International Journal of Psychophysiology 95 (2015) 333–337

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,d], Trent Nicol [b], Nina Kraus [b,c]

[a] State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, 100875, China
[b] Auditory Neuroscience Laboratory, Department of Communication Sciences, Northwestern University, Evanston, IL, 60208, United States
[c] Departments of Neurobiology and Physiology and Department of Otolaryngology, Northwestern University, Evanston, IL, 60208, United States
[d] 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 22 January 2015

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 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

* 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: nany@bnu.edu.cn (Y. Nan).
URL's: http://skoe.slhs.uconn.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/

"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 protocol 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 phonological differences. In that case, it would not be necessary to access talker information to complete the task.

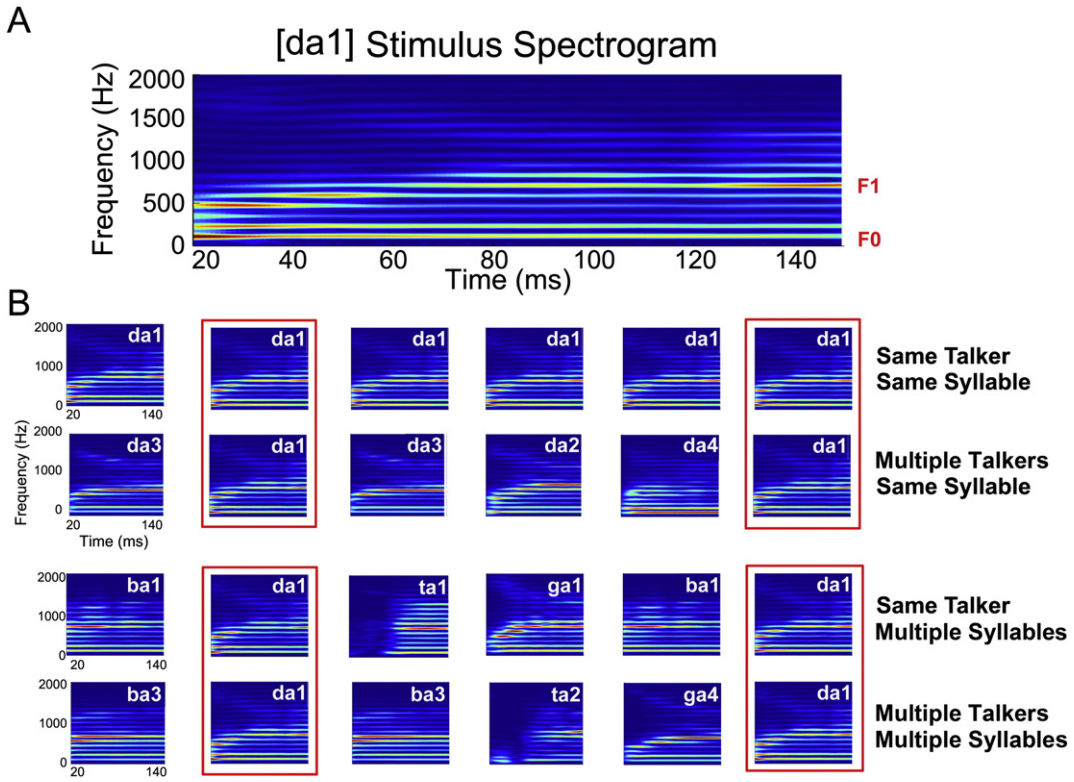

Fig. 1. Stimulus characteristics and experimental design. (A) The spectrogram of the target stimulus [da1]. The fundamental frequency (F0) (118 Hz) and the first formant (F1: 460–720 Hz) of the stimulus target are labeled in red. (B) The target stimulus [da1] was presented in the same-syllable (top two panels) and multiple-syllable (bottom two panels) contexts. The numerals indicate the talker; thus [da1] and [ga1] are spoken by the same talker. 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]). Within each condition, the syllables were either produced by the same talker or multiple talkers. This two (talker: same vs. multiple) by two (linguistic context: same syllable vs. multiple syllables) design resulted in four different conditions for the target [da1]. As shown in the spectrograms, stimuli in these four conditions differed from each other both temporally and spectrally. [da1] was event matched between the conditions (red boxes), such that it occurred within the same relative position across the stimulus sequence.

