tones or noise bands has precluded the possibility of dissociating pitch from brightness. Our results suggest a distributed code, with no clear anatomical distinctions between auditory cortical regions responsive to changes in either pitch or timbre, but also reveal a population code that can differentiate between changes in either dimension within the same cortical regions.

Introduction

Pitch and timbre play central roles in both speech and music. Pitch allows us to hear intonation in a language, and notes in a melody. Timbre allows us to distinguish the vowels and consonants that make up words, as well as the unique sound qualities of different musical instruments. Combinations of pitch and timbre enable us to identify a speaker's voice, as well as a particular piece of music.

Several studies have been devoted to elucidating the cortical code for pitch; less attention has been paid to timbre. Bendor and Wang (2005) identified pitch-selective neurons in the marmoset cortex, located in a region near the anterolateral border of primary auditory cortex (A1) and the rostral field, and anterolateral and middle lateral non-primary belt areas. These neurons responded selectively to a specific fundamental frequency (F0 – the physical correlate of pitch), independent of the sound's overall spectral content. Several fMRI studies in humans have identified an anatomically analogous region in anterolateral Heschl's Gyrus (HG) that also seems particularly responsive to pitch (Gutschalk et al., 2002; Patterson et al., 2002; Penagos et al., 2004; Norman-Haignere et al., 2013), while posterior regions of HG, superior temporal sulci (STS), and insula have been found to be active in timbre processing (Menon et al., 2002). A PET study by Platel et al. (1997) examining pitch, timbre, and rhythm responses during active tasks, found hemispheric differences between pitch and timbre. However, only two different timbres were used (an oboe that was either "bright" or "dull"), and the differences were found outside of the auditory cortex in the right frontal lobe. Other studies have failed to observe distinct, or modular, processing of pitch (e.g., Bizley et al., 2009; Hall and Plack, 2009). A combined MEG/EEG study by Gutschalk and Uppenkamp (2011), looking at cortical processing of pitch and vowels (which have different timbres due to variation in spectral shape) found overlapping responses in the anterolateral HG, suggesting a lack of spatial distinction across these dimensions.

However, conclusions regarding spatial location using MEG or EEG are necessarily limited, given their generally poor spatial resolution. In a single-unit physiology study, Bizley et al. (2009) used stimuli that varied in F0 (corresponding to pitch), spectral envelope peak (corresponding to brightness, an important dimension of timbre), and spatial location, in order to identify neurons in the ferret auditory cortex that were selective for one or more of these dimensions. They found a distributed population code in the auditory cortices of ferrets with over two-thirds of the units responding to at least two dimensions. Most often, those dimensions were pitch and brightness. Further, a study by Hall et al. (2005) suggested lateral HG may be more of a perceptual processing site than a region that encodes temporal acoustic structure (an underlying structure inherent to both pitch and spatial location). In summary, the degree to which representations of pitch and timbre are spatially separated in the auditory cortex remains unclear.

Here we investigated whether variations in pitch and brightness elicit activity in distinct regions of the temporal lobes during a passive listening task, using functional magnetic resonance imaging (fMRI). A similar question was posed by Warren et al. (2005). They found overlapping bilateral regions of activation in the temporal lobes to sounds that varied in either F0 or spectral envelope shape, but found additional activation when spectral envelope shape was varied along with alternations between harmonic and noise stimuli. Based on their results, Warren et al., (2005) suggested that the mid portion of the right STS contains a specific mechanism for processing spectral envelopes, the acoustic correlate of brightness, which extended beyond the regions responsive to pitch or spectro-temporal fine structure. However, Warren et al. (2005) did not attempt to equate their changes in pitch or spectral shape in terms of perceptual salience, making the direct comparisons difficult to interpret. In our paradigm, inspired by the experimental design of Zatorre and Belin (2001), we generated sound sequences that varied in either F0 (pitch) or spectral peak position (brightness), where the changes in either dimension were equated for perceptual salience. The variance of the sequence in the dimension of interest

(pitch or brightness) was parametrically varied and the BOLD responses were measured. Our hypothesis was that regions selective for pitch or brightness should show increases in activation with increases in the variance or range of pitch or timbre within each sequence, and that modular processing of the two dimensions would be reflected by spatially distinct regions of the temporal lobe being selectively responsive to changes in the two dimensions.

Materials and Methods

Participants

Ten right-handed subjects (mean age of 23.8 years, standard deviation 2.0; five females and five males) were included in the analysis. An eleventh subject was discovered to have been left-handed and his data were subsequently excluded from analysis. All subjects had normal hearing, defined as audiometric pure-tone thresholds of 20 dB hearing level (HL) or better at octave frequencies between 250 Hz and 8 kHz, and were recruited from the University of Minnesota community. Musical experience of the subjects ranged between 0 and 23 years. Three subjects had musical experience of two years or less, while seven had at least nine years of experience.

Stimuli and procedure

Tone sequences were 30 s in duration, containing 60 notes each. Each tone had a total duration of 300 ms, including 10-ms raised-cosine onset and offset ramps and consecutive tones were separated by 200-ms silent gaps. Stimuli were presented binaurally (diotically) at 85 dB SPL. The 30-s tone sequences were interspersed with 15 s silence to provide a baseline condition. Sequences were generated from scales created with steps that were multiples of the average F0 difference limen (DL) of 1.3% for pitch or the average spectral centroid DL of 4.5% for timbre,

as established in an earlier study (Allen and Oxenham, 2014). This approach was used to equate for perceptual salience across the two dimensions. All harmonics of the complex tone up to 10,000 Hz were generated and scaled to produce slopes of 24 dB/octave around the center frequency (CF), or spectral centroid, with no flat bandpass region. The F0s and spectral centroids in each sequence were geometrically centered around 200 and 900 Hz, respectively (Fig. 1A). In each sequence the scale stepsize was selected to be 1, 2, 5, or 10 times the average DL. Each scale consisted of 5 notes spaced apart by one scale step. The note sequence on each trial was created by randomly selecting notes (with replacement) from the 5-note scale, with the constraint that consecutive repeated notes were not permitted. Each level of variation (i.e., stepsize) was presented once per scan in random order (Fig. 1B). Each scan contained all stepsizes across both dimensions. The presentation order of the dimensions and stepsizes was generated randomly for each scan and for each subject separately. The scans were 6 minutes in duration, and a total of 6 scans were run consecutively for each subject (See Fig. 1C).

Subjects listened passively to the stimuli while watching a silent video. MATLAB (Mathworks, Natick, MA) and the Psychophysics Toolbox (www.psychtoolbox.org) were used to generate the stimuli and control the experimental procedures. Sounds were presented via MRI-compatible Sensimetrics (Malden, MA) S14 model earphones with custom filters.

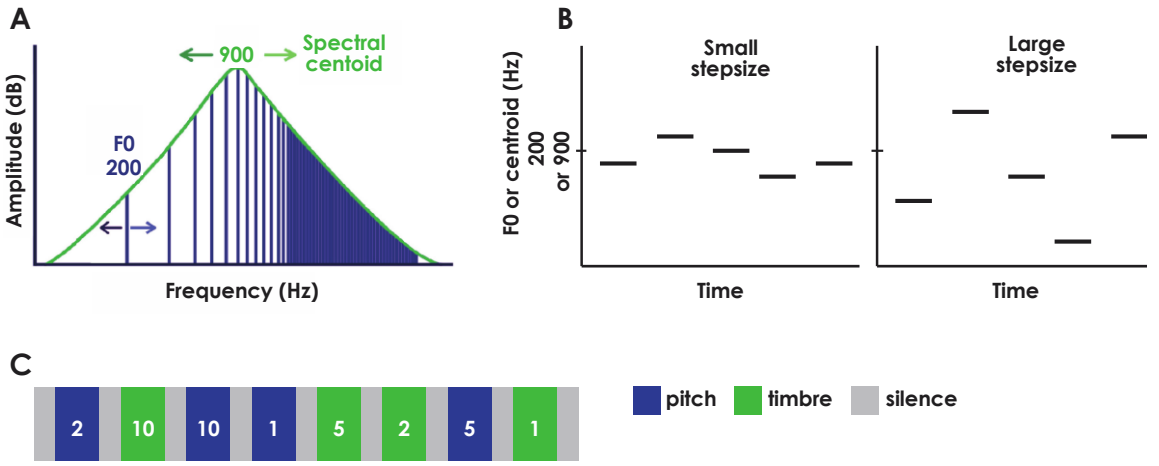

Fig. 1. Schematic diagrams of the stimuli. A. Spectral representation of the stimuli used in this study (plotted on log–log axes). Changing the F0 results in changes in the frequencies of the harmonics (represented by the vertical lines). Changing the center frequency of the filter results in changes in the spectral centroid of the sound and hence changes in the amplitudes (but not frequencies) of the harmonics. Lighter-colored arrows indicate that shifting in the rightward direction results in a sound with a higher pitch (increase in F0) or a brighter timbre (increase in spectral centroid). B. Tone sequences with small and large stepsizes. For the pitch sequences, the y-axis is F0, centered around 200 Hz; for the timbre sequences, the y-axis is spectral centroid, centered around 900 Hz. C. Experimental block design layout. Thirty-second pitch- and timbre-varying sequences are indicated in blue and green, respectively. Fifteen-second silent gaps for a baseline measure are indicated in grey. The presentation order of stepsizes, indicated in white text, was randomized. All possible stepsizes across both dimensions were included in each scan.

Data acquisition

The data were acquired at a 3T scanner (Siemens Prisma) at the Center for Magnetic Resonance Research (CMRR, University of Minnesota). Anatomical T_1-weighted images and field maps were acquired. The MPRAGE T1-weighted anatomical image parameters were: TR =

2600 ms; TE = 3.02 ms; matrix size = 256 x 256; 1 mm isotropic voxels. The pulse sequence for the functional scans used slice-accelerated multiband echo planar imaging (EPI) (Xu et al., 2013), and sparse temporal acquisition (Hall et al., 1999). The acquisition parameters for the functional scans were: TR = 6000 ms; time of acquisition (TA) = 2000 ms; silent gap = TR – TA = 4000 ms; TE = 30 ms; multiband factor = 2; number of slices = 48; partial Fourier 6/8; matrix size = 96 × 96; 2 mm isotropic voxels. A total of 60 volumes were collected in each of the six scans. Slices were angled in effort to avoid some of the motion from eye movement, and covered the majority of the brain. However, for most subjects the top of the parietal and part of the frontal cortices were excluded.

Data analysis

Data were preprocessed using the Analysis of Functional NeuroImages (AFNI) software package (Cox, 1996) and FSL 5.0.4 (http://fsl.fmrib.ox.ac.uk/). Statistical analyses and visualization were performed with AFNI and SPSS (IBM, New York, NY). Preprocessing included distortion correction using FSL's FUGUE, six-parameter motion correction, spatial smoothing (3 mm FWHM Gaussian blur), and pre-whitening.

For each subject, a general linear model (GLM) analysis was performed that included regressors for each experimental condition (i.e., each of the four stepsizes for pitch and timbre), six motion parameters, and Legendre polynomials up to the fourth order to account for baseline drift (modeled separately for each run). Each subject's brain was transformed into Montreal Neurological Institute (MNI) space (Mazziotta et al., 1995). Beta weights (regression coefficients) for individual voxels were estimated by the GLM for each condition for each subject, as were contrasts comparing pitch, timbre, and stepsize conditions, and a contrast comparing all sounds to baseline.

Group level analyses with subject as a random effect included a one-sample t-test

performed on the unmasked, unthresholded beta weights for each dimension (i.e., separately for pitch and timbre, averaged across all stepsizes) using the AFNI function 3dttest++. A paired t-test was performed in the same manner, comparing the pitch condition to the timbre condition.

To determine whether BOLD response increased linearly with increasing stepsize, the Pearson product-moment correlation between BOLD response to stepsize and a linear trend were computed in each voxel for each subject, separately for pitch and timbre. These correlation coefficients were then Fisher z-transformed and submitted to a one-sample t-test compared to zero, within a mask created by the union of all subjects' individual regions of interest (iROIs), to test the average correlation for significance across subjects.

For all analyses in AFNI, in light of the inflated false-positive findings by Eklund et al. (2016), smoothness values were obtained using AFNI's 3dFWHMx spherical autocorrelation function (acf) parameters at the individual level, and then averaged for the group level. These acf values were then used in AFNI's 3dClustSim function (AFNI 16.1.27) to obtain nearest-neighbor, faces touching, two-sided cluster thresholds via a Monte Carlo simulation with 10,000 iterations. This determined the probability of clusters of a given size occurring by chance if each voxel has a 1% chance displaying a false positive. Based on these probabilities, clusters smaller than those that would occur by chance more than 5% of the time were filtered out of the results to achieve a cluster-level alpha = .05.

Multi-voxel pattern analysis (MVPA) was performed using Princeton's MVPA toolbox for MATLAB with the backpropagation classifier algorithm for analysis (http://code.google.com/p/princeton-mvpa-toolbox/). In order to restrict the number of voxels in our analyses, we added a functionally defined mask, based on our univariate analysis results, containing voxels that were active for a particular subject during the sound conditions (pitch or timbre). We then thresholded this starting voxel set to contain only the 2,000 most responsive voxels across both hemispheres for each subject, making the number of voxels in each mask

consistent across subjects as well as reducing the number of voxels used for classification, in an attempt to improve classifier performance (e.g., De Martino et al., 2008; Schindler et al., 2013). Functional volumes sampled within 5 seconds of a transition between conditions were eliminated, to account for the lag in the hemodynamic response. Functional volumes during rest conditions were also eliminated in order for the classifier to be trained exclusively on the pitch and timbre conditions. Data were z-scored, and each run was treated as a separate time course in order to eliminate any between-run differences caused by baseline shifts. An n-minus-one (leave-one-out) cross-validation scheme was used, with six iterations, accounting for the six runs. Each iteration trained a new classifier on five of the six runs and tested it on the remaining run. A feature selection function was used to discard uninformative voxels, with a separate ANOVA run for each iteration.

Results

Whole-brain analyses of pitch and timbre

Figure 2 shows BOLD activity at the group level separately for pitch- and timbre-variation conditions contrasted with silence with single-sample t-tests. Similar bilateral activation can be seen, with the strongest activation occurring in and around HG for both dimensions. A paired t-test revealed no significant differences (no surviving voxels) between the pitch and timbre conditions at the group level, with a cluster threshold of 1072 microliters (134 voxels). At the individual level, only two of the ten subjects showed any significant differences between the pitch and timbre conditions (pitch-timbre), and neither of them had any significant clusters within the auditory cortex. There was no connection between these two subjects in terms of musicianship, as one had two years of musical training, while the other had 16.

ROI analysis

Two auditory ROIs in the temporal lobes were functionally defined in individual subjects (iROIs) based on the contrast of all sound conditions vs. baseline (silence), one in each hemisphere. The average (± SEM) cluster size of these iROIs was 2507 voxels ± 135.4 [left hemisphere (LH): 2451 ± 171.1; right hemisphere (RH): 2564 ± 217.7]. A two-tailed paired t-test revealed no significant difference in cluster size between hemispheres ($t_{(9)} = 0.60$, p = 0.565).

Within each iROI, the subject' beta weights for acoustical dimension (pitch and timbre) at each stepsize were averaged across voxels. A repeated-measures 2x2x4 ANOVA with average BOLD response within each subject's iROIs as the dependent variable and factors of acoustical dimension (pitch and timbre), hemisphere (right and left) and stepsize (1, 2, 5, and 10 times the DL) showed no main effect of hemisphere ($F_{(1,9)} = 1.2$, p = 0.3) or dimension ($F_{(1,9)} = 2.2$, p = 0.172), indicating that the overall level of activation in the ROIs was similar across hemispheres and across the pitch and timbre conditions. There was, however, a main effect of stepsize ($F_{(3,27)} = 14.7$, p = 0.0001), as well as a significant linear trend ($F_{(1,9)} = 31.\ 5$, p = 0.0001), indicating increasing activity with increasing stepsize. No significant interactions were observed, indicating that the effect of stepsize was similar in both hemispheres ($F_{(3,27)} = 1.3$, p = 0.302) and for both dimensions ($F_{(3,27)} = 1.2$, p = 0.346). Figure 3 depicts the mean beta weight for each stepsize for pitch and timbre within each of the left and right hemisphere ROIs.

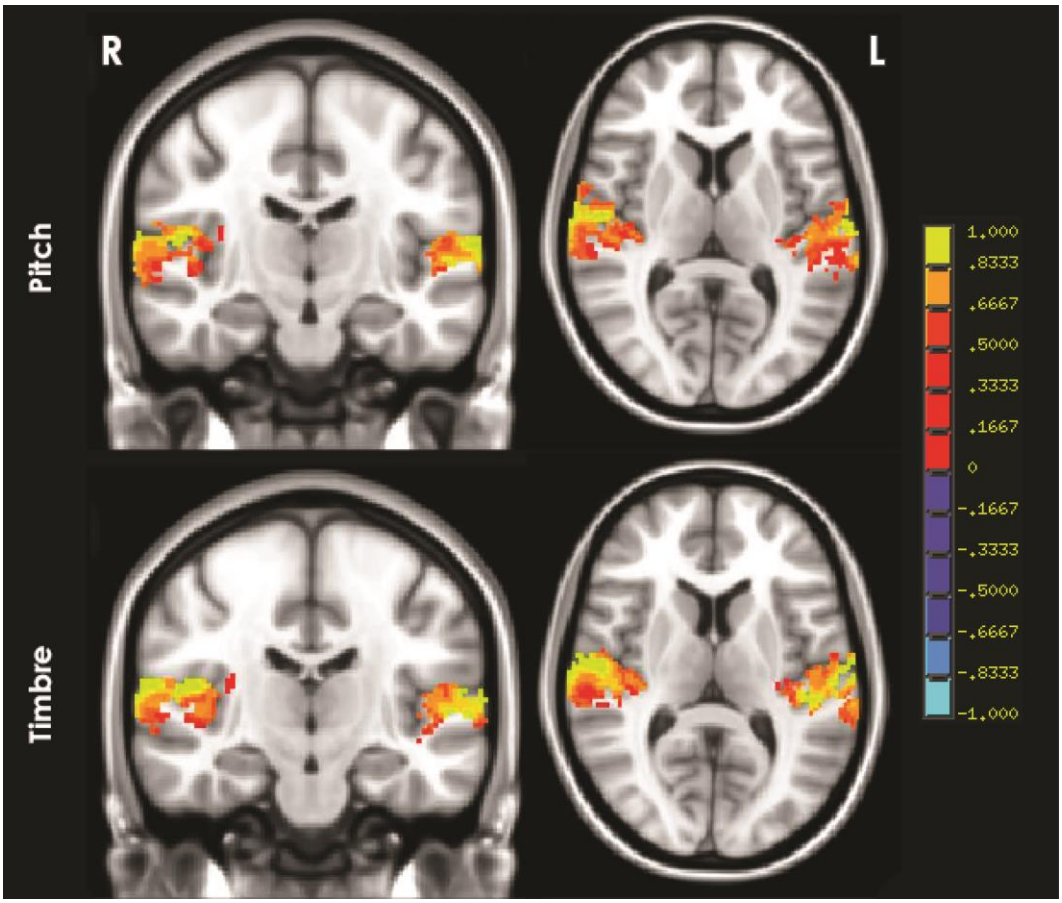

Fig. 2. Group-level statistical maps of pitch (top) and timbre (bottom), pooled across all stepsizes, both contrasted with silence. A cluster in each of right and left superior temporal gyri for pitch (center of mass: R: 56, -16, 8; L: -53, -22, 9) and timbre (center of mass: R: 56, -18, 9; L: -53, -24, 9) conditions, respectively. Color scale values range from -1 to 1, in units of percentage change relative to baseline. No voxels survive the contrast of pitch and timbre (pitch-timbre).

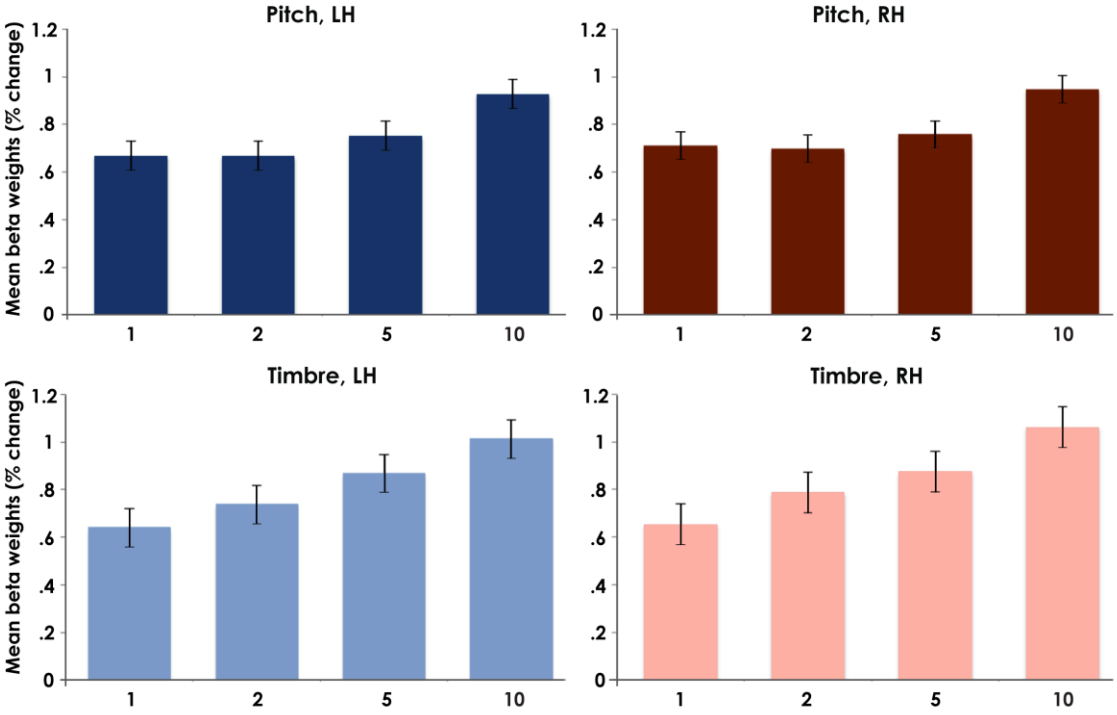

Fig. 3. Bar graphs showing mean beta weights in percentage change across all subjects' iROIs at each stepsize (1, 2, 5, and 10 DL) for pitch (top row) and timbre (bottom row) in each hemisphere (left and right). Error bars indicate +/- one standard error of the mean across subjects.

Correlations between BOLD and stepsize in pitch and timbre

The main purpose of the experiment was to identify regions that were selectively sensitive to either pitch or timbre variations. We reasoned that such regions would show increased activation with increasing stepsize (and hence sequence range and variance) in the relevant dimension. Results of the single-sample t-test of Fisher z-transformed r coefficients compared to 0 within the union of iROI masks, with a cluster threshold of 464 microliters (58 voxels), are shown in Fig. 4A. Results are limited to voxels within the MNI template. In line with the linear trends in activation with increasing stepsize observed in the analysis of iROI means, the heatmap shows that voxels within the union mask were positively correlated with stepsize in both

the pitch and timbre dimensions. In addition, there was no clear spatial separation between the regions most sensitive to pitch changes and those most sensitive to timbre changes, either within or between hemispheres. This point is illustrated further with binary versions of each map in Fig. 4A overlaid to show which voxels the two maps have in common (Fig. 4B). Previous studies found pitch to be represented in the anterior-lateral portion of Heschl's Gyrus (Patterson et al., 2002; Penagos et al., 2004; Norman-Haignere et al., 2013); however, the large degree of spatial overlap we found across these dimensions does not support strongly modular processing of pitch or timbre within this region.

