The voice-specific process revealed by neuromagnetic responses

Atsuko Gunji^{1,2,3}, Daniel Levy^{4}, Ryouhei Ishii^{1}, Ryusuke Kakigi^{3} and Christo Pantev^{1}

^{1} The Rotman Research Institute for Neuroscience, Baycrest Centre for Geriatric Care, Toronto, Canada.
^{2} Research Fellow of the Japan Society for the Promotion of Science, Tokyo, Japan.
^{3} Dept. of Integrative Physiology, National Institute for Physiological Sciences, Myodaiji, Okazaki, Japan.
^{4} Department of Psychology, The Hebrew University of Jerusalem, Mt. Scopus, Jerusalem, Israel.

Abstract

Neuroimaging studies may provide evidence for perceptual specificity elicited by human voice. We tried to identify the voice specific activities with high spatial and temporal resolution using whole-head magnetoencephalography (MEG). Volunteers were instructed to listen to sung and corresponding instrumental sounds matched in fundamental-frequency. The stimuli were 16 acoustically different sounds, comprising eight types, which consisting of sounds produced by four singers and four musical instruments at each of two fundamental frequencies: 220 Hz (musical note A3) and 261.9 Hz (C3). In both types of stimuli, two components of evoked responses were recorded at approximately 100 and 400 ms after the stimulus onset, respectively. The source locations of equivalent current dipoles (ECDs) for both components were estimated around the superior temporal sulcus in both hemispheres. Compared with the instrument, the RMS and source strength for the voice were significantly larger at approximately 100 ms ($p<0.05$). The operation of a gating system directing human voice stimuli might be processed differently as compared with other auditory stimuli.

1 Introduction

The perception of speaker-related features of the voice plays a major role in human communication. Functional magnetic resonance imaging studies (fMRI) [1, 2] reported bilateral voice-selective activities along the upper bank of the superior temporal sulcus (STS). These regions showed greater activation to vocal sounds than to non-vocal environment sounds, when subjects passively listened to the sound stimuli.

Although these neuroimaging studies have demonstrated brain regions engaged in the voice specific processes, they cannot characterize the temporal relationship of the activities in these regions. Using electroencephalography (EEG) measurements, Levy et al. [3] compared event related potential (ERP) to human voices with those of instrumental sounds. The authors reported an ERP component at 320 ms after the stimulus, specific for the human voice. This positive EEG component was larger for the human voice than for corresponding instrument sound. Therefore, the authors speculate that like face recognition in the visual system, this component might be reflecting a process specific to human voice.

Since MEG has better spatial resolution as compared to EEG we asked the question *where* and *when* is the response to human voice coded in the brain? The objective of the present study was to evaluate the spatio-temporal activity related to the voice-sensitive region in the STS. By means of whole-head MEG we measured the changes of brain activity related to the perception of voice and non-voice sounds, which were matched for pitch, duration and amplitude.

2 Methods

2.1 Subjects

The subjects were five normal right-handed volunteers (three males and two females; mean age 30 years, range 27-35 years).

2.2 Stimuli

To make a direct comparison with results reported by Levy et al. [3], the same exact stimuli have been used. They were 16 acoustically different sounds, comprising stimuli from two categories: (1) human voice sounds produced by four singers of different genders (mezzo soprano, alto, bass and baritone), and (2) musical instrument sounds produced by four strings (violin, viola, cello and bass), at each of two fundamental frequencies: A3 (220Hz) and C4 (261.9Hz). All stimuli were edited to yield equivalent root mean square (RMS) power using the CoolEdit 2000 sound editor (Syntrillium Software Corporation). The duration of the stimuli was 500 ms including an envelope of 10 ms rise and fall times.

