Phonological experience modulates voice discrimination: Evidence from functional brain networks analysis

Xueping Hu a,1, Xiangpeng Wang a,1, Yan Gu a, Pei Luo a, Shouhang Yin a, Lijun Wang a, Chao Fu c, Lei Qiao a, Yi Du b*, Antao Chen a*

a Key Laboratory of Cognition and Personality of Ministry of Education, Faculty of Psychology, Southwest University, Chongqing 400715, China
b CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
c School of Psychology, Northwest Normal University, Lanzhou 730070, China

A R T I C L E   I N F O

Article history:
Received 14 October 2016
Revised 19 April 2017
Accepted 3 June 2017
Available online 27 June 2017

Keywords:
Voice discrimination
Language familiarity effect
Phonological experience
Cognitive control
Large-scale brain network

A B S T R A C T

Numerous behavioral studies have found a modulation effect of phonological experience on voice discrimination. However, the neural substrates underpinning this phenomenon are poorly understood. Here we manipulated language familiarity to test the hypothesis that phonological experience affects voice discrimination via mediating the engagement of multiple perceptual and cognitive resources. The results showed that during voice discrimination, the activation of several prefrontal regions was modulated by language familiarity. More importantly, the same effect was observed concerning the functional connectivity from the fronto-parietal network to the voice-identity network (VIN), and from the default mode network to the VIN. Our findings indicate that phonological experience could bias the recruitment of cognitive control and information retrieval/comparison processes during voice discrimination. Therefore, the study unravels the neural substrates subserving the modulation effect of phonological experience on voice discrimination, and provides new insights into studying voice discrimination from the perspective of network interactions.

© 2017 Published by Elsevier Inc.

1. Introduction

Human voice discrimination is defined as the ability to judge whether two voices are similar or different based on matching of basic auditory parameters (Van Lancker & Kreiman, 1987). Developed early in newborns, the ability is essential for human communication and social interactions (Belin, Fecteau, & Bedard, 2004; Belin, Zatorre, Lafaille, Ahad, & Pike, 2000;Latinus, McAleer, Bestelmeyer, & Belin, 2013). As a voice-based process, voice discrimination is influenced by speaker variability, and in addition, language familiarity (Fleming, Giordano, Caldara, & Belin, 2014). In particular, listeners are better at identifying voices in their familiar language than in an unfamiliar one. This phenomenon is captured by the language familiarity effect (LFE) (Fleming et al., 2014; Thompson, 1987; Winters et al., 2008). Since firstly described by Thompson (1987), the LFE has attracted a large amount of attention and been confirmed in bilinguals (Fleming et al., 2014; Thompson, 1987; Winters et al., 2008), infants (Johnson, Westrek, Nazzi, & Cutler, 2011), and even in dyslexic, receptive aphasic and phonagnosia patients (Perrachione, Del Tufo, & Gabrieli, 2011; Van Lancker, Cummings, Kreiman, & Dobkin, 1988).

Concerning the LFE, a main point comes into controversy is whether the effect is based on comprehension of linguistic information or familiarity with the phonological structure (Fleming et al., 2014). While no conclusion has been made, the study by Fleming et al. (2014) indicated that the phonological aspects of language ability only could influence voice discrimination. In particular, the authors employed time-reversed speech that excludes the influence of linguistic comprehension but retains acoustic cues important for voice identification (Johnson et al., 2011; Van Lancker, Kreiman, & Emmorey, 1985). They found that both English and Chinese groups were more sensitive to voice in their native language than in their non-native language. The observation is obviously interesting; however, since behavioral data alone could do little to provide further evidence, the mechanism subserving

http://dx.doi.org/10.1016/j.bandl.2017.06.001
0093-934X/© 2017 Published by Elsevier Inc.
the modulation of phonological experience on voice discrimination remains unclear.

In a typical voice discrimination process, listeners should shift attention from irrelevant information to goal-directed information and map the acoustic patterns onto phonetic idiosyncrasies of the talker (Francis, Baldwin, & Nusbaum, 2000; Magnuson & Nusbaum, 2007; Nusbaum & Magnuson, 1997). Then listeners get accesses to the “acoustical voice space” to track the acoustical-phonetic imprint when hearing the first voice stimuli (Andics et al., 2010; Latinus, Crabbe, & Belin, 2011). After that, by detecting the differences between the first and second voices, listeners make a judgment about whether the two voices are produced by the same speaker or not (Van Lancker & Kreiman, 1987). Therefore, in addition to acoustic processing involved in voice discrimination, various cognitive processes (e.g., working memory, selective attention) should also be recruited.

