NeuroImage 142 (2016) 602–612

Contents lists available at ScienceDirect

NeuroImage

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

Whispering - The hidden side of auditory communication

Sascha Frühholz [a,b,c,d,*], Wiebke Trost [d], Didier Grandjean [d,e]

[a] Department of Psychology, University of Zurich, Zurich 8050, Switzerland
[b] Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich 8057, Switzerland
[c] Center for Integrative Human Physiology (ZIHP), University of Zurich, 8057, Switzerland
[d] Swiss Center for Affective Sciences, University of Geneva, Geneva 1202, Switzerland
[e] Department of Psychology, University of Geneva, Geneva 1205, Switzerland

ARTICLE INFO

Article history:
Received 28 May 2016
Accepted 11 August 2016
Available online 12 August 2016

Keywords:
Voice
Whispering
Emotion
Auditory system
Neural network
fMRI

ABSTRACT

Whispering is a unique expression mode that is specific to auditory communication. Individuals switch their vocalization mode to whispering especially when affected by inner emotions in certain social contexts, such as in intimate relationships or intimidating social interactions. Although this context-dependent whispering is adaptive, whispered voices are acoustically far less rich than phonated voices and thus impose higher hearing and neural auditory decoding demands for recognizing their socio-affective value by listeners. The neural dynamics underlying this recognition especially from whispered voices are largely unknown. Here we show that whispered voices in humans are considerably impoverished as quantified by an entropy measure of spectral acoustic information, and this missing information needs large-scale neural compensation in terms of auditory and cognitive processing. Notably, recognizing the socio-affective information from voices was slightly more difficult from whispered voices, probably based on missing tonal information. While phonated voices elicited extended activity in auditory regions for decoding of relevant tonal and time information and the valence of voices, whispered voices elicited activity in a complex auditory-frontal brain network. Our data suggest that a large-scale multidirectional brain network compensates for the impoverished sound quality of socially meaningful environmental signals to support their accurate recognition and valence attribution.

Introduction

Francis Ford Coppola's film "The Godfather" opens with the famous whispering scene where Bonasera, after some introductory conversation, fearfully whispers a displeasing request to Don Corleone. The scene is a classic example of how individuals switch their vocalization mode to whispering when deeply moved by their emotions in certain social contexts. Whispering has both biological and social functions. Socially, whispering is used to confine communication to listeners in immediate proximity to signal closeness and secrecy, which supports in-group-cohesion (Cirillo and Todt, 2002) and prevents eavesdropping (Morrison and Reiss, 2013), respectively. Sometimes whispering is also used to completely hide one's own affective states and feelings from being recognized by others. Its biologic functions are demonstrated in some animals, which, for example, switch to vocal whispering when confronted with a predator (Morrison and Reiss, 2013) or to initiate sexual behavior (Ladich, 2007). This unique mode of vocal communication is specific to the auditory modality, since no other sensory modality allows a similar kind of qualitative switch in communication.

During auditory communication speakers often adaptively switch from voiced to whispered vocal expression depending on the context. By "voiced" expressions we here refer to the usual or common mode of vocalizations, which to a large extend are based on vocal cord vibrations. Concerning unvoiced whispering, which does not include vocal cord vibrations, speakers not only whisper during normal speech to confine communication to nearby listeners, but they mainly switch from a voiced to a whispering mode to vocally express their inner affective states in particular contexts (Bachorowski and Owren, 2001). For example, although fearful individuals sometimes scream loudly when facing immediate danger, they most often whisper fearfully and vigilantly in threatening situations when the source of threat is not clearly detectable. Furthermore, although individuals usually express aggressive vocalizations in a voiced mode in order to intimidate another individual, they sometimes express their aggression in a low and whispered manner, especially in unfamiliar social contexts.

The switch in expression mode thus has an adaptive function, depending on certain conditions and emotional states. Alhough voiced and whispered vocalizations can express the same emotional states, they are considerably different in terms of their acoustic profile, especially in their spectral acoustic properties (Jovicic, 1998). Whispered vocalizations result from aperiodic and turbulent airflow in the vocal tract, leading to reduced salience of the vocal pitch in whispered voices (i.e.