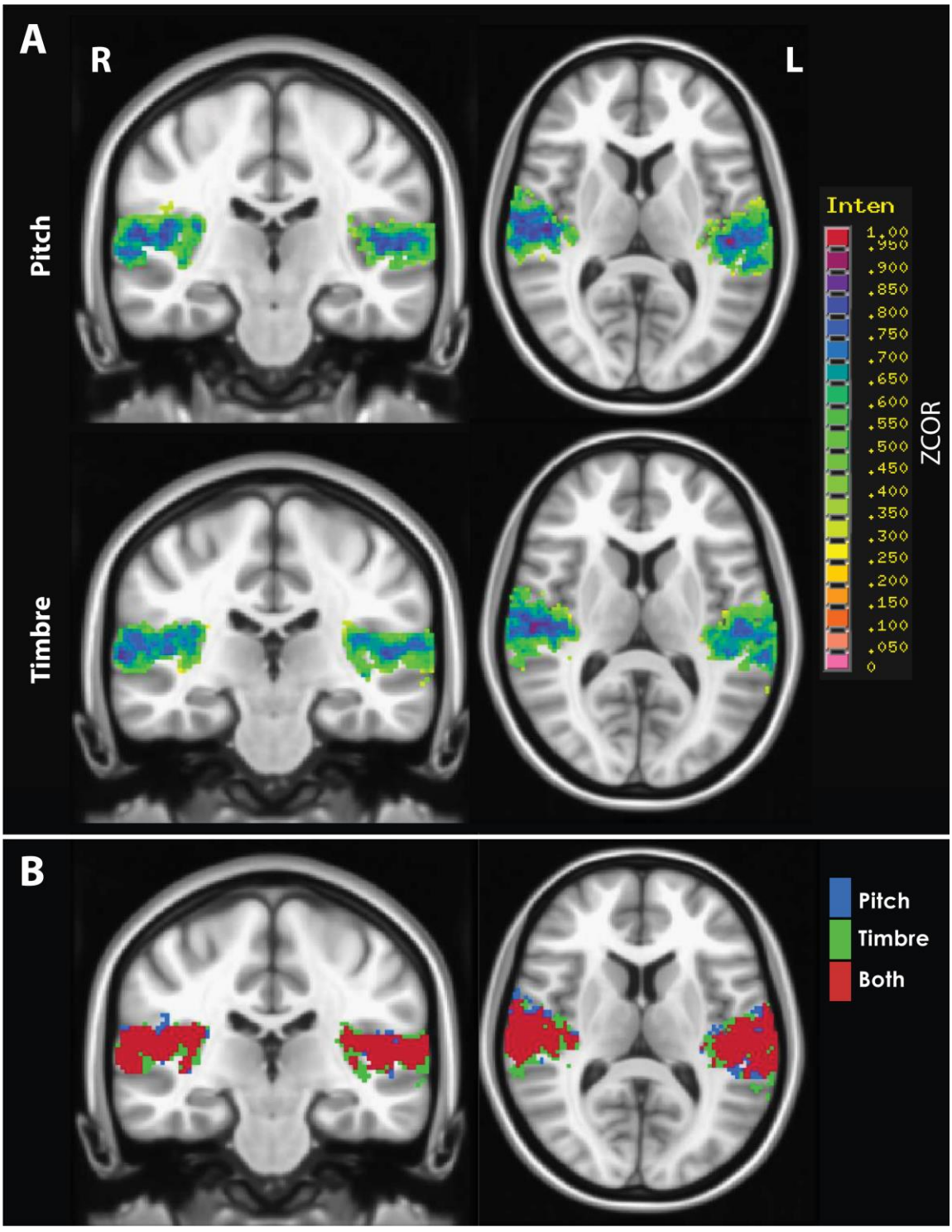

Fig. 4. Group-level correlation coefficient maps A. & B. A. Heat maps of positive mean Fisher's z-transformed correlation coefficients (ZCOR) for pitch (top) and timbre (bottom), limited to

voxels within a union of all subjects' iROI masks. No significant negative correlations were found. A cluster is shown in each hemisphere for pitch (peak: R: 52, -10, 6; L: -46, -24, 10) and timbre (peak: R: 48, -20, 12; L: -52, -18, 6) conditions, respectively. B. Maps indicating which voxels the maps in A. have in common. The significant correlation coefficients within the pitch map (blue), the significant correlation coefficients within the timbre map (green), and the voxels these two maps have in common (red).

Surface-based analyses

In order to determine whether there were any significant differences between the spatial distributions of these correlation coefficients, we identified the anterior-lateral and posterior-medial coordinates of HG on a flattened patch of auditory cortex in each hemisphere for each subject (Fig. 5). Right hemisphere coordinate systems were mirrored in the medial-lateral dimension to align with the left hemisphere. Fisher z-transformed correlations coefficients and iROI masks were transformed to the cortical surface (using AFNI's 3dVol2Surf), using the "median" sampling option to assign the median of the volume values found along the surface normal to each surface vertex, and aligned for each subject to this new coordinate system.

Surface maps of the contrast between pitch and timbre illustrate that there was no systematic difference between representations of the two dimensions in the left (Fig. 6A) or right (Fig. 6B) hemisphere. Contrast was computed as $(r^2_{pitch}-r^2_{timbre})/(r^2_{pitch}+r^2_{timbre})$, where each r represents the average (across subjects) correlation between the BOLD signal and stepsize. Projections of the data onto axes parallel to and orthogonal to HG also reveal nearly complete overlap of pitch and timbre correlations.

Histograms of pitch/timbre contrast for left (Fig. 6C) and right (Fig. 6D) hemispheres show that strong correlations with timbre were more common than strong correlations with pitch. This finding is also reflected in the steeper slopes for timbre relative to pitch in Fig. 3. Therefore,

while the spatial distribution of pitch and timbre responses is largely overlapping, the BOLD response shows stronger correlation with timbre scales, in spite of the fact that the stepsizes were perceptually matched to the pitch stepsizes.

As a final test of the spatial distribution of responses, a weighted center of mass (COM) was calculated for each subject, weighting each surface vertex by the square of the correlation coefficient (i.e., accounted variance) for either pitch stepsizes or timbre stepsizes (Fig. 6E). After Bonferroni correction for multiple comparisons, paired t-tests indicated that the left hemisphere showed a significant difference in the direction running along (parallel to) HG, going from anterior-lateral to posterior-medial in the cortex ($t_{(9)}$ = -3.9, p = 0.016 (p = 0.004, uncorrected)), but no difference in the direction running across (perpendicular to) HG ($t_{(9)}$ = 2.3, p = 0.18 (p = 0.045, uncorrected)). The right hemisphere showed no significant differences in either direction (along HG, $t_{(9)}$ = -1.9, p = 0.36 (p = 0.09, uncorrected); across HG, $t_{(9)}$ = 2.4, p = 0.172 (p = 0.043, uncorrected)). The slight divergence between the location of strong pitch and timbre correlations is also evident in the projection of the pitch/timbre contrast running parallel to HG (Fig. 6A). The weighted COM of timbre responses was more anterior and lateral than pitch responses, but the overall spatial similarity of the pitch and timbre responses and the very small difference between the COMs suggest caution in interpreting this outcome. Overall, the results do not provide support for the idea of a pitch region in the anterior portion of the auditory cortex that is not responsive to changes in other dimensions.

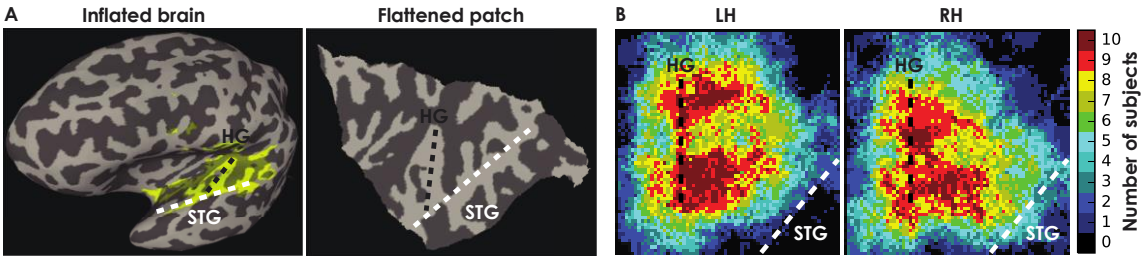

Fig. 5. Spatial distribution of the iROI masks in the auditory cortex in each hemisphere with respect to Heschl's gyrus. A. Individual subject's inflated brain (left panel) with iROI mask, and a flattened patch (right panel) of the auditory cortex. Heschl's gyrus (black dashed line) and superior temporal gyrus (white dashed line) are labeled for this subject. B. Summation of iROI masks across all subjects in the left hemisphere (left panel) and right hemisphere (right panel), color-coded to indicate the number of subjects for which each surface vertex was inside their iROI.

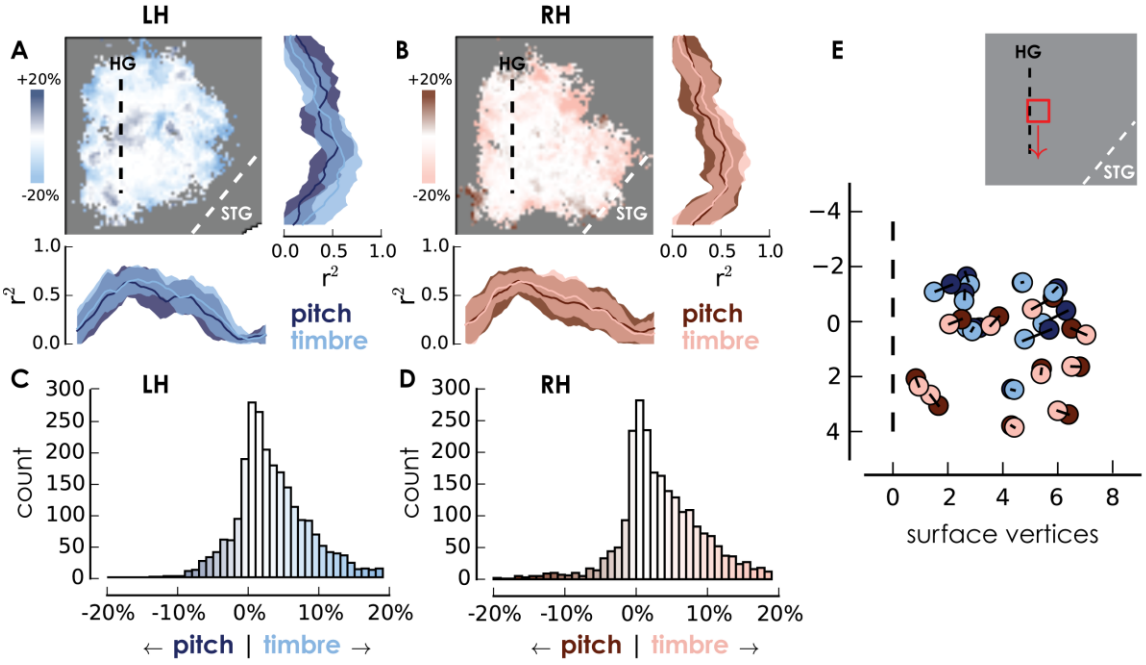

Fig. 6. Spatial distribution of correlation coefficients for pitch and timbre. A. & B. Left hemisphere (blues) and right hemisphere (reds) contrast maps within the sound mask (vertices inside the auditory ROI of at least 5 subjects), with darker colors indicating pitch had a higher correlation coefficient in a given voxel. To the right and bottom are projections of the mean (SD) proportion of variance explained, parallel and perpendicular to Heschl's gyrus. C & D. Distribution of the contrast between variance explained by pitch and timbre stepsize across all voxels within the mask in each hemisphere. E. Variance-weighted center of mass (COM) for each

subject for each dimension in each hemisphere. Black lines connect the center of mass for each condition within a hemisphere for each subject. Inset above demonstrates how small the spatial range is for the COMs.

Excitation-pattern analysis

The general similarity in responses to variations in pitch and timbre suggested the possibility of a single representation, perhaps based on the tonotopic organization within the auditory pathways that begins in the cochlea. Changes in both the F0 and the spectral centroid produce changes in the tonotopic representation of sound. It may be that the activation differences measured by our fMRI study reflect tonotopy, rather than the extraction of higher-level features, such as pitch or timbre. We tested this hypothesis by deriving the predicted changes in tonotopic representation, based on the differences in the auditory excitation pattern between successive notes produced by the pitch and timbre sequences. The predicted changes in excitation were derived using the recent model of Chen et al. (2011), which itself is based on the earlier model of Moore et al. (1997); see Moore (2014) for a review. An example of the excitation patterns generated by notes that differ in either F0 or spectral centroid is shown in Fig. 7A.

The change in excitation from one note to the next (ΔE) was quantified as the sum of the absolute differences in specific loudness across frequency. The average change in excitation ($\overline{\Delta E}$) between successive notes in the melody for each stepsize was estimated by running simulations of sequences containing 1000 notes per stepsize. This enabled us to predict the average changes in excitation at different stepsizes for both dimensions.

The predictions show that the changes in excitation are larger and vary more with stepsize for changes in spectral centroid than for changes in F0 (Fig. 7B). If BOLD responses simply reflected average changes in excitation based on tonotopy, rather than a response to the

features of pitch and timbre (where stepsizes were equated for perceptual salience across the two dimensions), then there should be a monotonic relationship between the BOLD response and the predicted excitation change ($\overline{\Delta E}$). The fact that the data do not fall on a single line, and instead separate based on whether pitch or timbre was varying, suggests that the BOLD responses are not simply a reflection of the tonotopic changes in activation produced by the stimuli.

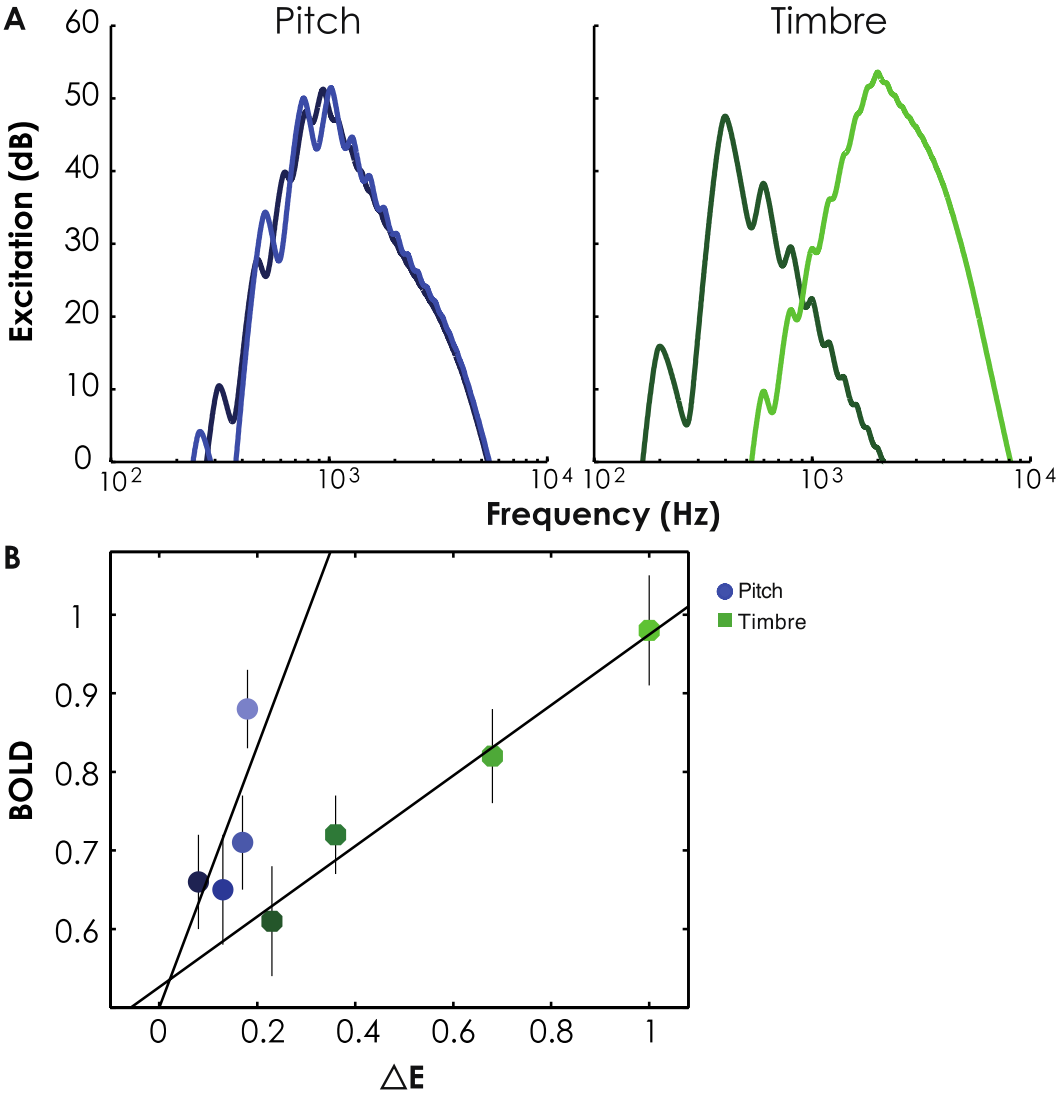

Fig. 7. A. Excitation patterns for the highest and lowest steps of the largest stepsize (10xDL) for the pitch and timbre conditions, respectively. Lighter colors indicate the higher pitch and brighter timbre, respectively. B. Scatter plot showing mean beta weight across all ten subjects at each

stepsize, averaged across hemispheres as a function of ΔE with a linear regression line for each dimension. Lighter colors indicate higher stepsizes. Error bars indicate +/- one standard error of the mean across subjects.

Multi-voxel pattern analysis

Although the univariate analyses do not support the existence of anatomically distinct pitch and timbre processing within auditory cortex, this finding does not rule out the possibility that the patterns of activity across the regions can still code for variations in the two dimensions. As suggested in the single-unit study of ferrets by Bizley et al., (2009), the same population of neurons could be used to code for both dimensions (or more). To explore this possibility, we employed MVPA (see Methods for procedural details).

Average classifier performance for predicting pitch versus timbre conditions was 61.6% across subjects, which was significantly above chance (50%), based on a two-tailed t-test ($p = 0.015$). For eight of the ten subjects, the classifier performed significantly above chance ($p < 0.0001$) for accurately discriminating pitch from timbre conditions, with performance from individual subject data ranging from 55-86% correct. These results suggest that there is a distinguishable difference in activity patterns across voxels for these conditions.

In order to determine if our results were strongly affected by the masks used, we compared our functionally defined ROI mask, based on our univariate analysis results, which was cluster-thresholded and limited to the 2000 most responsive voxels, to results using other masks types: (1) an ROI mask not limited to 2,000 voxels, but thresholded at $p = 0.01$ and cluster thresholded (resulting in a greater number of voxels), (2) a mask containing voxels strongly correlated with stepsize (created with correlation coefficient data from the *Correlations between BOLD and stepsize in pitch and timbre* section) ($p = 0.01$, cluster thresholded), and (3) a mask containing voxels strongly correlated with stepsize, intersected with the 2,000-voxel mask

(further reducing the number of voxels in each subjects' mask). Classifier performance results across masks can be seen in Table 1. Paired t-tests revealed no significant differences across mask types, suggesting the differences between voxels included in each mask type did not have a strong effect on classifier performance, and that classifier performance remained reasonably consistent within subjects.

Subject	ROI mask thresholded to 2000 vox.	ROI mask	Step size correlated vox. mask intersected with 2000 vox. mask	Step size correlated voxel mask	Mean classifier performance (%) for each subject	SD
1	**73**	70	71	68	**70.5**	2.08
2	45	43	**47**	38	**43.25**	3.86
3	49	**53**	50	47	**49.75**	2.50
4	55	52	51	**57**	**53.75**	2.75
5	65	63	**66**	64	**64.5**	1.29
6	69	**72**	69	63	**68.25**	3.77
7	56	56	**61**	60	**58.25**	2.63
8	**86**	80	70	74	**77.5**	7.00
9	**56**	**56**	55	**56**	**55.75**	0.50
10	62	**66**	60	63	**62.75**	2.50
Mean classifier performance (%) for each mask	**61.6**	**61.1**	**60**	**59**	**60.4**	1.16
SD	12.17	11.11	8.91	10.34	10.3	

Table 1. Princeton's MVPA toolbox classifier performance (in percent) distinguishing pitch from timbre conditions using four different masks. Blue fill indicates mask used in our analyses. Values in bold indicate best classifier performance for each subject.

Finally, we examined classifier performance when comparing only the largest stepsizes. Given that the largest stepsizes produce the most salient perceptual changes, these may be the easiest conditions for the classifier to differentiate. A repeated-measures 3x2 ANOVA comparing the stepsizes (all, 5 & 10, or 10), and mask type (2000 voxel mask or standard mask) showed no main effect of stepsizes or mask type, and no interactions (see Table 2), indicating that including only the stepsizes with the greatest perceptual variation did not improve classifier performance, perhaps due to the reduced amount of data when only a subset of stepsizes was considered.

	ALL STEP SIZES	ALL STEP SIZES	STEP SIZES 5 & 10	STEP SIZES 5 & 10	STEP SIZE 10	STEP SIZE 10
Subject	**ROI mask thresholded to 2000 vox.**	**ROI mask**	**ROI mask thresholded to 2000 vox.**	**ROI mask**	**ROI mask thresholded to 2000 vox.**	**ROI mask**
1	**73**	70	71	71	71	71
2	45	43	49	46	**50**	42
3	49	53	49	**56**	53	54
4	55	52	**59**	56	56	53
5	65	63	64	65	**66**	64
6	69	**72**	69	71	70	70
7	56	56	**59**	53	55	58
8	86	80	**88**	79	**88**	80
9	**56**	**56**	54	55	**56**	49
10	62	66	60	**69**	61	68
Mean classifier performance (%) for each mask	**61.6**	**61.1**	**62.2**	**62.1**	**62.6**	**60.9**
SD	12.17	11.11	11.71	10.37	11.45	11.68

Table 2. Classifier performance comparing all stepsizes to stepsize five and ten only, and ten only, across two ROI masks (ROI mask thresholded to 2000 voxels, and the standard functional mask). Bold values indicate best performance for a given subject.

Discussion

In this study, we compared human cortical processing of the auditory dimensions of pitch and timbre. Conventional univariate analyses revealed no significant differences in terms of the regions dedicated to processing variations in these two dimensions, with the exception of a slight difference in the weighted center of mass of the clusters of voxels whose responses were correlated with stepsize in the direction parallel to the HG (anterior-lateral to posterior-medial) in

the LH. These results provide no evidence for modular and exclusive processing of the two dimensions in separate regions of auditory cortex, at least on the coarse level of analysis available with fMRI.

While previous studies of pitch found active regions in the anterior portion of HG, bilaterally, providing converging evidence that these regions are important for pitch processing, we found broader bilateral regions throughout the auditory cortices that were responsive to pitch as well as timbre variation. It is possible, however, that had we contrasted our periodic stimuli with aperiodic stimuli, such as noise, we would have found elevated activation in anterior regions for pitch and timbre, consistent with dipole locations found by Gutschalk and Uppenkamp (2011) using MEG. Instead, our results focus exclusively on the contrast between pitch and timbre, and suggest that the pitch-sensitive regions in the aforementioned studies may not be uniquely dedicated to pitch processing.

Although our univariate results indicate that pitch and timbre processing takes place in common anatomical regions of the auditory cortices, their decodability using MVPA suggests that they may engage distinct circuitries within these regions. In this respect, our results are consistent with the conclusions of the single-unit study in the auditory cortex of ferrets, which also suggested population-based codes for pitch and timbre, with many neurons showing sensitivity to changes in both dimensions (Bizley et al., 2009).