In the current study, we sought to explore the modulation of phonological experience on voice discrimination from the perspective of neural substrates. The materials we used were time-reversed speech sentences similar to those used in Fleming et al. (2014). Meanwhile, phonological experience in the current study was indexed by language familiarity, and we employed three different languages: native language (Mandarin Chinese) and non-native languages (English, second language; Korean, unfamiliar language) to construct a gradient change. During fMRI scanning, participants were asked to perform a same/different voice discrimination task across different language contexts. The current study first examined the regional activity in response to the task. In addition, since voice discrimination is based on the processing of multi-source information, we hypothesized that network-based neural interactions are key elements subserving the phenomenon. In particular, we examined several networks that might be recruited in voice discrimination in an exploratory way. To specify, there were a voice acoustic network and a voice identity network (Andics et al., 2010; Blank, Wieland, & von Kriegstein, 2014; Latinus et al., 2011) sensitive to a general voice-identity processing; a default mode network associated with person-identity information (Arnott, Heywood, Kendtridge, & Goodale, 2008; Blank et al., 2014; Simmons, Reddish, Bellgowan, & Martin, 2010); a salience network engaged in detecting salience stimuli (Menon, 2011); and a frontoparietal network responsible for attentional control (Menon, 2011; Power et al., 2011). Neural interactions in the current study were calculated using a multi-regional psychophysiological interaction (PPI) modeling analysis. And this method has been demonstrated as effective in investigating dynamic functional connectivity between regions/networks (Cocchi, Zalesky, Fornito, & Mattingley, 2013; Cocchi et al., 2014; Gerchen, Bernal-Casas, & Kirsch, 2014). We predicted that typical brain regions of voice processing could be activated, along with control-related systems. Importantly, regional activity and neural interactions could operate as a function of phonological experience.

2. Methods

2.1. Participants

Thirty-six native Mandarin Chinese volunteers from Southwest University participated in the current study (four males; M = 21.11 ys, SD = 0.23). All participants are seniors of English majors without any experience of Korean, and have learned English for more than 10 years (M = 10.71, SD = 0.292). All participants were right-handed with normal hearing, and reported no neurological or psychiatric disorders. Written informed consents were obtained from all participants before scanning. To evaluate the language proficiency of English, all subjects finished an English language proficiency test (the Transparent Language Proficiency Test: http://www.transparent.com/) and the Bilingual Switching Questionnaire test (BSWQ) (Rodriguez-Fornells, Krämer, Lorenzo-Seva, Festman, & Münte, 2012). The mean score of language proficiency test was 84.92 ± 1.27 (% of correct responses), and that of the BSWQ was 9.91 ± 0.24 for L1S (switch to Chinese), 7.71 ± 0.29 for L2S (switch to English), 9.03 ± 0.31 for CS (contextual switch), 8.46 ± 0.31 for US (unintended switch) and 35.11 ± 0.80 for OS (overall switch), respectively. More information about the BSWQ could be find in the research of Rodriguez-Fornells et al. (2012). These scores suggest that the L2 proficiency of volunteers is above average. The study was approved by the Human Ethics Committee from Southwest University, China.

2.2. Stimuli

The testing stimuli were selected from two speech corpuses and then edited. The Mandarin Chinese and Korean stimuli were selected from OSCAAR (The Online Speech/Corpora Archive and Analysis Resource, https://oscaar.ci.northwestern.edu/index.html), the English speech stimuli were selected from the PN/NC corpus (McCloy et al., 2013). In order to avoid influence from paralinguistic information of voices, such as gender (Belin et al., 2004; Pisanski, Cartei, McGettigan, Raine, & Reby, 2016), the speech stimuli included six sentences from five native male speakers in each language respectively. Recordings were sampled at 16 bit and 22.05 kHz, time-reversed (Fig. 1B), and normalized for root mean square (RMS) amplitude to 70 dB SPL. The average duration of speech recordings is 1777.47 ms (SD = 115.79 ms) for Mandarin Chinese, 1856.17 ms (SD = 106.86 ms) for English and 1760.60 ms (SD = 123.98 ms) for Korean, respectively. All stimuli were edited using Adobe Audition 3.0 and Praat (Boersma, 2001; Boersma & Weenink, 2015). The detailed acoustic features of stimuli are listed in Table S1.

