ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 5aSCb: Production and Perception II: The Speech Segment (Poster Session)

5aSCb17. Cortical hemodynamic response patterns to normal and whispered speech
Gerard B. Remijn*, Mitsuru Kikuchi, Yuko Yosimura, Sanae Ueno, Kiyomi Shitamichi and Yoshio Minabe

*Corresponding author's address: International Education Center / Center for Applied Perceptual Research, Kyushu University, Fukuoka, 815-8545, Fukuoka, Japan, remijn@design.kyushu-u.ac.jp

Whispered speech is often used in direct person-to-person communication as a means to confidentiality. Compared with normally-vocalized speech, whispered speech is predominantly unvoiced, i.e., produced without vocal fold vibration, and has no clear fundamental frequency. By using near-infrared spectroscopy (NIRS), we assessed cortical hemodynamic response patterns to normally-vocalized and whispered speech in adult listeners (n=13). Stimuli consisted of 20-s strings of Japanese word associations spoken by a female voice. Average oxygenated hemoglobin values (oxy-Hb) were obtained over two regions of interest (ROIs). Oxy-Hb values during the perception of normally-vocalized speech were highest over the left temporal ROI, but not significantly different from values measured over other ROIs. Oxy-Hb values during whispered speech were highest over the right temporal ROI and significantly higher (p<0.05) than those obtained over the left temporal ROI. No significant differences, however, were found in oxy-Hb comparisons between normally-vocalized and whispered speech, although the right temporal ROI comparison bordered on significance, with whisper inducing the higher value. Together, the results seem to suggest that whispered speech is a potent catalyst of cortical hemodynamic activity, especially over the right temporal cortex, in spite of its relatively modest sound level as compared to normal speech.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Whisper is a deliberately degraded speech signal that is commonly used in highly personalized communication. Whispered speech, or unvoiced speech, is produced by forcing air through a constricted opening between the vocal folds and larynx (Vestergaard & Patterson, 2009; Swerdlin, Smith, & Wolfe, 2010). The air that passes through the vocal tract generates a turbulent noise that can be modified by the vocal tract resonances. In whispered speech, the natural spectral variations created by the vocal tract are thus preserved, but the lack of vocal fold vibration modifies the speech sound in a number of ways. Serving its main purpose in human communication, whisper is typically 15-20 dB softer than normally-vocalized speech (Traunmüller & Eriksson, 2000). It further has a spectral tilt of approximately +6 dB/octave (Schwartz, 1970), more elongated syllables than normal speech (Schwartz, 1967), and an absent fundamental frequency (Tartter, 1989).

The acoustic features of whispered speech influence its intelligibility. Whisper is easily masked by noise in daily life (Cirillo, 2004) and laboratory situations, although a robust release from masking is seen when the noise is amplitude-modulated, even at high signal-to-noise ratios (Freyman, Griffin, & Oxenham, 2012). In quiet, identification of whispered vowels has a 10-15% fall off compared to the identification of normally-vocalized vowels (Kallайл & Emanuel, 1984; Tartter, 1991), while correct identification of whispered consonants is also considerably low (Tartter, 1991). Correct identification of speaker gender is almost 20% worse for whispered speech as compared to normally-vocalized speech (Lass et al., 1976; although see Schwartz, 1968).

Few studies have focused on the cortical activity during the perception of whispered (or unvoiced) speech. So far, studies that investigated the neural correlates of unvoiced speech perception featured speech fragments of normally-vocalized (or voiced) speech. One such example is the study of voice-onset time (VOT). Electrophysiological studies have shown that the difference between voiced and unvoiced VOT is expressed in the amplitude of the N1 component in auditory evoked potentials, known to reflect the basic encoding of acoustic information in the auditory cortex (e.g., Näätänen & Picton, 1987). Voiced VOT syllables generate larger N1 amplitudes than unvoiced VOT syllables (Simos et al., 1998; Sharma et al., 2000; Zaehle et al., 2007). Neurophysiological studies on the perception of fully unvoiced words or sentences, however, have not yet been performed.

