Hearing Research 308 (2014) 60–70

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

Explaining the high voice superiority effect in polyphonic music: Evidence from cortical evoked potentials and peripheral auditory models

Laurel J. Trainor [a,b,c,*], Céline Marie [a,b], Ian C. Bruce [b,d], Gavin M. Bidelman [e,f]

[a] Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton, ON, Canada
[b] McMaster Institute for Music and the Mind, Hamilton, ON, Canada
[c] Rotman Research Institute, Baycrest Centre, Toronto, ON, Canada
[d] Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada
[e] Institute for Intelligent Systems, University of Memphis, Memphis, TN, USA
[f] School of Communication Sciences & Disorders, University of Memphis, Memphis, TN, USA

ARTICLE INFO

Article history:
Received 19 January 2013
Received in revised form
12 July 2013
Accepted 25 July 2013
Available online 3 August 2013

ABSTRACT

Natural auditory environments contain multiple simultaneously-sounding objects and the auditory system must parse the incoming complex sound wave they collectively create into parts that represent each of these individual objects. Music often similarly requires processing of more than one voice or stream at the same time, and behavioral studies demonstrate that human listeners show a systematic perceptual bias in processing the highest voice in multi-voiced music. Here, we review studies utilizing event-related brain potentials (ERPs), which support the notions that (1) separate memory traces are formed for two simultaneous voices (even without conscious awareness) in auditory cortex and (2) adults show more robust encoding (i.e., larger ERP responses) to deviant pitches in the higher than in the lower voice, indicating better encoding of the former. Furthermore, infants also show this high-voice superiority effect, suggesting that the perceptual dominance observed across studies might result from neurophysiological characteristics of the peripheral auditory system. Although musically untrained adults show smaller responses in general than musically trained adults, both groups similarly show a more robust cortical representation of the higher than of the lower voice. Finally, years of experience playing a bass-range instrument reduces but does not reverse the high voice superiority effect, indicating that although it can be modified, it is not highly neuroplastic. Results of new modeling experiments examined the possibility that characteristics of middle-ear filtering and cochlear dynamics (e.g., suppression) reflected in auditory nerve firing patterns might account for the higher-voice superiority effect. Simulations show that both place and temporal AN coding schemes well-predict a high-voice superiority across a wide range of interval spacings and registers. Collectively, we infer an innate, peripheral origin for the higher-voice superiority observed in human ERP and psychophysical music listening studies.

This article is part of a Special Issue entitled <Music: A window into the hearing brain>.

1. Introduction

In many musical genres, more than one sound is played at a time. These different sounds or *voices* can be combined in a *homophonic* manner, in which there is one main voice (*melody line* or *stream*) with the remaining voices integrating perceptually in a chordal fashion, or in a *polyphonic* manner in which each voice can be heard as a melody in its own right. In general, compositional practice is to place the most important melody line in the voice or stream with highest pitch. Interestingly, this way to compose is consistent with studies indicating that changes are most easily

Abbreviations: AN, auditory nerve; CF, characteristic frequency; EEG, electroencephalography; ERP, event-related potential; *F*0, fundamental frequency; ISIH, interspike interval histograms; MEG, magnetoencephalography; MMN, mismatch negativity

* Corresponding author. Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada. Tel.: +1 905 525 9140x23007.
E-mail address: ljt@mcmaster.ca (L.J. Trainor).

0378-5955/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.heares.2013.07.014

detected in the highest of several streams (Crawley et al., 2002; Palmer and Holleran, 1994; Zenatti, 1969). However, to date, no explanation has been offered as to how or where in the auditory system this high-voice superiority effect arises. In the present paper, we first review electroencephalographic (EEG) and magnetoencephalographic (MEG) evidence indicating that the high-voice superiority effect is present early in development and, although somewhat plastic, cannot easily be reversed by extensive musical experience. We then present new simulation results from a model of the auditory nerve (AN) (Zilany et al., 2009; Ibrahim and Bruce, 2010) that indicate that the effect originates in the peripheral auditory system as a consequence of the interaction between physical properties of musical tones and nonlinear spectrotemporal processing properties of the auditory periphery.

2. The high voice superiority effect in auditory scene analysis: event-related potential evidence for a pre-attentive physiological origin

It has been argued that musical processing, like language, is unique to the human species (e.g., McDermott and Hauser, 2005). Although some species appear able to entrain to regular rhythmic patterns (Patel et al., 2009; Schachner et al., 2009), and others can be trained to respond to pitch features such as consonance and dissonance (Hulse et al., 1995; Izumi, 2000), none appear to produce music with the features, syntactic complexity, and emotional connections of human music. At the same time, human music rests firmly on basic auditory perceptual processes that are common across a variety of species (e.g., Micheyl et al., 2007; Snyder and Alain, 2007), such that musical compositions using abstract compositional systems, not rooted in the perceptual capabilities of the auditory system, are very difficult to process (e.g., Huron, 2001; Trainor, 2008). Huron (2001), for example, has shown that many of the accepted rules for composing Western tonal music might have arisen based on fundamental, general features of human auditory perception (e.g., masking, temporal coherence). Here we argue that the high voice superiority effect is the direct consequence of properties of the peripheral auditory system.

The human auditory system evolved in order to perform complex spectrotemporal processing aimed at determining what sound sources (corresponding to *auditory objects*) are present in the environment, their locations, and the meanings of their output (Griffiths and Warren, 2004; Winkler et al., 2009). Typically, there are multiple simultaneously-sounding objects in the human environment (e.g., multiple people talking, airplanes overhead, music playing on a stereo). The sound waves from each auditory object (and their echoes) sum in the air and reach the ear as one complex sound wave. Thus, in order to determine what auditory objects are present, the auditory system must determine how many auditory objects are present, and which components of the incoming sound wave belong to each auditory object. This process has been termed auditory scene analysis (Bregman, 1990). Auditory scene analysis has a deep evolutionary history and appears to operate similarly across a range of species (Hulse, 2002) including songbirds (Hulse et al., 1997), goldfish (Fay, 1998, 2000), bats (Moss and Surlykke, 2001), and macaques (Izumi, 2002).

Because the basilar membrane in the cochlea in the inner ear vibrates maximally at different points along its length for different frequencies in an orderly tonotopic fashion, it can be thought of as performing a quasi-Fourier analysis. Inner hair cells attach to the basilar membrane along its length and tend to depolarize at the time and location of maximal basilar membrane displacement, thus creating a tonotopic representation of frequency channels in the auditory nerve that is maintained through subcortical nuclei and into primary auditory cortex. A complementary temporal representation, based on the timing of firing across groups of neurons, is also maintained within the auditory system. From this spectrotemporal decomposition, the auditory system must both integrate frequency components that likely belong to the same auditory object, and segregate frequency components that likely belong to different auditory objects. These processes of integration and separation must occur for both sequentially presented and simultaneously presented sounds. For example, the successive notes of a melody line or the successive speech sounds of a talker need to be grouped as coming from the same auditory source and form a single auditory object. Moreover, this object must be separated from other sequences of sounds that may also be present in the environment. With respect to simultaneously-occurring sounds, the harmonic frequency components of a complex tone must be integrated together and heard as a single auditory object whereas the frequency components of two different complex tones presented at the same time must be separated.

A number of cues are used for auditory scene analysis. For example, sequential sounds that are similar in pitch, timbre and/or location tend to be grouped perceptually (see Bregman, 1990 for a review). The closer together sounds are in time, the more likely they are to be integrated (e.g., Bregman and Campbell, 1971; Bregman, 1990; Darwin and Carlyon, 1995; van Noorden, 1975, 1977). Pitch provides one of the most powerful cues for sequential integration (e.g., see Micheyl et al., 2007). For example, successive tones that are close in fundamental frequency (*F*0) are easily integrated and are heard as coming from a single auditory object whereas tones differing in *F*0 remain distinct, and are difficult to integrate into a single auditory object (e.g., Dowling, 1973; Sloboda and Edworthy, 1981; van Noorden, 1975, 1977).

Sound frequency is also critical for auditory scene analysis in the context of simultaneous sounds. Sounds with well-defined pitch (e.g., musical tones) typically contain energy at an *F*0 and integer multiples of that frequency (harmonics or overtones). Thus, a tone with an *F*0 of 400 Hz will also contain energy at 800, 1200, 1600, 2000, … Hz and, consequently, the representation of that tone will be distributed across the basilar membrane. The perceived pitch typically corresponds to that of a puretone of the fundamental frequency, but the pitch is determined from the set of harmonics, as evidence by the fact that removal of the fundamental frequency does not alter the pitch appreciatively (i.e., case of the missing fundamental). If two tones are presented simultaneously, their harmonics will typically be spread across similar regions of the basilar membrane. As long as harmonic frequencies are more than a critical bandwidth apart, the auditory system is exquisitely able to detect subtle differences in intensity between simultaneously-presented harmonics (e.g., Dai and Green, 1992). The auditory system uses a number of cues to determine how many simultaneously presented tones are present and which harmonics belong to which tone. One of the most important cues is harmonicity. Integer related frequency components will tend to be grouped as coming from a single source, and will be segregated from the other frequency components given their common harmonicity. The operation of harmonicity in auditory scene analysis has been demonstrated in a number of ways (see Bregman, 1990). For instance, mistuning one harmonic in a complex tone causes that harmonic to be perceptually segregated from the complex tone, giving rise to the perception of two auditory objects, one at the pitch of the mistuned harmonic and the other at the fundamental frequency of the complex tone (Alain and Schuler, 2002).

The physiological processes underlying auditory scene analysis likely involve many levels of the auditory system (e.g., see Alain and Winkler, 2012; Snyder and Alain, 2007; for reviews). The participation of the auditory periphery (*channeling theory*) is strongly suggested from studies showing that streaming according to

frequency is strongest for stimuli with the least overlap between representations on the basilar membrane (e.g., Hartmann and Johnson, 1991) and from studies showing decreases in stream segregation with increases in intensity, which lead to greater overlap of representations along the cochlear partition (e.g., Rose and Moore, 2000). At the same time, fMRI studies strongly suggest cortical involvement (Deike et al., 2004; Wilson et al., 2007), and electrophysiological recordings from awake macaques indicate that sequential auditory streaming could be accomplished in primary auditory cortex (Fishman et al., 2001; Micheyl et al., 2007). The notion that auditory scene analysis involves a coordination of both innate bottom–up processes, learned relations, and top–down attentional processes has been proposed by a number of researchers (e.g., Alain and Winkler, 2012; Bregman, 1990; Snyder and Alain, 2007; van Noorden, 1975). Several EEG studies also indicate that sequential streams are formed in auditory cortex at a preattentive stage of processing (e.g., Gutschalk et al., 2005; Nager et al., 2003; Shinozaki et al., 2000; Snyder et al., 2006; Sussman, 2005; Winkler et al., 2005; Yabe et al., 2001).