[*] Corresponding author at: University of Zurich, Department of Psychology, Binzmühlestrasse 14, Box 18, 8050 Zurich, Switzerland.
E-mail address: sascha.fruehholz@uzh.ch (S. Frühholz).

http://dx.doi.org/10.1016/j.neuroimage.2016.08.023
1053-8119/© 2016 Elsevier Inc. All rights reserved.

the breathiness of whispered voices) (Higashikawa et al., 1996), which usually carries important acoustic information about their affective meaning (Banse and Scherer, 1996). Besides vocal pitch, whispered and voiced vocalizations also differ in temporal features (Schwartz, 1967). This reduced acoustic profile of whispered voices, especially a reduced pitch salience, usually imposes strong challenges on the brain of human listeners. We accordingly aimed at providing a combined description, first, of the acoustic properties and, second, of the perceptual and neural network dynamics during the decoding of natural whispered and voiced affective vocalizations. These descriptions are related to our two main research questions. First, can emotions be accurately perceived in whispered voices? Second, are there similar or different neural mechanisms for decoding emotions in voiced and whispered vocalizations given that especially the latter portray only limited acoustic information?

Concerning the latter question, the neural decoding of emotions conveyed by voices predominantly, but not exclusively, involves a neural network consisting of the lateral and medial frontal cortex as well as the auditory cortex (Frühholz and Grandjean, 2013a, b; Frühholz et al., 2014), with a strong structural (Frühholz et al., 2015a) and functional connectivity (Frühholz and Grandjean, 2012) between these regions (Ethofer et al., 2012). The functional role of the auditory cortex has been proposed to underlie the acoustic analysis of physical voice features (Frühholz et al., 2012; Wiethoff et al., 2008) and the perceptual integration of voice features into an acoustic percept (Frühholz et al., 2012). Given a sufficient amount of relevant acoustic voice information, such as the level and the temporal dynamics of vocal pitch, pitch salience, intensity, or the harmonics-to-noise ratio (Frühholz et al., 2016b; von Kriegstein et al., 2010), and given the extraction of these voice features in the auditory cortex (Lewis et al., 2012; Lewis et al., 2009; Patterson et al., 2002; Penagos et al., 2004), the auditory cortex might also perform some generic emotional analysis on these voice features and the voice percept without functional support from other brain regions (Frühholz et al., 2016a). This case might be expected for the neural processing of voiced emotional vocalizations given their wide range of distinctive and discriminative acoustic features (Banse and Scherer, 1996).

However, in case of the impoverished sound quality of whispered vocalizations, the extraction of the available acoustic feature information in the auditory cortex might not be sufficient for an accurate emotional classification of emotional vocalizations, and this more challenging decoding might be supported by additional activation in an extended brain network, especially consisting of frontal brain regions. The inferior frontal cortex (IFC) shows activity and functional connectivity to the auditory cortex if acoustic cue salience in emotional voices decreases (Leitman et al., 2010), pointing to an enhanced cognitive evaluation in the IFC under challenging acoustic conditions (Frühholz and Grandjean, 2013b). The IFC might also enrich the perception of impoverished sounds by retrieving acoustic memory information from long-term, stored prototype emotional vocalizations (Binder et al., 2009). In the present study, we accordingly aimed at investigating the neural dynamics of processing voiced and whispered emotional vocalizations. We specifically investigated the acoustic decoding of these vocalizations in auditory cortical regions that are sensitive to spectral and temporal information of sounds as well as in an extended neural network in the frontal cortex that might provide in-depth cognitive evaluation to compensate for the lack of sensory sound information especially in whispered voices.

Materials and methods

Participants

Fifteen healthy participants recruited from the Geneva University took part in the experiment (seven male; mean age 23.67 years, $SD = 3.87$, age range 18–33 years). All participants were right-handed, had normal or corrected-to-normal vision, and normal hearing abilities. No participant presented a neurologic or psychiatric history. All participants gave informed and written consent for their participation in accordance with ethical and data security guidelines of the University of Geneva. The study was approved by the local ethics committee of the University of Geneva.

