International Journal of Psychophysiology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Psychophysiology

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

Auditory brainstem's sensitivity to human voices

Yun Nan [a,*], Erika Skoe [b,c,f], Trent Nicol [b,c], Nina Kraus [b,c,d,e]

[a] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, 100875, China
[b] Auditory Neuroscience Laboratory, Australia
[c] Department of Communication Sciences, Northwestern University, Evanston, IL 60208, United States
[d] Department of Neurobiology and Physiology, Northwestern University, Evanston, IL 60208, United States
[e] Department of Otolaryngology, Northwestern University, Evanston, IL 60208, United States
[f] Department of Speech, Language and Hearing Sciences and Psychology, University of Connecticut, Storrs, CT 06209, United States

ARTICLE INFO

Article history:
Received 30 July 2014
Received in revised form 17 December 2014
Accepted 21 December 2014
Available online xxxx

Keywords:
Voice
Auditory brainstem
Frequency following response

ABSTRACT

Differentiating between voices is a basic social skill humans acquire early in life. The current study aimed to understand the subcortical mechanisms of voice processing by focusing on the two most important acoustical voice features: the fundamental frequency (F0) and harmonics. We measured frequency following responses in a group of young adults to a naturally produced speech syllable under two linguistic contexts: same-syllable and multiple-syllable. Compared to the same-syllable context, the multiple-syllable context contained more speech cues to aid voice processing. We analyzed the magnitude of the response to the F0 and harmonics between same-talker and multiple-talker conditions within each linguistic context. Results establish that the human auditory brainstem is sensitive to different talkers as shown by enhanced harmonic responses under the multiple-talker compared to the same-talker condition, when the stimulus stream contained multiple syllables. This study thus provides the first electrophysiological evidence of the auditory brainstem's sensitivity to human voices.

1. Introduction

Recognizing conspecific voices is a critical survival skill for many animal species, such as fur seals and macaques (Rendall et al., 1998; Insley, 2000; Petkov et al., 2008; Sliwa et al., 2011). For humans, voice not only constitutes the primary auditory identity of an individual, but also serves as a vehicle for speech.

Acoustically, voice is represented primarily by the fundamental frequency (F0) and the formant patterns (for a review, see Belin, 2006). Vocal features are constrained by the physical construct of an individual's vocal apparatus, which includes a source (the vocal folds in the larynx) and a filter (the vocal tract above the larynx) (Ghazanfar and Rendall, 2008; Latinus and Belin, 2011). The vocal F0 normally varies as a function of the size of an individual's vocal folds, whereas the formant pattern is determined by both the physical size and the dynamic configuration of an individual's vocal tract during articulation (Latinus and Belin, 2011).

Given its important role in social interaction, there has been a growing interest in exploring the neural mechanisms underlying human voice perception. Brain imaging data has shown that voice-specific brain regions are mostly localized in the superior temporal cortices (Belin et al., 2000, 2002) and emerge around 4 to 7 months after birth (Grossmann et al., 2010). However, where and how the primary acoustic voice features, including the F0 and the formant patterns (for a review, see Belin, 2006) are represented in the brain is still unclear. The brainstem frequency following response (FFR) offers a window into the brain's encoding of these two important voice features. The FFR originates from the inferior colliculus (Smith et al., 1975), reflecting the encoding of periodic information in auditory stimuli with high fidelity (Skoe and Kraus, 2010; Musacchia et al., 2007; Krizman et al., 2012; Krishnan et al., 2005).

The current study aims to investigate subcortical encoding of human voices using the FFR. We measured the FFR in a group of young adults by presenting the same acoustic token ([da] spoken by a male voice) under same-talker and multiple-talker conditions. We predicted that this target stimulus would elicit greater FFRs under a multiple-talker relative to a same-talker condition, owing to the neuronal facilitation effect reported in a previous study (Belin and Zatorre, 2003) in which heightened activation was found in the right anterior temporal lobe for a multiple-talker condition compared to a same-talker condition.