While auditory scene analysis applies to all sounds, music represents a somewhat special case in that to some extent, integration and segregation are desired at the same time. In homophonic music, it is desired that the melody line segregates from the other voices (and in polyphonic music that all lines segregate from each other), while at the same time the voices need to fit together harmonically and integrate to give sensations of different chord types (e.g., major, minor, dominant sevenths, diminished) that are defined by the pitch interval relations between their component tones.

Members of our group (Fujioka et al., 2005) presented the first evidence that two simultaneously-presented melodies with concurrent tone onsets form separate memory traces in auditory cortex at a preconscious level. They showed, further, that the higher-pitched melody formed a more robust memory trace than the lower-pitched melody. Specifically, they conducted an event-related potential (ERP) study in which they measured the amplitude of the mismatch negativity (MMN) component in response to deviant (changed) notes in either the higher or the lower of two simultaneous melodies. When measured at the scalp, MMN manifests as a frontally negative peak (reversing polarity at posterior sites consistent with a main generator in auditory cortex) occurring around 150–250 ms after the onset of an unexpected deviant sound in a stream of expected (predicable) standard sounds (see Näätänen et al., 2007; Picton et al., 2000; for reviews). Although affected by attention, MMN does not require conscious attention to be elicited and can be measured in young infants (Trainor, 2012). MMN only occurs when the deviant sound occurs less frequently than the standard sound and MMN increases in amplitude the rarer the deviant sounds, suggesting that MMN reflects a response to an unexpected event that the brain failed to predict. Fujioka et al. presented two simultaneous 5-note melodies with concurrent tone onsets. In different conditions, the two melodies (A and B) were transposed such that in half the conditions melody A was in the higher voice and in the other half melody B was in the higher voice. On 25% of trials, the final tone of the higher melody was changed (deviant) and on another 25% of trials the final tone of the lower melody was changed. Thus, 50% of trials were standard and 50% were deviant. If the two melodies were integrated into a single memory trace, a very small or non-existent MMN would be expected. However, if each melody was encoded in a separate memory trace, the deviance rate for each melody would be 25% and an MMN response would be expected. Fujioka et al. found that robust MMN was elicited, suggesting that separate memory traces were formed for each melody (Fig. 1). Furthermore, the MMN was much larger for deviants in the high than in the low voice, providing the first evidence that the high-voice superiority effect manifests preattentively at the level of auditory cortex.

We then investigated the high voice superiority effect further with simplified stimuli (Fujioka et al., 2008). In this case, the A and B melodies were each replaced by a single tone separated in pitch by 15 semitones (one semitone equals 1/12 octave), so that listeners heard a repeating high and a repeating low tone with simultaneous onsets and offsets. On 25% of trials (deviants) the higher tone was raised by two semitones. On another 25% of trials, the lower tone was lowered by two semitones. As in Fujioka et al. (2005), a high voice superiority effect was evident, with larger MMN to deviants in the higher than in the lower voice. Using the Glasberg and Moore (2002) loudness model, we estimated the short-term loudness level of the stimuli used in Fujioka et al. (2008) and found a very similar level of loudness across stimuli with mean = 85.2 phons and SD = 0.8 phon. Thus we infer that these MMN results cannot be due to differences in loudness between the high and low voices.

In order to better understand this effect, several control conditions were added as well, each containing only one voice (i.e., either the stream of high tones or the stream of low tones alone). In one control condition, both deviants (25% of trials each) were presented in the same voice. MMN was larger and earlier in the original condition when both voices were present than in this control condition when only a single voice was present, confirming that separate memory traces exist for the two simultaneous voices. In other control conditions, each again involving only one of the voices, only one of the deviants (25% of trials) was presented, so that responses to that deviant could be compared when the voice was presented on its own compared to when it was presented in the context of a higher or a lower simultaneous voice. The results indicated that MMN measured in the high voice in isolation was similar to MMN measured in that voice when it was presented in the context of a lower voice. However, MMN measured in the low voice in isolation was larger than when measured in that voice in the context of a higher voice. Taken together, these results provide support for the idea that the high voice superiority effect manifests preattentively at the level of auditory cortex for both tones and complex melodies.

Finding evidence for a high voice superiority effect in auditory cortex does not necessarily indicate that it is the origin of the effect. Indeed, it is quite possible that it has a more peripheral origin, and the effect simply propagates to more central regions. In fact, there is evidence that musicians show better encoding at the level of the brainstem for the harmonics of the higher of two simultaneously presented tones (Lee et al., 2009). Bregman (1990) proposed that many aspects of auditory scene analysis have a strong bottom–up component that is likely innate. Because cortex and, thus, top–down processing is very immature in young infants, one way to test this hypothesis is to examine whether young infants form auditory streams. There is evidence that infants can form separate streams from sequentially presented stimuli (Demany, 1982; Fassbender, 1993; McAdams and Bertoncini, 1997; Smith and Trainor, 2011; Winkler et al., 2003) and a recent study indicates that infants can also use harmonicity to segregate mistuned harmonics from a complex tone containing simultaneously presented frequency components (Folland et al., 2012). Finally, it should be noted that these auditory scene analysis abilities emerge prior to experience-driven enculturation to the rhythmic and pitch structure of the music in the infants' environment (see Trainor and Corrigall, 2010; Trainor and Hannon, 2012; Trainor and Unrau, 2012; for reviews).

Members of our group (Marie and Trainor, 2013) tested whether 7-month-old infants also show a high voice superiority effect by presenting them with stimuli similar to those of Fujioka et al. (2008) and measuring the MMN component of the ERP. Specifically, each of the two simultaneously presented streams (high and

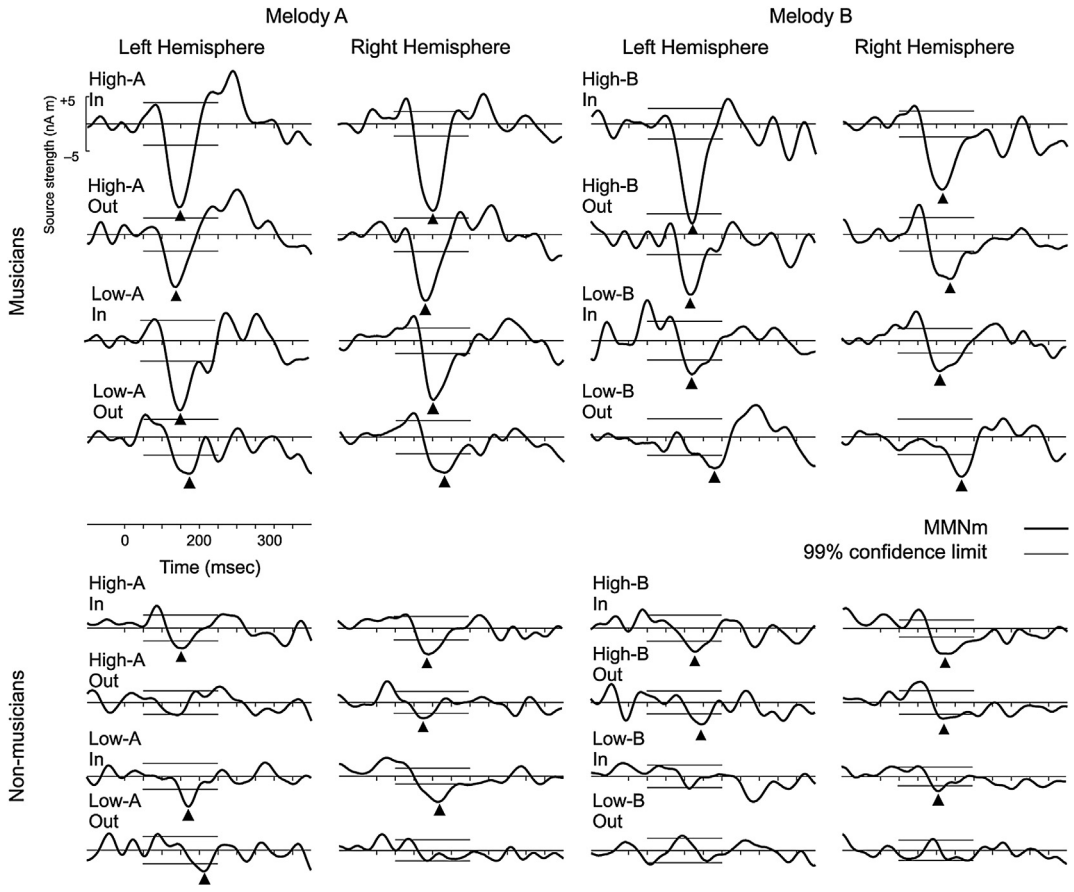

Fig. 1. The grand averaged ($n = 10$ subjects) difference (deviant − standard) waveforms from a source in auditory cortex showing MMN responses to deviants (arrows) in Melody A (left panel) and Melody B (right panel) when each melody was in the higher or the lower voice. Responses from musicians are shown in the upper panel and responses from non-musicians in the lower panel. Also shown separately are MMN responses when the deviant notes fell outside the key of the melody and when they remained within the key of the melody. Time zero represents the onset of the deviant note and thin lines show the upper and lower limits of the 99% confidence interval for the estimated residual noise. It can be seen that responses are larger for deviants in the higher than the lower voice, and also for musicians than nonmusicians. Reprinted with permission from Fujioka et al. (2005).

low tones separated by 15 semitones) contained deviants that either increased or decreased in pitch by a semitone. The two control conditions consisted of either the high stream alone or the low stream alone. MMN responses to deviants were larger and earlier in the higher than in the lower voice when both were presented simultaneously (Fig. 2). Furthermore, MMN to deviants in the higher voice were larger when the high voice was presented in the context of the lower voice than when presented alone. In contrast, MMN to deviants in the lower voice were smaller when the lower voice was presented in the context of the higher voice than when presented alone. These results indicate that the high voice superiority effect emerges early in development and therefore likely involves a strong, bottom–up aspect such that it might not be affected greatly by experience.