Results showed that these 15 participants performed well above chance, with performance for the same-syllable condition being 96.9% ± 2.3% (mean percentage of correct responses ± SD) and for the multi-syllable condition being 86.9% ± 6.9%. As expected, all the participants performed better in the same-syllable condition than the multi-syllable condition, $F_{(1,14)} = 30.4$, $p < 0.001$. The same-syllable condition also produced much shorter response times than the multi-syllable condition: $F_{(1,14)} = 65.2$, $p < 0.001$, same-syllable, 384.8 ± 79.1 ms, multi-syllable condition, 566.1 ± 100.4 ms. In sum, these behavioral results showed that the stimuli used in the main study are sufficient for the participants to successfully access talker information.

It is widely accepted that individuals with pitch expertise, either in the music or language domain, are able to encode pitch more effectively than their non-experienced peers (Bidelman et al., 2011a, 2011b; Wong et al., 2007). Here, one important question to ask is whether our Mandarin speakers, due to their extensive experiences with a tone language, were able to differentiate different voices based only on the F0, an acoustic cue contributing to pitch perception (Micheyl et al., 2010). Although all the voices used here were male and the F0 range was not wide, there were still F0 differences among all voices. In block one, the F0 of [da1] was 118 Hz, [da2] 110 Hz, [da3] 106 Hz, and [da4] 105 Hz. While the F0s of syllables [da3] and [da4] were quite close in frequency, compared to [da3], [da4] was miscategorized more often to talker 1 ($t_{(14)} = 2.43$, $p = 0.03$ (2-tailed): [da3], 98.7% ± 3.5% (mean percentage of correct responses ± SD); [da4], 94.0% ± 6.3%). Conversely, [da2] was much closer to [da1] than [da4] in terms of the F0 distance, but participants performed similarly well in distinguishing these talkers from talker 1 ($t_{(14)} = -0.68$, $p = 0.51$ (2-tailed): [da2], 92.0% ± 10.1%; [da4], 94.0% ± 6.3%). These analyses suggested that even in the same-syllable block, the difference in F0 between voices is not sufficient to predict the behavioral performance in voice perception. Nonetheless, we can rule out the possibility that F0 is the only factor contributing to voice perception for these Mandarin speakers.

2.5. Analysis of the subcortical responses

In the electrophysiological component of this study, analyses focused on the subcortical response to the sustained vowel portion of the stimulus [da1] (60–180 ms, see Fig. 2). This is because the vowel, relative to consonant, contributes more to voice perception (Owren and Cardillo, 2006), and thus is quite often the main focus in human voice research (for a review, see Latinus and Belin, 2011).

We measured the frequency following response (FFR, a component of the ABR (Moushegian et al., 1973)) across the four different conditions. To examine the response in the frequency domain, we performed a fast Fourier transform in the MATLAB programming environment (Mathworks, Inc.). The strength of spectral encoding was obtained by calculating the average spectral amplitudes within specific 10 Hz frequency bins surrounding F0 (118 Hz) and the subsequent four harmonics (H_2–H_5). The four harmonics were within the frequency range of the first formant of the target stimulus. In addition to F0, the first two harmonics (Kreiman and Gerratt, 2010) and the first formant (Latinus and Belin, 2011) are also very important features for voice perception. The average amplitudes of the H_2 to H_5 bins were summed as a composite score representing the overall strength of harmonic encoding (Parbery-Clark et al., 2009).

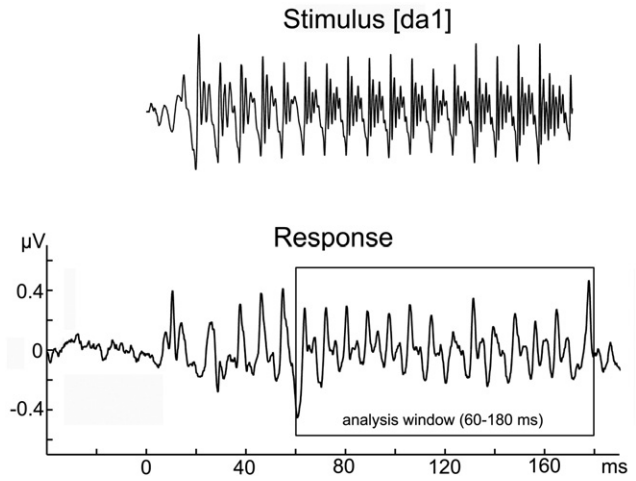

Fig. 2. Stimulus and the grand average brainstem responses. The stimulus [da1] and the grand average waveform across all four presentation conditions of [da1]. The analysis window was 60–180 ms, covering the FFR responses to the sustained vowel portion of the stimulus [da1].