2.3. Experimental procedures

For each language, the recording of the six sentences by five native male speakers resulted in 30 time-reversed speech stimuli. For each language, the same time-reversed sentence by the same speaker was repeated twice and formed the same condition (SC) of 30 speech pairs; while for the different condition (DC), the same time-reversed sentence by two distinct speakers was presented consecutively, resulting 60 kinds of speech pairs. To balance the number of the stimuli, the final materials included 60 pairs of “same” stimuli and 60 pairs of “different” stimuli for each language across three runs. Each run contained 40 pairs of stimuli with an equal proportion of the “same” and “different” trials. In total, the experiment consisted of 360 trials. The order of language runs was counterbalanced across subjects. To avoid the consecutive presentation of the same speaker’s speeches, the sequence of the stimuli within each language was pseudo-randomized. Stimuli were presented binaurally at a comfortable intensity level using MR-compatible headphones. Each trial started with a yellow speaker icon lasted for 2000 ms (during which the first speech was presented), followed by a 500 ms blank screen. After that, a second yellow speaker icon was displayed for 4000 ms, during which the second speech was presented and subjects were told to discriminate whether the two voices were the same or different and make responses by pressing the “left” key for “same” trials and the “right” key for “different” trials on a two-button pad using in the scanner. The inter-trial-interval (ITI) was jittered from 1 s to 3 s (2 s in average).
2.4. fMRI data acquisition and analyses

Functional and anatomical images were acquired with a Siemens 3 T scanner (Siemens Magnetom Trio TIM, Erlangen, Germany). A T2* -weighted continuous gradient echo-planar imaging (EPI) pulse sequence was applied to functional images (TR, 1.5 s; TE, 30 ms; flip angle, 90° / 176°; FoV, 960 × 960 mm²; matrix size, 64 × 64; 24 interleaved 5 mm-thick slices). Within a run, a total of 247 EPI images were acquired. Then, T1-weighted images were recorded with a total of 176 slices at a thickness of 1 mm (TR, 1900 ms; TE, 2.52 ms; flip angle, 90° / 176°; FoV, 250 × 250 mm²).

All scanning images were preprocessed and analyzed using SPM8 (The Wellcome Trust Centre for Neuroimaging, University College London, UK; http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). The first five images were discarded to achieve magnetically steady images. Then the images were corrected for differences in slice-timing, realigned, and then co-registered with the T1 anatomical image. Next, all images were spatially normalized to the standard MNI template brain (voxel size: 3 × 3 × 3 mm³). The first 0.5 seconds were discarded to minimize the impact of head movement. The resulting images were smoothed using Gaussian kernel with a full-width at half maximum (FWHM) parameter of 6 × 6 × 6 mm³. In the following analyses, data from three participants were excluded because of excessive movement (>3.0 mm relative motion), resulting in 33 participants’ data.

The BOLD response was modeled in a general linear model (GLM). For each run, three regressors (i.e., same, different and error) were modeled to create the design matrix. In addition, onsets of regressors were created and convolved with the canonical hemodynamic response function (HRF). To minimize the impact of head movement, the six realignment parameters for each participant were included as confounding factors in each design matrix. The duration of each event was set to 4 s, a default high-pass filter cutoff at 128 s was applied to remove slow signal drifts.

For each model, a gray mask with the possibility threshold of 0.4 in SPM8 was applied. Voxel-wise parameters were estimated using Restricted Maximum-Likelihood (ReML) for each regressor. The contrasts of interest were defined as stimuli types for each language (i.e., “English same stimuli pair”, “English different stimuli pair”, and “English different stimuli pair vs. “English same stimuli pair”) and were generated in the 1st-level analysis. In the group level analysis, a 3 × 2 design matrix was created to test the main effect of the language (English, Korean, Mandarin Chinese) and the condition (same, different), and their interaction effect. Contrast estimates for each condition were entered into a full factorial repeated measures analysis of variance design. Then all reported locations were based on standard SPM anatomical atlas.

For multiple comparison correction, the Gaussian Random Filed Theory (GRFT) in the DPABI software package (Yan, Wang, Zuo, & Zang, 2016) was exerted to define the cluster-extent threshold for the main and interaction effects. The brain mask employed in the analysis was created based on the gray matter template in SPM8, and only those voxels with a probability higher than 0.4 in the template were included. Then the gray matter mask was resized at the voxel size of 3 × 3 × 3 mm³. In total, there were 49464 voxels within the mask. Three different cluster-extent thresholds, that is k ≥ 29 (for the main effect of language), 81 (for the main effect of condition) and 33 (for the interaction effect), were applied at a voxel-level p < 0.001 and cluster-level p < 0.05 (Yan et al., 2016).