Fujioka et al. (2005) examined the effects of musical experience on high-voice superiority and found larger MMN responses overall in musicians compared to nonmusicians, but that both groups similarly showed larger responses to deviants in the higher than in the lower voice. Members of our group (Marie et al., 2012) tested the effects of experience further, asking whether the high voice superiority effect could be reversed by experience. They reasoned that musicians who play bass-range instruments have years of experience focusing on the lowest-pitched voice in music. Specifically, they hypothesized that if the high voice superiority effect is largely a result of experience with music, musicians who play soprano-range instruments should show a high-voice superiority effect, but it should be reversed in musicians who play bass-range musical instruments. Using the two 5-note melodies of Fujioka et al. (2005), they measured MMN to deviants in the higher and lower of the two voices. They found significant differences in MMN responses between musicians playing soprano-range instruments and musicians playing bass-range instruments. Specifically, musicians playing soprano-range instrument showed the expected high voice superiority effect, with significantly larger MMN to deviants in the higher than in the lower voice. In musicians playing bass-range instruments, MMN was also larger to deviants in the higher than in the lower voice, but this difference was attenuated and did not reach statistical significance. These results are consistent with the hypothesis that experience can affect the degree of high voice superiority, but suggest that even very extensive experience focusing on the lowest voice in music cannot reverse the high voice superiority effect.

In sum, the ERP results suggest that the high voice superiority effect manifests at a preattentive stage of processing, does not require top–down attentional control, is present early in development and, although it can be reduced, is not reversible by extensive experience. Together these results suggest that the high voice superiority effect in music may have an origin in more peripheral sites of auditory processing. This of course cannot be tested by measuring cortical potentials such as MMN, so to explore the possibility that high voice superiority in music emerges as the result of peripheral auditory neurophysiological processing, we examined response properties from an empirically grounded, phenomenological model of the auditory nerve (AN) (Zilany et al., 2009). In particular, because

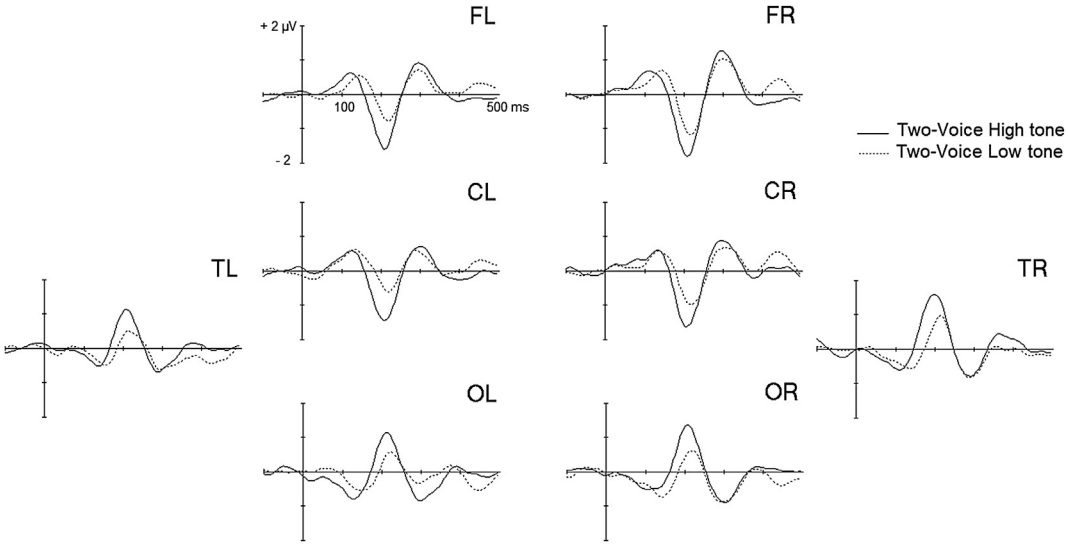

Fig. 2. Grand averaged ($n = 16$) MMN difference (deviant − standard) waveforms from left (L) and right (R) frontal (F), central (C), temporal (T) and occipital (O) scalp sites. Time zero represents the onset of the deviant tone. The polarity reversal from front to back of the scalp is consistent with a generator in auditory cortex. MMN is larger for deviants that occur in the high than in the low voice. Reprinted with permission from Marie and Trainor (2013).

we are interested in humans, we used the more recent generation of this model (Ibrahim and Bruce, 2010), which incorporates recent estimates of human cochlear tuning.

3. Neural correlates of the higher tone salience at the level of auditory nerve

Initial attempts to explain the high voice superiority effect focused on explanations involving peripheral physiology and constraints of cochlear mechanics. In these accounts, peripheral masking and/or suppression are thought to influence the salience with which a given voice is encoded at the auditory periphery yielding a perceptual asymmetry between voices in multi-voice music (Plomp and Levelt, 1965; Huron, 2001). However, as noted by recent investigators (e.g., Fujioka et al., 2005, 2008; Marie and Trainor, 2013), given the asymmetric shape of the auditory filters (i.e., peripheral tuning curves) and the well-known upward spread of masking (Egan and Hake, 1950; Delgutte, 1990a,b), these explanations would, on the contrary, predict a low voice superiority. As such, more recent theories have largely dismissed these cochlear explanations as they are inadequate to account for the high voice prominence reported in both perceptual (Palmer and Holleran, 1994; Crawley et al., 2002) and ERP data (Fujioka et al., 2008; Marie and Trainor, 2013).

In contrast to these descriptions based on conceptual models of cochlear responses to pure tones, single-unit responses from the AN have shown rather convincingly that peripheral neural coding of realistic tones and other complex acoustic stimuli can account for a wide range of perceptual pitch attributes (Cariani and Delgutte, 1996a,b). As such, we reexamine the role of peripheral auditory mechanisms in accounting for the high voice superiority using the realistic piano tones used in the MMN studies. Specifically, we aimed to determine whether or not neurophysiological response properties at the level of AN could account for the previously observed perceptual superiority of the higher voice in polyphonic music.

3.1. Auditory-nerve model architecture

Spike-train data from a biologically plausible, computational model of the cat AN (Zilany et al., 2009; Ibrahim and Bruce, 2010) was used to assess the salience of pitch-relevant information encoded at the earliest stage of neural processing along the auditory pathway. This phenomenological model represents the latest extension of a well-established model rigorously tested against actual physiological AN responses to both simple and complex stimuli, including tones, broadband noise, and speech-like sounds (Zilany and Bruce, 2006, 2007). The model incorporates several important nonlinearities observed in the auditory periphery, including cochlear filtering, level-dependent gain (i.e., compression) and bandwidth control, as well as two-tone suppression. Recent improvements to the model introduced power-law dynamics and long-term adaptation into the synapse between the inner hair cell and auditory nerve fiber, yielding more accurate responses to temporal features of complex sound (e.g., amplitude modulation, forward masking) (Zilany et al., 2009). Model threshold tuning curves have been well fit to the CF-dependent variation in threshold and bandwidth for high-spontaneous rate (SR) fibers in normal-hearing cats (Miller et al., 1997). The stochastic nature of AN responses is accounted for by a modified nonhomogenous Poisson process, which includes effects of both absolute and relative refractory periods and captures the major stochastic properties of AN responses (e.g., Young and Barta, 1986). Original model parameters were fit to single-unit data recorded in cat (Zilany and Bruce, 2006, 2007). However, more recent modifications (Ibrahim and Bruce, 2010)—adopted presently—have attempted to at least partially "humanize" the model, incorporating human middle-ear filtering (Pascal et al., 1998) and increased basilar membrane frequency selectivity to reflect newer (i.e., sharper) estimates of human cochlear tuning (Shera et al., 2002; Joris et al., 2011).

3.2. Rate-place representation of the ERP-study stimuli

It is instructive to look first at how the stimuli used in the ERP study of Marie and Trainor (2013) are expected to be represented by the auditory nerve. In this analysis, shown in Fig. 3, we look at the so-called *rate-place* representation of the acoustic stimuli, that is, the spike count as a function of the AN fiber characteristic frequency (CF). By comparing this rate-place neural representation (the green curves in Fig. 3) to the stimulus frequency spectrum (the dark blue curves in Fig. 3), it is possible to observe how the AN represents each of the individual harmonics of the low and high tones when presented

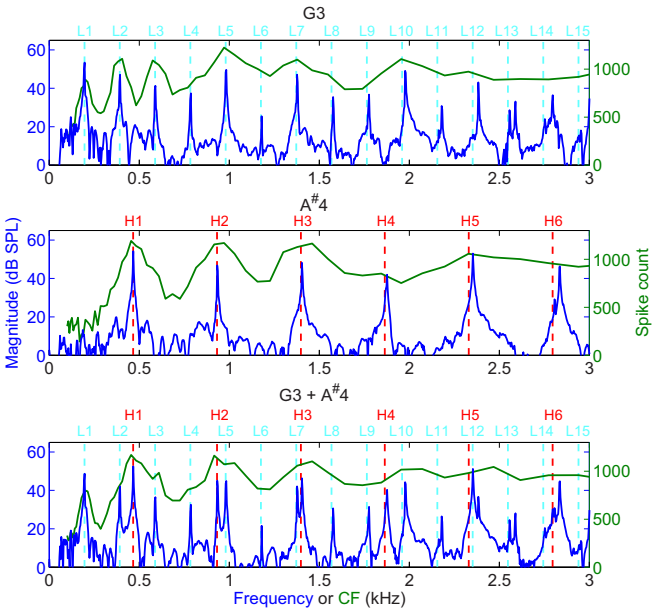

Fig. 3. Simulated neural rate-place representation of the standard low tone (G3; top panel), of the standard high tone ($A^{\#}4$; middle panel) and of the combined presentation of the low and high tones (G3 + $A^{\#}4$; bottom panel) from the ERP study of Marie and Trainor (2013). In each panel, the dark blue curve shows the frequency spectrum of the acoustic stimulus (with corresponding scale on the left) and the green curve shows the neural response (with the scale on the right) as a function of the AN fiber characteristic frequency (CF). The spike count is the summed response of fifty model AN fibers at each CF over a 150-ms period of the stimulus presentation. The fifty fibers at each CF have a physiologically-realistic mix of spontaneous discharge rates and corresponding thresholds (Liberman, 1978). The responses are calculated at 59 different CFs, logarithmically-spaced between 100 Hz and 3 kHz. The vertical dashed cyan lines indicate the nominal harmonic frequencies for the low tone (labeled as L1–L15), and the vertical dashed red lines those of the high tone (labeled as H1–H6).