In the present study, we used near-infrared spectroscopy (NIRS) to assess cortical hemodynamics in adult listeners in response to whispered and normally-vocalized word strings. NIRS enables assessment of changes in cerebral blood flow in response to sensory stimulation, and changes in oxygenated hemoglobin (oxy-Hb) values in particular provide an indirect indication of local cortical activation to a stimulus (Devor et al., 2005). The speech stimuli consisted of word associations, which were alternately presented in 20-s strings of either whispered or normally-vocalized speech (Figure 1). A frontal region and a temporal region were selected as regions of interest (ROIs). Both regions are involved in the processing of the acoustic properties of speech, as found in studies that assessed either electrophysiological or hemodynamic response patterns to speech masked by noise (e.g., Martin et al., 1997; Whitling et al., 1998; Muller-Gass et al., 2001; Kozou et al., 2005), or acoustically degraded speech (e.g., Liikanen et al., 2007; Miettinen et al., 2010). The main purpose of the experiment was to explore whether cortical processing of the differences in the acoustic properties between whispered, i.e., (deliberately) acoustically-degraded speech, and normally-vocalized speech could be expressed in the listener’s hemodynamic response. Since acoustically degraded speech tends to cause increased electrophysiological activity that is lateralized toward the right hemisphere, as compared to normally-vocalized speech (Liikanen et al., 2007), we tentatively expected that hemodynamic activity patterns for both speech modes used here would differ most profoundly in the right hemisphere as well.

METHOD

Participants

Thirteen males participated in the experiment. They were students of Kanazawa University, Japan, and employees of Yokogawa Electric Inc, Kanazawa, Japan, aged 22 – 45 years. All participants were right-handed and had normal hearing. Informed consent was obtained from all participants after an explanation about the workings of the NIRS equipment and the procedure of the experiment. The procedure was approved by the Ethics Committee of Kanazawa University Hospital and followed the Declaration of Helsinki.
Stimuli

The stimuli (Figure 1) consisted of Japanese word associations, taken from a rhyme book aimed at preschool children. The words were spoken by a Japanese-native female, either in a normal voice (N) or in a whispering voice (w). The same word associations were used for both speech modes. The word association strings were alternately presented in three epochs of 20 s per speech mode, comprising a 120-s stimulus in total (Figure 1D). The starting speech mode was counterbalanced among participants (either NwNwNw or wNwNwN).

Near-infrared Spectroscopy (NIRS) Measurements

NIRS measurements were made with a continuous wave system from Shimadzu Kyoto, Japan (Foire 3000), with two 4×3 sets of optode probes. The set over the left and right hemisphere each consisted of 6 light emitters and 6 detectors. Both sets were placed in a diadem-like, hard-plastic frame (Yokogawa Electric Inc / Shimadzu Kyoto). Each set comprised 17 channels and was placed symmetrically on either side of the head. Following the international 10-20 system for electro-encephalography and NIRS optode placement (Okamoto et al., 2004), placement was such that a) the fourth anterior channel column covered the imaginary line between electrode positions C3-T3 for the left hemisphere and C4-T4 for the right, and b) the bottom channel row covered the imaginary line over F7-T3-T5 for the left hemisphere and F8-T4-T6 for the right. Over each hemisphere, 7 channels were included in the frontal ROI, and 7 in the temporal ROI. One channel was included in both ROIs. NIRS measurements were made with a sampling rate of 10 Hz, and commenced after a resting period of two minutes to stabilize the initial hemodynamic response.

Stimulus Presentation and Procedure

The stimuli were presented to the participant in a magnetically-shielded room, otherwise used for magnetoencephalography measurements. The listener was asked to passively listen to the speech stimuli, and to maintain a fixed bodily position during NIRS measurements. The stimuli were presented through an amplifier (Denon PMA-
390IV) and headphones (Panasonic RP-HT260). The level of the normal-speech stimuli was 66 dBA (fast-peak, A-weighted) on average, within a range of 56-73 dBA, as measured with a Yokogawa M&C S-54 sound level meter. The level of the whisper stimuli was 57 dBA on average, within a range of 54-65 dBA. Average peak-intensity levels obtained at 30 consecutive time windows of 2 s showed that normally-vocalized speech (67.27±0.17 dB) was significantly more intense than whispered speech [62.42±0.21 dB; unpaired t(29)=18.24, p<0.001]. The background noise in the room was 39.3 dBA.

Data Treatment and Statistical Analyses

The raw oxy-Hb data were visually inspected and digitally low-pass filtered at 0.1 Hz to remove noise. Linear baseline correction of the raw oxy-Hb data was conducted using the mean of the 5-sec period before a target epoch. Oxy-Hb data for each ROI (frontal and temporal) on both hemispheres were averaged for each participant and speech mode. Averaging was done over the middle 10 s of each of the three 20-s epochs per speech mode, i.e., oxy-Hb values for the first and last 5 s were not included in the analyses. Repeated-measures Analysis of Variance (ANOVA) was performed for each of the ROIs, with speech mode (whisper or normal speech) and hemisphere (left or right) as factors.