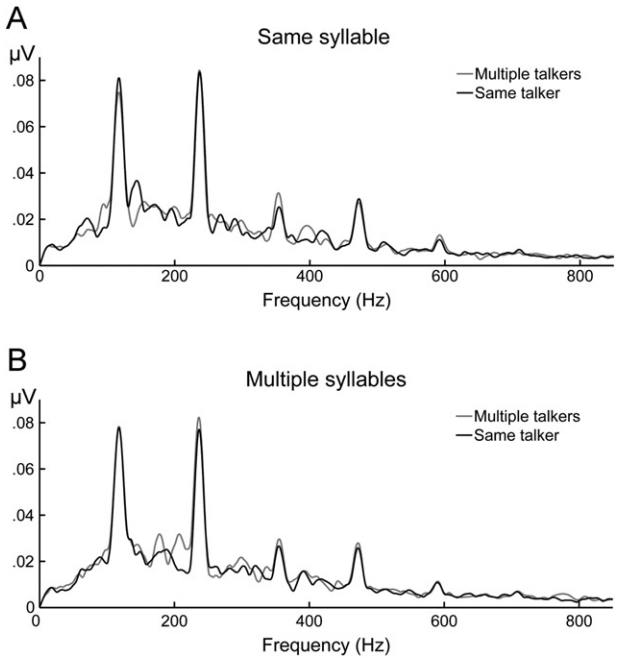

Fig. 3. Frequency-domain responses for the four experimental conditions. The response spectrum shows well-defined spectral peaks for F0 (118 Hz) and H2–H5 (in the range of 200 Hz through 600 Hz) for multi-talker condition relative to same-talker condition in the same-syllable (A) and multiple-syllable (B) contexts respectively.

The F0 and harmonic measures were each entered into a 2 (talker: same-talker vs. multiple-talker) × 2 (linguistic context: same-syllable vs. multiple-syllable) repeated measures ANOVA followed by subsequent post hoc tests when required. Bonferroni corrections were applied when appropriate.

3. Results

A 2 (talker) × 2 (linguistic context) repeated-measures ANOVA of the harmonic responses showed a significant interaction between talker and linguistic context, $F_{(1,11)} = 7.838$, $p = 0.017$ ($\eta^2 = 0.176$, medium effect). No main effect of talker or linguistic context was observed. Post-hoc comparisons indicated that participants yielded greater harmonic responses to the target stimulus under the multiple-talker than the same-talker condition ($t_{(11)} = 2.59$, $p = 0.025$, Bonferroni corrected) in multiple-syllable context (multi-talker, 0.055 ± 0.015 μV; same-talker, 0.050 ± 0.015 μV) (Figs. 3 and 4). No talker effect on the harmonic responses was observed under same-syllable condition ($t_{(11)} = 0.188$, $p = 0.855$, Bonferroni corrected). However, the 2 (talker) × 2 (linguistic context) repeated-measures ANOVA of the F0 responses produced no significant main effects or interaction.

In order to minimize the influence of unequal signal to noise ratios between harmonics or subjects on the results, we reran the analysis on log-transformed amplitudes. The results were consistent, with a significant interaction between talker and linguistic context for the log-transformed harmonic responses, $F_{(1,11)} = 9.865$, $p = 0.009$ ($\eta^2 = 0.221$, medium effect). No main effect of talker or linguistic context was observed. Post-hoc comparisons indicated that participants yielded greater harmonic responses to the target stimulus under the multiple-talker than the same-talker condition ($t_{(11)} = 2.78$, $p = 0.018$, Bonferroni corrected) in multiple-syllable context. No talker effect of the harmonic responses was observed under same-syllable condition. Moreover, the 2 (talker) × 2 (linguistic context) repeated-measures ANOVA of the log-transformed F0 responses obtained no significant main effects or interaction.

4. Discussion

The present study provides the first electrophysiological evidence of voice sensitivity at the subcortical level. This sensitivity to voice manifested as a change in how vocal harmonics were encoded depending on whether the target stimulus was presented in same-talker vs. multiple-talker conditions, with the multiple-talker condition producing enhanced responses. This observed multi-talker "subcortical enhancement" resembles greater activation found in the right superior temporal sulcus for multiple talkers relative to the same talker in the previous fMRI study (Belin and Zatorre, 2003), indicating a similar response facilitation to different voices at both cortical and subcortical levels.