separately and when presented together. As shown in Fig. 3, most of the lower harmonics of both the standard low tone (top panel) and the standard high tone (middle panel) are well represented (or *resolved*) in the shape of the rate-place profile of the model AN response, that is, spectral peaks at the harmonics are represented by higher spike counts in the AN fibers tuned to those frequencies and spectral dips between the harmonics are represented by lower spike counts for CFs in the dips. Note that some of the higher harmonics that are lower in intensity are less well resolved. One interesting feature of the AN response to the low tone (top panel) is that the peak spike count in response to the fundamental (L1) is less than that of many of the other harmonics (particularly L2, L3, L5, L7 and L10), even though the fundamental has the highest amplitude of all the harmonics in the spectral representation (as can be seen from the dark blue stimulus spectrum curve). This results from the bandpass filtering of the middle ear and the lower cochlear gain at low CFs, which together attenuate low frequency stimulus components. However, note that the loudness of the low tone, calculated with the model of Glasberg and Moore (2002), is only 2.8 phon quieter when the fundamental is completely absent from the complex tone. Thus, even if the middle-ear filtering reduces the representation of the low tone's fundamental (L1), its additional harmonics maintain the overall loudness level.

In contrast to isolated tone presentation, when the two tones are played simultaneously (Fig. 3, bottom panel), the first few harmonics of the high tone (H1–H3) are well resolved in the neural response (green curve), but only the fundamental of the low tone (L1) is well resolved; its other harmonics are not. This is evident by the peak in the AN response at a CF matching the low tone fundamental frequency (L1), but not at L2 and L3. This contrasts with when the low tone is presented in isolation (top panel). The fundamental (H1) and second harmonic (H2) of the high tone visibly suppress the neural responses to the second and third harmonics of the low tone (L2 and L3) in the bottom panel of Fig. 3. The interaction between each tone's harmonics can be explained by the well-known phenomena of "two-tone suppression" that occurs due to cochlear nonlinearities. When two nearby frequency components are presented, the one with higher intensity suppresses the one with lower intensity (see Delgutte, 1990a,b, as well as Zhang et al., 2001, for a review of tone-two suppression and how it is achieved in this computational model). In keeping with most natural sounds, in the tones from the MMN study of Marie and Trainor (2013), the intensity of the first few harmonics rolls off with increasing harmonic number such that when a harmonic from the low tone is close in frequency to a harmonic from the high tone, the latter will be of lower harmonic number and therefore more intense. Consequently, at most CFs, the high tone's components act to suppress those of the lower tone. As such, the high tone's harmonics are more faithfully represented in the neural response to the combined stimulus. This is evident in the pattern of neural response to the combined tones (green curve), which bears closer resemblance to that of the high (middle panel) than that of the low tone (top panel). The relatively small peak spike count at L1 can be explained by the filtering of the middle ear and the lower cochlear gain at low CFs.

In order to quantify the similarity between the neural responses to the combined tone and to each tone alone, we performed a linear regression between the pairs of spike count curves for CFs from 125 Hz to 1.75 kHz, a frequency region in which the harmonics of both tones are generally well resolved. Results confirmed a higher degree of correlation between the neural responses of the combined tones and the high tone alone (adjusted $R^2 = 0.79$) than between neural responses of the combined tone and the low tone alone (adjusted $R^2 = 0.74$). Note that we repeated these simulations with a version of the auditory-periphery model that has no middle-ear filter and fixed basilar-membrane filters (such that two-tone suppression is absent from the model). In this case, the result changes dramatically (see Supplemental Fig. S1). Indeed, without middle-ear filtering and two-tone suppression, the adjusted R^2 value for the high tone response drops to 0.74, while the adjusted R^2 value for the low tone response increases to 0.84. This indicates that in the absence of intrinsic peripheral filtering and nonlinearities, a low-voice superiority is actually predicted.

Finally, when the different deviant stimuli from Marie and Trainor (2013) are tested with the full auditory periphery model (i.e., including middle-ear filtering and two-tone suppression), the predicted neural response tends again to be dominated by the high tone for at least the first few harmonics (results not shown).

The roll off in intensity with increasing harmonic number is a common feature of natural tones, including the human voice, and therefore a high voice dominance might be expected for most pairs of natural tones presented at equal overall intensity. Presentation of a low-frequency tone at a sufficiently greater intensity would be expected to overcome the suppressive effects of a simultaneous high-frequency tone. Similarly, synthetic harmonic complexes with equal-amplitude harmonics (as are often used in psychophysical experiments) would not be expected to exhibit the same degree of high-voice superiority as natural tones, because the equal amplitude harmonics would not lead to as clear a pattern of dominance in the nonlinear interactions in the cochlea. In other words, two-tone suppression would not consistently work in favor of the harmonics of one tone or the other.

3.3. Temporal-representation pitch salience for tone pairs

The rate-based simulation results of the previous section not only help explain the results of the ERP studies but also prompt the question of how middle-ear filtering and cochlear two-tone suppression affect the neural representation of tone pairs over a range of musical intervals and in different registers. While computational models of pitch perception based on rate-place representations have been proposed (e.g., Cohen et al., 1995), they have not yet been validated with the physiologically-accurate AN model. Therefore, in the following simulations, we explore temporal measures of pitch encoding (which have been validated with the AN model) to examine the possibility that neural correlates of the high voice superiority exist in the fine timing information in AN firing patterns. Previous work has demonstrated that temporal-based codes (e.g., autocorrelation) provide robust neural correlates for many salient aspects relevant to music listening including sensory consonance, tonal fusion, and harmonicity (Bidelman and Heinz, 2011). Furthermore, previous studies have shown that cochlear two-tone suppression has similar effects on the rate-place and temporal representations of harmonic complexes (Bruce et al., 2003; Miller et al., 1997) so it is expected that these peripheral effects would again manifest in temporal characteristics of AN responses.

3.3.1. Stimuli

Musical dyads (i.e., intervals composed by two simultaneously presented notes) were synthesized using harmonic tone-complexes each consisting of 10 harmonics added in cosine phase. Component amplitudes decreased by −6 dB/octave to mimic the spectral roll off produced by natural instrumental sounds and voices. We ran simulations in three frequency ranges. In each range, the fundamental frequency (*F*0) of the lower tone was fixed (either C2, C3, C4). The higher *F*0 was varied to produce different musical (and nonmusical) intervals within a multi-octave range (variation of the higher tone *F*0: *low range*: C2–C6, 65–1046 Hz; *middle*: C3–C6, 130–1046 Hz; *high*: C4–C6, 261–1046 Hz). Within each range, the *F*0 of the higher tone was successively increased by ¼ semitone (cf. the smallest interval in music: 1 semitone) resulting in 48 intervals/octave. Stimulus waveforms were 300 ms in duration (including 10 ms rise–fall times) and were presented at an intensity of 70 dB SPL. Broadly speaking, intensity and spectral profile have minimal effects on temporal based AN representations of pitch (Cariani and Delgutte, 1996b; Cedolin and Delgutte, 2005; Bidelman and Heinz, 2011), consistent with the invariance of pitch perception to manipulations in these parameters (e.g., Houtsma and Smurzynski, 1990). Thus, in the present simulations, we limit our analysis to a single musical timbre (decaying harmonics) presented at moderate intensity. More extensive effects of stimulus intensity and spectral content on AN encoding of musical intervals have been reported previously (Bidelman and Heinz, 2011).

3.3.2. Neural pitch salience computed via periodic sieve template analysis of AN spike data

To quantify pitch-relevant information contained in AN responses, we adopted a temporal analysis scheme used previously to examine the periodicity information contained in an aggregate distribution of neural activity (Cedolin and Delgutte, 2005; Bidelman and Heinz, 2011). An ensemble of 70 high-SR (>50 spikes/s) auditory nerve fibers was simulated with CFs spread across the cochlear partition (80–16,000 Hz, logarithmic spacing). First-order interspike interval histograms (ISIH) were estimated for each CF (Fig. 4A) (for details, see Bidelman and Krishnan, 2009; Bidelman and Heinz, 2011). Individual ISIHs were then summed across CFs to obtain a pooled interval distribution for the entire neural ensemble representing all pitch-related periodicities contained in the aggregate AN response. To estimate the neural pitch salience of each musical interval stimulus, the pooled ISIH was then input to a "periodic sieve" analysis, a time-domain analog of the classic pattern recognition models of pitch which attempt to match response activity to an internal harmonic template (Goldstein, 1973; Terhardt et al., 1982). Sieve templates (each representing a single pitch) were composed of 100 μs wide bins situated at the fundamental pitch period and its multiples (Fig. 4B); all sieve templates with *F*0s between 25 and 1000 Hz (2 Hz steps) were used to analyze ISIHs.

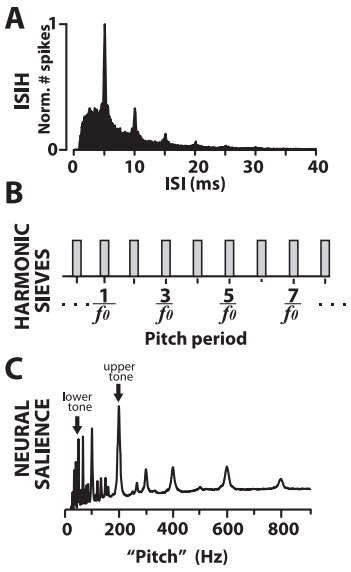

Fig. 4. Procedure for computing "neural pitch salience" from AN responses to a single musical interval. Single-unit responses were generated by presenting two-tone intervals (100 stimulus repetitions) to a computational model of the AN (Zilany et al., 2009; Ibrahim and Bruce, 2010) using 70 model fibers (CFs: 80–16,000 Hz.) (A) From individual fiber spike trains, interspike interval histograms (ISIHs) were first estimated to index pitch periodicities contained in individuals fibers. Fiber ISIHs were then summed to create a pooled, population-level ISIH indexing the various periodicities coded across the AN array. (B) Each pooled ISIH was then passed through a series of periodic sieves each reflecting a single pitch template (i.e., *F*0). The magnitude at the output of a single sieve reflects the salience of pitch-relevant information for the corresponding *F*0 pitch. (C) Analyzing the output across all possible sieve templates (*F*0 = 25–1000 Hz) results in a running salience curve for a particular stimulus. Salience magnitudes at the *F*0s corresponding to the higher and lower tone were taken as an estimate of neural pitch salience for each tone in a dyad (arrows). See text for details.