2.5. Psychophysiological interaction (PPI) analyses

In the current study, a multiregional PPI modeling approach (Cocchi et al., 2014; Gerchen et al., 2014; Hearne, Cocchi, Zalesky, & Mattingley, 2015) was implemented to examine the large-scale brain network interactions associated with voice discrimination in a whole-brain manner. Instead of the voxel-based method assessing functional coupling between a single region and the
other voxels in the brain, this novel method could effectively calculate effective connectivity between pairs of brain regions (Cocchi et al., 2014; Hearne et al., 2015).

In the current study, we defined 17 regions of interest pertaining to several brain networks. All regions were defined around the peak coordinates with a 6 mm radius based on the activation results in the current study, and then were categorized into 5 networks based on the functions of them in the literature. In general, these networks are often observed in the process of voice recognition/discrimination or cognitive control. Specifically, the frontoparietal network (FPN) included bilateral dorsolateral prefrontal cortices (DLPFC) and posterior parietal cortices (PPC) (Dosenbach et al., 2007), the salient network (SN) included bilateral anterior insula (AI) and the anterior cingulate cortex (ACC) (Seeley et al., 2007), the default mode network (DMN) included the ventromedial prefrontal cortex (VMPFC) and the posterior cingulate cortex (PCC) (Andrews-Hanna, Reider, Sepulcre, Poulin, & Buckner, 2010; Greicius, Krasnow, Reiss, & Menon, 2003), the voice-acoustic network (VAN) included bilateral superior temporal gyrui (STG) and inferior frontal gyrui (IFG) (Andics et al., 2010; Latinus et al., 2011) and the voice-identity network (VIN) included bilateral anterior middle temporal gyri (aMTG, a part of anterior temporo-parietal lobule (ATL)) (Andics et al., 2010; Arnott et al., 2008; Belin & Zatorre, 2003; Blank et al., 2014) and the angular gyrus (AG) (Andics et al., 2010; Arnott et al., 2008). Detailed information for the network definition and the seed regions is listed in Table 1.

For each seed region, we extracted the time course (first eigen-variate) to create the PPI term. The current study employed several psychological variables to detect the interactions under the SC and the DC respectively, as well as the differences between the SC and DC (contrast: different vs. same). For each language context, we got the PPI terms representing regional activity during the SC and DC separately. For each participant, a GLM model was used to evaluate condition-dependent influence of one certain region on another. The dependent variable was the activity within the target region, while the explanatory variable was the PPI term of the source region. The activity of the region (main effects of the physiological and the psychological factors) and the corresponding psychological regressor were included as confounding covariates. Then this step was repeated, resulting in a $17 \times 17$ connectivity matrix for each individual and condition. This matrix stored the parameter estimate ($\beta$) and directionality between regions. Especially, this resulted in an asymmetric matrix, half of which of same direction (e.g., X to Y), while the other half-contained connectivity estimates of the opposite direction (e.g., Y to X) (Hearne et al., 2015).

At the group level, we tested whether the neural interactions between regions significantly differed between conditions under distinct language context with a within-subject ANOVA analysis. Multiple comparisons were corrected using the network-based statistic (NBS) (Zalesky, Cocchi, Fornito, Murray, & Bullmore, 2012; Zalesky, Fornito, & Bullmore, 2010) with a primary F-statistic threshold of 3.5. For the network-level analysis, we averaged the $\beta$ values of all the regions pertaining to specific networks, regardless of whether each region showed significant connectivity to other regions, to obtain intra-network and inter-network connectivity patterns.

3. Results

3.1. Behavioral results

As shown in Fig. 1C, the main effect of language on reaction time was significant ($F_{(2, 64)} = 22.493, p < 0.001$), where the discrimination of Mandarin Chinese was faster than English (SC: $t = 5.036, p < 0.001$; DC: $t = 5.908, p < 0.001$, paired-sample t tests). However, neither the main effect of condition nor the interaction effect was significant ($F_{(1, 32)} = 0.177, p = 0.677; F_{(2, 64)} = 0.282, p = 0.972$).

According to signal detection theory, we also calculated the hit rate (accurate rate of the DC), the false alarm rate (error rate of the SC) and the signal detection sensitivity index ($d'$). The results showed a significant language effect on the hit rate ($F_{(2, 64)} = 35.72, p < 0.001$). Post-hoc analysis showed that the languages all differed with each other (Mandarin Chinese and English: $p = 0.005$; Mandarin Chinese and Korean: $p < 0.001$; English and Korean: $p < 0.001$, Fig. 1D). There was also a significant language effect on the false alarm rate ($F_{(2, 64)} = 3.166, p < 0.05$, Fig. 1E), where the Mandarin Chinese differed significantly with English ($p = 0.022$, Bonferroni post hoc tests). There was no difference between Mandarin Chinese and Korean ($p > 0.05$), English and Korean ($p > 0.05$). Furthermore, there was a significant main effect of language on the $d'$ index ($F_{(2, 64)} = 32.56, p < 0.001$, Fig. 1F). Post-hoc analysis revealed significant differences between Mandarin Chinese and English ($p < 0.001$), between Mandarin Chinese and Korean ($p < 0.001$), and between English and Korean ($p = 0.017$) respectively.