We found evidence supporting our hypothesis that regions selective for pitch or timbre show increases in activation with increases in the size of the range covered within each sequence. In other words, larger variations in either pitch or timbre within the sequences led to larger changes in BOLD in both dimensions, akin to Zatorre and Belin's (2001) findings for spectral and temporal variations.

It is worth considering how the use of melodies may have affected our results. Our stimulus sets for both pitch and timbre variations were presented in the form of tone sequences

that could be perceived as pitch melodies and timbre "melodies." It has been found that pitch, loudness, and brightness (i.e., timbre) can all be used to identify familiar melodies, which suggests a substrate for detecting and recognizing patterns of sound variations that generalizes beyond pitch (McDermott et al., 2008; Graves et al., 2014). If the recognition of pitch and timbre melodies is subserved by similar cortical circuits, it seems reasonable to expect similar regions of activation. Further, melody processing is considered a higher level of auditory processing, which may be represented in non-primary auditory cortical regions (e.g., Patterson et al., 2002b). Thus, it is possible that the regions active in this study include higher-level processing than basic pitch or timbre processing, which might explain the spread of activation along the superior temporal gyri. Contrary to expectations based on higher-level processing, the activation we found was relatively symmetric across hemispheres and covered large regions of Heschl's gyrus; other studies have found limited and more right-lateralized processing of pitch melodies (e.g., Zatorre et al., 1994; Griffiths et al., 2001).

In studies of auditory perception, pitch and timbre are often treated as separable dimensions (e.g., Fletcher, 1934; Kraus et al., 2009; McDermott et al., 2010). However, several studies have also shown that the two can interact (e.g., Krumhansl and Iverson, 1992; Warrier and Zatorre, 2002; Russo and Thompson, 2005; Marozeau and de Cheveigné, 2007). A recent psychoacoustic study showed that pitch and brightness variations interfered with the perception of the other dimension, and that the interference effects were symmetric; in other words, variations in pitch affected the perception of brightness as much as variations in brightness affected pitch perception (Allen and Oxenham, 2014). The finding held for both musically trained and musically naive subjects. The strong overlap in cortical activation of the two dimensions found in the present study may also reflect the perceptual difficulty in separating the two dimensions. Although our study was not designed to investigate potential differences between people with and without extensive musical training, comparing a subset of subjects with the most training (3

subjects with 15, 16, and 23 years of training) with a subset of subjects with the least training (3 subjects with 0, 1, and 2 years of training) did not reveal any significant differences or clear trends within these groups either in terms of the degree of activation or correlation with melody range in either dimension.

Finally, one potential limitation of the study is that it involved a passive listening task. It is possible that the results may have been different if subjects had been engaged in a task that involved either pitch or brightness discrimination. Auditory attention has also been found to modulate activity in the superior temporal gyrus (e.g., Jäncke et al., 1999). Attention to auditory stimuli has been found to produce stronger activity throughout large areas in the superior temporal cortex, compared to when attention is directed towards visual stimuli (Degerman et al., 2006). When subjects were instructed to discriminate between tones and identify the brighter timbre, Reiterer et al. (2007) found activity in a bilateral network including cingulate and cerebellum, as well as core and belt areas of the auditory cortices. This same network was active when subjects were performing loudness discrimination tasks, again highlighting the existence of overlapping neural networks for processing sound. However, for timbre, Broca's area was also active, resulting in a left-hemisphere dominance, highlighting the connection between timbre discrimination and processing of vowels in language. It may be that similar dissociations between pitch and timbre would become apparent in an active version of the task undertaken in this study.

Acknowledgements

This research was supported by NIH grant R01 DC005216 and by the Brain Imaging Initiative of the College Liberal Arts, University of Minnesota. Andrea Grant, Ingrid Johnsrude, Michelle Moerel, Juraj Mesik, Zeeman Choo, and Jordan Beim provided helpful advice and assistance. The authors declare no competing financial interests.

Chapter 4

Cortical Correlates of Attention to Auditory Features

Allen, E. J., Burton, P. C., Mesik, J., Olman, C. A., & Oxenham A. J. (2018). Cortical correlates of attention to auditory features. Manuscript submitted for publication.

Abstract

Pitch and timbre are two primary features of auditory perception that are generally considered independent. However, an increase in pitch (produced by a change in fundamental frequency) can be confused with an increase in brightness (an attribute of timbre related to spectral centroid), and vice versa. Previous work indicates that pitch and timbre are processed in overlapping regions of the auditory cortex, but are separable to some extent via multivoxel pattern analysis (MVPA). Here we tested whether attention to one or other feature increases the spatial separation of their cortical representations, and whether attention can enhance the cortical representation of these features in the absence of any physical change in the stimulus. Participants listened to pairs of tones varying in pitch, timbre, or both, and judged which tone had the higher pitch or brighter timbre. Variations in each feature engaged common auditory regions with no clear distinctions at a univariate level. Attending to one feature in the presence of irrelevant variations in the other led to differences in frontal activation, but did not improve the separability of the neural representations of pitch and timbre at the univariate level. At the multivariate level, the classifier performed above chance in distinguishing between conditions in which pitch or timbre was discriminated. The results confirm that the computations underlying pitch and timbre perception are subserved by strongly overlapping cortical regions, but reveal that attention to one or other feature leads to distinguishable activation patterns, even in the absence of physical differences in the stimuli.

Keywords: pitch, timbre, auditory cortex, attention, fMRI

Significance Statement

While pitch and timbre are generally thought of as independent auditory features of a sound, pitch height and timbral brightness can be confused for one another. This study shows that pitch and timbre variations are represented in overlapping regions of auditory cortex, but that they produce distinguishable patterns of activation. Most importantly, the patterns of activation can be distinguished based on whether participants attended to pitch or timbre, even when the stimuli remained physically identical. The results therefore show that variations in pitch and timbre are represented by overlapping neural networks, but that attention to different features of the same sound can lead to distinguishable patterns of activation.

Introduction

Pitch and timbre are two fundamental perceptual dimensions of sound. Variations in pitch carry information about intonation and melody, whereas timbre is closely related to sound quality and identity. Despite the importance of pitch and timbre in auditory and speech perception, it remains unclear how they are represented in the cortex. A recent fMRI study found that pitch and timbre variations were represented in largely overlapping regions of the auditory cortex, although the patterns of activation could be distinguished using multi-voxel pattern analysis (MVPA; Allen et al., 2017). However, this conclusion was based on a passive listening task. It is possible that the representations of pitch and timbre become more spatially distinct, and thus more separable, when attention is directed to them.

Auditory attention has been found to modulate activity in wide regions of the superior temporal gyrus (STG) (e.g., Degerman et al., 2006; Jäncke et al., 1999). A recent meta-analysis by Alho et al. (2014) compared neural representations of several sound dimensions and categories (pitch, spatial location, speech, and voice processing) during active and passive fMRI measurements. Although speech or voice processing loci were not found to change with attention, pitch was found to activate more posterior and lateral areas in STG during active tasks, while the passive listening loci were shifted more anteriorly, toward the lateral end of Heschl's gyrus (HG), the macroanatomical landmark that corresponds most closely to primary auditory cortex (PAC). Although some studies have examined cortical representations of timbral dimensions (Menon et al., 2002; Reiterer et al., 2008; Allen et al., 2018) none has yet examined the effects of modulating attention to timbre. It thus remains possible that an attentionally demanding task may enhance the spatial separability of the cortical representations of pitch and timbre.

In addition to possible differences between active and passive listening conditions, participants' attention can be directed to a particular sound feature. Recent studies have shown

that attention can enhance the representation of a specific sound within a mixture (e.g., Ding and Simon, 2012a, 2012b; Mesgarani and Chang, 2012) and there is some evidence for neuronal modulation in visual cortex, as a function of attention to different features within a visual stimulus (Saenz et al., 2002); however, it is unknown whether attention to a specific auditory feature selectively enhances the representation of that feature over others, rather than just enhancing the representation of the entire object.

This study examines whether task-based attention enhances the separation of the neural correlates of pitch and timbre, relative to that found in a passive-listening task (Allen et al., 2017), and asks whether neural correlates of attention to either pitch or timbre emerge even when the physical stimulus remains identical. The results suggest that the representations of pitch and timbre variation are subserved by strongly overlapping cortical regions, even in the active-task conditions, but reveal that attention to one or other dimension can lead to distinguishable activation patterns using MVPA, even in the absence of physical differences in the stimuli.

Materials and Methods

Ethics statement

The experimental procedures were approved by the Institutional Review Board (IRB) for human subject research at the University of Minnesota. Written informed consent was obtained from each participant before starting the measurements.

Participants

Twenty right-handed subjects (mean age: 28.3, standard deviation (SD): 6.5; 10 males, 10 females) from the University of Minnesota community participated in the experiment. All subjects had normal hearing, defined as audiometric pure-tone thresholds of 20 dB hearing level

(HL) or better at octave frequencies between 500 Hz and 8 kHz. The musical experience of the subjects ranged from 0 to 23 years, with 13 subjects reporting 8 or more years of formal musical training, three reporting between 3 and 7 years, and the remaining four reporting 2 years or less.

Pre-scan experimental design

Prior to being scanned, each subject's difference limens (DLs) were measured for pitch and timbre discrimination. For pitch discrimination, we measured the DL for fundamental frequency (F0), i.e., the periodicity of a sound, a physical variable most closely associated with pitch. For timbre, we measured the spectral centroid DL, a physical manipulation that leads to reported changes in the timbral dimension of "brightness" (e.g., Fastl and Zwicker, 2007). The paradigm was similar to that used by Allen and Oxenham (2014) and Allen et al. (2017). Stimuli were generated in MATLAB (Mathworks, Natick, MA) and presented using the AFC toolbox for auditory psychophysics (Ewert, 2013). Pairs of successive harmonic complex tones were presented diotically through HD 650 headphones (Sennheiser, Old Lyme, CT) at a sampling rate of 44,100 Hz. Each tone was 500 ms in duration, with a 300 ms interstimulus interval (ISI). The tones had 20 ms raised-cosine onset and offset ramps, and harmonics up to 10,000 Hz were added in sine phase and scaled independently, producing 24 dB/octave slopes around the center frequency (i.e., the spectral centroid). The 3-dB bandwidth of the filter was ¼ octave, with complexes having no flat bandpass region. Sounds were presented at an overall level of 66 dB sound pressure level (SPL).

Participants listened to pairs of tones presented sequentially, and on each trial selected the tone with the higher pitch or brighter timbre (i.e., a standard two-alternative forced-choice procedure). Stimuli were paired with boxes on the screen that would light up with each tone, with the question, "Which pitch was higher?" or "Which timbre was brighter?" depending on the task. Feedback was then given, indicating whether the response was correct or incorrect. For the pitch

condition, the spectral centroid of the filter remained unchanged at 900 Hz, and the F0 was roved +/- 10% uniformly around 200 Hz across trials. For the timbre task, the F0 remained unchanged at 200 Hz, and the spectral centroid roved +/- 10% uniformly around 900 Hz across trials. A two-down, one-up adaptive tracking rule was used to converge on a DL for both F0 and spectral centroid, corresponding to performance of 70.7% correct (Levitt, 1971). The starting value of **Δ**F0 or **Δ**CF was 200%, which was initially increased or decreased by a factor of 2. Following the first staircase reversal (i.e., the first direction change in the tracking variable from "up" to "down") this factor was decreased to 1.26, and then to the final step size of 1.12 after two more reversals. After six reversals at this step size, the run was terminated, and the DL in each run was calculated as the geometric mean of the **Δ** value at the last six reversals points. Each participant's final DL for each dimension was based on the geometric mean DL across six task blocks. All blocks of one dimension were completed before beginning measurements in the next dimension, and this ordering was counterbalanced across subjects.

After DLs were calculated, discrimination performance was measured using a method of constant stimuli with the F0 or spectral centroid difference set to five times the DL measured for each individual participant (5DL). The reason for multiplying the DL by 5 was threefold: (1) to allow participants to perform near ceiling, confirming that they are attending to the correct dimension on each task, (2) to allow for the fact that the task was presented in the acoustically noisy MRI scanner environment (as the DLs were originally measured in silence), and (3) to ensure that the changes in pitch in timbre remained roughly equally salient (Allen and Oxenham, 2014). Performance was based on 100 trials of each: pitch alone comparisons (PA; when only F0 is varying), timbre alone comparisons (TA; when only spectral centroid is varying), both pitch and timbre varying, but with subjects attending to only pitch (PwT), and both pitch and timbre varying, but with subjects attending only to timbre (TwP). In all cases, the participants were

instructed to select the tone with either the higher pitch or brighter timbre in the tone pair (see Fig. 1).

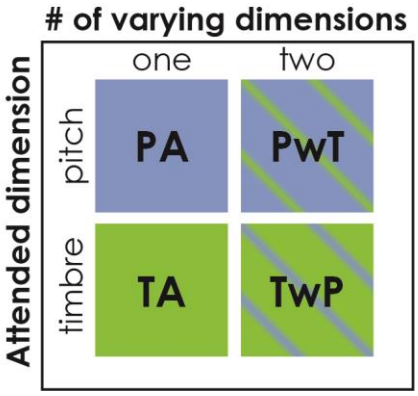

Fig. 1. Chart showing all combinations of attended dimensions (pitch or timbre) and number of varying dimensions (one or two) in the study, totaling four different experimental conditions: pitch alone (PA), pitch discrimination with timbre varying (PwT), timbre alone (TA), and timbre discrimination with pitch varying (TwP). Note that the stimuli in the PwT and TwP conditions are identical; the only difference is the dimension to which subjects were instructed to attend.

Experimental design during scan

The stimuli were presented at 63 dB SPL via MRI-compatible Sensimetrics S14 earphones with custom filters, designed to compensate for the frequency response of the hardware. Stimulus parameters were the same as those for pre-scanner behavioral testing. During each task scan, subjects completed 28 trials of one of the four discrimination tasks shown in Fig. 1 using differences that were set to be five times the DL for each subject. The direction of change for each dimension (up or down) was selected randomly and independently in each trial. These trials were evenly divided into 4 blocks, one for each condition, separated by rest periods to measure the baseline signal. Trials were presented during interaquisition intervals to reduce

acoustic contamination from the scanner. Following each trial, subjects had 2 s to respond. Subjects' responses were collected via a button-box. Stimuli were paired with boxes projected onto a screen and viewed through a mirror mounted on top of the head coil. The words "pitch," "timbre," "rest," and "end" appeared on the screen as task cues. Feedback for correct and incorrect responses appeared in the form of happy and sad emoticons, respectively. Missed responses were followed by a presentation of an asterisk symbol on the screen. Due to some technical difficulties with the button box, some subject responses were missed despite subjects having pressed the button during the allotted response window (mean number of missed responses across subjects: 3 out of 224, SD: 4.27). These missed responses were not included in the calculation of task performance.

Each condition was repeated in a pseudo-random counterbalanced order, for a total of 56 trials per condition. The two Alone conditions in counterbalanced order always preceded the two Varying conditions in counterbalanced order. For example, one scan session could be ordered as follows: PA, TA, PwT, and TwP, followed by TA, PA, PwT, and TwP (see Fig. 2). This pseudo-random counterbalancing was intended to remind subjects what pitch and timbre changes sounded like in isolation, prior to being tested on the more challenging task of attending to one when both dimensions varied.

Magnetic resonance imaging

The data were acquired at a 3T (Siemens Prisma) MRI scanner. To minimize the contamination of the functional data with scanner noise, we used a pulse sequence with sparse temporal acquisition (Hall et al., 1999). The pulse sequence used slice accelerated multiband (factor 2) echo planar imaging (EPI) (Xu et al., 2013) with a repetition time (TR) of 6 s (acquisition time of 2 s, and an inter-acquisition silent interval of 4 s), providing a voxel resolution of 2 mm isotropic. Each functional volume had 48 slices, angled upward to avoid the

eyes in an effort to reduce eye movement artifacts, while covering most of the brain. However, in many subjects, the posterior portion of the parietal cortex was not included. A total of 8 functional scans were acquired for each participant, each of which took about 4 minutes to complete and consisted of 39 volumes. To correct the spatial distortions from inhomogeneity in the B_0 magnetic field, we also collected each participant's field map. To localize functional activations, we additionally collected anatomical T1-weighted images which were co-registered with the EPI data.

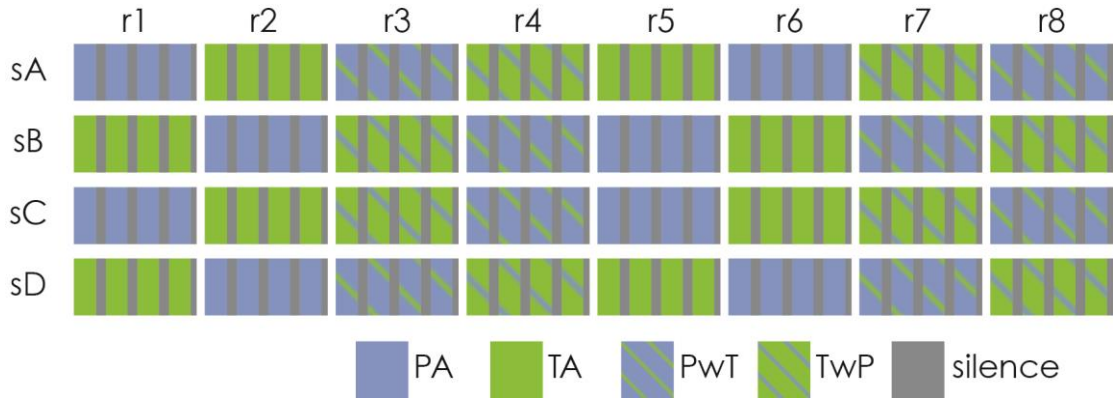

Fig. 2. Schematic diagram of the pseudo-random counterbalancing of eight functional runs within a scanning session for four different subjects. Abbreviations: r = run, s = subject, PA = pitch alone, TA = timbre alone, PwT = pitch with timbre varying, TwP = timbre with pitch varying. Note that the stimuli used in the PwT and TwP conditions were identical.

Statistical analysis. Data were preprocessed using the Analysis of Functional NeuroImages (AFNI) software package (Cox, 1996) and FSL 5.0.4 (http://fsl.fmrib.ox.ac.uk/). Statistical analyses and visualization were performed with AFNI. Preprocessing included distortion correction via FSL's FUGUE, six-parameter motion correction, spatial smoothing (4 mm FWHM Gaussian blur), and pre-whitening.

For each subject, an event-related general linear model (GLM) analysis was performed that included regressors for each of the four experimental conditions, six motion parameters, and Legendre polynomials up to the fourth order to account for baseline drift (modeled separately for each run). Each subject's brain was transformed into Montreal Neurological Institute (MNI) space (Mazziotta et al., 1995). Beta weights (regression coefficients) for individual voxels were estimated by the GLM for each condition for each subject, as were contrasts comparing conditions within pitch, within timbre, between pitch and timbre, and a contrast comparing all sounds to baseline. Cortical surface-based visualization was done in AFNI's SUMA (SUrface MApping) https://afni.nimh.nih.gov/Suma using the FreeSurfer brain surface MNI N27.

Group-level analyses with subject as a random effect included paired-sample t-tests performed on the unmasked, unthresholded beta weights for each contrast using the AFNI program 3dttest++. Voxels were thresholded at $p < 0.01$, uncorrected. Correction for multiple comparisons was achieved by determining the minimum significant cluster size. Taking into account increasing concerns over the risk of inflated false positives with this method, as reported by Eklund et al., (2016), a nonparametric permutation test was used. This permutation test randomized the signs of the residuals of the model among subjects, and then performed a t-test, with these steps iterated 10,000 times, to determine nearest-neighbor, faces touching, two-sided cluster thresholds. This method, implemented within 3dttest++ using the 'clustsim' option, determined the probability, with each voxel having a 1% chance of displaying a false positive, of clusters of a given size occurring by chance. Based on these probabilities, clusters smaller than those that would occur by chance more than 5% of the time were filtered out of the results to achieve a cluster-level $\alpha = 0.05$. The t-tests were conducted within a gray matter mask containing anatomically defined auditory cortices and frontal lobe regions (see Fig. 3). The mask, which was created on the cortical surface, was made up of the following gyri and sulci in the left and right hemispheres: superior temporal (including banks), Heschl's, supramarginal, precentral, superior

frontal, middle frontal (caudate and rostral), inferior frontal (opercularis, triangularis, orbitalis), orbitofrontal (lateral and medial), and anterior cingulate (caudal and rostral), as well as the insulae, temporal poles, and frontal poles. These regions were defined by the Desikan-Killiany Atlas (Desikan et al., 2006).

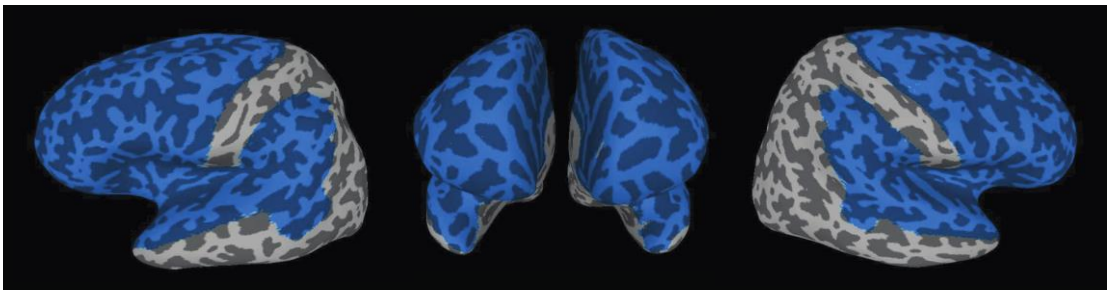

Fig. 3. Mask of the auditory cortex and frontal lobe regions. (From left to right) left hemisphere, front-facing view, and right hemisphere.

Multivoxel pattern analysis

In addition to univariate analyses, we employed multivoxel pattern analysis (MVPA) which has the advantage of being more sensitive to differences between conditions than the univariate approach, as it examines the patterns of activity across several voxels, as opposed to averaging across them (Norman et al., 2006) and may reveal differences at the voxel level that are not apparent via standard univariate analyses. MVPA was performed using Princeton's MVPA toolbox for MATLAB with a backpropagation classifier algorithm (http://code.google.com/p/princeton-mvpa-toolbox/). In order to restrict the number of voxels in our analyses, we added a functionally defined mask based on our univariate analysis results. This mask contained voxels that were most active during the auditory tasks (all sound conditions contrasted with the silence baseline), thresholded to the 2,000 most active (positive) voxels across both hemispheres for each subject. This cutoff was chosen so that the number of voxels in each

mask was the same across subjects, and to limit the number of voxels used for classification (e.g., De Martino et al., 2008; Schindler et al., 2013).

For baseline estimation, TRs immediately following a transition from a sound trial to rest were eliminated, to account for the lag in the hemodynamic response as it dropped back down to baseline during rest (silence). The preprocessed data were then normalized by dividing by the mean baseline signal to eliminate any between-run differences caused by baseline shifts and multiplied by 100, converting the data into percent signal change.