It should be noted that the subcortical talker effect was dependent on the linguistic context of the speech stimulus, such that it was only evident in the multiple-syllable but not the same-syllable context. This suggests that the automatic sensory encoding of human voices at the subcortical level relies on the presence of additional linguistic cues, corroborating previously observed linguistic context effects on behavioral voice perception (Goggin et al., 1991). Indeed, the ability to understand speech and the ability to recognize voices are often closely linked (Perrachione et al., 2011; Chandrasekaran et al., 2011). A recent study has identified the left posterior middle temporal gyrus as a crucial brain region sensitive to

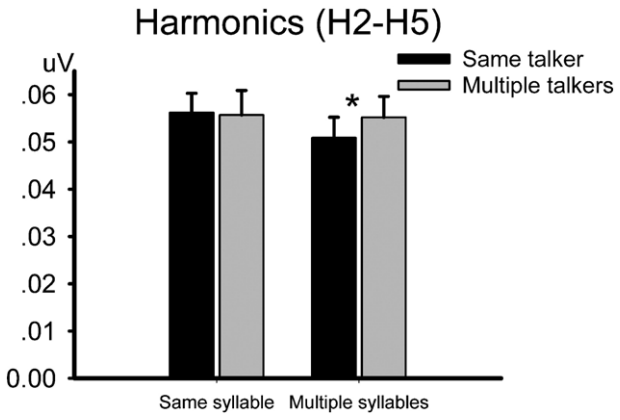

Fig. 4. Mean harmonic amplitudes for the four different experimental conditions. Mean amplitudes of harmonic encoding for the same-talker (black) and multi-talker (grey) conditions in the same-syllable and multiple-syllable contexts. Participants showed stronger harmonic encoding for the target stimulus [da1] in the multi-talker condition compared to the same-talker condition, only in the multiple-syllable context. Error bars indicate standard error. *, $p < 0.05$.

both voice and speech information (Chandrasekaran et al., 2011). Moreover, another study by Perrachione et al. (2011), showed that individuals with dyslexia had impaired voice recognition. However, because voice perception is tightly linked to speech processing, even normal controls demonstrated a similar behavioral deficit in voice recognition when encountering a novel language where speech cues were not readily accessible.

Our findings are consistent with the special roles that the formant frequencies play in voice processing (Latinus and Belin, 2011). The F0 carries prosodic information and in tone languages variations in F0 also convey lexical differences. However, in the current study, these linguistic functions of the F0 are greatly minimized due to the experimental design in which a level F0 was applied to all syllables. On the other hand, the formant pattern—encoded by harmonic responses—carries not only voice identity but also linguistic information. This may also explain why we only observed the FFR enhancement with harmonics but not the F0.

In conclusion, the current study offers the first electrophysiological evidence of voice sensitivity at the subcortical level. Future studies should continue to investigate whether this subcortical voice sensitivity is affected by specific auditory experiences such as linguistic and musical training. This line of research will shed light on the development of possible intervention programs that target groups with voice processing deficits such as those on the autism spectrum or with dyslexia (Russo et al., 2008; Gervais et al., 2004; Perrachione et al., 2011).

Acknowledgments

This work was supported by the 973 Program [2014CB846103], by the National Natural Science Foundation of China [31221003 and 31471066], by the 111 project [B07008], and by the Fundamental Research Funds for the Central Universities. The physiological data were collected when Y. Nan was a visiting scholar at Auditory Neuroscience Laboratory. We owe our thanks to members of the Auditory Neuroscience Lab for their valuable input. We thank Dr. Patrick C.M. Wong and Dr. Francis Wong for help in recording stimuli. We thank Dr. Ann Bradlow for her advice on the experimental design and Wuxia Yang for her help in the behavioral validation data collection. We also thank the reviewer for the insightful comments.