3.2. Phonological experience modulated regional activation in voice discrimination

The main effect of language revealed significant activation in the left IFG, the left superior temporal sulcus/gyrus (STS/G), the left pre-supplementary motor area (SMA), and the bilateral AI (Fig. 2A and Table S2) (uncorrected $p < 0.001$ with cluster > 29 voxels, two-sided test). Regional activation that showed a significant main effect of condition was observed in the prefrontal cortex (e.g., the IFG, the AI and the SMA), the parietal cortex (e.g., the PPC and the angular gyrus), the temporal cortex (e.g., the STS/G and the aMTG) and the cerebellum (for details, see Fig. 2B and Table S2). These results survived at a significant level of $p < 0.001$ (uncorrected, cluster > 81 voxels, two-sided test).

Most importantly, as shown in Fig. 3A and Table S3, a number of clusters in the left IFG, AI and ACC showed a language x condition interaction, indicating a significant modulation of the language familiarity on voice discrimination in these regions (uncorrected $p < 0.001$ with cluster > 33 voxels, two-sided test). For most of the regions, there was a significant language effect in the DC (but not in the SC), where the English and Korean elicited stronger activity than Mandarin Chinese, while English and Korean did not differ much in the BOLD signal change of SC or DC (for detail information, see Suppl. Results).

3.3. Phonological experience modulated functional connectivity in voice discrimination

Seventeen regions constituting 5 brain networks were included in the multi-PPI analyses (for details, see Table 1). Results showed that the neural interactions between a large set of brain regions were modulated by language familiarity during the processing of voice discrimination (Fig. 4A and B). However, the averaging functional connectivity between all those regions did not show a significant language x condition interaction ($F_{(2, 50)} = 0.936, p = 0.399$, Fig. 4C).

For the network-level analyses, a within-subject repeat-measures ANOVA showed that the inter-network connectivity from the FPN to the VIN manifested a significant interaction effect ($F_{(2, 50)} = 8.593, p < 0.001$, Fig. 4D). Follow-up paired t-tests revealed that in the DC, the connectivity strength was significantly stronger for Korean than for English ($t_{25} = -3.875, p = 0.001$), and...
marginally significant stronger for Korean than for Mandarin Chinese ($t_{25} = 2.049$, $p = 0.051$). Furthermore, there was also a marginal significant difference between English and Mandarin Chinese ($t_{25} = 1.933$, $p = 0.065$) for the connectivity strength. Whilst, the functional connectivity did not differ between the three languages in the SC ($p > 0.05$). In addition, a similar language condition interaction effect was observed in the neural coupling from the DMN to the VIN ($F(2, 50) = 6.422$, $p = 0.003$, Fig. 4E). Further paired $t$-tests suggested that there were significant differences between Mandarin Chinese and English ($t_{25} = 2.295$, $p = 0.03$), Mandarin Chinese and Korean ($t_{25} = 3.064$, $p = 0.005$), but not between English and Korean ($t_{25} = 1.369$, $p > 0.05$) in the DC. However, these differences were not observed in the SC.

Table 1
Regional activation results used for Psychophysiological interaction analyses.