Given that each run in our scan sessions consisted entirely of one condition type (and silent periods), as shown in Fig. 4A, our experimental design was incompatible with the traditional leave-one-run-out cross-validation procedure used in MVPA packages. To rectify this, we used our data to create a set of pseudo-runs, each containing an even sampling of all four conditions. To do this, we took our preprocessed data, in percent signal change, and divided each of our eight runs into four blocks, consisting of seven trials each. These block lengths were chosen because each series of seven consecutive trials in a run was followed by 18 seconds of silence. The first trial within each block was removed, again to account for the lag in the hemodynamic response. We also removed spikes in the timecourse that were more than four standard deviations from the mean of the run (excluding silence). The remaining trials within each block were then averaged together, resulting in one value per block per voxel. We then divided the resulting 32 activation patterns (8 runs x 4 blocks) into 8 pseudo-runs, each containing one activation pattern per condition. These pseudo-runs, depicted as columns in Fig. 4B, were then z-scored and subjected to eight-fold cross-validated MVPA analysis, where each training set consisted of 7 pseudo-runs, totaling 28 patterns, while the remaining pseudo-run (4 patterns) was used as the testing set.

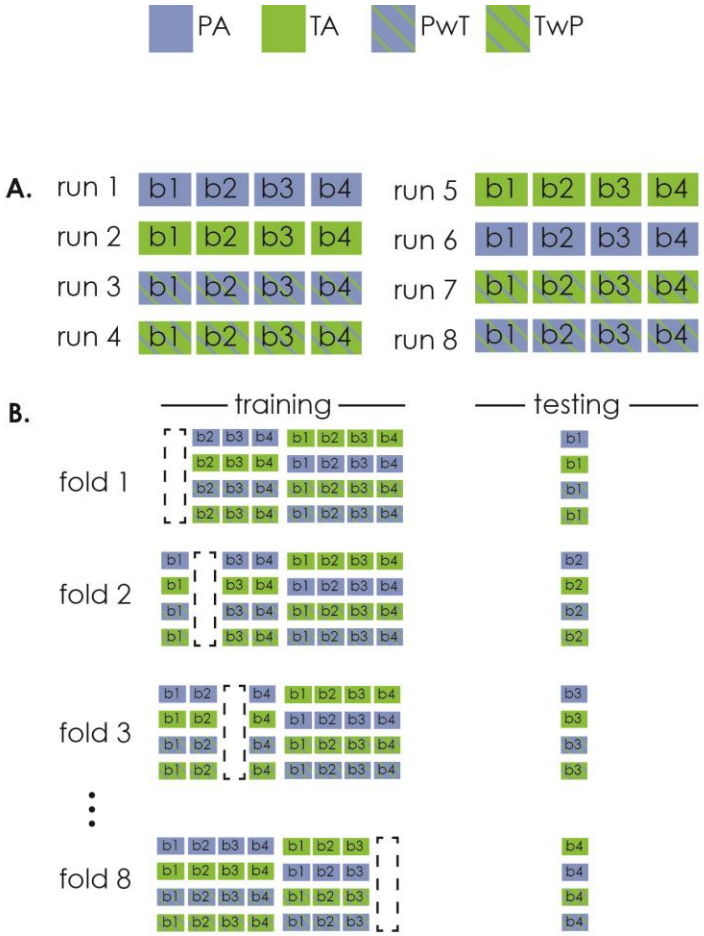

Fig. 4. Cross-validation procedure for MVPA. A. The eight functional runs are represented as rows, each parsed into four blocks of trials, totaling 32 blocks per subject. B. The pseudo runs are represented as columns. In each fold of the cross-validation, 28 of the blocks were used for training and four were used for testing, with one block of each condition type represented in the testing data. Abbreviation: b = block.

Results

Pre-scan behavioral task performance

The geometric mean F0 DL across participants was 1.06%, 95% CI [0.7 1.5], and the average spectral-centroid DL was 4.3%, 95% CI [3.5 5.1], in good agreement with earlier studies

using similar stimuli (e.g., Allen and Oxenham, 2014). As anticipated, performance on the constant-stimuli task utilizing the scaled 5*DL variation was high, with a mean proportion of correct responses of 95.6% (SD: 5.7%) across all conditions. Average performance within conditions is reported in Table 1A. Due to the near-ceiling performance in these tasks, a non-parametric Friedman test on the four conditions indicated a significant difference between conditions ($\chi^2 = 20.67$, $p < 0.0001$). We then ran Wilcoxon signed ranks tests to compare conditions. As reported in Table 1B, after a Bonferroni correction for multiple comparisons, no significant difference in performance between the pitch and timbre conditions was found. This was true when comparing PA and TA conditions, as well as the PwT and TwP conditions. There was, however, a significant difference between both alone (BA) and both varying (BV) conditions, and this difference existed between PA and PwT, as well as TA and TwP, indicating that, as expected, the conditions in which both dimensions were varying were more challenging than the conditions in which only one dimensions was varying.

A.

Condition(s)	Mean % (SD)
All Pitch	94.6 (6.9)
All Timbre	96.6 (3.9)
PA	97.8 (3.7)
TA	98.3 (1.5)
BA	98.0 (2.8)
BV	93.2 (6.7)
PwT	91.5 (7.9)
TwP	94.9 (4.8)

B.

Rank order	Z statistic	p-value (corrected)
All pitch < All timbre	-1.24	1.32
PA < TA	-0.21	5.04
PwT < TwP	-1.48	0.84
BV < BA	-4.73	0.0006*
PwT < PA	-3.50	0.0006*
TwP < TA	-3.09	0.012*

Table 1. Pre-scan behavioral task performance. A. Mean and standard deviation of all conditions and B. Wilcoxon signed ranks test Z statistics and Bonferroni corrected p-values comparing task performance between conditions. Asterisks indicate significant differences.

Behavioral task performance during scanning

Similar to performance in the pre-scanner session, performance in the scanner was near ceiling (mean: 95%, SD: 4.3). Average performance within conditions is reported in Table 2A. Again, a non-parametric Friedman test indicated a significant effect of condition ($\chi^2 = 16.94$, $p < 0.001$), so Wilcoxon signed ranks tests were used to compare pairs of conditions. As reported in Table 1B, after a Bonferroni correction, there were no significant differences between the pitch and timbre conditions, neither between the PA and TA conditions, nor between the PwT and TwP conditions, suggesting the perceptual salience and subsequent task difficulty remained relatively equivalent across dimensions at 5*DL. However, as in the pre-scanner task performance, the performance for BA conditions was significantly better than performance for BV conditions. When comparing the PA and PwT conditions, the difference was marginal after correction, but remained significant for the TA versus TwP conditions.

A.

Condition(s)	Mean % (SD)
All Pitch	94.1 (7.4)
All Timbre	95.8 (5.9)
PA	96.8 (5.5)
TA	98.1 (1.8)
BA	97.5 (4.1)
BV	92.4 (7.8)
PwT	91.4 (8.1)
TwP	93.4 (7.5)

B.

Rank order	Z statistic	p-value (corrected)
All pitch < All timbre	-0.79	2.58
PA < TA	-0.35	4.38
PwT < TwP	-0.77	2.64
BV < BA	-3.73	0.0006*
PwT < PA	-2.48	0.06
TwP < TA	-2.81	0.03*

Table 2. Behavioral task performance in the scanner. A. Mean and standard deviation of all conditions and B. Wilcoxon signed ranks test Z statistics and Bonferroni corrected p-values comparing task performance between conditions. Asterisks indicate significant differences.

Group-level analysis

For all conditions compared to baseline (silence), we found robust activation of the auditory cortices (see Fig. 5). Significant clusters were found within the combined auditory cortex and frontal lobe mask for the contrast between either dimension varying alone and both dimensions varying, while attending to one (BV – BA): BA conditions had significant clusters in the right medial orbitofronal gyrus, right cingulate gyrus, and right and left superior frontal gyri (SFG); the BV conditions had significant clusters in left inferior frontal gyrus (IFG), left middle frontal gyrus (MFG), and right and left anterior insulae (see Fig. 6).

Considering the same contrast, but for each dimension individually, the TwP – TA contrast revealed that TwP had a significant cluster in left IFG, a significant cluster in left medial frontal gyrus, and a significant cluster in right anterior insula, whereas no regions showed significantly greater activation for TA (see Fig. 7). The PwT – PA contrast showed no significant differences. Additionally, when contrasting any pitch condition to any timbre condition (i.e., PA – TA, PwT – TwP, or all pitch conditions – all timbre conditions), no significant clusters were found.

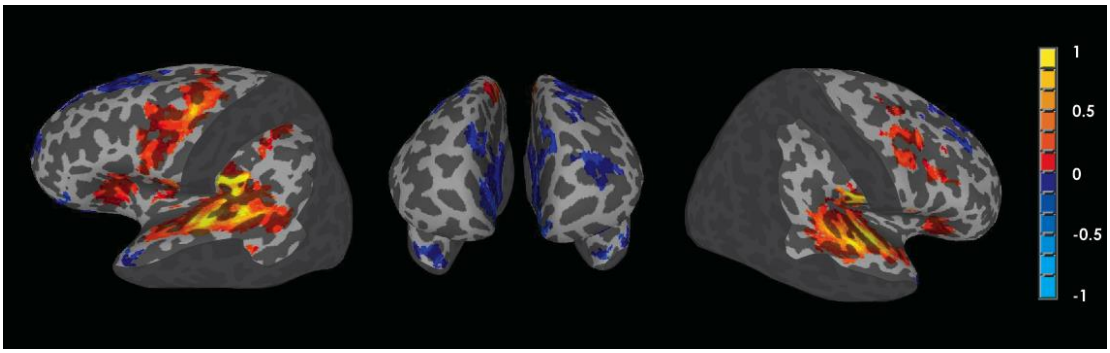

Fig. 5. Group-level statistical map on an inflated brain showing the mean of all sound conditions relative to baseline (masked, thresholded at the single voxel level [$p < 0.01$], and cluster

thresholded [p < 0.05]). Color scale values range from -1 to 1 units of percentage change. For reference, the grayed out areas, which are the regions not colored blue in Fig. 3, have been added to denote regions not included in analysis.

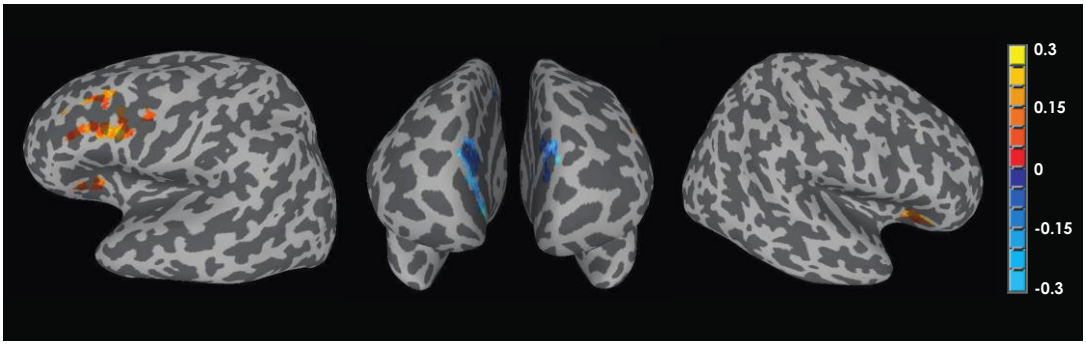

Fig. 6. Group-level statistical map for the BV – BA contrast (masked, thresholded at the single voxel level [p < 0.01], and cluster significance thresholded [p < 0.05]). Warm colors indicate voxels responding more strongly during the BV tasks and cool colors indicate voxels responding more strongly during the BA tasks. Color scale values range from -0.3 to 0.3 in units of percentage change.

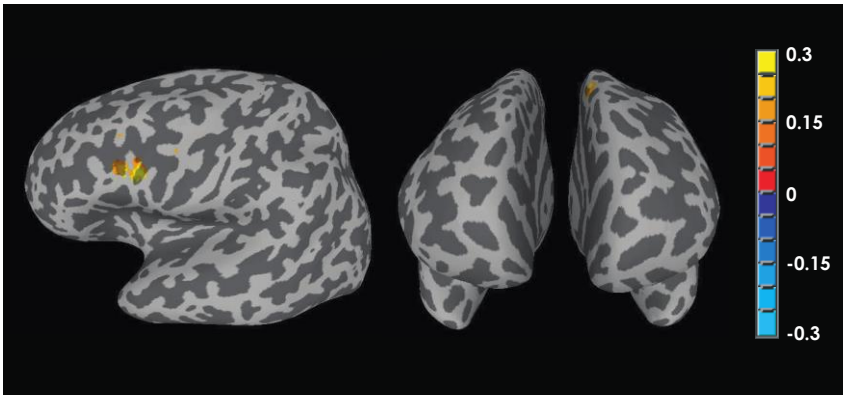

Fig. 7. Group-level statistical map for the TwP – TA contrast (masked, thresholded at the single voxel level [p < 0.01], and cluster significance thresholded [p < 0.05]). Warm colors indicate voxels responding more strongly during the TwP task and cool colors indicate voxels responding

more strongly during the TA task. Color scale values range from -0.3 to 0.3 in units of percentage change.

MVPA results

As shown in Fig. 8, average 4-way classifier performance for distinguishing between pitch alone, timbre alone, pitch varying, and timbre varying was 86.6% [SD = 8.0 percentage points], which was significantly above chance (25%), based on a one-tailed t-test ($t_{19} = 34.3$ $p < 0.0001$). Average classifier performance for pitch conditions versus timbre conditions was 88.2% [SD = 6.7], which was also significantly above chance (50%) ($t_{19} = 25.6$, $p < 0.0001$). Average classifier performance for BA conditions versus BV conditions was 86.6% [SD = 8.1], which was also significantly above chance (50%) ($t_{19} = 20.1$, $p < 0.0001$).

Additionally, we tested how well the MVPA classifier performed if it was trained on the BA conditions but tested on the BV conditions. This test determines whether the cortical representations of PA (and TA) can predict the differences in representation under conditions where both dimensions are varying (PwT and TwP) but participants are attending to either pitch or timbre. Classifier performance for each subject is shown in Fig. 9. While there was variability in the performance, average performance across subjects was 61.8% [17.1], which was significantly above chance, based on a one-tailed t-test ($t_{19} = 3.08$, $p < 0.003$). This outcome shows that attention to each dimension enhances the pattern corresponding to changes in that dimension, even when the physical stimulus is identical.

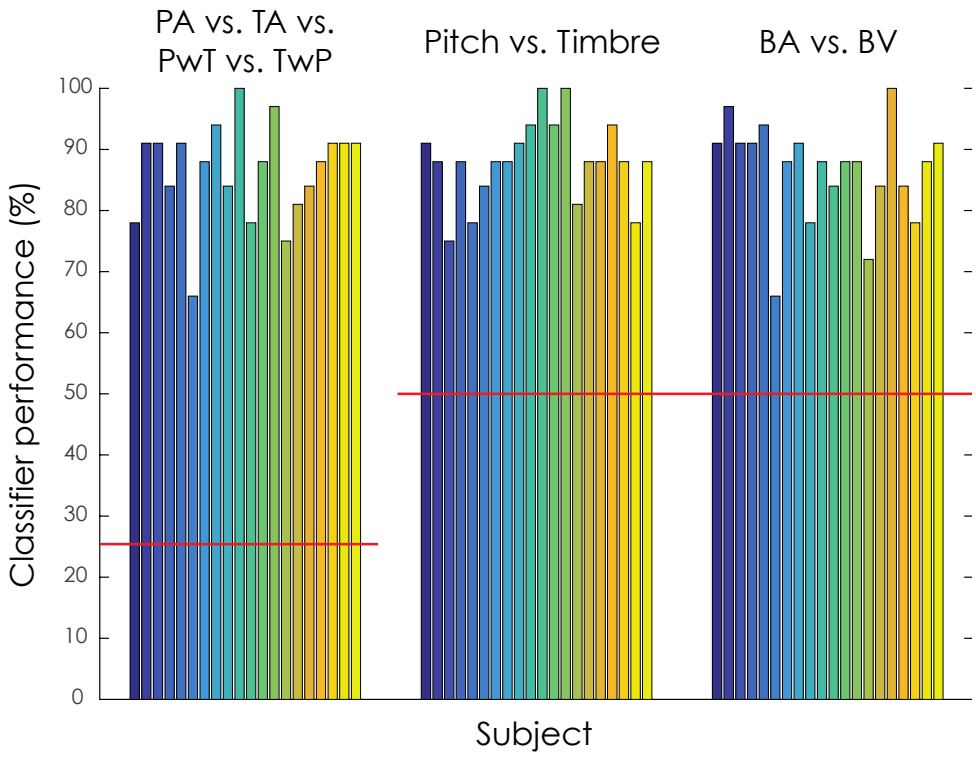

Fig. 8. MVPA average classifier performance for each of the 20 subjects (indicated by bar color) on three different classifications. Abbreviations: PA = pitch alone, TA = timbre alone, PwT = pitch with timbre, TwP = timbre with pitch, BA = both alone, BV = both varying. The horizontal red lines denote chance performance.

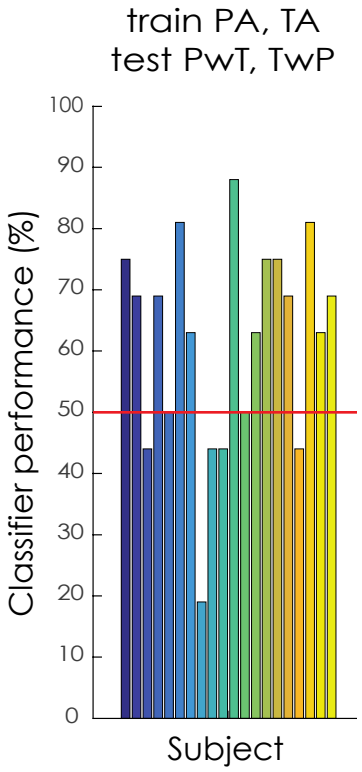

Fig. 9. MVPA average classifier performance for each of the 20 subjects (indicated by bar color) when training on alone conditions (PA and TA) and testing on the varying conditions (PwT and TwP). The red line denotes chance performance.

Exploratory MVPA and univariate analyses

In addition to the leave-one-run-out cross-validation using the pseudo runs, some additional exploratory MVPA analyses were performed. First, we analyzed the errors in classifier performance for the 4-way classifier to determine whether, despite its high performance, there were any conditions that were consistently confused with one another (e.g., when the correct condition was PA, was it classified more often as PwT than TA or TwP?). Results are shown in Fig. 10. While it is difficult to draw any strong conclusions, as there are no obvious indications of the classifier consistently confusing one condition for another, these results reveal that the classifier performed well across all four conditions.

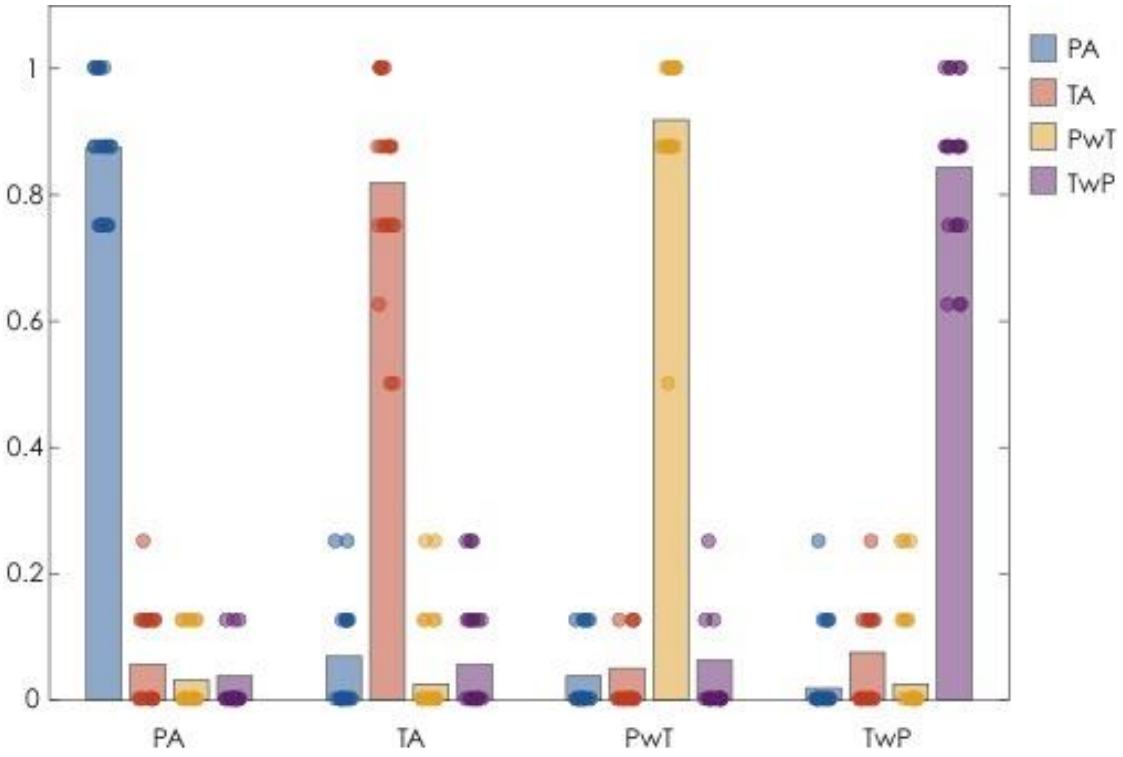

Fig. 10. Classifier confusion patterns for the 4-way classifier. Bar graphs showing average classifier guesses for each condition (PA, TA, PwT, and TwP). Individual subject classifier guesses are superimposed on the corresponding bars. X-axis labels indicate the correct condition.

Additionally, we performed MVPA on subset of the conditions, comparing just the PA with TA conditions, and comparing just the PwT with TwP conditions. In both cases, classifier performance was significantly above chance. We then performed MVPA on half of the conditions, comparing just the PA with PwT conditions, and comparing just the TA with TwP conditions. In both cases, again, classifier performance was significantly above chance. Results are reported in Table 3A.

Next, in order to ensure that our non-standard approach of extracting the blocks from our single-condition scans and arranging them in pseudo runs was not producing inflated results, we

randomized the condition labels to see if classifier performance would drop to chance. Indeed, with random labels, no classifier performed significantly above chance (Table 3B). Next, we trained the classifier on half of the data and tested it on the other half. In our first iteration, we split the data by training on the first two blocks of each run, and testing on the second two blocks of those runs (and vice versa). As expected, since the classifier had less training data than in the leave-one-run-out cross-validation (87.5% training, 12.5% testing), classifier performance was worse, but remained significantly above chance (Table 3C). Our second 50-50 split involved training on the first four runs and testing on the second four runs (and vice versa). In this case, classifier performance dropped further, in some cases no longer reaching significance (Table 3D).

One potential reason for this performance drop is that during training, the classifier may have overfit run-specific signals, leading to poor generalization in the test run. This possibility appears particularly likely when considering the sources of information that the classifier can rely on. When there is just one run per condition in the training set, classification can rely either on run-specific differences that are unrelated to the experimental condition, signals related to the experimental condition itself, or some combination of the two. If the former contribution is significant, then the performance in the test run should drop significantly, consistent with our results above. When, however, trials from multiple runs become intermixed in training, run-specific signals should become less reliable since they are inconsistent across trials from different runs. Consequently, the classifier is more likely to learn the condition-specific signatures of the BOLD signal, which are consistent across runs, and should maintain good performance with the test data, as we observed in the original analysis. As such, given that our experimental design contains only two runs per condition, our original approach to training the classifier with pseudo-runs containing trials from all runs should have been more sensitive to the condition-specific signals that this study is concerned with.

