NeuroImage 91 (2014) 375–385

Contents lists available at ScienceDirect

NeuroImage

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

A neural mechanism for recognizing speech spoken by different speakers

Jens Kreitewolf [a,*], Etienne Gaudrain [b,c], Katharina von Kriegstein [a,d]

[a] Max Planck Institute for Human Cognitive and Brain Sciences, Max Planck Research Group Neural Mechanisms of Human Communication, D-04103 Leipzig, Germany
[b] University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology/Head and Neck Surgery, 9700 RB Groningen, Netherlands
[c] University of Groningen, Graduate School of Medical Sciences, Research School of Behavioural and Cognitive Neurosciences, 9713 GZ Groningen, Netherlands
[d] Humboldt University of Berlin, Psychology Department, D-12489 Berlin, Germany

ARTICLE INFO

Article history:
Accepted 3 January 2014
Available online 13 January 2014

Keywords:
fMRI
Glottal fold
Heschl's gyrus
Linguistic prosody
Voice

ABSTRACT

Understanding speech from different speakers is a sophisticated process, particularly because the same acoustic parameters convey important information about both the speech message and the person speaking. How the human brain accomplishes speech recognition under such conditions is unknown.

One view is that speaker information is discarded at early processing stages and not used for understanding the speech message. An alternative view is that speaker information is exploited to improve speech recognition. Consistent with the latter view, previous research identified functional interactions between the left- and the right-hemispheric superior temporal sulcus/gyrus, which process speech- and speaker-specific vocal tract parameters, respectively. Vocal tract parameters are one of the two major acoustic features that determine both speaker identity and speech message (phonemes). Here, using functional magnetic resonance imaging (fMRI), we show that a similar interaction exists for glottal fold parameters between the left and right Heschl's gyri. Glottal fold parameters are the other main acoustic feature that determines speaker identity and speech message (linguistic prosody).

The findings suggest that interactions between left- and right-hemispheric areas are specific to the processing of different acoustic features of speech and speaker, and that they represent a general neural mechanism when understanding speech from different speakers.

Introduction

The same speech message can be acoustically very different depending on who is speaking (e.g., Peterson and Barney, 1952). Nevertheless, the human brain shows remarkable robustness to speaker-related variations despite the fact that the same acoustic parameters convey important information for speech understanding as well as for speaker recognition (reviewed in Obleser and Eisner, 2009; Pisoni, 1997). Glottal pulse rate (GPR) (Figs. 1A/B, green), for instance, which is the result of movements of the glottal folds, signals whether an utterance is a statement or a question (i.e., linguistic prosody) and determines the voice height of a speaker. To date, it is an open question how the human brain accomplishes robust speech recognition under conditions where information about speech and speaker is encoded in the same parameter (like it is the case for GPR).

For many years, neuroscientific research on speech recognition has been performed separately from work on speaker recognition, either implicitly or explicitly assuming that these are two independent processes (reviewed in Belin et al., 2004; Hickok and Poeppel, 2007; Pisoni, 1997; Scott and Johnsrude, 2003). However, several findings from behavioral (reviewed in Cutler et al., 2010; Nusbaum and Magnuson, 1997; Nygaard, 2005) and neuroimaging studies (e.g., Chandrasekaran et al., 2011; Kaganovich et al., 2006; Wong et al., 2004) showed that there are strong interdependencies between speech and speaker recognition and that even non-speech contexts can shift phoneme categorization (Laing et al., 2012). Recent fMRI work has suggested that speech recognition in the context of changing speakers relies on functional interactions between left- and right-hemispheric areas processing specific acoustic features of speech and speaker (von Kriegstein et al., 2010). In that study, speech stimuli were resynthesized to evoke speaker changes by variations of vocal tract parameters (Fig. 1A, blue), which, similar to glottal fold parameters, affect both the perceived identity of the speaker (Fig. 1B, bottom right) and parts of the speech message (i.e., phonemes in the case of vocal tract parameters) (Fig. 1B, top right) (Gaudrain et al., 2009; Lavner et al., 2000; Smith and Patterson, 2005). However, it remained unclear whether interactions between specific areas in the right and left hemispheres are restricted to vocal tract parameters and to the task of phoneme recognition. Here, we investigated whether such interactions also occur when speakers differ in their glottal fold parameters and during a task that involves recognizing aspects of the speech message that are determined by glottal fold parameters (i.e., linguistic prosody). Finding a similar interaction for speech- and speaker-specific glottal fold parameters would

* Corresponding author at: Max Planck Institute for Human Cognitive and Brain Sciences, MPRG Neural Mechanisms of Human Communication, Stephanstr. 1a, 04103 Leipzig, Germany. Fax: +49 341 9940 2499.
E-mail address: kreitewolf@cbs.mpg.de (J. Kreitewolf).