<table>
<thead>
<tr>
<th>Region</th>
<th>H</th>
<th>BA</th>
<th>MNI coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$x$</td>
</tr>
<tr>
<td>Voice-Acoustic Network</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferior Frontal Gyrus</td>
<td>L</td>
<td>45</td>
<td>−51</td>
</tr>
<tr>
<td>Inferior Frontal Gyrus</td>
<td>R</td>
<td>45</td>
<td>54</td>
</tr>
<tr>
<td>Superior Temporal Gyrus</td>
<td>L</td>
<td>22</td>
<td>−57</td>
</tr>
<tr>
<td>Superior Temporal Gyrus</td>
<td>R</td>
<td>22</td>
<td>54</td>
</tr>
<tr>
<td>Voice-Identity Network</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anterior Middle Temporal Gyrus</td>
<td>L</td>
<td>21/38</td>
<td>−57</td>
</tr>
<tr>
<td>Anterior Middle Temporal Gyrus</td>
<td>R</td>
<td>21/38</td>
<td>60</td>
</tr>
<tr>
<td>Angular Gyrus</td>
<td>L</td>
<td>39</td>
<td>−39</td>
</tr>
<tr>
<td>Angular Gyrus</td>
<td>R</td>
<td>39</td>
<td>45</td>
</tr>
<tr>
<td>Salience Network</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anterior Insula</td>
<td>L</td>
<td>13</td>
<td>−33</td>
</tr>
<tr>
<td>Anterior Insula</td>
<td>R</td>
<td>13</td>
<td>36</td>
</tr>
<tr>
<td>Anterior Cingulate Gyrus</td>
<td>L</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>Fronto-Parietal Network</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dorsolateral Prefrontal Cortex</td>
<td>L</td>
<td>46</td>
<td>−48</td>
</tr>
<tr>
<td>Dorsolateral Prefrontal Cortex</td>
<td>R</td>
<td>46</td>
<td>48</td>
</tr>
<tr>
<td>Posterior Parietal Cortex</td>
<td>L</td>
<td>40</td>
<td>−33</td>
</tr>
<tr>
<td>Posterior Parietal Cortex</td>
<td>R</td>
<td>40</td>
<td>36</td>
</tr>
<tr>
<td>Default Mode Network</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ventromedial Prefrontal Cortex</td>
<td>L</td>
<td>10/11</td>
<td>−3</td>
</tr>
<tr>
<td>Posterior Cingulate Gyrus</td>
<td>L</td>
<td>31</td>
<td>−9</td>
</tr>
</tbody>
</table>

Note: H, hemisphere; BA, Brodmann area.

Fig. 2. Main effects of language (A) and condition (B) in regional activation. Color bar represents a scale of $F$ values. The maps are thresholded at a significant level of $p < 0.001$, uncorrected with cluster > 29 (for panel A) or 81 (for panel B) voxels.

4. Discussion

By manipulating language familiarity across Mandarin Chinese, English and Korean, the current study observed a significant influence of phonological experience on voice discrimination. That is, individuals showed better performance (e.g., shorter reaction time and higher signal detection sensitivity) in discriminating voices in their familiar language relative to unfamiliar ones. Meanwhile, regional activity sensitive to this effect was observed in voice- and cognitive-related regions, such as the bilateral IFG and the ACC. Further, the functional couplings between several networks were modulated by phonological experience, indicating that voice discrimination depends on dynamic large-scale network interactions. Specifically, directed connectivity from the FPN and the DMN to the VIN were two major neural pathways subserving phonological experience-based cognitive control on voice discrimination.
Coincided with existing findings, the present study observed increased reaction time and decreased sensitivity of voice discrimination when the language was unfamiliar, indicating significant LFE in voice discrimination (Fleming et al., 2014; Johnson et al., 2011; Winters et al., 2008). In the level of fMRI data, this effect was reflected in regional changes of several brain regions responsible for voice discrimination as a function of language familiarity. That is, decreased language familiarity was accompanied by increased recruitment of voice- and cognitive-related regions in telling the distinctions between the same and different speakers. According to the literature, the ACC engages in detecting incongruent stimuli (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Carter et al., 1998; Chen et al., 2008) and provides updated prediction of expected cognitive demand to exert adaptive control (Sheth et al., 2012). The AI maps salient events to recruit control-related processes (Menon & Uddin, 2010; Nelson et al., 2010). In addition, the IFG is sensitivity to voice-acoustic encoding (Andics et al., 2012). The AI maps salient events to recruit control-related processes (Menon & Uddin, 2010; Nelson et al., 2010). In addition, the IFG is sensitivity to voice-acoustic encoding (Andics et al., 2012). In common voice-identity processing, those regions could get accesses to the identity information of voices and integrate multiple resources to support cognitive operations (Wong, Nusbaum, & Small, 2004). Here the enlarged activation differences between the same and different speaker discrimination were accompanied by decreased phonological experience, suggesting that the influence of language familiarity on voice discrimination is achieved by biasing the engagement of cognitive resources. In particular, for native language that is most familiar for individuals, matching between stored voice characteristics and the ongoing voice information would be facilitated; while for unfamiliar languages (English and Korean), individuals should call for more cognitive resources to accomplish the retrieval and comparison processes that depend on memory systems (Arnott, Grady, Hevenor, Graham, & Alain, 2005; Overath et al., 2007). This is also consistent with the argument that language experience could shape the neural processing of speech and further modulate the involvement of cognitive control areas in multilinguals (Abutalebi et al., 2013; Intartaglia et al., 2016; Pierce, Chen, Delcenserie, Genesee, & Klein, 2015).