A.	**Subset of conditions**	**Mean % (SD)**	**t_{19}**	**p-value**
	PA vs. TA	88.0 (9.4)	18.0	0.0001*
	PwT vs. TwP	92.4 (8.9)	21.3	0.0001*
	PA vs. PwT	89.1(9.5)	18.4	0.0001*
	TA vs. TwP	90.0 (8.2)	21.7	0.0001*
B.	**Randomized Labels**	**Mean % (SD)**	**t_{19}**	**p-value**
	4-way (chance: 25%)	28.1 (10.6)	1.31	0.102
	pitch vs. timbre	44.8 (11.7)	1.98	0.97
	BA vs. BV	47.7 (12.5)	0.84	0.79
C.	**50% training (1/2 the blocks) 50% testing (1/2 the blocks)**	**Mean % (SD)**	**t_{19}**	**p-value**
	4-way (chance: 25%)	64.4 (12.1)	14.5	0.0001*
	pitch vs. timbre	76.3 (12.8)	9.17	0.0001*
	BA vs. BV	72.5 (11.5)	8.72	0.0001*
D.	**50% training (1/2 the runs) 50% testing (1/2 the runs)**	**Mean % (SD)**	**t_{19}**	**p-value**
	4-way (chance: 25%)	30.2 (15.7)	1.47	0.08
	pitch vs. timbre	59.2 (14.1)	2.90	0.005*
	BA vs. BV	53.3 (20.3)	0.72	0.24

Table 3. Exploratory MVPA. A. Classifier performance on subsets of the conditions. B. Classifier performance with randomized labels (should be at chance). C. Classifier performance when training on half of the blocks in each run and testing on the other half. D. Classifier performance when training on half of the runs and testing on the other half. Asterisks indicate significant differences.

In light of the high classifier performance in the leave-one-run-out cross-validation, we examined whether masking individual subjects' results by using their 2,000 most active voxel masks would also reveal significant contrasts at the univariate level. At a threshold of $p < 0.01$, and a cluster threshold of 17, based on a Monte Carlo simulation, the majority of subjects did not have significant contrasts: 8 of the 20 subjects showed significant clusters for the pitch vs. timbre contrast, 7 subjects showed significant clusters for the BA vs. BV contrast, 4 subjects showed

significant clusters for the PA vs. TA contrast, and 7 subjects showed significant clusters for the PwT vs. TwP contrast. The two-tailed point-biserial correlation between MVPA classifier performance and the existence of a significant univariate contrast for the same comparison (i.e., pitch vs. timbre, BA vs. BV, PA vs. TA, or PwT vs. TwP) was not significant for pitch vs. timbre, BA vs. BV, PA vs. TA, or PwT vs. TwP. Results are reported in Table 4. This suggests the MVPA classifier performance was not driven by significant contrasts at the univariate level.

	***R*pb**	**p-value**
pitch vs. timbre	0.23	0.34
BA vs. BV	-0.27	0.26
PA vs. TA	0.21	0.37
PwT vs. TwP	0.14	0.56

Table 4. Exploratory correlations. A. Point-biserial correlations between MVPA classifier performance and univariate results.

Discussion

The present study aimed to determine whether task-related attention to one dimension when listening to sounds varying in pitch height and/or brightness would lead to more spatially distinct representations of pitch and timbre. Univariate analyses suggest that, both at the group level and for the majority of participants at the individual level, this was not the case. No significant differences in activation were observed between the pitch-varying and timbre-varying conditions, regardless of whether variations in the other dimension were present or not. Thus, it seems that the spatial overlap between representations of pitch and brightness is observed under both passive (Allen et al., 2017) and active listening conditions.

Despite the lack of significant univariate differences between pitch and timbre conditions,

differences did emerge in frontal regions as a function of the number of dimensions varying (one or two). All frontal regions identified appear to be part of the attentional network (for a review, see Petersen and Posner, 2012), and suggest that these regions were being differentially recruited for conditions in which only one dimension was varying compared to conditions in which both dimensions were varying. Specifically, when only one dimension was varying, it elicited a significantly stronger activation in the right and left superior frontal gyri, right medial orbitofrontal cortex, and right cingulate gyrus, while the varying conditions had significantly stronger activation in the left IFG and the anterior portion of the insulae in both hemispheres. The orbitofrontal cortex, known to be involved in decision-making (e.g., Wallis, 2007), has been shown to respond to sound, have direct connections to the auditory cortex in animal studies, and may be able to modulate sound processing (Romanski et al., 1999; Winkowski et al., 2017). The superior frontal gyrus has been found to be involved in working memory (e.g., Rypma and D'Esposito, 1999), and activation of the anterior cingulate has been linked to attentional demand in the auditory domain (Benedict et al., 1998). It remains unclear, however, why regions associated with decision-making, working memory, and attention would respond more strongly during the easier tasks, in which only one dimension is varying. The left IFG is where Broca's area, a language processing region, is known to be located (e.g., Binder et al., 1997). However, this region has been shown to respond to non-speech sounds as well (for a review, see Fadiga et al., 2009), so it is not surprising that this region would be responsive to variations in pitch and timbre. The insula has been shown to be involved in many types of auditory processing (for a review, see Bamiou et al., 2003), and may be involved in the integration of bottom-up detection of salient stimuli and top-down attentional control (Menon and Uddin, 2010). While no voxels survived the contrast between PA and PwT conditions, the TwP – TA contrast did reveal differences. Specifically, significant clusters were found in frontal regions left IFG, superior frontal gyrus, and anterior insula for the TwP condition, but no clusters were significant for the

TA condition, suggesting the more demanding timbre tasks were linked to the recruitment of additional frontal lobe resources.

While the univariate analyses could not differentiate between condition pairs, surprisingly high MVPA classification performance was obtained with the four-way classifier (PA vs. PwT vs. TA vs. TwP), as well as both two-way classifiers tested (alone vs. BV, pitch vs. timbre, PA vs. TA, and PwT vs. TwP). Despite there being no significant overall spatial differences between pitch and timbre representation, it is clear that the patterns of activation within these regions are distinct. This outcome is consistent with earlier findings obtained under passive listening conditions (Allen et al., 2017), and extends them by showing that classification remains possible even under conditions where there are no physical differences between the stimuli. This distinction suggests that representations within auditory cortex may already reflect the perception of features more strongly than they represent the physical stimuli themselves. In this way, the outcomes extend earlier work using EEG, MEG, and ECoG, which has shown that attention to entire auditory objects or streams (e.g., one voice in the presence of another) can profoundly alter cortical activity (Hillyard et al., 1973; Ding and Simon, 2012a; Zion Golumbic et al., 2013). The current study suggests that this modulation of attention extends to features within auditory objects, and not just entire objects. There is evidence to suggest such feature-based attentional modulation also exists in visual cortex for visual objects (Saenz et al., 2002).

A possible future direction would be to use encoding models to explicitly characterize pitch and timbre selectivity throughout the auditory cortex, and explore how these populations are modulated by attention. These approaches have been successfully used to characterize suppressive stimulus interactions in the visual (Brouwer and Heeger, 2011) and somatosensory systems (Brouwer et al., 2015), and could be similarly useful in understanding the interactions between representations of pitch and timbre.

Overall, these results show that actively attending to either dimension does not result in a

spatial separation in their representations that is detectable via conventional univariate analyses, but that the patterns of activation within these regions appear to be distinct for pitch and timbre. In addition, attending to one dimension results in patterns of activation that can be predicted by the patterns of activation recorded when just one dimension is varied, suggesting that attention to one auditory dimension can enhance that dimension's cortical representation, in the absence of any physical change in sound.

Acknowledgements

This research was supported by NIH grant R01 DC005216, the High Performance Connectome Upgrade for Human 3T MR Scanner 1S10OD017974-01, and by the Brain Imaging Initiative of the College Liberal Arts, University of Minnesota. Andrea Grant and Jordan Beim provided helpful assistance. The authors declare no competing financial interests.

Chapter 5

Encoding of Natural Timbre Dimensions in Human Auditory Cortex

Allen, E. J., Moerel, M., Formisano, E., & Oxenham A. J. (2018), Encoding of natural timbre dimensions in human auditory cortex, NeuroImage, 60-70.

Abstract

Timbre, or sound quality, is a crucial but poorly understood dimension of auditory perception that is important in describing speech, music, and environmental sounds. The present study investigates the cortical representation of different timbral dimensions. Encoding models have typically incorporated the physical characteristics of sounds as features when attempting to understand their neural representation with functional MRI. Here we test an encoding model that is based on five subjectively derived dimensions of timbre to predict cortical responses to natural orchestral sounds. Results show that this timbre model can outperform other models based on spectral characteristics, and can perform as well as a complex joint spectrotemporal modulation model. In cortical regions at the medial border of Heschl's gyrus, bilaterally, and regions at its posterior adjacency in the right hemisphere, the timbre model outperforms even the complex joint spectrotemporal modulation model. These findings suggest that the responses of cortical neuronal populations in auditory cortex may reflect the encoding of perceptual timbre dimensions.

Keywords: auditory cortex, encoding models, music, perception, timbre

Introduction

Timbre, the perceptual quality or color of a sound, is defined as everything by which a listener can distinguish between two sounds with the same loudness, pitch, spatial location, and duration (ANSI, 2013). For instance, it is differences in timbre that allow us to distinguish a violin from a guitar, or one vowel sound from another. Among the typical adjectives that fall under the category of timbre are “brightness”, “clarity”, “harshness”, “fullness”, and “noisiness” (Stepanek, 2006). Efforts have been made to identify and quantify the most salient aspects of timbre through the use of multidimensional scaling (MDS) techniques (e.g., Grey, 1977; Elliott et al., 2013). MDS utilizes subjective measures to determine how perceptually similar a selection of sounds are to one another, thereby creating a geometric representation that derives the subjective distances between a diverse set of stimuli using as few dimensions as possible (Grey, 1977). After collecting similarity ratings for musical instrument sounds with unique timbres, Grey (1977) used MDS to identify three dimensions that best represented the distribution of timbres. The first dimension was related to the spectral energy distribution of the sounds (ranging from a low to high spectral centroid, corresponding to timbral descriptors ranging from dull to bright), and the other two related to temporal patterns, such as whether the onset was rapid (like a struck piano note or a plucked guitar string) or slow (as is characteristic of many woodwind instruments) and the synchronicity of higher harmonic transients.

Grey’s influential study contained only sixteen instrumental sounds from three instrument families, placing some limits on the generalizability of the outcomes, and used sounds that may not have all had exactly the same fundamental frequency (F0), which itself may have affected some aspects of timbre judgments (e.g., Moore and Glasberg, 1990; Warrier and Zatorre, 2002; Allen and Oxenham, 2014). Elliott et al. (2013) extended Grey’s approach by using 42 natural orchestral instruments from five instrument families, all with the same F0 (311 Hz, the E♭ above

middle C). After collecting similarity and semantic ratings, they performed multiple analyses, including MDS. They consistently found five dimensions to be both necessary and sufficient for describing the timbre space of these orchestral sounds.

The aim of the current study was to determine whether similar dimensions can be identified in the cortical representations of timbral differences. Although the literature on the neural representations of timbre is limited, there is some evidence to suggest it is processed in both primary and secondary auditory cortical regions including superior temporal sulcus (STS), posterior Heschl's gyrus (HG), and planum temporale (PT), bilaterally, with possible hemispheric asymmetries (Casey, Thompson, Kang, Raizada, & Wheatley, 2012; Halpern, Zatorre, Bouffard, & Johnson, 2004; Menon et al., 2002; Staeren, Renvall, De Martino, Goebel, & Formisano, 2009; Warren et al., 2005). However, previous studies have not attempted to differentiate the neural representations of different timbral dimensions, and have not explored the possibility that a subjectively based model of timbre could predict patterns of cortical activation in response to sound. In the present study, we use fMRI encoding (Kay, Naselaris, Prenger, & Gallant, 2008; Moerel, De Martino, & Formisano, 2012; Santoro et al., 2014) to determine whether neural populations in the cortex can represent the timbre dimensions identified by Elliott et al. (2013), and compare this model's performance with that of models based on the spectral and temporal characteristics of the sounds.

Materials and Methods

Ethics statement

The experimental procedures were approved by the Institutional Review Board (IRB) for human subject research at the University of Minnesota. Written informed consent was obtained from each participant before starting the measurements.

Participants

Ten right-handed subjects (mean age of 28.6 years, standard deviation [STD] = 8.6 years; five females, five males) participated in this study. All subjects had normal hearing, defined as audiometric pure-tone thresholds of 20 dB hearing level (HL) or better, at octave frequencies between 250 Hz and 8 kHz, and were recruited from the University of Minnesota community. Musical experience of subjects ranged from zero to 18 years, with eight of the 10 subjects having at least 10 years of musical experience.

Stimuli and procedure

The stimulus set consisted of 42 professionally recorded natural Western orchestral instrument sounds, taken from the study of Elliott et al. (2013). The sounds were originally obtained from the McGill University Master Samples collection (Opolko & Wapnick, 2006) and were manipulated to all have the same F0 of 311 Hz (E♭), and a subjective duration of one second, as described in Elliott et al. (2013). Spectrograms for a subset of these sounds are shown in Fig. 1. Instrument families included strings, flutes, brass, single reeds, and double reeds. When the rms of the stimuli was normalized, the perceptual loudness of the sounds at the level of 75 dB SPL varied noticeably. In order to equalize the perceived loudness of the stimuli, we processed them using a loudness model (Chen et al., 2011; Moore, 2014b), and scaled the sounds to produce roughly equal predicted loudness for each sound. This resulted in perceptually equal loudness for 41 of the 42 sounds. One of the sounds, a muted C trumpet, required manual adjustment to subjectively match the perceptual loudness of the other sounds, presumably because certain aspects of the sound (e.g., sharp attack and broad spectrum) were not adequately captured by the

loudness model. The adjusted level was selected by four raters (inter-rater differences were no more than 2 dB).

After the loudness adjustments, the average level of the sounds was 74 dB SPL and the range was 62 to 81 dB SPL (STD = 3.2 dB). Sounds were presented via MRI-compatible Sensimetrics (Malden, MA) S14 earphones with custom filters.

Magnetic resonance imaging

Images were acquired in a 3T MR scanner (Siemens Prisma) at the Center for Magnetic Resonance Research (CMRR, University of Minnesota) using a 32-channel head coil. For each subject, we collected anatomical images and a functional dataset. The MPRAGE T1-weighted anatomical image parameters were: repetition time (TR) = 2600 ms; echo time (TE) = 3.02 ms; matrix size = 256 x 256; 1 mm isotropic voxels. The acquisition parameters for the functional scans were: TR = 2400 ms; time of acquisition (TA) = 1000 ms; silent gap = TR – TA = 1400 ms; TE = 30 ms; multiband factor = 4; number of slices = 44; matrix size = 672 × 672; 2 mm isotropic voxels. Slices were angled to align with the Sylvian Fissure, and covered the majority of the brain. However, for most subjects the top of the parietal and frontal lobes were excluded, along with the bottom of the occipital lobe.

The functional dataset followed an event-related design, where the sounds were presented in the silent gaps between acquisitions. Six functional runs were collected per subject. In each run, a unique subset of seven of the 42 sounds was repeated four times in pseudo-random order. The division of sounds into separate sets of seven was important for maintaining independence between training and testing datasets in the fMRI encoding analysis (see below). The stimuli within each sound set were manually selected to include a variety of instruments across multiple instrument families. These sound sets remained consistent across subjects, but the presentation order of the stimuli within each set was randomized, and the order of the sets throughout the

scanning session was counterbalanced across subjects in a Latin-square design. The presentation times of the sound trials were pseudo-randomly jittered with an interstimulus interval of 2, 3, 4, or 5 TRs. Three silent trials (with no stimuli present) and three catch trials were also included in each run. For the catch trials, intended to keep subjects alert, they were instructed to perform a one-back task in which they pressed a button any time a successive repeat of the same sound was presented. This one-back task never occurred for the same sound more than once in a given run. For the one-back task repeats, the maximum jitter was set to 4 TRs (9.6 s). The one-back task catch trials were excluded from analysis. With the 28 test sounds (four repetitions of seven sounds from the collection) and 3 catch-trial sounds, a total of 31 sounds were presented per run, along with 3 silent trials. Including about 10 s of silence preceding each run and about 5 s following each run, the total duration of one run was approximately 5 minutes.

The data were preprocessed in BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). Preprocessing included slice scan time correction (using cubic spline), 3D motion correction (using trilinear/sinc interpolation) aligned to the first volume of the first run, and a high-pass filter (GLM-Fourier) cutoff of 3 cycles per run. Distortion correction was performed using the Correction based on Opposite Phase Encoding (COPE) plugin in BrainVoyager QX, which estimated distortions based on volumes from a posterior-anterior (PA) phase-encoding (PE) direction and volumes from an AP PE direction (Fritz et al., 2014), and applied corrections to the functional data. Functional slices were coregistered to the anatomical data, and then normalized to Talairach space (Talairach & Tournoux, 1988). Automatic segmentation with manual corrections of the gray matter (GM) - white matter (WM) boundary was performed using the anatomical data. Using this boundary, each hemisphere for each subject was then inflated and brought to Cortex Based Aligned (CBA) space (Goebel, Esposito, & Formisano, 2006). CBA-averaged group-level GM-WM meshes were also generated in BrainVoyager QX.

Sound representation by the encoding models

We used fMRI encoding to test several hypotheses for how the brain represents the timbre of natural orchestral instruments. Under the fMRI encoding approach, each hypothesis is defined as an encoding model. We can distinguish between hypotheses by comparing the accuracy with which each of the trained models is able to predict the fMRI response patterns to novel testing sounds. We tested the performance of four encoding models, described below.

First, the subjective *timbre* model represents the hypothesis that responses to the sounds are well described by the five dimensions of timbre identified by Elliott et al. (2013) (see Fig. 1). The first dimension, D1, was semantically described as 'hard, sharp, high-frequency energy balance'. The second dimension, D2, was described as 'varying level, dynamic, vibrato, ringing release'. D3 was characterized as 'noisy, small instrument, unpleasant'. Sounds scoring high on D4 were described as 'compact, steady pitch, pure'. Finally, D5 had no significant correlates among semantic descriptor pairs. Figure 2A shows the sounds' representation in the space of the *timbre* model. The values of each sound on each of the five dimensions were taken from Elliott et al. (2013). As they were obtained using MDS, the five timbral dimensions were not correlated (Figure 2B).

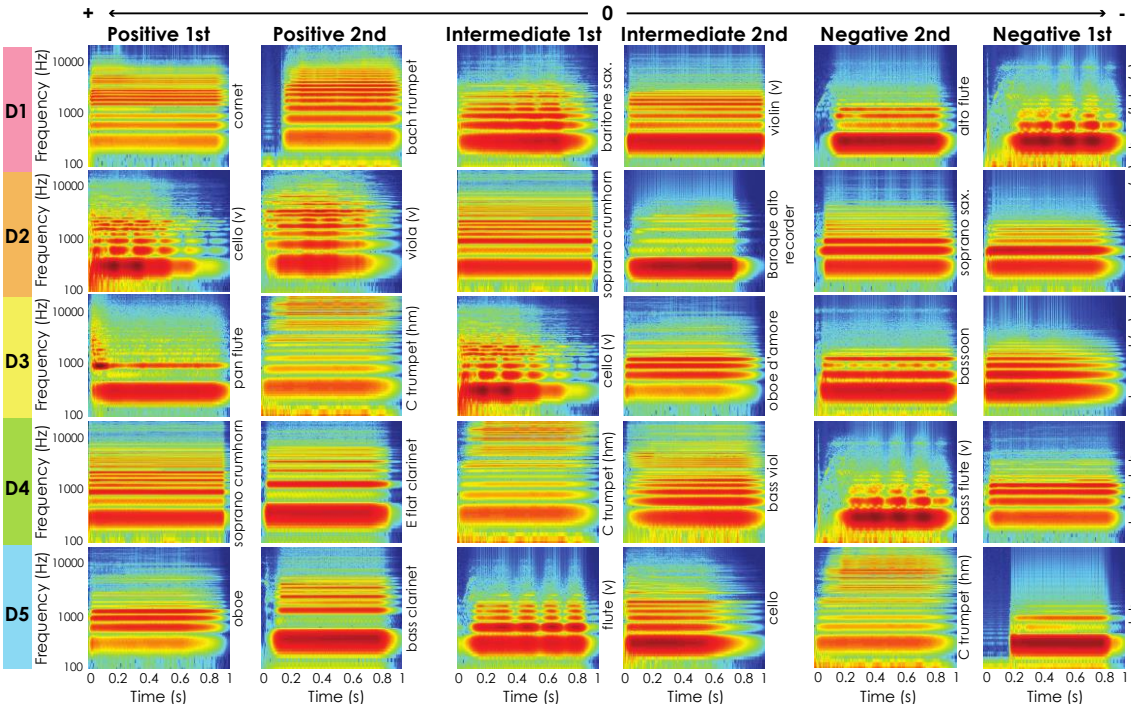

Fig. 1. Spectrograms of the sounds with (columns from left to right) the two most positive, two intermediate, and the two most negative values on each of the five timbre dimensions (rows). Abbreviations: v = vibrato, m = muted, h = harmonic.

Second, the *joint spectrotemporal modulation* (*STM*) model represents the hypothesis that cortical sound processing is well represented by the frequency-specific spectrotemporal modulation tuning of neuronal populations. Sounds are expressed by their frequency-specific spectrotemporal modulation content, obtained as the output of a two-stage biologically inspired model of auditory processing (Chi et al., 2005; Santoro et al., 2014; NSL Tools package, available at http://www.isr.umn.edu/Labs/NSL/Software.htm). This model is similar to the *timbre* model in that it takes into account both spectral and temporal properties of sound, but relies solely on the physical description of sound (transformed via simulated auditory processing), and not on any human subjective judgments. The first stage of this model mimics 'early' auditory processing, and consists of 128 overlapping bandpass filters equally spaced along a logarithmic

frequency axis (180-7040 Hz; range of 5.3 octaves). The output of this 'early' stage is a spectrogram, which serves as input to the second 'cortical' stage of the model. This stage uses a set of modulation filters (temporal modulation center frequencies, ω) and spectral modulation center frequencies (cycles/octave, Ω) to extract the spectrotemporal modulation content from the spectrograms. The modulation filters are applied at each time-frequency bin, and the absolute value of the complex-valued model output is then averaged over time. The full STM model contained ω = 30 features, and Ω = 15 features. We divided the frequency axis into 128 bins with equal bandwidth in octaves, and averaged the modulation energy within each frequency bin, resulting in 57,600 features (128×30×15). The sounds' frequency-specific spectrotemporal modulation characteristics as represented by this full model are shown in Fig. 2D-F. This full model was then reduced to 36 features in order to fit it to the fMRI data. The 36 features were: ω = [3, 9, 27] Hz × Ω = [0.5, 1, 2] cycles/octave, with the frequency axis divided into 4 bins with equal bandwidth in octaves. The spectral and temporal modulation filters had Q_{3dB} values of 1.2 and 1.8, respectively. The 36-feature limit was chosen on account of having 42 unique sounds in our stimulus set and wanting to ensure that the number of features in the model was less than the number of unique sounds in our stimulus set. Correlations between the model's 36 features are shown in Figure 2C.

