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

A R T I C L E I N F O

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

A B S T R A C T

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 ($F0$) are easily integrated and are heard as coming from a single auditory object whereas tones differing in $F0$ 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 $F0$ and integer multiples of that frequency (harmonics or overtones). Thus, a tone with an $F0$ 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