Neural pitch salience for a single *F*0 template was estimated by dividing the mean density of ISIH spike intervals falling within the sieve bins by the mean density of activity in the whole interval distribution. Activity falling within sieve "windows" adds to the total pitch salience while information falling outside the "windows" reduces the total pitch salience. By compounding the output of all sieves as a function of *F*0 we examine the relative strength of all possible pitches present in AN which may be associated with different perceived pitches as well as their relative salience (Fig. 4C). Salience magnitudes at *F*0s corresponding to both the higher and lower note were taken as an estimate of neural pitch salience for each tone in a given dyad (Fig. 4C, arrows). When considering a range of dyads, this procedure allows us to trace the relative strengths between individual tone representations at the level of AN and assess how such representations are modulated dependent upon the relationship between simultaneously sounding musical pitches.

3.3.3. Temporal-representation modeling results and discussion

AN neural pitch salience is shown for individual tones within dyadic intervals in low, medium, and high registers (Fig. 5, left panels). Generally speaking, we observe consistent patterns of local variation in salience functions. Notably, the salience of the lower tone peaks when the two pitches achieve a harmonic relationship (e.g., octave, fifth), intervals which maximize the perceived consonance of the musical sonority. These findings are consistent with previous results demonstrating a role of pitch salience and "neural harmonicity" in the perceived consonance (i.e., pleasantness) of musical dyads (McDermott et al., 2010; Bidelman and Heinz, 2011). This increased pitch salience for the lower tone at more consonant intervals is achieved because in these cases, some harmonics are shared between the lower and higher tones. Consequently, there is an absence of suppression and, rather, reinforcement, which acts to increase the salience of the overall pitch representation. This result is directly related to the work of DeWitt and Crowder (1987) who showed that two tones are more likely to fuse and be perceived as a single tone when they stand in a consonant relation. Here, we demonstrate that these perceptual effects occur as a result of characteristics of peripheral and AN firing properties. These findings corroborate our recent work demonstrating increased salience/fusion in neural responses for consonant, relative to dissonant pitch relationships (Bidelman and Heinz, 2011).

Comparing AN salience across both tones shows a systematic bias; higher pitches are consistently more robust than their lower tone counterpart across nearly all interval pairs tested. Computing the ratio between higher and lower tone salience provides a visualization of the relative strength between tones in each musical interval where values greater than unity reflect a higher tone dominance (Fig. 5, right panels). Consistent with single tone patterns (Fig. 5) and human behavior (Palmer and Holleran, 1994; Crawley et al., 2002), higher tone superiority (i.e., ratio >1) is observed across the range of intervals tested (C2–C6: 65–1046 Hz) but is generally stronger in lower relative to higher registers (cf. top vs. bottom panels). Indeed, in the highest register, reinforcement of the pitch salience of the lower tone at consonant (octave, perfect fifth) intervals can actually result in greater pitch salience of the lower tone at these intervals (Fig.5, bottom panels) (see also, Bidelman and Heinz, 2011). The increased higher tone dominance in lower registers suggests that neural representations, and hence the resulting musical percept, might be more distinct when the soprano melody voice is supported by a low, well-grounded bass. Indeed, compositional practice in the Western tradition supports this notion. The register in which the melody voice is carried is

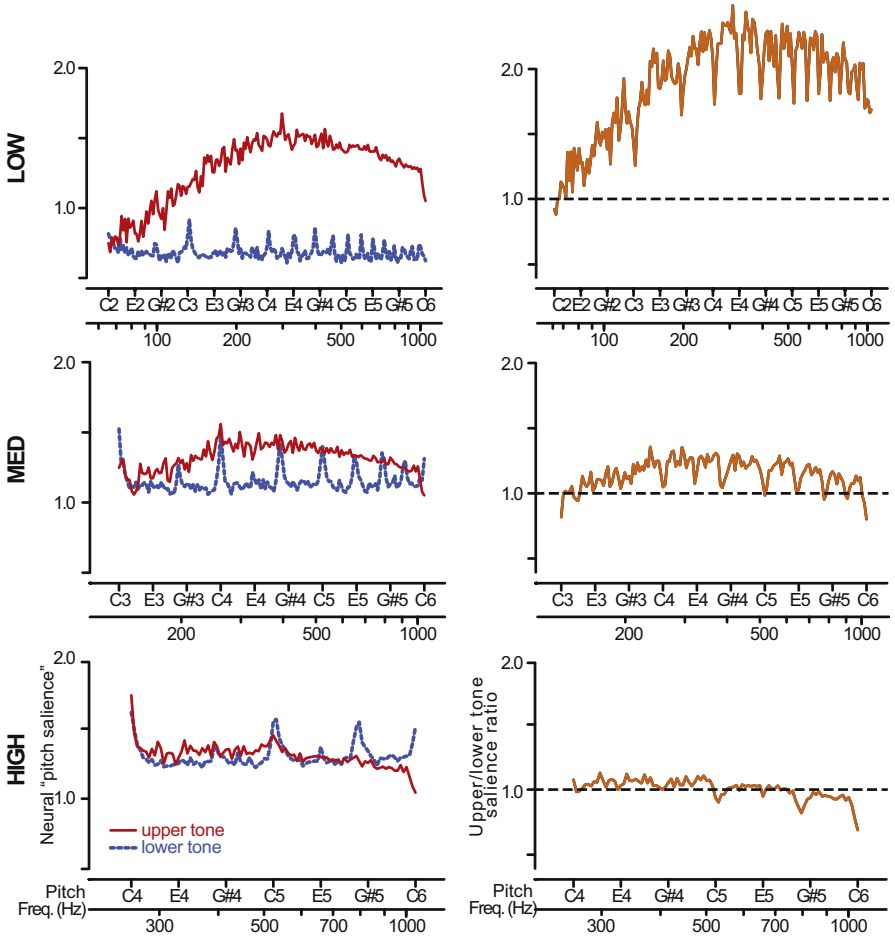

Fig. 5. AN neural pitch salience predicts higher tone superiority for musical dyads (i.e., intervals composed by two simultaneously presented notes). The lower *F*0 of the two tones was fixed at C2 for the upper panels, C3 for the middle panels and C4 for the lower panels, while the higher tone was allowed to vary. AN neural pitch salience is shown as a function of the spacing (¼ semitone steps) between the *F*0s of the lower and higher tones for low (C2–C6: 65–1046 Hz), middle (C3–C6: 130–1046 Hz), and high (C4–C6: 261–1046 Hz) registers of the piano (left panels). As indicated by the positive ratio of higher to lower tone salience (i.e., >1; dotted line), the representation of each pitch at the level of AN shows a systematic bias toward the higher tone, mimicking the perceptual higher voice superiority reported behaviorally (right panels). Two additional effects can be seen. The high voice superiority effect diminishes with increasing register and the pitch salience of the lower tone increases when the two tones form a consonant interval (e.g., octave [C in the higher tone], perfect fifth [G], perfect fourth [F], major third [E]).

usually selected so as to maximize the separation between the low bass and melody (soprano) while maintaining the salience of the melodic line (Aldwell and Schachter, 2003). Alternatively, the decrease in high voice dominance with increasing register may reflect the fact that musical pitch percepts are both weak and more ambiguous at higher frequencies (Moore, 1973; Semal and Demany, 1990). A weakening in pitch percept would ultimately tend to reduce the internal contrast between multiple auditory streams thereby normalizing the salience between simultaneous sounding pitches (e.g., Fig. 5, lower right panel).

If these simulations are repeated with pure tones, instead of the realistic harmonic complexes (as in Fig. 5), then the high voice superiority is lost for the middle and high registers (see Supplemental Fig. S2). In fact for a middle register, low frequency pure tones actually have higher predicted salience than high frequency pure tones. This result is consistent with the upward spread of masking and asymmetry of two-tone suppression for pure tones (Egan and Hake, 1950; Delgutte, 1990a,b). That high-voice superiority is seen in AN responses to harmonic complexes rather than pure-tones (compare Fig. 5 and S2) suggests that suppression plays an important role in establishing this effect for realistic musical sounds. However, we note that the temporal-pitch model does predict a high-voice superiority for pure tones in the lowest register. Future investigations are warranted to determine if this effect is caused by differences in the behavior of two-tone suppression at very low CFs or by the structure of the temporal-pitch model itself.

To further examine the potential influence of neurophysiological peripheral coding on more ecologically valid musical stimuli, we examined AN pitch salience profiles generated in response to a prototypical chorale from the repertoire of J.S. Bach. The Bach Chorales are largely regarded as definitive exemplars of the polyphonic music style and as such, offer the opportunity to extend our analysis to more realistic examples of music listening. The opening measures of the chorale "*Christ lag in Todes Banden*" are shown in Fig. 6. The soprano and bass voices were first isolated by extracting them from the four-part texture. A MIDI version of the two-voice arrangement was then used as a controller for a sampler built into Finale 2008 (MakeMusic, Inc.), a professional grade music notation program, to output an audio file of the excerpt played by realistic piano instrumental samples (Garritan Instruments). The audio clip was then passed to the AN model as the input stimulus waveform. Neural pitch salience profiles were then computed individually for each voice based on the aggregate output of the AN response on every quarter note beat of the chorale. Tracing individual note salience over time provides a running neurometric profile of the relative strengths of both voices in the Bach piece as represented in AN.