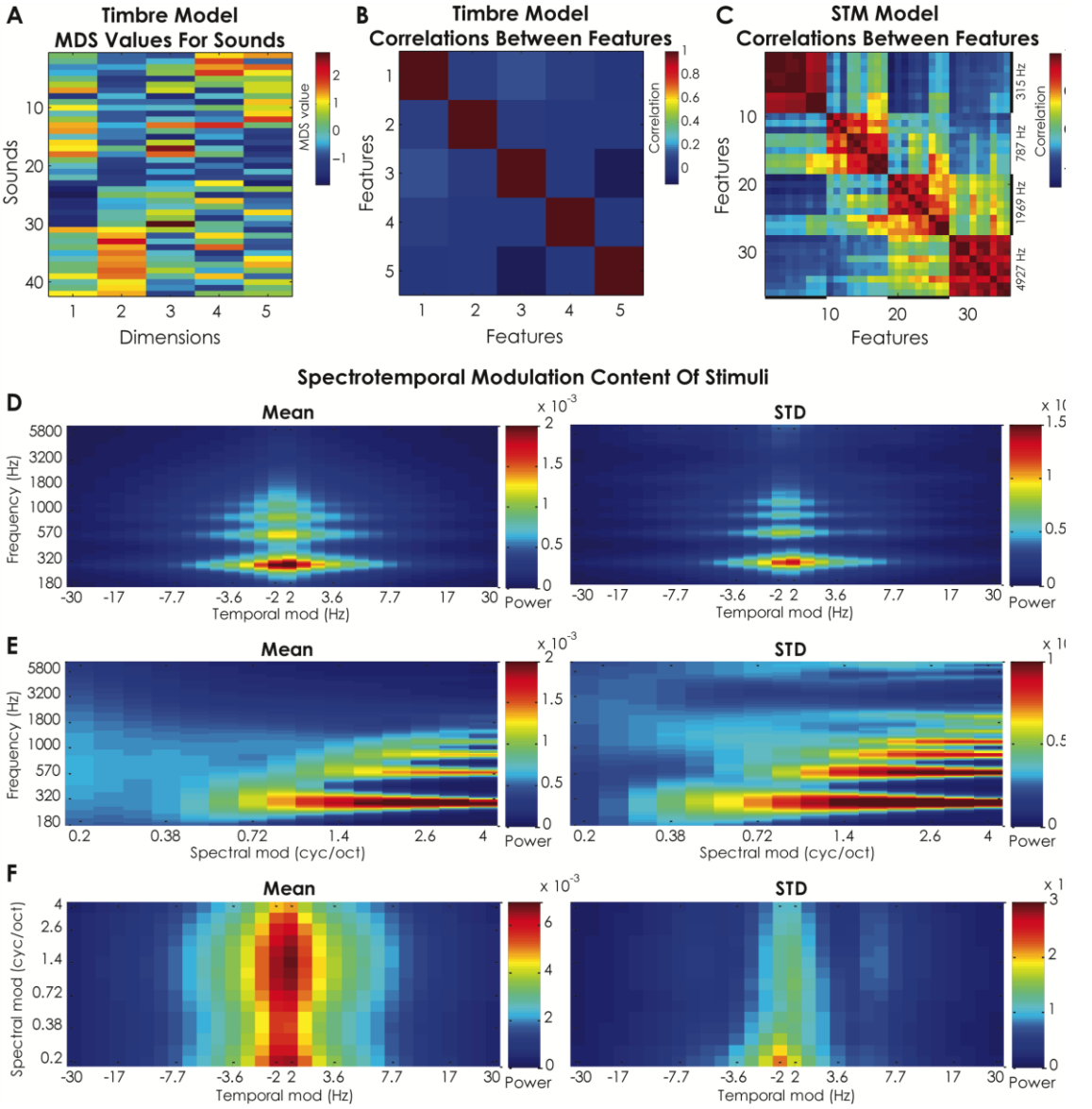

Fig. 2. Sound representation by the *timbre* and *STM* models and frequency-specific spectrotemporal modulation content of the sounds. (A) MDS values for all 42 sounds across the five dimensions (i.e., features) of the *timbre* model, taken from Elliott et al. (2013). (B) Correlation between each of the five *timbre* model features. (C) Correlation between each of the 36 *STM* model features reflecting a high correlation between spectrotemporal modulation features within the same frequency bin. Frequency bins are labeled on the right y-axis. (D) The distribution of temporal modulations across frequency, (E) the distribution of spectral modulation

across frequency, and (F) spectral modulations as a function of temporal modulations. The mean and standard deviation (STD) across sounds are shown in the left and right column, respectively.

Third, the *cochlear filter mean* model represents the hypothesis that responses to the sounds are well described by the spectral content of the sounds and the frequency tuning exhibited in the cochlea. This model therefore postulates that the cortical responses reflect primarily the long-term spectral profile of sounds, as filtered by the cochlea, without regard to their temporal properties. The representation of the sounds in the space of this model was obtained based on the output of the first stage of the model underlying the *STM* model. The resulting "cochleograms" were averaged over time, and the frequency axis was divided equally into 36 logarithmic frequency bins (resulting in 36 model features).

Finally, the *spectral centroid* model represents the hypothesis that cortical coding of timbre is dominated by the spectral centroid of a sound, corresponding to the perception of "brightness" or "sharpness" (e.g., von Bismarck, 1974), as represented by Grey's (1977) first dimension, and reflected by cortical tuning to the sounds' spectral centroids. This is essentially a simplified version of the cochlear filter mean model, in that it postulates that the spectral centroid of the sound dominates the representation over other spectral features. The spectral center of gravity c, for each sound was identified by taking the sum of the frequencies f_i, weighted by their normalized amplitudes a_i:

$$c = \frac{\Sigma(f_i a_i)}{\Sigma a_i}$$

The sounds' representation in the model space was then obtained by creating a [1 x f] vector of zeros for each sound (where f represents the center frequencies of the frequency bins of the

cochlear filter mean model), and assigning the frequency bin that contained that sound's spectral centroid with a value of one. Frequency bins that did not contain the centroid for any of the 42 sounds were removed. A total of 17 frequency bins remained, resulting in 17 features for this model.

Model training and testing

Model training and testing was done using MATLAB (Mathworks, Natick, MA). We performed the analysis independently for the training and testing runs, which contained completely distinct sets of sounds. That is, model training and testing were performed with 6-fold cross-validation. For each cross-validation, 5 runs (i.e., 35 sounds) served for model training and one run (i.e., 7 sounds) was left out for model testing.

The fMRI responses to the 42 natural orchestral instrument stimuli were estimated as follows. For each cross-validation, the training data were used to compute noise regressors using the GLMdenoise technique (Kay et al., 2013; GLMdenoise available at: http://kendrickkay.net/GLMdenoise/), and to estimate the hemodynamic response function (HRF) of each voxel across all sounds. This HRF was fixed, and was used in a regression analysis that included the regressors as estimated by GLMdenoise, to estimate the amplitude of the voxel's response (i.e., the beta weight) to each of the training and testing sounds. Next, we identified the voxels that responded significantly to the sounds ($T > 3.5$, $p < 0.001$, uncorrected). For these voxels, regularized linear regression (ridge regression; see Santoro et al., 2014, for details) was used to compute the relationship between the measured fMRI responses and the stimulus features of each model. This relationship (i.e., the trained model) represented how much each feature contributed to a given voxel's response, referred to as the voxel's population response function.

The trained model was evaluated by its ability to predict the fMRI responses to the set of testing sounds that were not used for model training. First, to gain insight into overall model

performance (across all regions with a significant response to the sounds) we computed a sound identification *prediction accuracy score*. Activity patterns for each of the test sounds were used to predict the sound identity based on its correlations with the predicted patterns of activity for each of the seven test sounds. These correlations were then sorted and assigned a rank score between one and seven (seven being the lowest rank). In the case of perfect performance, the correlation between the predicted and actual patterns would always be ranked higher for comparisons within the same sound than across different sounds, so the correlation rank, r_i, would always be 1. In the case of chance performance, the expected correlation rank would be in the middle, i.e., 4. Prediction accuracy P_i was then computed for each sound *i* using the following formula:

$$P_i = 1 - \left(\frac{r_i - 1}{N_{test} - 1}\right)$$

where r_i is the rank across the N_{test} = 7 sounds in the test set. The overall prediction accuracy was then computed as the mean of *P* across all sounds (i.e., averaging across the 6 cross-validation folds), yielding a value between zero and one (perfect prediction score = 1; chance = 0.5). This method for calculating prediction accuracy, while less common than forced-choice accuracy measures that look exclusively at stimuli that are accurately classified (i.e., those that ranked first), has the advantage of taking into account the whole distribution of ranks (beyond those ranked first) to assess the model performance (see e.g., Kay et al., 2008; Moerel et al., 2012; Santoro et al., 2014).

Second, in order to gain insight into the variations in model performance throughout brain areas, we evaluated model accuracy per voxel. For each voxel, we computed the correlation between predicted and measured responses to the testing sounds. Resulting correlations were Fisher's z transformed, and averaged across cross-validations to obtain a map of prediction accuracy per subject for each encoding model.

Group map generation and analysis. Group maps of model prediction accuracy were computed by smoothing single subject prediction accuracy maps, with local averaging up to a distance of four vertices (repeat value = 4) that were then brought into CBA space. For each vertex that was included in at least eight individual subject maps, a one sample t-test was performed to test if the observed prediction accuracy (i.e., the correlation between predicted and observed responses to testing sounds) was significantly greater than 0. Following the correction for multiple comparisons using False Discovery Rate (FDR), resulting maps were thresholded at $q(FDR) < 0.05$.

In order to compare the prediction accuracy of two encoding models, single subject prediction accuracy maps were smoothed (repeat value = 4) and brought into CBA space. For each vertex that was included in at least eight individual maps, a paired samples t-test was performed to test if there was a significant difference between the prediction accuracies of the two encoding models. If more than eight subjects were available for a given vertex, paired t-tests were run on a random selection of eight subjects out of all available subjects (this step was taken to ensure equal degrees of freedom and equal number of possible permutations across vertices, see below). To correct for multiple comparisons we used a cluster size thresholding method based on nonparametric permutations. That is, for each vertex we applied the paired t-test to all possible permutations of the eight subjects across the two models ($2^8 = 256$ permutations), resulting in 256 permuted maps. We then generated a null distribution of cluster size, considering a single-voxel threshold of $t > 1.8$. Cluster sizes that occurred less frequently than in 5% in the null distribution were considered significant.

Finally, we created group maps for each dimension of the trained *timbre* model. This was an exploratory analysis, with the aim of gaining insight into the cortical representation of the timbre dimensions. For each timbre dimension, we obtained the single subject map as the voxels' weights under the trained *timbre* model and smoothed the maps with a Gaussian kernel of 2 mm

full-width at half-maximum (FWHM). We converted the individual subject maps to binary maps, by setting the voxel to -1 or 1 if the weight was smaller or greater than zero, respectively. Next, the individual subject binary maps were brought to CBA space. Probability maps were created by assigning each voxel with the proportion of subjects that showed the same sign in their weight map (chance = 0.5; perfect congruency among subjects = 1; map threshold set to 0.75).

Results

We observed significant responses to the sounds throughout the superior temporal cortex bilaterally (see Fig. 3). The temporal auditory responsive regions included Heschl's gyrus (HG), and adjacent regions on Heschl's sulcus (HS), planum polare (PP), planum temporale (PT), superior temporal gyrus (STG), and superior temporal sulcus (STS). Beyond the auditory cortices, we observed responses to the sounds in the inferior frontal gyrus, the inferior frontal sulcus, the postcentral gyrus, and the intraparietal sulcus.

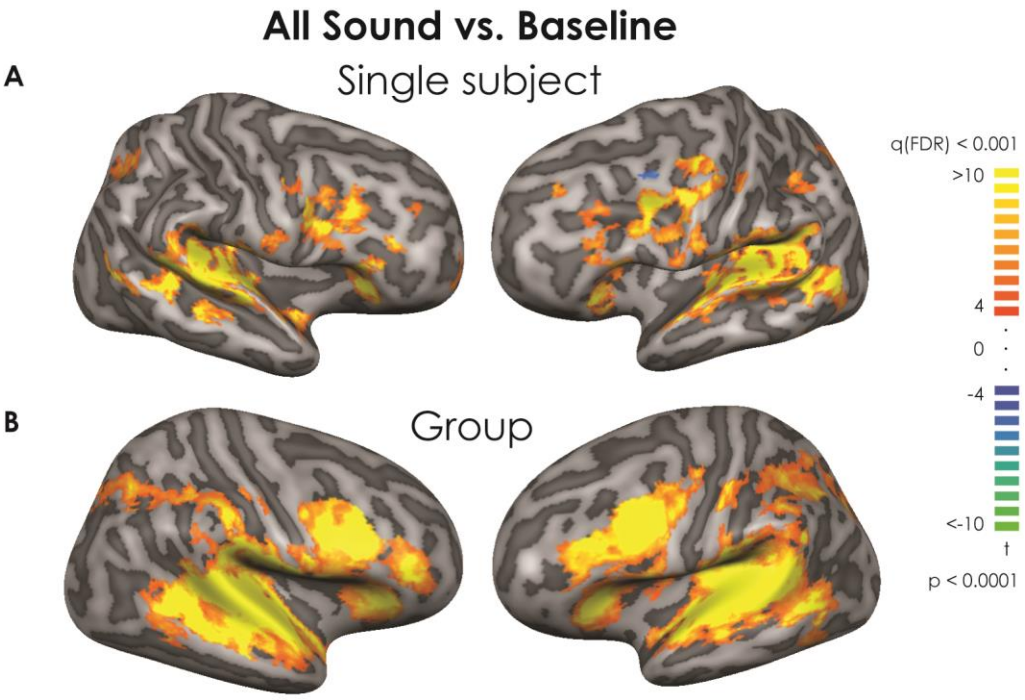

Figure 3. Brain maps showing average activation across all runs and across all sounds compared to baseline. (A) fMRI response of a single subject to sound stimuli. (B) Group-level fixed-effects GLM maps. Both the single subject and group maps are thresholded at $P < 10^{-4}$ (corresponding to q(FDR) < 0.001), cluster thresholded (cluster size = 25), with nearest-neighbor interpolation.

Prediction accuracies for the four encoding models are shown in Figure 4. All models except for the *spectral centroid* model performed significantly above chance (0.5) in a one-tailed t-test (mean [SE]; *timbre*: 63% [0.02], $t_9 = 5.97$, $P = 0.0001$, $d = 1.89$; *STM*: 60% [0.01], $t_9 = 6.72$, $P < 0.0001$, $d = 2.12$; *cochlear filter mean*: 56% [0.02], $t_9 = 3.39$, $P = 0.004$, $d = 1.07$). The *timbre* model performed significantly better than the *cochlear filter mean* model ($t_9 = 2.93$, $P = 0.02$, $d = 0.93$), and the *spectral centroid* model ($t_9 = 3.89$, $P = 0.004$, $d = 1.21$). The *STM* model also performed significantly better than the *spectral centroid* model ($t_9 = 3.70$, $P = 0.005$, $d = 1.13$). There was no significant difference between the *timbre* model and the *STM* model ($t_9 =$

1.26, $P = 0.24$, $d = 0.40$), nor between the *cochlear filter mean* model and the *spectral centroid* model ($T_{(9)} = 0.41$, $P = 0.69$, $d = 0.49$).

To test whether the *STM* model's prediction accuracy might improve with the inclusion of more features, we also ran a version of the model that contained 576 features (36 frequency bins X 4 spectral modulations [0.5 1 2 4] X 4 temporal modulations [1 3 9 27]. The average prediction accuracy [SE] in this case was: 59% [0.02], which was not significantly different from the 36-feature version ($t_9 = 0.48$, $P = 0.64$, $d = 0.15$), nor did it outperform the *timbre* model ($t_9 = 1.59$, $P = 0.15$, $d = 0.50$).

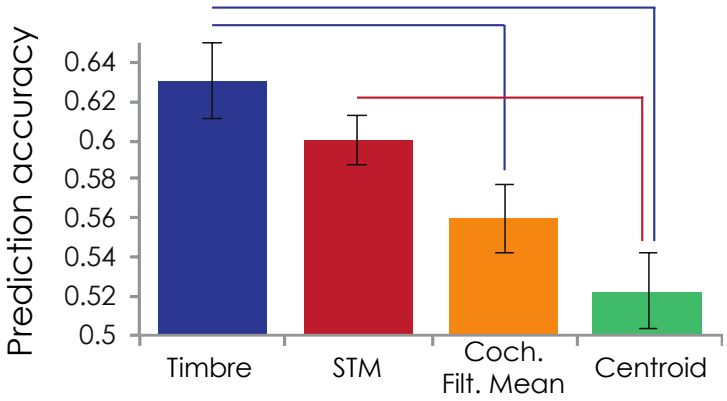

Fig. 4. Mean prediction accuracy across the encoding models. Average model performance across ten subjects for the *timbre*, *STM*, *cochlear filter mean*, and *spectral centroid* models. Error bars represent +/- 1 standard error of the mean. Blue lines indicate which models performed significantly worse than the *timbre* model, and red lines indicate which models performed significantly worse than the *STM* model. No other significant differences were found across models.

Cortical variation in encoding model prediction accuracy

Figure 5A shows variations in model performance throughout the cortex. These maps indicate how well the measured responses from individual voxels to sounds were represented by the different models. Given that the *spectral centroid* model did not perform significantly above chance, we excluded it from further analysis. Although all models displayed the highest prediction accuracies around the superior temporal plane (STP) and STG, significantly above-chance accuracy was also observed in frontal regions. Note that differences in the performance of a single model across the brain could result from location-specific differences in noise level (for a review, see Schoppe et al., 2016), and therefore the differences within each panel of Fig. 5A should be interpreted with caution.

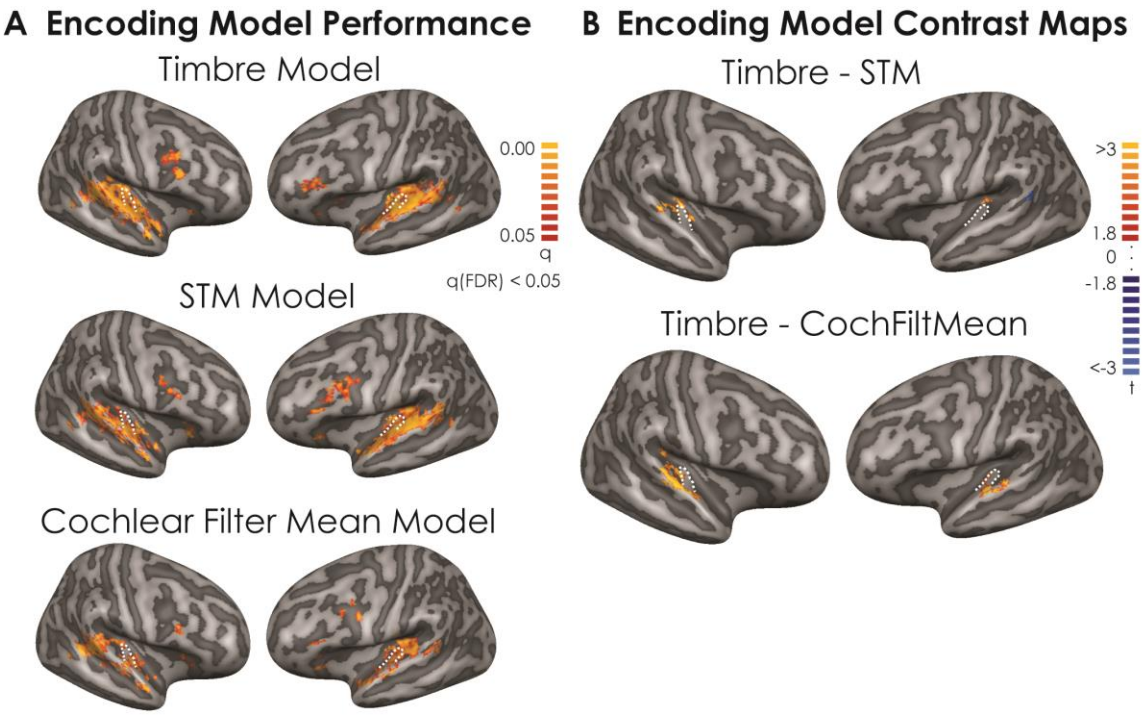

Fig. 5. Group-level model performance. (A) The maps show the cortical regions with a significant (q[FDR] < 0.05) correlation between measured and predicted responses to sounds. From top to bottom, performance of the *timbre*, *STM*, *cochlear filter mean*, and *spectral centroid* model are shown. (B) Group-level differences between models. Positive values (warmer colors)

indicate voxels for which the *timbre* model performed significantly better, and negative values (cooler colors) indicate voxels for which the *STM* or *cochlear filter mean* (in the top and bottom panel, respectively) performed significantly better. White dotted lines indicate HG.

Contrast maps

In order to compare the models in terms of the significant regional differences in their performance, we contrasted each model with the *timbre* model (see Fig. 5B). Warmer colors indicate regions in which the *timbre* model has significantly better performance compared to the other models, and cooler colors indicate regions where the other models have significantly better performance than the *timbre* model. Overall, the maps show more warm colors than cool colors, reflecting the overall higher performance of the *timbre* model (i.e., higher sound identification score). The timbre model outperformed all other models in representing processing in right hemispheric regions posterior to HG (covering HS and anterior PT). A comparison of the two best-performing models, the *STM* and *timbre* models, revealed considerable overlap, but also some regional differences. Specifically, the *timbre* model's representation is superior to that of the *STM* model in regions at the medial end of HG bilaterally, and at the posterior and anterior adjacency of HG (i.e., HS and first transverse temporal sulcus (FTS), respectively) in the right hemisphere. These areas may reflect either primary or belt regions of auditory cortex (Moerel et al., 2014). The *timbre* model also outperforms the *STM* model in a small region on the STG of the right hemisphere, likely reflecting a belt region of auditory cortex. Conversely, the *STM* model outperforms the *timbre* model in a small region at the posterior end of the STG in the left hemisphere, potentially corresponding to the parabelt region of the auditory cortex (Moerel et al., 2014). Furthermore, compared to the *cochlear filter mean* model, the timbre model performs better in regions along the HG and STG bilaterally, and HS in the right hemisphere. The superior performance seen in lateral HG may correspond to a difference in core auditory regions, while the

differences observed in HS of the right hemisphere and the STG bilaterally may correspond to belt and parabelt regions, respectively (Moerel et al., 2014).

Analysis of the timbre dimensions

According to Elliott et al. (2013), around 90% of the perceptual variance in the acoustic stimuli is explained by these five dimensions, and the dimensions are ordered by the amount of variance explained, with D1-D3 explaining the most variance. In order to explore a possible correspondence between this perceptual variance and the neural variance, we tested each dimension of the timbre model separately. The mean prediction accuracy results were: D1: 56%, D2: 60%, D3: 58%, D4: 49%, D5: 52%. In a one-tailed t-test, the first three dimensions were significantly above chance ($t_9 = 4.00$, $P = 0.003$, $d = 1.26$; $t_9 = 5.28$, $P = 0.001$, $d = 1.67$; and $t_9 = 4.21$, $P = 0.002$, $d = 1.33$, respectively), suggesting the first three dimensions best predict responses to novel test sounds.

We explored the overlap in the sound representations captured by the *timbre* and *STM* models by using canonical correlation analysis (CCA) (Hotelling, 1936) and linear regression. CCA was used to identify two new sets of features that share the largest amount of information (i.e., the maximum correlation), and linear regression was used to compute the transformation that best describes the features of one model in terms of the features of the other. We describe each approach and report the results below.

CCA and linear regression procedures. CCA was performed in a four-fold cross-validation loop (where a random 75% of the sounds and their representation in the models' space were used for training, and the remaining 25% for testing on an independent data set), repeated 1000 times, to evaluate the canonical correlation using an independent data set. Overfitting of the *STM* model (36 features, 42 sounds) was prevented by using the first 14 principal components (PCs) of the model. These 14 PCs explained 99.8% of the variance in the training data and 98.2%

of the variance in the test data. The PC decomposition was performed on the training data and the test data was projected on the PC space. Since the *timbre* model contains only five features, dimensionality reduction on the *timbre* model was not needed. For each cross-validation, the CCA was run on the training data. Next, we computed the proportion of variance in the original *STM* model that could be explained by the canonical covariates of the *timbre* model, and likewise, the proportion of variance in the original *timbre* model that could be explained by the canonical covariates of the *STM* model. For each cross-validation, this was computed by projecting the test data of each model to the canonical covariate space obtained on the training data. On the test data sets, a linear regression between the full set of canonical covariates of one model to the set of original features of the other model was performed. Performing this analysis on test data independent from the (training) data used to compute the canonical covariates avoids overfitting.