RESULTS

The oxy-Hb values for n=13, as averaged over the middle 10 s of the stimuli, are shown in Figure 2. Overall, normally-vocalized speech induced the most pronounced increase in oxy-Hb over the left temporal ROI. By contrast, whisper induced the most pronounced oxy-Hb increase over the right temporal ROI. ANOVA for oxy-Hb obtained over the temporal ROI, however, showed no significant main effect of speech mode \([F(1,12)=0.001, p=0.99]\). The main effect of hemisphere was also not significant \([F(1,12)=1.149, p=0.31]\). The speech mode \(\times\) hemisphere interaction, however, was significant \([F(1,12)=4.871, p=0.05]\), in that whispered speech induced a higher oxy-Hb magnitude over the right temporal ROI than over the left. ANOVA for oxy-Hb obtained over the frontal ROI revealed no significant effects. The effects of speech mode \([F(1,12)=0.667, p=0.43]\), hemisphere \([F(1,12)=0.144, p=0.71]\), and their interaction \([F(1,12)=0.146, p=0.71]\) were not significant.

\(\text{FIGURE 2.}\) Average oxy-Hb values obtained in the experiment \((n=13)\). The oxy-Hb difference between whispered speech (gray) and normally-vocalized speech (white) neither reached significance over the temporal ROI (left half of the figure), nor over the frontal ROI (right half). Whispered speech induced significantly higher oxy-Hb \((\ast; p<0.05)\) over the right temporal ROI as compared to the left temporal ROI. Oxy-Hb obtained over the frontal ROI did not show any significant within-speech mode lateralization. (ROI, Region of Interest; oxy-Hb, oxygenated hemoglobin; mMol, millimole; mm, millimeter).
DISCUSSION

Some electrophysiological studies have shown that compared to normal speech, cortical responses to acoustically degraded speech tend to be more pronounced over the right hemisphere (Liikkanen et al., 2007). We found a similar trend in the cortical hemodynamic data obtained in the present experiment, although the oxy-Hb difference between both speech modes merely bordered on significance for the right temporal ROI (p=0.07), where whisper induced higher oxy-Hb values than normal speech. In a related study on speech processing in preschool children, the oxy-Hb difference between whisper and normal speech did reach significance over the frontal ROI (Remijn et al., submitted), with the oxy-Hb response to whisper being the larger response. We hypothesized that the hemodynamic responsiveness to whisper may have partly reflected increased effort in processing its “degraded” acoustic properties. With regard to the present data, it is possible that the processing of whispered speech required relatively less effort in adult participants, who have a more fully-developed automaticity in their processing of speech than preschoolers. Hence, the oxy-Hb difference between speech modes did not reach significance here.

A significant lateralization difference was found, however, within whisper-induced oxy-Hb. Whispered speech caused a significantly higher oxy-Hb over the right temporal ROI than over the left temporal ROI, which showed rather atypically high oxy-Hb values below zero. It is plausible that the rightward oxy-Hb increase during whisper reflects differences in hemispheric processing, especially with regard to the heavy distortion of the fine temporal structure in whispered speech. Electrophysiological studies have shown that processing of speech with a degraded amplitude resolution generally is rightward biased (e.g., Miettinen et al., 2011; Miettinen et al., 2012). To investigate such potential acoustics-related influences on whisper-induced lateralization, if validated, neurophysiological assessment with finer temporal acuity than NIRS is necessary.

An alternative possibility is that the rightward lateralized response to whisper concerns processing of the speaker’s intent. Whisper is not often used in daily life and, if so, often in highly personalized communication that requires “mind-reading” of the speaker’s confidentiality. Although word associations with neutral content were employed in the present study, the male participants may have shown increased attentional engagement during the whispering female voice. Not only attentional engagement to whisper’s acoustic properties, but also to the special role of whisper in human interaction thus could have influenced the hemodynamic response, since increased attentional engagement to sound in general increases cortical oxy-Hb (e.g., Remijn & Kojima, 2010). This explanation may also fit in with the studies showing that cortical frontotemporal activity during theory-of-mind tasks tends to be lateralized toward the right (e.g., Griffin et al., 2006). Future studies that control for participant gender and/or affective content might be employed. In any regard, the present study shows an example of a relatively prominent cortical response toward a relatively weak sensory stimulus. Since physically weak stimuli tend to induce weak cortical responses (e.g., for speech signals see Tremblay et al., 2006), the present rightward laterality may not be solely caused by the processing of whisper’s acoustic properties.

ACKNOWLEDGMENTS

Supported by grant 22591276 from the Japan Society for the Promotion of Science to GR, and the Hokuriku Innovation Cluster for Health Science (MEXT Program for Fostering Regional Innovation from the Japanese Ministry of Education, Culture, Sports, Science, and Technology.

REFERENCES