As shown in Fig. 6B, neural pitch salience derived from AN responses reveals a higher tone superiority for the Bach excerpt extending the results we observed for simple synthetic two-tone intervals (Fig. 5) to more realistic instrumental timbres and composition. Maximal high tone superiority was observed with the soprano and bass voice farthest apart (Fig. 6C). In addition, the magnitude of the higher tone superiority covaried well with the semitone distance between voices (Pearson's $r = 0.85$, $p < 0.001$). These results suggest that while the neurophysiological representation of the higher tone is often more salient than that of the lower tone in realistic musical textures, higher voice superiority also depends on the relative spacing between musical voices. Notably, we find that this effect is not simply monotonic. Rather, our simulations for both simple two-tone intervals (Fig. 5) and the Bach chorale (Fig. 6B) suggest, at least qualitatively, that the melody voice is most prominent against the bass (i.e., highest salience ratio contrast) when they are separated by ~2–2.5 octaves (24–30 semitones) (cf. peak in Fig. 5, upper left panel vs. Fig. 6B, beat #7); moving the voices closer or farther apart tends to decrease the neural salience contrast between higher and lower notes. It is interesting to note that the majority of writing in this and other Bach chorales tend to show soprano/bass voice spacing of about 2–2.5 octaves. We find that this compositional practice is closely paralleled in the neural pitch salience profiles extracted from AN responses.

The AN simulations presented here demonstrate peripheral correlates of the high-voice superiority effect at the level of AN. Interestingly, the effect does not seem to be driven by loudness *per se*, as the higher voice remains more salient even when the loudness between lower and higher tones is similar. Nevertheless, future work should examine the particular acoustic parameters which might contribute to the persistent dominance of the higher (soprano) voice in multi-voice music. A more comprehensive investigation of model responses could also be used to test and validate how changes in specific acoustic parameters such as sound intensity and spectral profile (timbre) manifest in human ERP responses, and how these neural correlates ultimately relate to the perceptual salience between auditory streams in music.

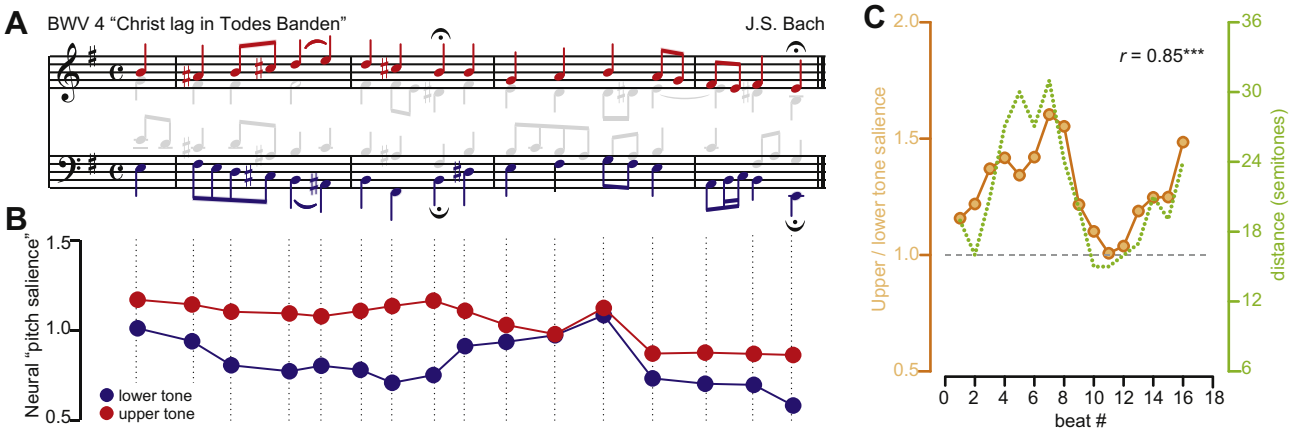

Fig. 6. AN neural pitch salience predicts higher voice superiority in natural Western music. (A) Opening measures of J.S. Bach's four-part chorale, "*Christ lag in Todes Banden (BWV 4)*". The soprano and bass voices are highlighted in red and blue, respectively. (B) Neural pitch salience derived from AN responses on each quarter note beat (demarcated by dotted lines) shows higher voice superiority across the excerpt. (C) Ratio of higher to lower tone neural pitch salience across the excerpt (solid lines) shows maximal higher voice superiority (i.e., ratio >1) with soprano-bass separation of ~2–2.5 octaves (24–30 semitones). The magnitude of the higher voice superiority covaries with the semitone distance between voices (dotted lines). ***$p < 0.001$.

4. Conclusions

Behavioral and event-related potential studies indicate that the higher of two simultaneously sounding tones and the highest melody line in polyphonic music are better encoded in sensory memory. Furthermore, this effect is present in young infants and, although modifiable by extensive experience attending to the lowest voice in music, is difficult if not impossible to reverse, suggesting a peripheral origin. Our modeling work supports this idea, and suggests that middle-ear filtering and cochlear nonlinearities tend to result in suppression of the harmonics of the lower of two simultaneously presented tones, giving rise to greater pitch salience for the higher compared to lower tone. Furthermore, the effect is greater in lower than higher pitch registers, with maximal suppression occurring for intervals of around two octaves. The ubiquitous placement of melody in the higher voice in music and the choice of spacing between voices likely results from the desire to maximize the perceptual salience of a musical motif against a background of accompanying pitches. Tracing neural responses to musical pitch at the earliest stages of the auditory pathway, we found that perceptual salience is optimized when underlying neural representations of single musical voices are maximally contrastive. While speculative, our results imply that the choice of register and intervallic spacing between the voices in polyphonic compositional practice are rooted in physiological constraints and response properties found within the peripheral auditory system.

Acknowledgments

This research was supported by grants from the Canadian Institutes of Health Research (CHIR) to LJT and from the Natural Sciences and Engineering Research Council of Canada (NSERC) to LJT and ICB. CM was supported by a postdoctoral fellowship from the NSERC CREATE grant in Auditory Cognitive Neuroscience.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.heares.2013.07.014.

References

Alain, C., Schuler, B.M., 2002. Neural activity associated with distinguishing concurrent auditory objects. J. Acoust. Soc. Am. 111, 990–995.
Alain, C., Winkler, I., 2012. Recording event-related brain potentials: applications to auditory perception. In: The Human Auditory Cortex, Springer Handbook of Auditory Research, vol. 43. Springer, Rueil-Malmaison, pp. 69–96.
Aldwell, E., Schachter, C., 2003. Harmony & Voice Leading. Thomson/Schirmer, United States.
Bidelman, G.M., Krishnan, A., 2009. Neural correlates of consonance, dissonance, and the hierarchy of musical pitch in the human brainstem. J. Neurosci. 29, 13165–13171.
Bidelman, G.M., Heinz, M.G., 2011. Auditory-nerve responses predict pitch attributes related to musical consonance-dissonance for normal and impaired hearing. J. Acoust. Soc. Am. 130, 1488–1502.
Bregman, A.S., 1990. Auditory Scene Analysis: the Perceptual Organization of Sounds. The MIT Press, Cambridge, Massachusetts.
Bregman, A.S., Campbell, J., 1971. Primary auditory stream segregation and perception of order in rapid sequences of tones. J. Exp. Psychol. 89, 244–249.
Bruce, I.C., Sachs, M.B., Young, E.D., 2003. An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. J. Acoust. Soc. Am. 113, 369–388.
Cariani, P.A., Delgutte, B., 1996a. Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch. J. Neurophysiol. 76, 1717–1734.
Cariani, P.A., Delgutte, B., 1996b. Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. J. Neurophysiol. 76, 1698–1716.
Cedolin, L., Delgutte, B., 2005. Pitch of complex tones: rate-place and interspike interval representations in the auditory nerve. J. Neurophysiol. 94, 347–362.
Cohen, M.A., Grossberg, S., Wyse, L.L., 1995. A spectral network model of pitch perception. J. Acoust. Soc. Am. 98, 862–879.
Crawley, E.J., Acker-Mills, B.E., Pastore, R.E., Weil, S., 2002. Change detection in multi-voice music: the role of musical structure, musical training, and task demands. J. Exp. Psychol. Hum. Percept. Perform. 28, 367–378.
Dai, H., Green, D.M., 1992. Auditory intensity perception: successive versus simultaneous, across-channel discriminations. J. Am. Stat. Assoc. 91, 2845–2854.
Darwin, C.J., Carlyon, R.P., 1995. Auditory grouping. In: Moore, B.C.J. (Ed.), Hearing: the Handbook of Perception and Cognition, vol. 6. Academic, London, pp. 387–424.
Deike, S., Gaschler-Markefski, B., Brechmann, A., Scheich, H., 2004. Auditory stream segregation relying on timbre involves left auditory cortex. Neuroreport 15, 1511–1514.
Delgutte, B., 1990a. Physiological mechanisms of psychophysical masking: observations from auditory-nerve fibers. J. Acoust. Soc. Am. 87, 791–809.
Delgutte, B., 1990b. Two-tone rate suppression in auditory-nerve fibers: dependence on suppressor frequency and level. Hear. Res. 49, 225–246.
Demany, L., 1982. Auditory stream segregation in infancy. Infant Behav. Dev. 5, 261–276.
DeWitt, L.A., Crowder, R.G., 1987. Tonal fusion of consonant musical intervals: the oomph in Stumpf. Percept. Psychophys. 41, 73–84.
Dowling, W.J., 1973. The perception of interleaved melodies. Cognitive Psychology 5, 322–337.
Egan, J.P., Hake, H.W., 1950. On the masking pattern of a simple auditory stimulus. J. Acoust. Soc. Am. 1950, 622–630.
Fassbender, C., 1993. Auditory Grouping and Segregation Processes in Infancy. Doctoral dissertation. Kaste Verlag, Norderstedt, Germany.
Fay, R.R., 1998. Auditory stream segregation in goldfish (*Carassius auratus*). Hear. Res. 120, 69–76.
Fay, R.R., 2000. Spectral contrasts underlying auditory stream segregation in goldfish (*Carassius auratus*). J. Assoc. Res. Otolaryngol. 1, 120–128.
Fishman, Y.I., Reser, D.H., Arezzo, J.C., Steinschneider, M., 2001. Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hear. Res. 151, 167–187.
Folland, N.A., Butler, B.E., Smith, N.A., Trainor, L.J., 2012. Processing simultaneous auditory objects: infants' ability to detect mistunings in harmonic complexes. J. Acoust. Soc. Am. 131, 993–997.
Fujioka, T., Trainor, L.J., Ross, B., 2008. Simultaneous pitches are encoded separately in auditory cortex: an MMNm study. Neuroreport 19, 361–366.
Fujioka, T., Trainor, L.J., Ross, B., Kakigi, R., Pantev, C., 2005. Automatic encoding of polyphonic melodies in musicians and nonmusicians. J. Cogn. Neurosci. 17, 1578–1592.
Glasberg, B.R., Moore, B.C.J., 2002. A model of loudness applicable to time-varying sounds. J. Audio Eng. Soc. 50, 331–342.
Goldstein, J.L., 1973. An optimum processor theory for the central formation of the pitch of complex tones. J. Acoust. Soc. Am. 54, 1496–1516.
Griffiths, T.D., Warren, J.D., 2004. What is an auditory object? Nat. Rev. Neurosci. 5, 887–892.
Gutschalk, A., Micheyl, C., Melcher, J.R., Rupp, A., Scherg, M., Oxenham, A.J., 2005. Neuromagnetic correlates of streaming in human auditory cortex. J. Neurosci. 25, 5382–5388.
Hartmann, W.M., Johnson, D., 1991. Stream segregation and peripheral channeling. Music Percept. 9, 155–184.
Houtsma, A., Smurzynski, J., 1990. Pitch identification and discrimination for complex tones with many harmonics. J. Acoust. Soc. Am. 87, 304–310.
Hulse, S.H., 2002. Auditory scene analysis in animal communication. In: Slater, P.J.B., Rosenblatt, J.S., Snowdon, C.T., Roper, T.J. (Eds.), Advances in the Study of Behavior, vol. 31. Academic Press, San Diego, pp. 163–200.
Hulse, S.H., Bernard, D.J., Braaten, R.F., 1995. Auditory discrimination of chord-based spectral structures by European starlings (*Sturnus vulgaris*). J. Exp. Psychol. Gen. 124, 409–423.
Hulse, S.H., MacDougall-Shackleton, S.A., Wisniewski, A.B., 1997. Auditory scene analysis by songbirds: stream segregation of birdsong by European starlings (*Sturnus vulgaris*). J. Comp. Psychol. 111, 3–13.
Huron, D., 2001. Tone and voice: a derivation of the rules of voicel-leading from perceptual principles. Music Percept. 19, 1–64.
Ibrahim, R.A., Bruce, I.C., 2010. Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues. In: Lopez-Poveda, E.A., Palmer, A.R., Meddis, R. (Eds.), The Neurophysiological Bases of Auditory Perception. Springer, New York, pp. 429–438.
Izumi, A., 2000. Japanese monkeys perceive sensory consonance of chords. J. Acoust. Soc. Am. 108, 3073–3078.
Izumi, A., 2002. Auditory stream segregation in Japanese monkeys. Cognition 82, B113–B122.
Joris, P.X., Bergevin, C., Kalluri, R., McLaughlin, M., Michelet, P., van der Heijden, M., Shera, C.A., 2011. Frequency selectivity in Old-World monkeys corroborates sharp cochlear tuning in humans. Proc. Natl. Acad. Sci. U. S. A. 108, 17516–17520.
Lee, K.M., Skoe, E., Kraus, N., Ashley, R., 2009. Selective subcortical enhancement of musical intervals in musicians. J. Neurosci. 29, 5832–5840.
Liberman, M.C., 1978. Auditory nerve response from cats raised in a low noise chamber. J. Acoust. Soc. Am. 63, 442–455.
Marie, C., Fujioka, T., Herrington, L., Trainor, L.J., 2012. The high-voice superiority effect in polyphonic music is influenced by experience: a comparison of musicians who play soprano-range compared to bass-range instruments. Psychomusicol. Music Mind Brain 22, 97–104.
Marie, C., Trainor, L., 2013. Development of simultaneous pitch encoding: infants show a high voice superiority effect. Cereb. Cortex 23, 660–669.