References

Belin, P., 2006. Voice processing in human and non-human primates. Philos. Trans. R. Soc. Lond. B Biol. Sci. 361, 2091–2107.
Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14, 2105–2109.
Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.
Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal-lobe response to vocal sounds. Brain Res. Cogn. Brain Res. 13, 17–26.
Bidelman, G.M., Gandour, J.T., Krishnan, A., 2011a. Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. J. Cogn. Neurosci. 23, 425–434.
Bidelman, G.M., Gandour, J.T., Krishnan, A., 2011b. Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. Brain Cogn. 77, 1–10.
Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Int. 5, 341–345.
Chandrasekaran, B., Kraus, N., 2010. The scalp-recorded brainstem response to speech: neural origins and plasticity. Psychophysiology 47, 236–246.
Chandrasekaran, B., Chan, A.H., Wong, P.C., 2011. Neural processing of what and who information in speech. J. Cogn. Neurosci. 23, 2690–2700.
Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthelemy, C., Brunelle, F., Samson, Y., Zilbovicius, M., 2004. Abnormal cortical voice processing in autism. Nat. Neurosci. 7, 801–802.
Ghazanfar, A.A., Rendall, D., 2008. Evolution of human vocal production. Curr. Biol. 18, R457–R460.
Goggin, J.P., Thompson, C.P., Strube, G., Simental, L.R., 1991. The role of language familiarity in voice identification. Mem. Cogn. 19, 448–458.
Grossmann, T., Oberecker, R., Koch, S.P., Friederici, A.D., 2010. The developmental origins of voice processing in the human brain. Neuron 65, 852–858.
Insley, S.J., 2000. Long-term vocal recognition in the northern fur seal. Nature 406, 404–405.
Kreiman, J., Gerratt, B.R., 2010. Perceptual sensitivity to first harmonic amplitude in the voice source. J. Acoust. Soc. Am. 128, 2085–2089.
Krishnan, A., Xu, Y., Gandour, J., Cariani, P., 2005. Encoding of pitch in the human brainstem is sensitive to language experience. Brain Res. Cogn. Brain Res. 25, 161–168.
Krizman, J., Marian, V., Shook, A., Skoe, E., Kraus, N., 2012. Subcortical encoding of sound is enhanced in bilinguals and relates to executive function advantages. Proc. Natl. Acad. Sci. U. S. A. 109, 7877–7881.
Latinus, M., Belin, P., 2011. Human voice perception. Curr. Biol. 21, R143–R145.
Micheyl, C., Divis, K., Wrobleski, D.M., Oxenham, A.J., 2010. Does fundamental-frequency discrimination measure virtual pitch discrimination? J. Acoust. Soc. Am. 128, 1930–1942.
Moushegian, G., Rupert, A.L., Stillman, R.D., 1973. Laboratory note. Scalp-recorded early responses in man to frequencies in the speech range. Electroencephalogr. Clin. Neurophysiol. 35, 665–667.
Musacchia, G., Sams, M., Skoe, E., Kraus, N., 2007. Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. Proc. Natl. Acad. Sci. U. S. A. 104, 15894–15898.
Owren, M.J., Cardillo, G.C., 2006. The relative roles of vowels and consonants in discriminating talker identity versus word meaning. J. Acoust. Soc. Am. 119, 1727–1739.
Parbery-Clark, A., Skoe, E., Kraus, N., 2009. Musical experience limits the degradative effects of background noise on the neural processing of sound. J. Neurosci. 29, 14100–14107.
Perrachione, T.K., Del Tufo, S.N., Gabrieli, J.D., 2011. Human voice recognition depends on language ability. Science 333, 595.
Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nat. Neurosci. 11, 367–374.
Rendall, D., Owren, M.J., Rodman, P.S., 1998. The role of vocal tract filtering in identity cueing in rhesus monkey (*Macaca mulatta*) vocalizations. J. Acoust. Soc. Am. 103, 602–614.
Russo, N.M., Skoe, E., Trommer, B., Nicol, T., Zecker, S., Bradlow, A., Kraus, N., 2008. Deficient brainstem encoding of pitch in children with Autism Spectrum Disorders. Clin. Neurophysiol. 119, 1720–1731.
Skoe, E., Kraus, N., 2010. Auditory brain stem response to complex sounds: a tutorial. Ear Hear. 31, 302–324.
Sliwa, J., Duhamel, J.R., Pascalis, O., Wirth, S., 2011. Spontaneous voice-face identity matching by rhesus monkeys for familiar conspecifics and humans. Proc. Natl. Acad. Sci. U. S. A. 108, 1735–1740.
Smith, J.C., Marsh, J.T., Brown, W.S., 1975. Far-field recorded frequency-following responses: evidence for the locus of brainstem sources. Electroencephalogr. Clin. Neurophysiol. 39, 465–472.
Stapells, D.R., 2000. Frequency-specific evoked potential audiometry in infants. In: Seewald, R.C. (Ed.), A Sound Foundation Through Early Amplification. honak AG, Basel, pp. 13–31.
Starr, A., Picton, T.W., Sininger, Y., Hood, L.J., Berlin, C.I., 1996. Auditory neuropathy. Brain 119 (Pt 3), 741–753.
Tice, R., Carrell, T., 1998. Level16 v. 2.0.3. University of Nebraska.
Wong, P.C., Skoe, E., Russo, N.M., Dees, T., Kraus, N., 2007. Musical experience shapes human brainstem encoding of linguistic pitch patterns. Nat. Neurosci. 10, 420–422.