1053-8119/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2014.01.005

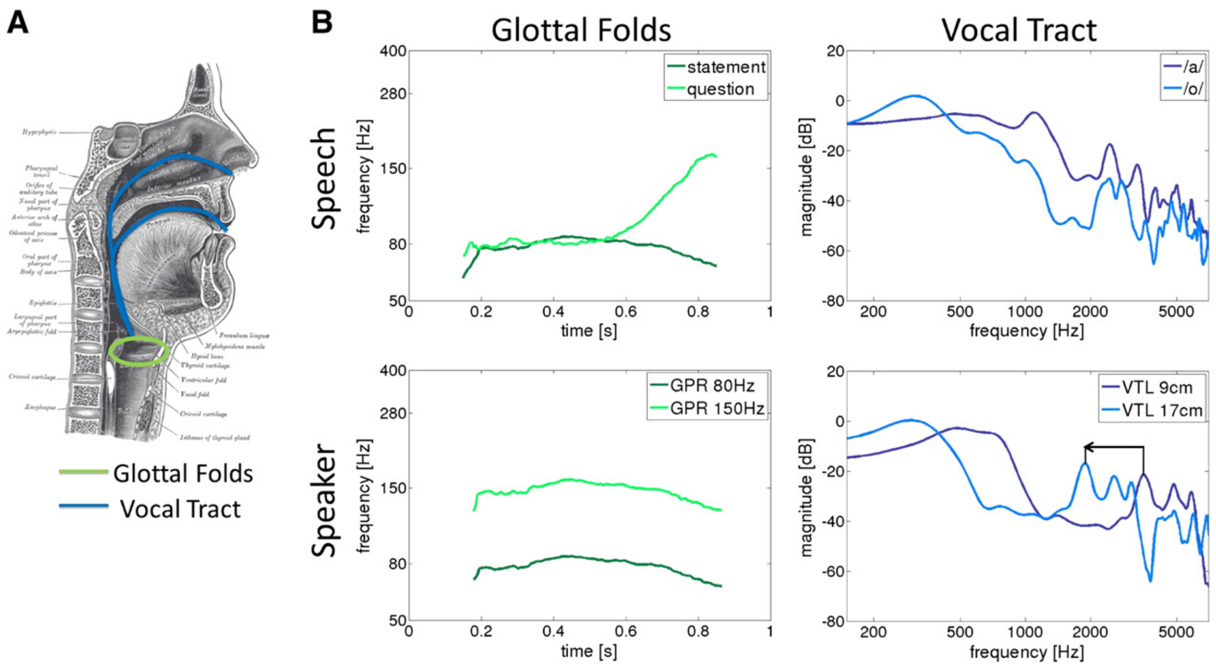

Fig. 1. A. Sagittal section through a human head and neck. Green circle, glottal folds; blue lines, extension of the vocal tract from glottal folds to tip of the nose and lips. B. The plots represent the contribution of glottal fold (left column) and vocal tract parameters (right column) to speech (top row) as well as speaker recognition (bottom row). For glottal fold parameters (left column), frequency is plotted against time on a semi-logarithmic scale. Dynamic variations of glottal pulse rate (GPR) over the course of an utterance determine linguistic prosody (such as whether the speech signal is a question or a statement) (top left). The fundamental frequency (*f*0) contour (i.e., pitch trajectory) of a question is rising at the end of the utterance, whereas the *f*0 contour of a statement is falling. The average GPR over time (bottom left), in contrast, provides information about the voice height (i.e., voice pitch) of the speaker which can be used for speaker recognition (Gaudrain et al., 2009; Lavner et al., 2000). For a higher-pitched voice, the *f*0 contour shifts towards higher frequencies. For vocal tract parameters (right column), magnitude is plotted against frequency; frequency is plotted on a logarithmic scale. Dynamic variations of the vocal tract (i.e., movement of the articulators) determine which speech sound is uttered by producing a different pattern of formants (i.e., peaks) in the spectral envelope (top right). In contrast, the anatomic features of the vocal tract, such as the vocal tract length, determine the timbre of the voice. For a longer vocal tract, formant positions are shifted towards lower frequency values (as indicated by the arrow; bottom right).