Additionally, linguistic context also affects voice perception. Compared to an unfamiliar language, a voice presented in a familiar language

* Corresponding author at: State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, 19 Xin-Wai St., Hai-Dian District, Beijing 100875, China. Tel./fax: +86 10 58802742.
E-mail address: dr.yunnan@gmail.com (Y. Nan).
URL's: http://www.brainvolts.northwestern.edu (E. Skoe), http://www.brainvolts.northwestern.edu (T. Nicol), http://www.brainvolts.northwestern.edu (N. Kraus).

http://dx.doi.org/10.1016/j.ijpsycho.2014.12.013
0167-8760/

is easier to recognize, due to the convergence of prosodic and phonetic cues in a familiar linguistic context (Goggin et al., 1991). In the current study, there were two linguistic contexts: in one the target stimulus [da] was presented within a stream of other [da] tokens (hereafter "same-syllable context"), and in the other the target stimulus was presented within a stream of other syllables (hereafter "multiple-syllable context"). More speech cues are available in the multiple-syllable context, which we predict would result in facilitated voice processing. Therefore, as compared to the same-syllable context, the multiple-syllable context was expected to show a larger talker effect (enhanced FFR responses to the same [da] in the multiple-talker relative to the same-talker condition).

2. Material and methods

2.1. Stimuli

Four native male speakers of American English were asked to produce [da], [ba], [ta] and [ga] with a steady fundamental frequency (F0). Recordings took place in a sound attenuated chamber using a Marantz digital audio recorder at a sampling rate of 44.1 kHz. In total, 10 syllables were employed in this study: four produced by talker 1 ([da1], [ta1], [ba1], and [ga1]), and two by each of the other three talkers ([da2], [ta2], [da3], [ba3], [da4], and [ga4]). These recordings were then duration-normalized to 170 ms using Praat software (Boersma, 2001). Using Praat, a level pitch contour was superimposed onto all the duration-normalized syllables without changing the original individual mean F0. Thus all 10 syllables had a level pitch contour, although the exact F0 differed (mean: 113 Hz; range: 105–119 Hz). All stimuli were then RMS normalized using Level 16 software (Tice and Carrell, 1998) to 70 dB. As a result, the target stimulus [da1] spoken by talker 1 was 170 ms long with a level fundamental frequency (F0: 118 Hz), a 15 ms voice-onset time, and four dynamic formants (F1: 460–720 Hz, F2: 1670–1240 Hz, F3: 2655–2520 Hz, F4: 2970–3910 Hz) over the duration of the stimulus. Further acoustic analysis showed that the target stimulus [da1] differed from the other speech sounds on several talker and/or phonetic features such as voice-onset time (/ta/), formant trajectory (/ba/ and /ga/) and F0.

2.2. Participants

Twelve young adults (9 females) with ages ranging from 18 to 23 years (mean, 20.4 ± 1.7 years) from Northwestern University participated in this study. Participants had no more than 3 years (mean: 0.5 years) of musical training and were not currently playing any instrument. All participants were right-handed, and reported no audiologic or neurologic deficits. Their self-reported normal hearing was confirmed with binaural audiometric thresholds at or below 20 dB HL for octaves from 250 to 8000 Hz, and normal ABRs to a click (Starr et al., 1996; for a review, see Stapells, 2000). Informed written consent was obtained from all participants. This research was approved by the Institutional Review Board of Northwestern University.

2.3. Procedure

Participants watched a silent captioned movie during the whole recording session and were instructed to remain wakeful but still (Skoe and Kraus, 2010). Stimuli were presented binaurally in alternating polarities at 70 dB sound pressure level (SPL) with an inter-stimulus interval of 87.14 ms (Neuroscan Stim 2; Compumedics) via insert earphones (ER-3, Etymotic Research, Elk Grove Village, IL, USA).