McAdams, S., Bertoncini, J., 1997. Organization and discrimination of repeating sound sequences by newborn infants. J. Acoust. Soc. Am. 102, 2945–2953.
McDermott, J.H., Hauser, M.D., 2005. The origins of music: innateness, uniqueness, and evolution. Music Percept. 23, 29–59.
McDermott, J.H., Lehr, A.J., Oxenham, A.J., 2010. Individual differences reveal the basis of consonance. Curr. Biol. 20, 1035–1041.
Micheyl, C., Carlyon, R.P., Gutschalk, A., Melcher, J.R., Oxenham, A.J., Rauschecker, J.P., Tian, B., Courtenay Wilson, E., 2007. The role of auditory cortex in the formation of auditory streams. Hear. Res. 229, 116–131.
Miller, R.L., Schilling, J.R., Franck, K.R., Young, E.D., 1997. Effects of acoustic trauma on the representation of the vowel /ɛ/ in cat auditory nerve fibers. J. Acoust. Soc. Am. 101, 3602–3616.
Moore, B.C., 1973. Frequency difference limens for short-duration tones. J. Acoust. Soc. Am. 54, 610–619.
Moss, C.F., Surlykke, A., 2001. Auditory scene analysis by echolocation in bats. J. Acoust. Soc. Am. 110, 2207–2226.
Näätänen, R., Paavilainen, P., Rinne, T., Alho, K., 2007. The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin. Neurophysiol. 118, 2544–2590.
Nager, W., Teder-Salejarvi, W., Kunze, S., Münte, T.F., 2003. Preattentive evaluation of multiple perceptual streams in human audition. Neuroreport 14, 871–874.
Palmer, C., Holleran, S., 1994. Harmonic, melodic, and frequency height influences in the perception of multivoiced music. Percept. Psychophys. 56, 301–312.
Patel, A.D., Iversen, J.R., Bregman, M.R., Schulz, I., 2009. Experimental evidence for synchronization to a musical beat in a nonhuman animal. Curr. Biol. 19, 827–830.
Pascal, J., Bourgeade, A., Lagier, M., Legros, C., 1998. Linear and nonlinear model of the human middle ear. J. Acoust. Soc. Am. 104, 1509–1516.
Picton, T.W., Alain, C., Otten, L., Ritter, W., Achim, A., 2000. Mismatch negativity: different water in the same river [Review]. Audiol. Neurootol. 5, 111–139.
Plomp, R., Levelt, W.J., 1965. Tonal consonance and critical bandwidth. J. Acoust. Soc. Am. 38, 548–560.
Rose, M.M., Moore, B.C.J., 2000. Effects of frequency and level on auditory stream segregation. J. Acoust. Soc. Am. 108, 1209–1214.
Schachner, A., Brady, T.F., Pepperberg, I.M., Hauser, M.D., 2009. Spontaneous motor entrainment to music in vocal mimicking animals. Curr. Biol. 19, 831–836.
Semal, C., Demany, L., 1990. The upper limit of "musical" pitch. Music Percept. 8, 165–176.
Shera, C.A., Guinan Jr., J.J., Oxenham, A.J., 2002. Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. Proc. Natl. Acad. Sci. U. S. A. 99, 3318–3323.
Shinozaki, N., Yabe, H., Sato, Y., Sutoh, T., Hiruma, T., Nashida, T., Kaneko, S., 2000. Mismatch negativity (MMN) reveals sound grouping in the human brain. Neuroreport 11, 1597–1601.
Sloboda, J., Edworthy, J., 1981. Attending to two melodies at once: The effect of key relatedness. Psychology of Music 9, 39–43.
Smith, N.A., Trainor, L.J., 2011. Auditory stream segregation improves infants' selective attention to target tones amid distracters. Infancy 16, 655–668.
Snyder, J.S., Alain, C., 2007. Toward a neurophysiological theory of auditory stream segregation. Psychol. Bull. 133, 780–799.
Snyder, J.S., Alain, C., Picton, T.W., 2006. Effects of attention on neuroelectric correlates of auditory stream segregation. J. Cogn. Neurosci. 18, 1–13.
Sussman, E., 2005. Integration and segregation in auditory scene analysis. J. Acoust. Soc. Am. 117, 1285–1298.
Terhardt, E., Stoll, G., Seewann, M., 1982. Algorithm for the extraction of pitch and pitch salience from complex tonal signals. J. Acoust. Soc. Am. 71, 679–687.
Trainor, L.J., 2008. The neural roots of music. Nature 453, 598–599.
Trainor, L.J., 2012. Musical experience, plasticity, and maturation: issues in measuring developmental change using EEG and MEG. Ann. N. Y. Acad. Sci. 1252, 25–36.
Trainor, L.J., Corrigall, K.A., 2010. Music acquisition and effects of musical experience. In: Riess-Jones, M., Fay, R.R. (Eds.), Springer Handbook of Auditory Research: Music Perception. Springer, Heidelberg, pp. 89–128.
Trainor, L.J., Hannon, E.E., 2012. Musical development. In: Deutsch, D. (Ed.), The Psychology of Music, third ed. Academic Press, San Diego, pp. 423–498.
Trainor, L.J., Unrau, A.J., 2012. Development of pitch and music perception. In: Werner, L., Fay, R.R., Popper, A.N. (Eds.), Springer Handbook of Auditory Research: Human Auditory Development. Springer, New York, pp. 223–254.
van Noorden, 1977. Effects of Frequency Separation and Speed on Grouping are Discussed in "ASA-90", 48-73. For a description of the effects of grouping on perception see "ASA-90", pp. 131–172, pp. 17–21.
van Noorden, L.P.A.S., 1975. Temporal Coherence in the Perception of Tone Sequences. Unpublished doctoral dissertation. Eindhoven University of Technology, Eindhoven, Netherlands.
Wilson, E.C., Melcher, J.R., Micheyl, C., Gutschalk, A., Oxenham, A.J., 2007. Cortical fMRI activation to sequences of tones alternating in frequency: relationship to perceived rate and streaming. J. Neurophysiol. 97, 2230–2238.
Winkler, I., Kushnerenko, E., Horváth, J., Ceponiene, R., Fellman, V., Huotilainen, M., Näätänen, R., Sussman, E., 2003. Newborn infants can organize the auditory world. Proc. Natl. Acad. Sci. U. S. A. 100, 11812–11815.
Winkler, I., Denham, S.L., Nelken, I., 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn. Sci. 13, 532–540.
Winkler, I., Takegata, R., Sussman, E., 2005. Event-related brain potentials reveal multiple stages in the perceptual organization of sound. Cogn. Brain Res. 25, 291–299.
Yabe, H., Winkler, I., Czigler, I., Koyama, S., Kakigi, R., Sutoh, T., Tomiharu, H., Kaneko, S., 2001. Organizing sound sequences in the human brain: the interplay of auditory streaming and temporal integration. Brain Res. 897, 222–227.
Young, E.D., Barta, P.E., 1986. Rate responses of auditory nerve fibers to tones in noise near masked threshold. J. Acoust. Soc. Am. 79, 426–442.
Zenatti, A., 1969. Le développement génétique de la perception musicale. Monogr. Francaises Psychol. 17. CNRS, Paris.
Zhang, X., Heinz, M.G., Bruce, I.C., Carney, L.H., 2001. A phenomenological model for the responses of auditory-nerve fibers. I. Nonlinear tuning with compression and suppression. J. Acoust. Soc. Am. 109, 648–670.
Zilany, M.S., Bruce, I.C., 2006. Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. J. Acoust. Soc. Am. 120, 1446–1466.
Zilany, M.S., Bruce, I.C., 2007. Representation of the vowel /ɛ/ in normal and impaired auditory nerve fibers: model predictions of responses in cats. J. Acoust. Soc. Am. 122, 402–417.
Zilany, M.S., Bruce, I.C., Nelson, P.C., Carney, L.H., 2009. A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. J. Acoust. Soc. Am. 126, 2390–2412.