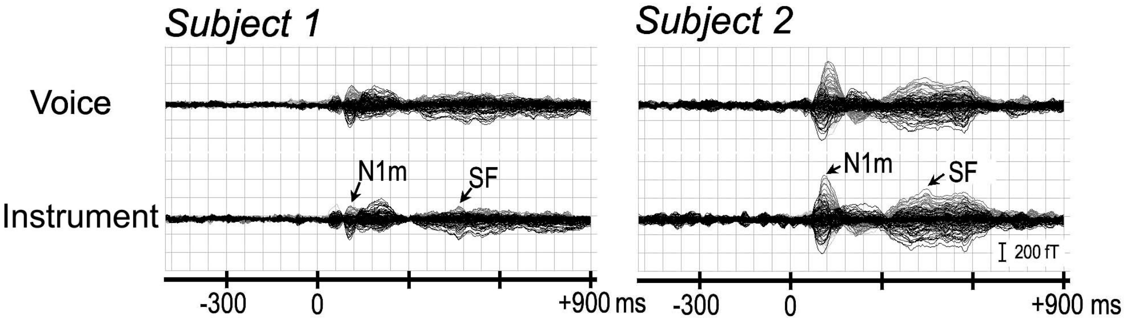

Fig. 1 Magnetic responses recorded for the voice and the instrument in two subjects. The waveforms recorded by 151 channels were superimposed. Two peaks of reversed-phase detection were recognized approximately 100 ms (N1m) and 400 ms (SF) after the stimulus onset for both sounds.

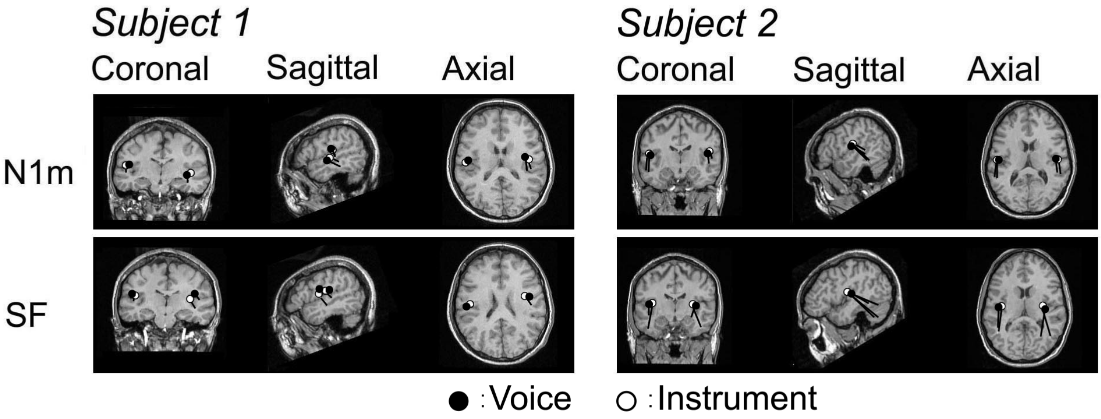

Fig. 2 Location of ECDs of N1m and SF components for the voice (filled circle) and the instrument (opened circle) in two subjects. Magnetic resonance imaging (MRI) scans were overlaid with the ECD sources detected by the components related to the voice and instrument sounds. The same anatomical landmarks (nasion and center points of the entrance to the bilateral ear canal) were used to create the MEG head-based 3D coordinate system. The landmarks were visualized in the MRI by affixing marker coils to these points.

2.3 Procedures

The stimuli were delivered to the subject’s ears at 60 dB HL through a pair of plastic tubes and ear-pieces with random inter-stimulus interval between 1000 and 2000 ms. In order to prevent the perception of a pseudo-melody, the stimuli were presented in separate blocks for the same fundamental frequency. They were presented in a random order in each block. The number of trials was 200 for each stimulus category (voice and instrument). There were 25 trials of four voices and 25 trials of four instruments in each of the two fundamental frequencies. Subjects were instructed to watch a silent film and do not pay attention to the auditory stimulation.