The linear regression between the two models was also performed in a four-fold cross validation loop, repeated 1000 times, and the average values of the explained variance on the test data were reported. Each feature of one model was described as a linear function of all of the features in the other model. The total variance in one model that could be explained by the other was computed as the sum of the explained variances of each feature. When the *STM* model was used as the independent dataset, overfitting was prevented by means of principal components regularization. For consistency with CCA analysis, the linear regression was performed on the subspace spanned by the first 14 PCs of the *STM* model. When the *timbre* model was used as the independent variable, no regularization was required and ordinary least squares (OLS) regression was used.

CCA and linear regression results. For the CCA we found, on average, across cross-validations and 1000 repetitions, that the canonical covariates of the *timbre* model explained 34.4% of the variance of the original *STM* model, while the canonical covariates of the *STM*

model explained 41.6% of the variance of the *timbre* model. For the linear regression we found, on average, across cross-validations and 1000 repetitions, that a linear combination of the features of the *timbre* model explained 37.1% of the variance of the original *STM* model, while a linear combination of the features of the *STM* model explained 38.2% of the variance of the *timbre* model. The CCA results are in overall accordance with the linear regression results and suggest that while there is a clear overlap between the two models, offering the possibility of (partially) understanding the *timbre* model in terms of basic acoustic features, there remains a substantial amount of variance in the *timbre* model that cannot be explained by the *STM* model and vice versa.

Linking the timbre dimensions to acoustic features. To further explore the acoustic basis of each of the timbre dimensions, we display 3D correlation heat maps between the *STM* model features and each of the five *timbre* model dimensions (Fig. 6A). Additionally, to explore the neurobiological correlate of each of the five timbre dimensions and quantify the consistency across subjects, we conduct an exploratory analysis of the trained *timbre* model, displaying those voxels for which the sign of the voxel's weight in the trained timbre model is consistent across the majority of subjects (Fig. 6B).

The first timbre dimension, D1, is semantically associated with "hard, sharp, high-frequency energy balance" (Elliott et al., 2013), and correlates most strongly with a combination of high frequencies and slow temporal modulations (Fig. 6A). The positive weights on medial HG suggest that these regions respond more strongly to sounds that score high on D1. In contrast, negative weights are distributed along STG, indicating that these cortical locations respond more strongly to sounds that score low on D1 (Fig. 6B). This may reflect the tonotopic organization of the auditory cortex, with a high frequency preference at the medial border of HG, and a low frequency preference along the STG (Langers, Backes, & van Dijk, 2007; Moerel et al., 2012), suggesting this dimension, at least in part, reflects the frequency content of sounds.

D2 is semantically associated with “varying level, dynamic, vibrato, and ringing release”, and is positively correlated with fast temporal modulations, especially in combination with intermediate frequency features (Fig. 6A). These characteristics seem appropriate for the semantic descriptor “ringing release”. In contrast, negative correlations with low temporal modulations are seen at low- to mid-range frequencies and low spectral modulations. D2 weights were consistently positive across a large number of voxels on the supratemporal plane (STP), indicating that these regions respond more strongly to faster temporal modulations. This is in accordance with previous studies that showed a strong bilateral activation of the auditory cortices for sounds with fast temporal modulations (e.g., Zatorre and Belin, 2001; Joanisse and DeSouza, 2014).

D3, which is semantically associated with “noisy, small instrument, and unpleasant”, correlates positively with high frequency features of the *STM* model especially when combined with fast temporal modulations (possibly corresponding to greater spectral irregularity and roughness), and negatively to low frequency features (Fig. 6A). In contrast, the strongest negative correlations were found for slow temporal modulations at low- to mid-range frequencies. This suggests that high frequency sounds with fast modulations may be perceived as more noisy and unpleasant. Like D2 weights, D3 weights were consistently positive across the STP. This is in accordance with previous work, which found unpleasant sounds to be associated with increased bilateral activation throughout auditory cortex (Plichta et al., 2011).

D4 corresponds to “compact, steady pitch, pure”, and correlates positively with the lowest *STM* frequency features, and negatively with mid-range frequency features. Positive D4 weights appear on primary auditory cortical regions centered on HG, suggesting that these regions respond more strongly to more compact and pure sounds. In contrast, negative weights are situated along the STG, which may respond more strongly to broader, more complex sounds. This organization is consistent with hierarchical auditory processing, with simple tones being

processed in early auditory cortical areas and more complex sounds undergoing greater processing in secondary or tertiary auditory regions (Patterson, Uppenkamp, Johnsrude, & Griffiths, 2002b; Tian & Rauschecker, 2004).

D5 is difficult to interpret, as the previous work by Elliott et al. (2013) did not reveal a semantic association with this dimension. D5 has strong positive correlations with features that combine mid-range frequencies, slow temporal modulations (~3 Hz), and middle spectral modulations (~1 cycle/octave; Fig. 6A). Furthermore, the anterolateral portion of HG displays positive D5 weights, bilaterally. This may point toward a lower-level dimension in the processing hierarchy, potentially associated with pitch strength (Penagos et al., 2004).

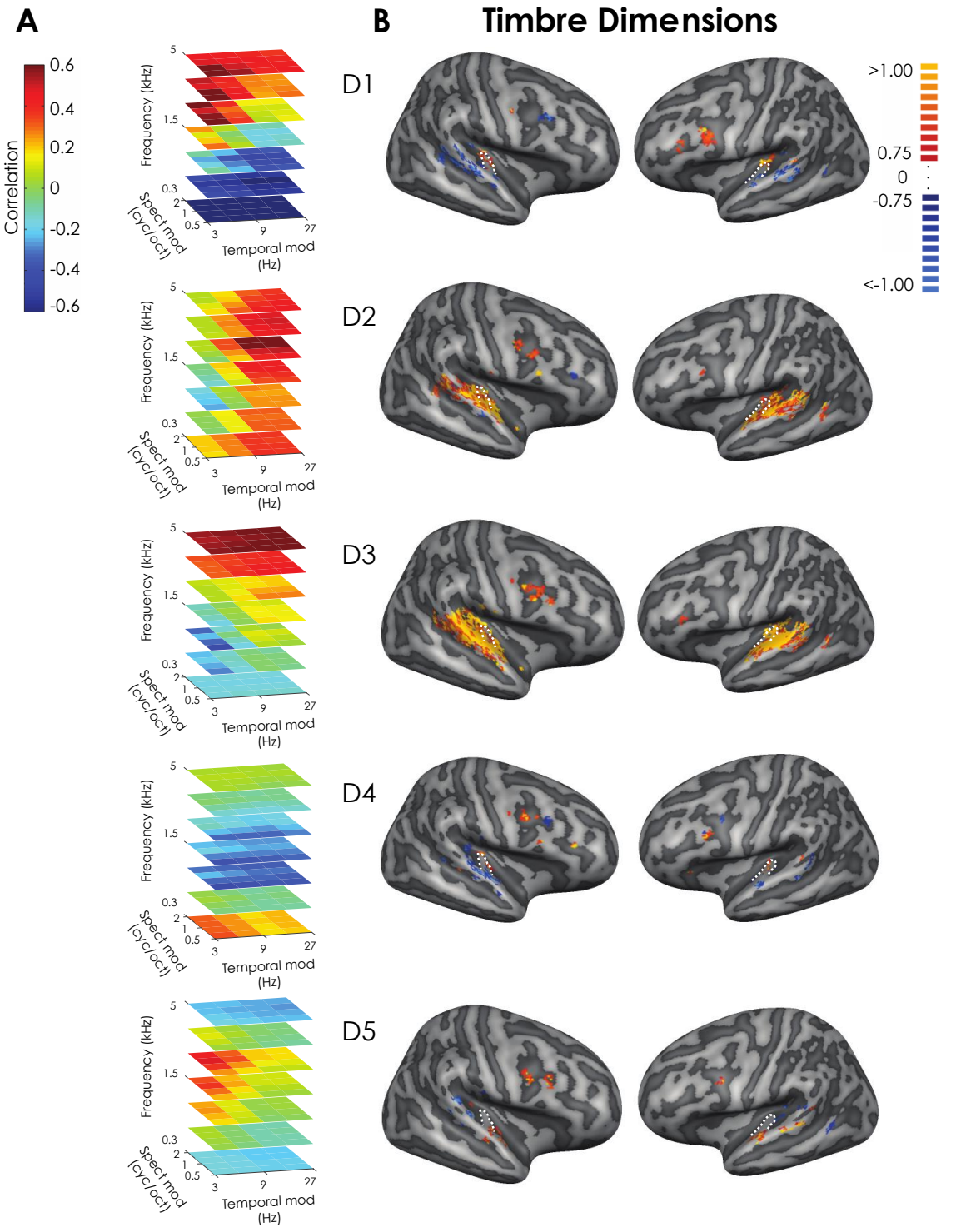

Fig. 6. Exploratory analyses of the timbre dimensions. (A) Slice plots showing marginal correlations between each of the five *timbre* dimensions and the features in the *STM* Model at

several different frequencies. (B) Group-level maps of the five dimensions of the *timbre* model. For each timbre dimension, warm and cool colors reflect across-subject consistently positive and consistently negative scores, respectively. A positive or negative weight reflects that as sounds scored higher or lower on that dimension, respectively, the BOLD response in the voxel increased. White dotted lines indicate HG.

Discussion

In this study, we used fMRI encoding to compare a timbre model derived from listeners' ratings of the sounds with acoustic models based on physical sound characteristics. We observed that the *timbre* model was able to predict a significant portion of the variance in the sound-evoked cortical activation. Furthermore, it performed significantly better than the other models tested, with the exception of a complex joint *spectrotemporal modulation* model. This finding, along with the observation that the two models shared a large part of the variation in the stimulus domain and the inferior performance of the uniquely spectral encoding models, supports the idea that joint spectrotemporal features are critical for capturing timbre perception (Patil, Pressnitzer, Shamma, & Elhilali, 2012).

However, we observed that the *timbre* model outperformed the joint *STM* model in a subset of the auditory cortical locations. Specifically, the *timbre* model performed significantly better in regions medial and posterior to HG, particularly in the right hemisphere. This suggests that while the *timbre* model only contains five features, it may be capturing some semantic or perceptual tuning properties of the auditory cortex that extend beyond those captured by the *spectrotemporal* model. Specifically, the differences observed in terms of the amount of shared variance between the *timbre* and *STM* models identified via CCA and linear regression may be a result of the *timbre* model capturing some nonlinear combination of physical features not

represented in the *STM* model. This may be a distinguishing component of higher-level semantic processing (Kay & Yeatman, 2017). In light of this, it would be tempting to combine these two models in hopes of achieving better model performance. However, concatenation of these models is suboptimal as the *timbre* model is made of features that are orthogonal to each other and the *STM* model has many collinear features. As a result, the regularization to be applied to each model separately differs substantially and concatenation would result in over-penalizing the *timbre* model. Therefore, an area that warrants future research is the development of methods to optimally combine models that explain different parts of the variance (see e.g., de Heer et al., 2017).

In addition to auditory regions, responses to sounds in frontal regions, such as the inferior frontal gyrus (IFG), were consistently predicted above chance across models. This may indicate that timbre features are also represented in frontal regions, but could also reflect higher-level auditory processing that is correlated with the features of the employed encoding models. One possible explanation is that model accuracy in frontal regions could be driven by sound recognition, since our stimuli were common musical instruments. Maeder et al. (2001) found certain regions to be more active for sound recognition compared to sound localization, including the left posterior IFG. Further, Broca's area may be included in the well-predicted cortical regions. While Broca's region is typically thought to be a higher-level language processing area, it has been suggested to also play a role in music processing (for a review, see Fadiga et al., 2009).

Timbre is a notoriously elusive acoustic feature to define and to investigate experimentally. In this study, the use of fMRI encoding (Naselaris & Kay, 2015) allowed us to explicitly test the representation of timbre-varying sounds throughout cortical neuronal populations. Employing natural sounds, this approach furthermore ensured that timbre varied across sounds in an ecologically valid manner. While many earlier studies have used encoding

models that represented the physical characteristics of natural images (Kay et al., 2008; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009) or sounds (Santoro et al., 2014), our work along with more recent studies (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Kay & Yeatman, 2017) demonstrates the utility of incorporating higher-level perceptual features into the encoding models. This represents a next evolution in fMRI encoding, where the method can be used to tackle those aspects of perception and cognition that are extremely challenging to capture using classical approaches.

The *timbre* model provides an efficient representation of processing in human auditory cortex via a compact model whose features are based on subjective ratings of timbre. Our results suggest that the distributed neural representation of timbre in the cortex may align with perceptual categorizations of timbre. Consequently, it may be possible to assign semantic labels to the multidimensional tuning of neuronal populations. Since the employed *timbre* model was customized for this particular set of orchestral instruments, studies that test a broader range of stimuli (i.e., more musical instruments, speech, and other natural sounds) are recommended in order to determine the extent of this model's generalizability.

Acknowledgements

This work was supported by the National Institute of Deafness and other Communication Disorders at the National Institutes of Health (grant number R01 DC005216), the Brain Imaging Initiative of the College Liberal Arts, University of Minnesota, the Erasmus Mundus Student Exchange Network in Auditory Cognitive Neuroscience (ACN), the Netherlands Organization for Scientific Research (NWO; VENI grant 451-15-012, and VICI grant 453-12-002), and the Dutch Province of Limburg. Juraj Mesik, Philip Burton, Cheryl Olman, Jordan Beim, and Taffeta Elliott provided helpful advice and assistance. The authors declare no competing financial interests.

Chapter 6

General Discussion

This dissertation research was motivated by a drive to better understand how we perceive and process two key components of sound – pitch and timbre. We approached this research using a combination of behavioral studies, neuroimaging, and computational modeling. Through these methods we have made several discoveries.

Our percepts of pitch and timbre can interact

From our behavioral studies we confirmed that pitch and timbre can interact with each other, such that when subjects are instructed to attend to one dimension and ignore the other, thresholds increase when both dimensions are varied, relative to when just one dimension is being manipulated. The fact that the interference is directional (with increases in pitch confused for increases in brightness) suggests that the dimensions of pitch and timbre are not completely orthogonal.

Moreover, we learned that when variations in these dimensions are equated for perceptual salience, their interactions are relatively symmetric. Somewhat surprisingly, we found this symmetry in both musicians and non-musicians, despite musicians having better F0 (pitch) difference limens than non-musicians, suggesting that, even with their acoustic expertise, musicians are susceptible to the interference effects that can occur when both the pitch and timbre of sounds are varying.

Cortical representations of pitch and timbre are similar

As an extension of our behavioral findings, we aimed to explore the neural substrates of the perceptual interaction between pitch and timbre. The debate about whether a “pitch center” exists in the auditory cortex is a hot topic that has yet to be fully resolved, and very little research has been done on the cortical processing of timbre. Thus, it was unknown whether distinct

neuronal populations processed the two dimensions or whether similar or overlapping populations in the human brain might be responsive to both. The lack of spatial separation in the cortical representations of variation in these two dimensions identified by our studies suggest there is cortical real estate in the auditory cortex, bilaterally, that is sensitive to both dimensions. However, thanks to the increased sensitivity of MVPA, we discovered that the patterns of responses within these shared regions appear to be unique, making it possible for classifiers to tease them apart. While the shift from passive listening to active tasks did not affect our univariate results, as pitch and timbre remained difficult to parse, spatially, the MVPA classifiers were again able to distinguish between the patterns of activation for these dimensions. What was particularly compelling was that attention to a given feature within a sound, in the absence of any physical differences across conditions, was sufficient for the classifier to successfully distinguish pitch from timbre discrimination.

Neuronal populations may be sensitive to the various dimensions of timbre

Up to this point, the thesis concentrated on a single aspect of timbre – brightness. However, the concept of timbre is known as a highly complex and ill-defined "multi-dimensional dimension" (Licklider, 1951). Exploring more natural manipulations of timbre (e.g., in the form of orchestral instruments), while more ecologically valid, also becomes more challenging in terms of teasing apart the multitude of dimensions that may be varying from one sound to the next. This was the final piece of the dissertation puzzle. We attempted to model how timbre is encoded in the auditory cortex. By using multi-dimensional scaling values of five timbre dimensions identified by Elliott, Hamilton, and Theunissen (2013) to develop a perceptually derived encoding model, we were able to successfully predict the cortical responses to these orchestral sounds.

Moreover, this model performed better than other models based on the spectral characteristics of the sounds, and performed as well as a complex joint spectrotemporal modulation model. These results suggest there may, in fact, be neuronal populations in the auditory cortex that are sensitive to the various perceptual dimensions of timbre.

Future directions

As a few questions get answered, many more follow. For example, we are interested in exploring the topographic mapping of these dimensions using ultra-high resolution (7T) fMRI. Traditionally, the stimuli used for tonotopic mapping of the auditory cortex are narrowband tones or noises, which simultaneously increase in pitch and brightness, with no means for dissociating these two dimensions. Our aim is to determine whether these cortical maps indeed reflect the center frequency of the stimulus (as in traditional tonotopic mapping), other whether they reflect pitch, based on the stimulus F0.

Additionally, while pitch and timbre were shown to have similar cortical representations when variations in F0 were being compared to variations in spectral centroid, we once again must acknowledge that this is just one manipulation of pitch being compared to one manipulation of timbre. Further, we do not know precisely how many dimensions of timbre there are. While we were able to use a five-feature encoding model to predict the cortical representations of 42 orchestral sounds, we have not explored how well this five-feature model generalizes to other natural sounds. Can these same five dimensions be used to describe speech, environmental sounds, or even *other* musical instruments that were not tested? Perhaps five dimensions would not be sufficient for a broader array of natural sounds, or perhaps the five dimensions deemed necessary and sufficient for describing these orchestral sounds are not the same dimensions that would be necessary and sufficient to describe the timbre space of a different subset of sounds.

At a more basic level, we are curious whether all audible frequencies get similar amounts of cortical real estate, or whether we allocate a larger portion of cortex to certain frequencies (e.g., lower, resolved frequencies that are commonly represented in speech compared to high, unresolved frequencies) creating a sort of "acoustic fovea". Further, it remains unclear the best way to distinguish primary auditory regions from secondary and tertiary regions in humans. There does not appear to be a one-to-one mapping between the macroanatomical structure of the auditory cortex and the functionally-defined regions. Perhaps it would be more appropriate to define these regions based on their degree of myelination, as there is evidence to suggest that primary regions have a greater myelination density than non-primary regions.

Lastly, with more multimodal fMRI research, perhaps we can also discover more commonalities that exist between the auditory and visual cortices, and more evidence for a general system that modulates both auditory and visual processing. Though our understanding of the auditory cortex remains murky relative to our understanding of the visual cortex, we are actively working to bring more clarity to this region. It is feasible that, in the not-too-distant future, the amount that is known about the auditory cortex will be comparable to that which is known about the visual cortex. While there is much left to be learned about the auditory system, especially in the human cortex, we are enticingly close to making many significant breakthroughs in this area.

References

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *J. Acoust. Soc. Am.*, *135*(3), 1371–1379.

ANSI. (1994). S1.1-1994, American National Standard Acoustical Terminology (R2004) (American National Standards Institute, New York, 1960).

ANSI. (2013). S1.1-2013, American National Standard Acoustical Terminology (American National Standards Institute, New York, 1960).

Attneave, F., & Olson, R. K. (1971). Pitch as a Medium: A New Approach to Psychophysical Scaling. *Am. J. Psychol.*, *84*(2), 147–166.

Barker, D., Plack, C. J., & Hall, D. a. (2012). Reexamining the evidence for a pitch-sensitive region: a human fMRI study using iterated ripple noise. *Cereb. Cortex*, *22*(4), 745–53. doi:10.1093/cercor/bhr065

Beal, A. L. (1985). The skill of recognizing musical structures. *Mem. Cognit.*, *13*(5), 405–412. doi:10.3758/BF03198453

Bendor, D. (2012). Does a pitch center exist in auditory cortex? *J. Neurophysiol.*, *107*(3), 743–6. doi:10.1152/jn.00804.2011

Bendor, D., & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature*, *436*(7054), 1161–5. doi:10.1038/nature03867

Bilsen, F. A. (1966). Repetition Pitch: Monaural Interaction of a Sound with the Repetition of the Same, but Phase Shifted, Sound. *Acustica*, *17*(5), 295–300.

Bizley, J. K., Walker, K. M. M., Silverman, B. W., King, A. J., & Schnupp, J. W. H. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *J. Neurosci.*, *29*(7), 2064–75. doi:10.1523/JNEUROSCI.4755-08.2009

Blackman, G. A., Hall, D. A., & Kingdom, U. (2014). During Auditory Functional Magnetic Resonance, *54*(April 2011), 693–704. doi:10.1044/1092-4388(2010/10-0143)Ziarati

Borchert, E. M. O., Micheyl, C., & Oxenham, A. J. (2011). Perceptual grouping affects pitch judgments across time and frequency. *J. Exp. Psychol. Hum. Percept. Perform.*, *37*(1), 257–69.

Burns, E. M. (1999). Intervals, Scales, and Tuning. In *Psychol. Music* (2nd ed., pp. 215–264). San Diego: Academic Press.

Burns, E. M., & Viemeister, N. F. (1976). Nonspectral pitch*. *J. Acoust. Soc. Am.*, *60*(4), 863–869.

Burns, E. M., & Viemeister, N. F. (1981). Played-again SAM : Further observations on the pitch of amplitude-modulated noise. *J. Acoust. Soc. Am.*, *70*(6), 1655–1660.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *J. Acoust. Soc. Am.*, *118*(1), 471. doi:10.1121/1.1929229

Cariani, P. a, & Delgutte, B. (1996). Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *J. Neurophysiol.*, *76*(3), 1698–716.

Casey, M., Thompson, J., Kang, O., Raizada, R., & Wheatley, T. (2012). Population Codes Representing Musical Timbre for High-Level fMRI Categorization of Music Genres. In *Mach. Learn. Interpret. Neuroimaging* (pp. 34–41). Berlin Heidelberg: Springer.

Chen, Z., Hu, G., Glasberg, B. R., & Moore, B. C. J. (2011). A new method of calculating auditory excitation patterns and loudness for steady sounds. *Hear. Res.*, *282*(1–2), 204–215. doi:10.1016/j.heares.2011.08.001

Chi, T., Ru, P., & Shamma, S. a. (2005). Multiresolution spectrotemporal analysis of complex

sounds. *J. Acoust. Soc. Am.*, *118*(2), 887. doi:10.1121/1.1945807

Clarke, S., & Morosan, P. (2012). Architecture, Connectivity, and Transmitter Receptors of Human Auditory Cortex. In D. Poeppel, T. Overath, A. N. Popper, & R. R. Fay (Eds.), *Hum. Audit. Cortex* (pp. 11–38). New York: Springer Science+Business Media, LLC.

Cox, R. W. (1996). AFNI : Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Comput. Biomed. Res.*, *29*, 162–173.

Crummer, G. C., Walton, J. P., Wayman, J. W., Hantz, E. C., & Frisina, R. D. (1994). Neural processing of musical timbre by musicians, nonmusicians, and musicians possessing absolute pitch. *J. Acoust. Soc. Am.*, *95*(5 Pt 1), 2720–7.