Supplemental Material

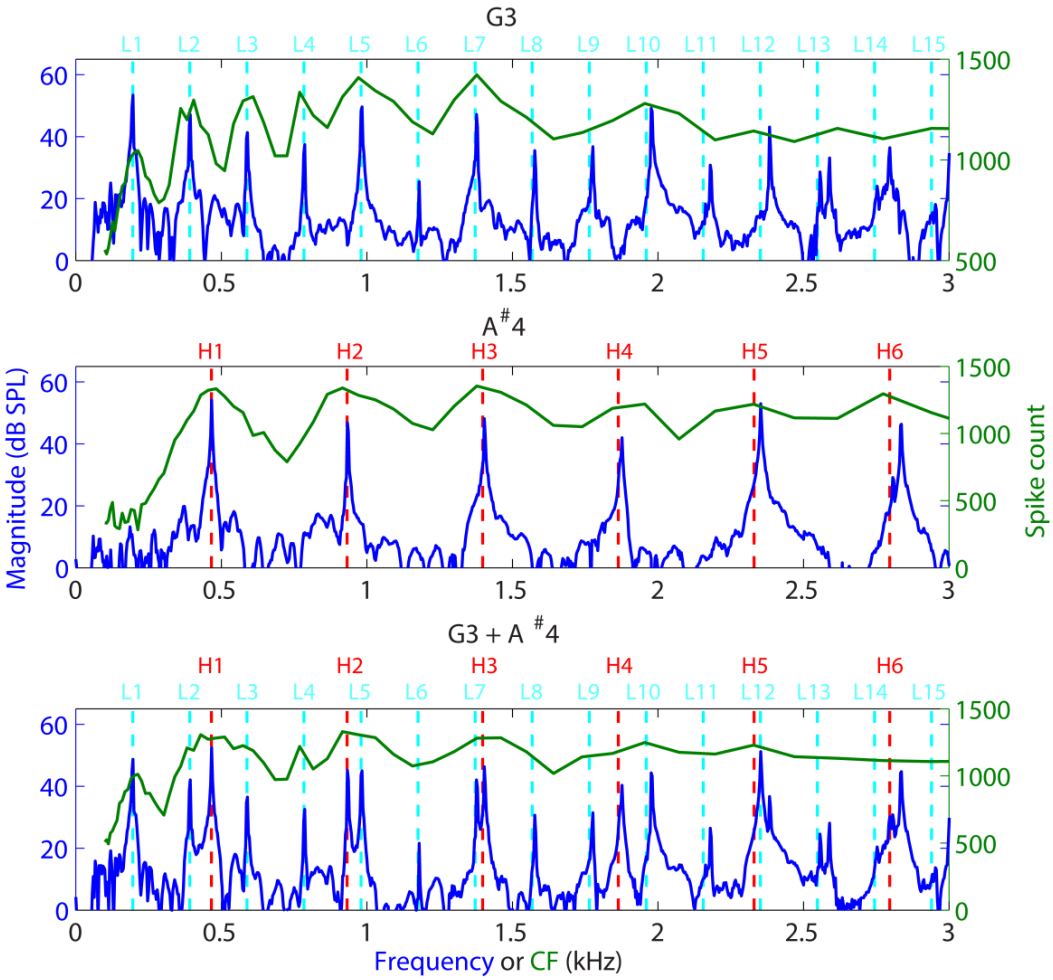

Figure S1: These plots replicate the model simulations and analysis shown in Fig. 3 with a version of the auditory-periphery model that has i) no middle-ear filter and ii) fixed basilar-membrane filters, such that two-tone suppression is absent from the model. In the top panel, it can be observed in the neural representation of the low tone (G3) that if two-tone suppression is removed then the responses to some of the lower-intensities harmonics (particularly the 4th harmonic – labeled L4) are increased, such that these individual harmonics are better resolved. For the high tone (A#4) representation shown in the middle panel, the lack of two-tone suppression primarily broadens the peaks of the responses to the harmonics; an increased representation of the 4th harmonic (labeled H4) is also seen in this case. In the bottom panel where the response to the combined tones (G3 + A#4) is shown, it can be observed that removing middle-ear filtering and two-tone suppression increases the responses to the first five harmonics of the low tone (labeled L1 – L5). Subsequently, when middle-ear filtering and two-tone suppression is absent, the response to the combined tones is closer to that of the low tone rather than the high tone, in contrast to the high-voice superiority that is demonstrated in Fig. 3 (where middle-ear filtering and two-tone suppression is present).

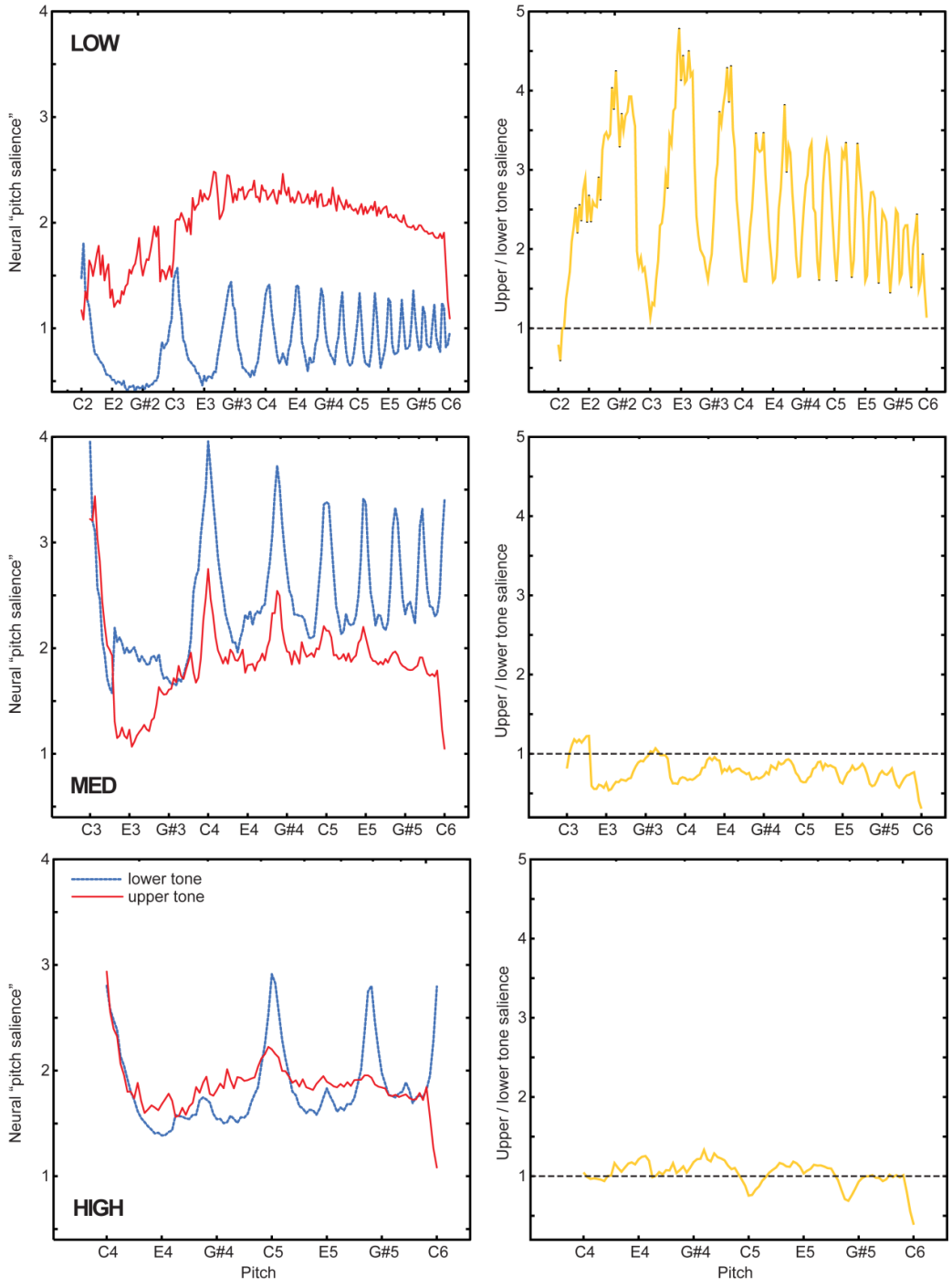

Figure S2: AN neural pitch salience (as in Fig. 5) but for loudness equated pure tones. Pure tone stimuli were equated in loudness (70 phons) according to the ISO 266 standard using the model described by Glasberg and Moore (2002). This essentially removes any frequency-dependent loudness effects due to middle ear filtering. Aside from the condition with unusually low bass (top row), loudness matched pure tones elicit much weaker or no high-voice superiority effect. In medium and high registers, the salience of the lower voice tends to dominate that of the upper. A lack of high-voice superiority for pure tones relative to complex tones can be attributed to the fact that for pure tones presented at these sound pressure levels, two-tone suppression tends to favor the lower-frequency tone (e.g., Delgutte, 1990b). This result is in contrast to the more

complicated patterns of suppression generated by the harmonics of realistic piano tones (see Fig. 5).