2.4 MEG recordings

Magnetic response was measured in a magnetically shielded room using a helmet-shaped 151-channel SQUID sensor array (Omega 151CTF Systems Inc.), equipped with axial first-order gradiometers.

The magnetic responses were filtered using 60 Hz notch filter and 100 Hz low pass filter and digitized at 312.5 Hz.

2.5 Data analysis

The baseline was corrected (DC offsets) for each channel according to the mean value of the signal before the stimulus onset. Epochs with eye movement and other artifacts were rejected before averaging. Stimulus related epochs of 500 ms before and 900 ms after the stimulus onset were averaged for each category.

The RMS value has been calculated for the voice and the instrument sounds in the peak latency for N1m component. The corresponding equivalent current dipoles (ECDs) were estimated at the RMS peak latency. For the sustained field (SF), the mean RMS has been calculated between 350 ms and 500 ms after the stimulus onset, and a moving ECD dipole model was used to estimate the source in the same latency range. Using Wilcoxon singed-ranks test, the RMS and the dipole moment values were compared between the voice and the instrument for each hemisphere.

3 Results

More than 165 epochs free of artifacts were collected for each stimulus condition and each subject. For mental stimuli for each subject all subjects, two clear components of the auditory evoked field were obtained at approximately 100 ms (N1m) and 400 ms (SF) after the stimulus onset, respectively (Figure 1).

The maximum RMS value of N1m component was 90.0 ± 51.7 fT (mean ± SD) for the voice and 83.0 ± 47.5 fT (mean ± SD) for the instrument. The RMS value was larger for the human voice than for the instrument sound ($p<0.05$). For the RMS value of the SF, there was no difference between the voice and the instrumental sounds.

The ECDs of both components were located in the vicinity of the superior temporal sulcus (STS) in each hemisphere (Figure 2). The mean of the residual variance was 8.39%. Compared to the instrumental sound, the N1m source strength for the voice was significantly larger ($p<0.05$). For the SF there was no significant difference between the voice and instrumental sounds.

4. Discussion

In this study we presented initial evidence that N1m amplitude of AEF might reflect the two sounds stimuli categories. The amplitude if human voice was larger when compared to instrumental sound. This result corresponds to the fMRI findings [1, 2] and might represent the increased neuronal activity at about 100 ms after the stimulus onset.

Eulitz et al. [4] reported that RMS of SF relating to human vowel sound was larger than that to pure tone. In our study, however, the RMS and source strength values seen for SF are similar between the voice and instrumental sounds. There are probably two reasons for this difference. The first one is that both, the voice and instrumental sounds in our study were complex tones, matched in fundamental frequency, having much higher degree of similarity than the spoken vowels and the pure tones in the study of Eulitz et al [4]. The second one is that the voice stimuli in our study have a less pronounced language character than the spoken vowels in the Eulitz et al. [4] study.

The results previously obtained with electric recordings [3] suggested that there are voice-specific processes about 320 ms after the stimulus onset. We cannot confirm this from our initial MEG data, however, the difference of N1m was remarkably pronounced between the voice and instrument sounds. Simultaneous MEG/EEG measurements, running currently in our laboratory, may help to answer the question whether MEG and EEG do not reflect different brain events.

5 References

[1] Belin, P., Zatorre RJ., Lafaille, P. et al.: Voice-selective areas in human auditory cortex sounds. Nature 403 (2000) 309-312.

[2] Binder, JR., Frost JA., Hammeke TA. et al.: Human temporal lobe activation by speech and non-speech sounds. Cerebr Cortex 10 (2000) 512-528.

[3] Levy, DA., Granot, R., Bentin, S.: Processing specificity for human voice stimuli: electrophysiological evidence. NeuroReport 12 (2001) 2653-2657.

[4] Eulitz, C., Diesch, E., Pantev, C. et al.: Magnetic and electric brain activity evoked by the procesing of tone and vowel stimuli. J Neurosci 15 (1995) 2748-2755.