De Angelis, V., De Martino, F., Moerel, M., Santoro, R., Hausfeld, L., & Formisano, E. (2017). Cortical processing of pitch: Model-based encoding and decoding of auditory fMRI responses to real-life sounds. *Neuroimage*, (November), 1–10. doi:10.1016/j.neuroimage.2017.11.020

de Cheveigné, A. (2010). Pitch perception. In C. J. Plack (Ed.), *oxford Handb. Audit. Sci. Hear.* (pp. 71–104). Oxford: Oxford University Press.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *J. Neurosci.*, *37*(27), 6539–6557. doi:10.1523/JNEUROSCI.3267-16.2017

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage*, *43*(1), 44–58. doi:10.1016/j.neuroimage.2008.06.037

Degerman, A., Rinne, T., Salmi, J., Salonen, O., & Alho, K. (2006). Selective attention to sound location or pitch studied with fMRI. *Brain Res.*, *1077*(1), 123–34. doi:10.1016/j.brainres.2006.01.025

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci.*, 201602413. doi:10.1073/pnas.1602413113

Elliott, T. M., Hamilton, L. S., & Theunissen, F. E. (2013). Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *J. Acoust. Soc. Am.*, *133*(1), 389–404. doi:10.1121/1.4770244

Engel, S. a, Glover, G. H., & Wandell, B. a. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb. Cortex*, *7*(2), 181–92.

Fadiga, L., Craighero, L., & D'Ausilio, A. (2009). Broca's area in language, action, and music. *Ann. N. Y. Acad. Sci.*, *1169*, 448–458. doi:10.1111/j.1749-6632.2009.04582.x

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics: Facts and Models*. Verlag Berlin Heidelberg New York: Springer.

Faulkner, A. (1985). Pitch discrimination of harmonic complex signals: residue pitch or multiple component discriminations? *J. Acoust. Soc. Am.*, *78*(6), 1993–2004.

Fletcher, H. (1934). Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *J. Acoust. Soc. Am.*, *6*(2), 59–69.

Formisano, E., Kim, D. S., Di Salle, F., van de Moortele, P. F., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, *40*(4), 859–69.

Fritz, L., Mulders, J., Breman, H., Peters, J., Bastiani, M., Roebroeck, A., … Goebel, R. (2014). Comparison of EPI distortion correction methods at 3T and 7T! In *OHBM Annu. Meet.*

Gfeller, K., Witt, S., Adamek, M., Mehr, M., Rogers, J., Stordahl, J., & Ringgenberg, S. (2002). Effects of training on timbre recognition and appraisal by postlingually deafened cochlear implant recipients. *J. Am. Acad. Audiol.*, *13*(3), 132–45.

Goebel, R., Esposito, F., & Formisano, E. (2006). Analysis of Functional Image Analysis Contest (FIAC) data with BrainVoyager QX: From single-subject to cortically aligned group General Linear Model analysis and self-organizing group Independent Component Analysis. *Hum. Brain Mapp.*, *27*(5), 392–401. doi:10.1002/hbm.20249

Graves, J., Micheyl, C., & Oxenham, A. J. (2014). Preferences for melodic contours transcend pitch. *J. Acoust. Soc. Am.*, *133*(6), 3366. doi:10.1121/1.4805757

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics (Vol. 1)*. (Wiley, Ed.). New York.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.*, *61*(5), 1270–7.

Griffiths, T. D., Kumar, S., Sedley, W., Nourski, K. V, Kawasaki, H., Oya, H., … Howard, M. a. (2010). Direct recordings of pitch responses from human auditory cortex. *Curr. Biol.*, *20*(12), 1128–32. doi:10.1016/j.cub.2010.04.044

Griffiths, T. D., Uppenkamp, S., Johnsrude, I., Josephs, O., & Patterson, R. D. (2001). Encoding of the temporal regularity of sound in the human brainstem. *Nat. Neurosci.*, *4*(6), 633–637. doi:10.1038/88459

Gutschalk, A., Patterson, R. D., Rupp, A., Uppenkamp, S., & Scherg, M. (2002). Sustained Magnetic Fields Reveal Separate Sites for Sound Level and Temporal Regularity in Human Auditory Cortex. *Neuroimage*, *15*(1), 207–216. doi:10.1006/nimg.2001.0949

Gutschalk, A., & Uppenkamp, S. (2011). Sustained responses for pitch and vowels map to similar sites in human auditory cortex. *Neuroimage*, *56*(3), 1578–1587. doi:10.1016/j.neuroimage.2011.02.026

Hacker, M. J., Ratcliff, R., Tables, P. B., & Ed, J. A. S. (1979). A revised table of d' for M-alternative forced choice. *Percept. Psychophys.*, *26*(2), 168–170.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., … Bowtell, R. W. (1999). "Sparse" temporal sampling in auditory fMRI. *Hum. Brain Mapp.*, *7*(3), 213–23.

Hall, D. A., & Plack, C. J. (2009). Pitch processing sites in the human auditory brain. *Cereb. Cortex*, *19*(3), 576–85. doi:10.1093/cercor/bhn108

Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, *42*(9), 1281–1292. doi:10.1016/j.neuropsychologia.2003.12.017

Hotelling, H. (1936). Relations Between Two Sets of Variables. *Biometrica*, *28*(3/4), 321–377.

Houtsma, A. J. M. (1997). Pitch and Timbre: Definition, Meaning and Use. *J. New*, *26*, 104–115.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. doi:10.1038/nature17637

Hyde, K. L., Lerch, J., Norton, A., Forgeard, M., Winner, E., Evans, A. C., & Schlaug, G. (2009). Musical training shapes structural brain development. *J. Neurosci.*, *29*(10), 3019–25. doi:10.1523/JNEUROSCI.5118-08.2009

Idiyatullin, D., Corum, C., Park, J.-Y., & Garwood, M. (2006). Fast and quiet MRI using a swept radiofrequency. *J. Magn. Reson.*, *181*(2), 342–9. doi:10.1016/j.jmr.2006.05.014

Jäncke, L., Mirzazade, S., & Shah, N. J. (1999). Attention modulates activity in the primary and the secondary auditory cortex: a functional magnetic resonance imaging study in human subjects. *Neurosci. Lett.*, *266*(2), 125–8.

Joanisse, M. F., & DeSouza, D. D. (2014). Sensitivity of human auditory cortex to rapid frequency modulation revealed by multivariate representational similarity analysis. *Front. Neurosci.*, *8*(September), 306. doi:10.3389/fnins.2014.00306

Johnsrude, I. S., Penhune, V. B., & Zatorre, R. J. (2000). Functional specificity in the right human

auditory cortex for perceiving pitch direction. *Brain, 123 (Pt 1*, 155–63.
Kanwisher, N., McDermott, J. H., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, *17*(11), 4302–11.
Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. doi:10.1038/nature06713.Identifying
Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013). GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.*, (7 DEC). doi:10.3389/fnins.2013.00247
Kay, K. N., & Yeatman, J. D. (2017). Bottom-up and top-down computations in word- and face-selective cortex. *Elife*, *6*(3), 191–195. doi:10.7554/eLife.22341
Kraus, N., Skoe, E., Parbery-Clark, A., & Ashley, R. (2009). Experience-induced malleability in neural encoding of pitch, timbre, and timing: Implications for language and music. *Ann. N. Y. Acad. Sci.*, *1169*, 543–557. doi:10.1111/j.1749-6632.2009.04549.x
Krumbholz, K., Patterson, R. D., & Pressnitzer, D. (2000). The lower limit of pitch as determined by rate discrimination. *J. Acoust. Soc. Am.*, *108*(3), 1170–1180.
Krumhansl, C. L., & Iverson, P. (1992). Perceptual interactions between musical pitch and timbre. *J. Exp. Psychol. Hum. Percept. Perform.*, *18*(3), 739–51.
Langers, D. R. M., Backes, W. H., & van Dijk, P. (2007). Representation of lateralization and tonotopy in primary versus secondary human auditory cortex. *Neuroimage*, *34*(1), 264–73. doi:10.1016/j.neuroimage.2006.09.002
Langers, D. R. M., Krumbholz, K., Bowtell, R. W., & Hall, D. A. (2014). Neuroimaging paradigms for tonotopic mapping (I): The influence of sound stimulus type. *Neuroimage*, *100*, 650–662. doi:10.1016/j.neuroimage.2014.07.044
Langers, D. R. M., & van Dijk, P. (2012). Mapping the tonotopic organization in human auditory cortex with minimally salient acoustic stimulation. *Cereb. Cortex*, *22*(9), 2024–38. doi:10.1093/cercor/bhr282
Leal, M. C., Shin, Y. J., Laborde, M., Calmels, M., Verges, S., Lugardon, S., … Fraysse, B. (2003). Music Perception in Adult Cochlear Implant Recipients. *Acta Otolaryngol.*, *123*(7), 826–835. doi:10.1080/00016480310000386
Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, *49*, 467–477.
Licklider, J. C. R. (1951). Basic correlates of the auditory stimulus. In *Handb. Exp. Psychol.* (pp. 985–1039). New York: Wiley.
Maeder, P., Meuli, R., Adriani, M., Bellmann, A., Fornari, E., Thiran, J.-P., … Clarke, S. (2001). Distinct Pathways Involved in Sound Recognition and Localization: A Human fMRI Study. *Neuroimage*, *14*(4), 802–16. doi:10.1006/nimg.2001.0888
Marozeau, J., & de Cheveigné, A. (2007). The effect of fundamental frequency on the brightness dimension of timbre. *J. Acoust. Soc. Am.*, *121*(1), 383. doi:10.1121/1.2384910
Marozeau, J., de Cheveigné, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *J. Acoust. Soc. Am.*, *114*(5), 2946. doi:10.1121/1.1618239
Mazziotta, J. C., Toga, A. W., Evans, A. C., Fox, P., & Lancaster, J. (1995). A probabalisitc atlas of the human brain: theory and rationale for its development. *Neuroimage*, *2*(January), 89–101. doi:10.1006/nimg.1995.1012
McAdams, S., & Bregman, A. (1979). Hearing Musical Streams. *Comput. Music J.*, *3*(4), 26–43, 60.
McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychol. Res.*, *58*(3), 177–92.
McDermott, J. H., Keebler, M. V, Micheyl, C., & Oxenham, A. J. (2010). Musical intervals and

relative pitch: frequency resolution, not interval resolution, is special. *J. Acoust. Soc. Am.*, *128*(4), 1943–51. doi:10.1121/1.3478785
McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2008). Is relative pitch specific to pitch? *Psychol. Sci.*, *19*(12), 1263–71. doi:10.1111/j.1467-9280.2008.02235.x
McDermott, J. H., & Oxenham, A. J. (2008). Music perception, pitch, and the auditory system. *Curr. Opin. Neurobiol.*, *18*(4), 452–63. doi:10.1016/j.conb.2008.09.005
Melara, R. D., & Marks, L. E. (1990). Interaction among auditory dimensions: timbre, pitch, and loudness. *Percept. Psychophys.*, *48*(2), 169–78.
Melara, R. D., & Mounts, J. R. (1993). Selective attention to Stroop dimensions: effects of baseline discriminability, response mode, and practice. *Mem. Cognit.*, *21*(5), 627–45.
Melara, R. D., & Mounts, J. R. (1994). Contextual influences on interactive processing: effects of discriminability, quantity, and uncertainty. *Percept. Psychophys.*, *56*(1), 73–90.
Menon, V., Levitin, D. J., Smith, B. K., Lembke, A., Krasnow, B. D., Glazer, D., … McAdams, S. (2002). Neural Correlates of Timbre Change in Harmonic Sounds. *Neuroimage*, *17*(4), 1742–1754. doi:10.1006/nimg.2002.1295
Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hear. Res.*, *219*(1–2), 36–47. doi:10.1016/j.heares.2006.05.004
Miller, E. K., Cohen, J. D., Logothetis, N. K., Bobbert, M., Gómez Álvarez, C., van Weeren, R., … Weishaupt, M. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–78. doi:10.1038/nature06976
Moerel, M., De Martino, F., & Formisano, E. (2012). Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. *J. Neurosci.*, *32*(41), 14205–14216. doi:10.1523/JNEUROSCI.1388-12.2012
Moerel, M., De Martino, F., & Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Front. Neurosci.*, *8*(July), 1–14. doi:10.3389/fnins.2014.00225
Moore, B. C. J. (2014a). Development and Current Status of the "Cambridge" Loudness Models. *Trends Hear.*, *18*(0), 1–29. doi:10.1177/2331216514550620
Moore, B. C. J. (2014b). Development and Current Status of the "Cambridge" Loudness Models. *Trends Hear.*, *18*, 233121651455062. doi:10.1177/2331216514550620
Moore, B. C. J., & Glasberg, B. R. (1990). Frequency discrimination of complex tones with overlapping and non-overlapping harmonics. *J. Acoust. Soc. Am.*, *87*(5), 2163–77.
Moore, B. C. J., Glasberg, B. R., & Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc.*, *45*(4), 224–240.
Naselaris, T., & Kay, K. N. (2015). Resolving Ambiguities of MVPA Using Explicit Models of Representation. *Trends Cogn. Sci.*, *19*(10), 551–554. doi:10.1016/j.tics.2015.07.005
Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, *63*(6), 902–915. doi:10.1016/j.neuron.2009.09.006
Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *J. Neurosci.*, *33*(50), 19451–69. doi:10.1523/JNEUROSCI.2880-13.2013
Opolko, F., & Wapnick, J. (2006). The McGill University Master Samples Collection on DVD. Montreal: McGill University.
Oxenham, A. J., Fligor, B. J., Mason, C. R., & Kidd, G. (2003). Informational masking and musical training. *J. Acoust. Soc. Am.*, *114*(3), 1543–1549. doi:10.1121/1.1598197
Oxenham, A. J., Micheyl, C., Keebler, M. V, Loper, A., & Santurette, S. (2011). Pitch perception

beyond the traditional existence region of pitch. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(18), 7629–34. doi:10.1073/pnas.1015291108

Pantev, C., Oostenveld, R., Engelien, a, Ross, B., Roberts, L. E., & Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature*, *392*(6678), 811–4. doi:10.1038/33918

Parbery-Clark, A., Skoe, E., & Kraus, N. (2009). Musical experience limits the degradative effects of background noise on the neural processing of sound. *J. Neurosci.*, *29*(45), 14100–7. doi:10.1523/JNEUROSCI.3256-09.2009

Patil, K., Pressnitzer, D., Shamma, S., & Elhilali, M. (2012). Music in Our Ears: The Biological Bases of Musical Timbre Perception. (F. E. Theunissen, Ed.)*PLoS Comput. Biol.*, *8*(11), e1002759. doi:10.1371/journal.pcbi.1002759

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002a). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, *36*(4), 767–76.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002b). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, *36*(4), 767–776. doi:10.1016/S0896-6273(02)01060-7

Penagos, H., Melcher, J. R., & Oxenham, A. J. (2004). A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *J. Neurosci.*, *24*(30), 6810–5. doi:10.1523/JNEUROSCI.0383-04.2004

Pitt, M. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *J. Exp. Psychol. Hum. Percept. Perform.*, *20*(5), 976–86.

Plack, C. J., Barker, D., & Hall, D. a. (2014). Pitch coding and pitch processing in the human brain. *Hear. Res.*, *307*, 53–64. doi:10.1016/j.heares.2013.07.020

Plack, C. J., & Oxenham, A. J. (2005). Overview : The Present and Future of Pitch. In *Pitch* (pp. 1–6). New York: Springer.

Plack, C. J., Oxenham, A. J., & Fay, R. R. (2006). *Pitch: Neural Coding and Perception (Vol. 24)*. New York: Springer.

Plichta, M. M., Gerdes, A. B. M., Alpers, G. W., Harnisch, W., Brill, S., Wieser, M. J., & Fallgatter, A. J. (2011). Auditory cortex activation is modulated by emotion: A functional near-infrared spectroscopy (fNIRS) study. *Neuroimage*, *55*(3), 1200–1207. doi:10.1016/j.neuroimage.2011.01.011

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. and G. F. Plomp & G. F. Smoorenburg (Eds.), *Freq. Anal. Period. Detect. Hear.* (Vol. 397, pp. 397–411). Leiden.

Poeppel, D., Overath, T., Popper, A. N., & Fay, R. R. (2012). *The human auditory cortex*. (A. N. Popper & R. R. Fay, Eds.) (Volume 43.). New York: Springer Science+Business Media, LLC.

Pressnitzer, D., Patterson, R. D., & Krumbholz, K. (2001). The lower limit of melodic pitch. *J. Acoust. Soc. Am.*, *109*(5), 2074–2084. doi:10.1121/1.1359797

Rademacher, J., Morosan, P., Schormann, T., Schleicher, a, Werner, C., Freund, H. J., & Zilles, K. (2001). Probabilistic mapping and volume measurement of human primary auditory cortex. *Neuroimage*, *13*(4), 669–83. doi:10.1006/nimg.2000.0714

Ravicz, M. E., & Melcher, J. R. (2000). Acoustic noise during functional magnetic resonance imaging. *J. Acoust. Soc. Am.*, *108*(4), 1683–1696.

Recanzone, G. H., Guard, D. C., & Phan, M. L. (2000). Frequency and Intensity Response Properties of Single Neurons in the Auditory Cortex of the Behaving Macaque Monkey. *J. Neurophysiol.*, *83*(4), 2315–2331.

Reiterer, S., Erb, M., Grodd, W., & Wildgruber, D. (2008). Cerebral Processing of Timbre and Loudness: fMRI Evidence for a Contribution of Broca's Area to Basic Auditory Discrimination. *Brain Imaging Behav.*, *2*(1), 1–10. doi:10.1007/s11682-007-9010-3

Risset, J., & Wessel, D. L. (1999). Exploration of timbre by analysis and synthesis. In D. Deutsch (Ed.), *Psychol. Music* (2nd ed., pp. 113–169). Academic Press Series in Cognition and Perception.
Robinson, K. (1993). Brightness and octave position: are changes in spectral envelope and in tone height perceptually equivalent? *Contemp. Music Rev.*, *9*(1–2), 83–95.
Ruggles, D. R., Freyman, R. L., & Oxenham, A. J. (2014). Influence of musical training on understanding voiced and whispered speech in noise. *PLoS One*, *9*(1), e86980. doi:10.1371/journal.pone.0086980
Russo, F. A., & Thompson, W. F. (2005). An interval size illusion: the influence of timbre on the perceived size of melodic intervals. *Percept. Psychophys.*, *67*(4), 559–68.
Saenz, M., & Langers, D. R. M. (2014). Tonotopic mapping of human auditory cortex. *Hear. Res.*, *307*, 42–52. doi:10.1016/j.heares.2013.07.016
Samson, S., & Zatorre, R. J. (1994). Contribution of the right temporal lobe to musical timbre discrimination. *Neuropsychologia*, *32*(2), 231–40.
Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput. Biol.*, *10*(1), e1003412. doi:10.1371/journal.pcbi.1003412
Schindler, A., Herdener, M., & Bartels, A. (2013). Coding of melodic gestalt in human auditory cortex. *Cereb. Cortex*, *23*(12), 2987–93. doi:10.1093/cercor/bhs289
Schneider, P., Scherg, M., Dosch, H. G., Specht, H. J., Gutschalk, A., & Rupp, A. (2002). Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. *Nat. Neurosci.*, *5*(7), 688–94. doi:10.1038/nn871
Schönwiesner, M., Rübsamen, R., & von Cramon, D. Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *Eur. J. Neurosci.*, *22*(6), 1521–8. doi:10.1111/j.1460-9568.2005.04315.x
Schoppe, O., Harper, N. S., Willmore, B. D. B., King, A. J., & Schnupp, J. W. H. (2016). Measuring the Performance of Neural Models. *Front. Comput. Neurosci.*, *10*(February), 10. doi:10.3389/fncom.2016.00010
Semal, C., & Demany, L. (2006). Individual differences in the sensitivity to pitch direction. *J. Acoust. Soc. Am.*, *120*(6), 3907–3915. doi:10.1121/1.2357708
Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychol. Rev.*, *89*(July), 305–333.
Shofner, W. P., & Yost, W. A. (1997). Discrimination of rippled-spectrum noise from flat-spectrum noise by chinchillas: evidence for a spectral dominance region. *Hear. Res.*, *110*(1–2), 15–24. doi:10.1016/S0378-5955(97)00063-4
Silbert, N. H., Townsend, J. T., & Lentz, J. J. (2009). Independence and separability in the perception of complex nonspeech sounds. *Atten. Percept. Psychophys.*, *71*(8), 1900–1915. doi:10.3758/APP
Singh, P. G., & Hirsh, I. J. (1992). Influence of spectral locus and F0 changes on the pitch and timbre of complex tones. *J. Acoust. Soc. Am.*, *92*, 2650–2661.
Staeren, N., Renvall, H., De Martino, F., Goebel, R., & Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. *Curr. Biol.*, *19*(6), 498–502. doi:10.1016/j.cub.2009.01.066
Stepanek, J. (2006). Musical sound timbre : Verbal descriptions and dimensions. In *Proc. 9th Int. Conf. Digit. Audio Eff.* (pp. 121–126). Montreal.
Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain.* New York: Thieme Medical.
Tervaniemi, M., Peretz, I., Brattico, E., Ja, M., & Järvenpää, M. (2009). The amusic brain: in

tune, out of key, and unaware. *Brain*, *132*, 1277–1286. doi:10.1093/brain/awp055
Tian, B., & Rauschecker, J. P. (2004). Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *J. Neurophysiol.*, *92*(5), 2993–3013. doi:10.1152/jn.00472.2003
Turner, R. S. (2009). The Ohm-Seebeck Dispute, Hermann von Helmholtz, and the Origins of Physiological Acoustics. *Br. J. Hist. Sci.*, *10*(1), 1. doi:10.1017/S0007087400015089
von Bismarck, G. (1974). Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acustica*, *30*(3), 146–159.
Walliser, K. (1969). Zusammenhänge zwischen dem Schallreiz und der Periodentonhöhe [Relationships between the sound stimulus and the pitch]. *Acustica*, *21*, 319–328.
Warren, J. D., Jennings, A. R., & Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *Neuroimage*, *24*(4), 1052–7. doi:10.1016/j.neuroimage.2004.10.031
Warren, J. D., Uppenkamp, S., Patterson, R. D., & Griffiths, T. D. (2003). Separating pitch chroma and pitch height in the human brain. *Proc. Natl. Acad. Sci. U. S. A.*, *100*(17), 10038–42. doi:10.1073/pnas.1730682100
Warrier, C. M., & Zatorre, R. J. (2002). Influence of tonal context and timbral variation on perception of pitch. *Percept. Psychophys.*, *64*(2), 198–207.
Warrier, C., Wong, P., Penhune, V., Zatorre, R. J., Parrish, T., Abrams, D., & Kraus, N. (2009). Relating structure to function: Heschl's gyrus and acoustic processing. *J. Neurosci.*, *29*(1), 61–9. doi:10.1523/JNEUROSCI.3489-08.2009
Xu, J., Moeller, S., Auerbach, E. J., Strupp, J., Smith, S. M., Feinberg, D. a, … Uğurbil, K. (2013). Evaluation of slice accelerations using multiband echo planar imaging at 3 T. *Neuroimage*, *83*, 991–1001. doi:10.1016/j.neuroimage.2013.07.055
Yang, H., Ma, W., Gong, D., Hu, J., & Yao, D. (2014). A Longitudinal Study on Children's Music Training Experience and Academic Development. *Sci. Rep.*, *4*, 5854. doi:10.1038/srep05854
Yost, W. A. (1996). Pitch strength of iterated rippled noise. *J. Acoust. Soc. Am.*, *100*(5), 3329–35.
Zaehle, T., Wüstenberg, T., Meyer, M., & Jäncke, L. (2004). Evidence for rapid auditory perception as the foundation of speech processing: a sparse temporal sampling fMRI study. *Eur. J. Neurosci.*, *20*(9), 2447–56. doi:10.1111/j.1460-9568.2004.03687.x
Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cereb. Cortex*, *11*(10), 946–53.
Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends Cogn. Sci.*, *6*(1), 37–46.
Zatorre, R. J., Evans, A. C., & Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *J. Neurosci.*, *14*(4), 1908–19.