Auditory brainstem responses were collected from the scalp (Cz) using Scan 4.3 (Compumedics, Charlotte, NC) with Ag-AgCl electrodes in a vertical, ipsilateral montage, with contact impedance below 2 kΩ for all electrodes. Four different conditions were collected and the order of conditions was counterbalanced across participants. These four conditions represented a 2 talker (same vs. multiple) by 2 linguistic context (same-syllable vs. multiple-syllable) factorial design (Fig. 1). For the same-talker, same-syllable condition, 6000 sweeps of [da1] were presented. In the multiple-talker, same-syllable condition, 1500 sweeps of [da1] were presented randomly in the context of [da]s ([da2], [da3], [da4]) produced by the other three speakers. For the same-talker, multiple-syllable condition, 1500 sweeps of [da1] were presented randomly in the context of other syllables ([ga1], [ta1], [ba1]) produced by talker 1. In the multiple-talker, multiple-syllable condition, 1500 sweeps of [da1] were presented randomly among other syllables produced by the other three speakers ([ta2], [ba3], [ga4]). Across the four conditions, the target stimulus [da1] was trial-matched, such that it occurred at the same point in time relative to the start of the condition. Each condition lasted between 24 and 28 min. Participants were allowed to take short breaks between conditions.

Using Neuroscan Edit, brainstem responses were processed offline by bandpass filtering from 70 to 2000 Hz (12 dB roll-off, zero phase-shift), epoching from −40 to 190 ms (stimulus onset occurring at 0 ms), and baseline correcting according to the pre-stimulus period. Sweeps with amplitude greater than ±35 μV were rejected. The final average responses were based on the same number of trials across the four conditions (700). The filtering parameters as well as the fast stimulus presentation rate minimized the influence of cortical activity in the final waveforms (Chandrasekaran and Kraus, 2010). We compared the response to [da1] across the four conditions.

2.4. Behavioral validation of the stimuli

It should be noted that the main study measuring subcortical responses to voices was conducted in Northwestern University (U.S.). To validate that the speech syllables used in the current study were ecologically plausible, such that the participants were able to differentiate talker 1 from the other talkers simply based upon these stimuli, we subsequently conducted a complementary behavioral test at Beijing Normal University (approved by the Institutional Review Board of Beijing Normal University, China). For this follow-up study, another group of young adults ($n = 15$, 6 males; ages 19 through 26, mean 22.7 ± 2.1 years) were recruited. They were all Mandarin-speaking students from Universities in Beijing. Informed written consent was obtained from all participants.

The participants were asked to listen to a list of syllables and to indicate for each single syllable whether talker 1 or a different talker produced it. There were two blocks, each containing 90 syllables. This validation study used the same set of stimuli as the main study; however, the stimuli were presented differently. The first block represented the same-syllable condition, including 60 [da1]s, 10 [da2]s, 10 [da3]s, and 10 [da4]s. The second block represented the multiple-syllable condition, comprising 20 [ba1]s, 20 [ga1]s, 20 [ta1]s, 10 [ta2]s, 10 [ba3]s, and 10 [ga4]s. The order of the syllables was randomized within each block. At the beginning of each block, participants were first trained to recognize syllables produced by talker 1 (block one: [da1]; block two: [ba1], [ga1], and [ta1]), then they were required to press a button each time talker 1 was presented or press another button when it was not talker 1. The response buttons were counterbalanced across participants. There was a 100 ms fixation time before the onset of each syllable and the participants were told to respond as quickly as possible.

It should be noted that the second block, i.e. the multi-syllable condition, was different from the multi-syllable multi-talker condition in the main study. In the main study, the multi-syllable multi-talker condition contained only [da1], but not the other syllables produced by talker 1, whereas here the multi-syllable condition comprised all syllables produced by talker 1 except [da1]. The multi-syllable condition was organized differently in the behavioral study to avoid the possibility that the participants could easily differentiate [da1] from other syllables produced by other talkers ([ta2], [ba3], and [ga4]) based merely on the

Please cite this article as: Nan, Y., et al., Auditory brainstem's sensitivity to human voices, Int. J. Psychophysiol. (2015), http://dx.doi.org/10.1016/j.ijpsycho.2014.12.013