Moreover, voice discrimination in current study recruited a large set of distributed brain regions pertaining to several large-scale networks, and interactions between those networks were differentially modulated by phonological experience. Interestingly, the modulation effect did not occur at the global scale, but was specific to the pathways from the FPN to the VIN, and from the DMN to the VIN. The FPN, anchored by the DLPFC and the PPC, has been demonstrated as engaged in higher-level cognitive functions (Dosenbach et al., 2007; Menon, 2011; Menon & Uddin, 2010; Wang et al., 2015). For voice-identity processing, the FPN plays an important role in information retrieval and comparison (Overath et al., 2007) and supports selecting among multiple items held in memory (Rowe, Toni, Josephs, Frackowiak, & Passingham, 2000). In particular, the activity of the FPN could exert influence over the cortical activity of the perceptual/sensory regions, such as those responsible for voice or speech processing (Obleser, Wise, Dresner, & Scott, 2007; Stevens, 2004; von Kriegstein & Giraud, 2004), by initiating control signals to accomplish task goals (Egger & Hirsch, 2005; Zanto, Rubens, Thangavel, & Gazzaley, 2011). In the current study, the connectivity from the FPN to the VIN changed as a function of language familiarity, suggesting that the phonological experience could modulates the top-down control signals over the VIN. This modulation effect was particularly obvious as the unfamiliarity of the language increased, where individuals had to make more efforts to discriminate voices under both the SC and DC. That is, increased language unfamiliarity was associated with stronger cognitive control that resulted in weaker activity in the VIN, reflecting in enhanced negative connectivity from the FPN to the VIN. Surprisingly, the connectivity turned to a positive one under the DC for Korean. This may indicate that the FPN lost control over the VIN, which results in inefficient down-regulation of voice identity processing, though further studies are needed to test this hypothesis. Combined, our findings suggest that the functional connectivity from the FPN to the VIN could serve as a cognitive neural basis for LFE in voice discrimination.

For the connectivity from the DMN to the VIN, there was a significant interaction effect between language familiarity and speaker discrimination conditions. This interaction effect was attributed to increased connectivity strength in the SC (although not statistically significant) but significantly decreased connectivity strength in the DC as a function of language unfamiliarity. The literature has proposed that the DMN is engaged in spontaneous cognition such as internal mentation, self-referential and emotional processing (Buckner, Andrews-Hanna, & Schacter,
In particular, the PCC and the VMPFC, key nodes of the DMN (Greicius et al., 2003), are suspected to be sensitive to person-identity recognition (Arnott et al., 2008; Blank et al., 2014; Simmons et al., 2010), and support encoding of person-related information (Kosaka et al., 2003). Additionally, the regions pertaining to the VIN (especially the ATL) have been demonstrated as core neural substrates subserving person/voice-identity based on the memory of person-related knowledge (Andics et al., 2010; Olson, Plotzker, & Ezzyat, 2007; Simmons et al., 2010). Simmons et al. (2010) reported significant functional connectivity between the ATL and the DMN nodes, and suggested that person identity process in the ATL was mediated by storage or retrieval processes of the DMN (Greicius et al., 2003; Maguire & Mummery, 1999). Remarkably, the current study observed a similar functional coupling from the DMN to the VIN, which indicated that the DMN could be active during the encoding and retrieval of speaker-specific information in case to support matching process in the VIN. During the comparison and discrimination of different speakers, this functional coupling was particularly active when facing native language that is familiar to individuals; however, when facing unfamiliar language, there was a significant decline of the connectivity strength. A plausible interpretation is that as a matter of information ambiguity, language familiarity could influence the encoding or retrieval processes in the DMN; however, obscure information encoded in or retrieved from the DMN could hardly facilitate voice discrimination during the processing of unfamiliar language because of lacking phonological experience.