Stimulus material and trial sequence

The stimulus material consisted of four speech-like but semantically meaningless two-syllable words ("*belam*", "*nolan*", "*minad*", "*namil*") spoken either in a neutral, fearful or aggressive tone (factor emotion) by two male and two female speakers including both normally voiced vocalizations and whispered vocalizations (factor phonation type), resulting in 96 different stimuli. The whispered vocalizations were produced only by exhalation from the lungs. Auditory stimuli had a mean duration of 633 ms ($SD = 172$ ms) with similar duration for neutral, aggressive and fearful voices (repeated measures ANOVA (rmANOVA); $F_{1,31} = 1.48$, *n.s.*). Stimuli were equated for mean energy ($M_{\text{erg}} = 2.98 \times 10^{-3}$ Pa^2/s, $SD_{\text{erg}} = 9.81 \times 10^{-4}$), scaled to mean sound pressure level of 70 dB, and had a linear fade-in/fade-out of 15 ms. A pre-evaluation of the stimuli by 21 participants (ten male; mean age 25.57 years, $SD = 3.58$, age range 22–34 years) revealed that neutral, aggressive and fearful voices were significantly rated as neutral (rmANOVA; $F_{4,124} = 114.86$, $p < 0.001$), aggressive ($F_{4,124} = 92.30$, $p < 0.001$), and fearful ($F_{4,124} = 192.53$, $p < 0.001$), respectively. Each stimulus had to be rated on five continuous scales (range 0–100, corresponding to low to high) on how much they expressed the respective emotion (fearful, aggressive, happy, sad, and neutral). Each stimulus was also rated for its arousal level (range 0–100, corresponding to low to high), and aggressive and fearful voices did not differ in arousal ratings (paired *t*-test; $t_{31} = 0.76$, *n.s.*), but were significantly higher in arousal ratings compared with neutral voices (aggressive: $t_{31} = 10.13$, $p < 0.001$; fearful: $t_{31} = 15.47$, $p < 0.001$).

During the main experiment, each vocalization was repeated twelve times, and they were presented in short blocks of six voices separated by 500 ms within these blocks, while the short blocks were followed by 4000 ms of no auditory stimulation. Blocks were preceded by a fixation cross of 1000 ± 175 ms duration, which cued the onset of a new block and remained on the screen until the offset of the last auditory stimulus in a block. The short blocks had a mean duration of 6359 ms (*SD* 589) and consisted of randomly chosen vocalizations of the same valence with no more than two times the same speaker or word presented in one block. After each short block participants had to indicate the valence of the vocalizations by a three alternative forced choice decision ("neutral", "aggressive", or "fearful"; response buttons counterbalanced across participants) using the index, middle, and ring finger of their right hand. The experiment was divided in two runs, and each run started either with 48 short blocks of voiced vocalizations followed by 48 blocks of whispered vocalizations, or vice versa (the order was counterbalanced across participants). Voiced and whispered vocalizations were not presented intermixed in order to avoid carry-over effects of perceptual processing from voiced vocalizations that might have influenced the perception of subsequent whispered vocalizations, or vice versa. During scanning, both voiced and whispered vocalizations were presented binaurally with magnetic resonance imaging-compatible headphones (Sensimetrics® insert earphones; http://www.sens.com/products/model-s14/) at a sound pressure level (SPL) of approximately 70 dB.

To localize human voice-sensitive regions in the bilateral superior temporal cortex (STC), we used 500 ms sound clips consisting of 70 nonhuman vocalizations and sounds (animal vocalizations, artificial sounds, natural sounds) and 70 human speech and nonspeech vocalizations presented at 70 dB SPL. The stimuli were the same as used by Capilla and colleagues (Capilla et al., 2013). Each sound clip was presented randomly once with a fixation cross preceding the onset by