be important since it would suggest that such interactions are not only restricted to one acoustic parameter in speech but represent a general feature of how the brain deals with acoustic speaker variability during speech processing.

We employed an fMRI design in which participants recognized linguistic prosody from speakers who differed only in their average GPR (Fig. 2A; 'prosody task/GPR change'). We used syllables spoken by a single speaker and selectively manipulated their average GPR to induce a perception of speaker change (Gaudrain et al., 2009; Lavner et al., 2000). We will call this 'GPR change' in the following. Furthermore, the syllables were resynthesized with pitch trajectories typical of either question or statement intonation (i.e., with rising or falling pitch) to test recognition of linguistic prosody. We used sophisticated vocoder software (Kawahara et al., 2008) to ensure that the speaker changes as well as the linguistic prosody was determined by GPR information only, while controlling for all other acoustic parameters. Stimuli were concatenated into sequences of six syllables, and after each syllable sequence, blood oxygen level-dependent (BOLD) responses were measured using fMRI. Participants were asked to report whether or not a presented syllable had a different linguistic prosody than the previous syllable (1-back prosody task); concomitantly, speakers changed in average GPR (GPR change) (Fig. 2A). In this condition, both prosody information and speaker information were encoded by the same anatomically defined acoustic parameter, namely GPR. In order to differentiate between questions and statements in this condition, participants had to disentangle speech- and speaker-specific GPR information; that is, GPR variation over the course of the syllable for prosody and average GPR for speaker identity. As control conditions, the experiment also included syllable sequences in which speaker changes were induced by a manipulation of vocal tract length instead of GPR (VTL change) (Fig. 1B, bottom right; Fig. 2B), and a control task in which participants had to report whether or not a presented syllable was spoken by a different speaker than the previous syllable (1-back speaker task) (Fig. 2). Importantly, the same syllable sequences were presented in the prosody and control tasks. In summary, the experiment had a 2 × 2 factorial design with the factors task (prosody vs. speaker task) and speaker change (GPR change vs. VTL change). This means that the prosody task was performed while speakers changed in either average GPR (prosody task/GPR change; Fig. 2A, top left) or VTL (prosody task/VTL change; Fig. 2A, top right). The speaker task required to focus on changes in speaker identity that were either solely induced by changes in average GPR (speaker task/GPR change; Fig. 2A, bottom left) or changes in VTL (speaker task/VTL change; Fig. 2A, bottom right). Since the aim of this study was to localize brain regions involved in recognition of GPR-based linguistic prosody from speakers who differ in average GPR, the contrast of interest was defined by the task × speaker change interaction ([(prosody task / GPR change − speaker task / GPR change) − (prosody task / VTL change − speaker task / VTL change)]; Fig. 2A). The rationale behind this procedure was to ensure that the observed BOLD response is specific to the recognition of GPR-based linguistic prosody when speakers differ in average GPR. We employed another type of speaker change (i.e., VTL change) and another task (i.e., speaker task) to control for the possibility that the BOLD response reflects a general activity increase only due to GPR-induced speaker changes or only due to the prosody task.

We hypothesized that (i) right Heschl's gyrus, which is known to process glottal fold parameters, deals with GPR-induced speaker changes during recognition of linguistic prosody; and that (ii) right Heschl's gyrus is functionally connected to its homologous area in the left hemisphere when participants recognize linguistic prosody from speakers who differ in their glottal fold parameters. These hypotheses were based on two strands of evidence from previous work. First, a previous study (von Kriegstein et al., 2010) showed that right posterior STG/STS deals with speaker changes during speech recognition when both types of information are determined by vocal tract parameters. Additionally, functional connectivity analyses showed that this right posterior STG/STS region interacted with a homologous area in the left posterior