Intriguingly, the connectivity between the STG, one part of VAN, and the rest brain regions was not modulated by the language familiarity modulated neural interactions between regions and networks. The significant interaction effect between language familiarity and condition displays in the panel A and panel B. In panel A, the pink nodes represent the seed regions of voice-acoustic network (VAN), cyan nodes represent the seed regions of voice-identity network (VIN), green nodes represent the regions of the FPN (fronto-parietal network), orange nodes represent the regions of the SN (salient network), and blue nodes represent the regions of the DMN (default mode network). In panel B, ACC, anterior cingulate cortex; AGL, left angular gyrus; AGR, right angular gyrus; AIL, left anterior insula; AIR, right anterior insula; aMTGL, left anterior middle temporal gyrus; aMTGR, right anterior middle temporal gyrus; DLPFCL, left dorsolateral prefrontal cortex; DLPFCR, right dorsolateral prefrontal cortex; IFGL, left inferior frontal gyrus; IFGR, right inferior frontal gyrus; PCC, posterior cingulate cortex; PPCL, left posterior parietal cortex; STGL, left superior temporal gyrus; STGR, right superior temporal gyrus; VMPFC, ventromedial prefrontal cortex. In panel C, D and E depicts changes in functional connectivity patterns under different condition and different language context for global means, connectivity from the FPN to the VIN and connectivity from the DMN to the VIN, respectively. SC, same conditions; DC, different conditions.
experience (Fig. 4B). As we known, the right STG is one of the most prominent human voice-sensitive regions (Andics et al., 2010; Latinus et al., 2011; von Kriegstein & Giraud, 2004). Thus, one would expect that close connections between the STG and other voice-sensitive regions should be observed in the voice discrimination process. However, the analysis we employed here focused on whether the phonological experience modulated the network interactions in voice discrimination instead of the network interactions themselves. Our data show that the strength of the connectivity between the STG and other voice-sensitive regions was constant across language contexts, suggesting that the early stage voice-acoustic processing in the STG is necessary for voice discrimination regardless of language familiarity.

In the present study, the materials we used involved two conditions: the SC, where the same sentence was repeated by one speaker; and the DC, where the same sentence was repeated by two different speakers. Although the time-reversed manipulation would largely weaken the sensation of sentence repetition and impede participants from making decision merely based on sentence’s acoustic and structural information, the condition sets still led to the difficulty to totally tell apart voice identity processing from acoustic processing, especially for the SC. However, as the ROIs we employed in the PPI analyses were defined in the combination of literatures and GLM results from the current study, and as our main aim was to investigate the modulation effect of phonological experience on voice discrimination, we therefore suggest that this limitation would not affect our main findings. We suggest that further studies are needed to examine the specific roles of voice identity and acoustic processing respectively in voice discrimination.

In conclusion, the present study unravels the neural substrates subserving voice discrimination that was mediated by language familiarity. The findings suggest that voice discrimination is achieved by incorporating cognitive and perceptual resources and recruiting a set of distributed fronto-parieto-temporal regions. In addition, the VIN is a key network associated with voice discrimination, whose activity is modulated by the FPN and the DMN. Specifically, phonological experience could bias the cognitive recruitment of the FPN and the information encoding and retrieval in the DMN to support voice discrimination as a function of language familiarity. Therefore, the present study highlights the importance of the neural interactions in voice discrimination and provides new avenues for studying the influence of phonological experience on voice processing from the perspective of large-scale brain networks. Furthermore, our study provides the empirical bases for the cognitive remediation of those individuals showing deficits in voice discrimination and recognition, i.e., phonagnosia (Garrido et al., 2009; Rosswadowitz et al., 2014; Van Lancker & Canter, 1982; Van Lancker et al., 1988) through cognitive training or non-invasive brain stimulation (Berkman, Kahn, & Merchant, 2014; Fregni & Pascual-Leone, 2007; Rossi & Rossini, 2014).

Statement of significance

The current study firstly uncovered the brain regional activation and functional coupling patterns that were modulated by language familiarity during voice discrimination. The findings are obviously important in interpreting the neural bases of voice discrimination on the context of language familiarity, and provide biomarkers for probing further studies.

Acknowledgements

This work was supported by grants from the Fundamental Research Funds for the Central Universities (SWU1609106, SWU1709107) and the National Natural Science Foundation of China (61431013) to A. T. Chen, and the “Thousand Young Talent Plan” to Y. Du, and the Fundamental Research Funds for the Central Universities of PR China (SWU1609314) to X. P. Wang. We thank David Fleming, Bettina Braun, Xiaodong Wang and In-Ki Jin for their speech technical assistance, Luca Cocchi for providing the script code of multi-regional PPI method. We also thank Chun Liang Chan for providing Mandarin Chinese and Korean Recordings. Recordings come from the “Diapix” and “Korean-English intelligibility” collection available from the Online Speech/Corpora Archive and Analysis Resource (OSCAAR). OSCAAR (https://oscaar.cn.northwestern.edu/index.html) is a project of the Speech Communication Research Group in the Department of Linguistics at Northwestern University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bandl.2017.06.001.

References
