Perceptual categorization of sound spectral envelopes reflected in auditory-evoked N1m

Tomomi Mizuochi,[1,CA] Masato Yumoto,[2] Shotaro Karino,[3] Kenji Itoh,[4] Keiko Yamakawa[4] and Kimitaka Kaga[1,3]

[1]Departments of Sensory and Motor Neuroscience; [2]Laboratory Medicine; [3]Otolaryngology-Head and Neck Surgery; [4]Cognitive and Speech Science, Graduate School of Medicine, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8655, Japan

[CA]Corresponding Author: mizuochi@m.u-tokyo.ac.jp

Received 17 February 2005; accepted 18 February 2005

Magnetic responses to periodic complex sounds with equivalent acoustic parameters except for two different fundamental frequencies (F_0) and 12 different spectral envelopes of vocal, instrumental, and linear shapes were recorded to determine the cortical representation of timbre categorization in humans. Responses at approximately 100 ms (N1m) to vocal and instrumental (nonlinear) sounds were localized significantly anterior to linear sound responses. N1m source strength for nonlinear sounds was significantly larger than that for linear sounds, and this difference was more marked in the left hemisphere than in the right. N1m peak latency only for vocal sounds was not affected by F_0. Perceptual categorization was reflected in N1m source strength and location (linear or nonlinear), and in N1m latency (vocal or nonvocal). *NeuroReport* 16:555–558 © 2005 Lippincott Williams & Wilkins.

Key words: Auditory-evoked magnetic fields; Formant; Linear spectral sound; Magnetoencephalography; Musical instruments; N1m; Spectral envelope; Timbre; Voice; Vowels

INTRODUCTION

A sound is composed of three major classes of perceptual aspects: pitch, loudness, and timbre [1]. Of these three, timbre is considered to be the most complex and multidimensional aspect, which strongly reflects the fine structure and physics of the sound source. For example, resonance in the vocal organs (formant), which uniquely corresponds to the physical geometry and acoustical properties of the vocal tract, the laryngeal, pharyngeal, and oral cavities, conveys cue information for individual identification (e.g. voiceprint) [2]. In other sound sources, such as musical instruments, similar acoustics are also present and confer unique spectral identities and characteristics on the harmonics of complex sounds, which are otherwise monotonous. The shaping of the spectral envelope by resonance is one of the principal components of timbre in natural sound sources. Timbre is also characterized by the long-term temporal fluctuations of amplitudes or frequencies. These two aspects of timbre need to be differentiated in cognitive research, because the neural substrates for processing of different acoustic features are considered to be different [3].

Several psychological, neurophysiological, and neuroimaging studies on timbre perception have been carried out. Helmholtz [4] recognized that an important index for the identification of spoken vowels and musical instruments is the spectral envelope. The perception of sounds with various spectral envelopes has been studied extensively in the field of psychology [5]. In some neurophysiological studies on the effect of acoustic parameters, such as center frequencies and their tonotopicity, stimuli with a simplified spectral structure were used [6,7]. Conversely, most studies using stimuli adopted from natural sounds [8] were conducted without differentiating the two different aspects of timbre described above. The results of such studies are controversial. Recently, some researchers stated that both the left and right hemispheres are evenly involved in timbre processing [8,9], although other researchers indicated right hemispheric dominance [10,11]. In neurophysiological studies, the latency of cortical activity associated with timbre processing is also controversial. Some researchers found related activities in a latency range of N1 [10,11]; others, however, found the activities in only a latency range of 300–500 ms [8].

Although several studies on timbre processing in humans have been conducted, little is known about how the spectral envelope of periodic complex sounds with and without resonance, which approximately correspond to natural and unnatural sounds, respectively, are represented in the auditory cortex. In the present study, we recorded auditory-evoked N1m (the magnetic counterpart of the N1 potential) response to periodic complex sounds with the same set of spectral frequencies and an identical temporal envelope. The only difference across the stimuli was the shape of the spectral envelope, which is categorized into three groups: vocal, instrumental, and linear sounds. The purpose of this study was to clarify the cortical representation of perceptual boundaries not only between vocal and nonvocal but also between natural (vocal and instrumental) and unnatural (linear) sounds. To this end, temporally degraded vowels were used as vocal sound stimuli, considering that the most frequently encountered voice in

daily life shared spectral contents with vowels. The stimulus duration was defined considering that prolonged sounds without temporal variation sound unnatural even when their spectral envelopes are adopted from nature.

MATERIALS AND METHODS

Four men and four women volunteers aged 24–57 years with no reported hearing deficits took part in the study. All of them were right-handed. They were fully informed of the methods and technique of noninvasive magnetoencephalographic (MEG) recordings before signing a written consent to participate in this study. The study protocol was approved by the Ethics Committee of the University of Tokyo.

Periodic complex sounds used in this study were 12 sounds with a fundamental frequency (F_0) of 110 Hz and 12 sounds with an F_0 of 220 Hz, which were categorized into three groups on the basis of their spectral envelopes: vocal, instrumental, and linear. The vocal sounds were reproduced by a speech synthesizer SMARTTALK (Oki, Tokyo, Japan) to prepare four different spectra with each F_0: the Japanese vowels, /a/, /i/, /o/, and /u/ in male and female voices. The instrumental sounds were sampled from a reed instrument section of a software synthesizer VSC-MP1 (Roland, Osaka, Japan) to prepare four different spectra with each F_0: saxophone, bassoon, clarinet, and oboe on the notes A_2 and A_3, considering their structural similarity to the human vocalization mechanism. The linear spectral sounds were constructed by the Fourier additive synthesis of sine waves with each F_0 to mimic the vocal and instrumental sounds with four different spectral slopes: −6, −8, −10, and −12 dB/oct. The slope range was adopted from a previous psychological study [5]. Using a sound editor, Audition (Adobe Systems, San Jose, California, USA), the vocal and instrumental sounds were transformed into a repeated succession of a single temporal pattern with a period of F_0 extracted from the central portion of the sampled waveforms. This procedure degraded vocal and instrumental sounds to pure tone complexes without temporal fluctuations. Thus, these 24 periodic complex sounds exactly matched the two different F_0's, and their harmonic frequencies varied only in spectral envelope (Fig. 1). These sounds were molded into stimuli with a temporal envelope of 140 ms duration: 10 ms linear rising, 120 ms plateau, and 10 ms linear falling. The intensity of the stimuli was adjusted to yield equivalent root-mean-square (RMS) acoustic power, considering the overall electroacoustic transfer function of the stimulation system used in this study.

The stimuli with the same F_0 (110 and 220 Hz) were presented separately in two blocks. The experimental order of the two blocks was counterbalanced across participants. For each block, the 12 sound stimuli were presented with equal probability in a pseudorandom order, with the stipulation that stimuli in the same category could not appear consecutively. The stimuli were sequenced by the STIM2 system (Neuroscan, El Paso, Texas, USA) and delivered binaurally to the participant's ears at 60 dB sound pressure level through the earphones ER-3A (Etymotic Research, Elk Grove Village, Illinois, USA) at a stimulus onset asynchrony of 990 ms. To maintain a constant vigilance level, the participants were instructed to watch a silent movie projected onto a screen in front of them and not to pay attention to the auditory stimuli. Magnetoencephalography data were recorded in a magnetically shielded

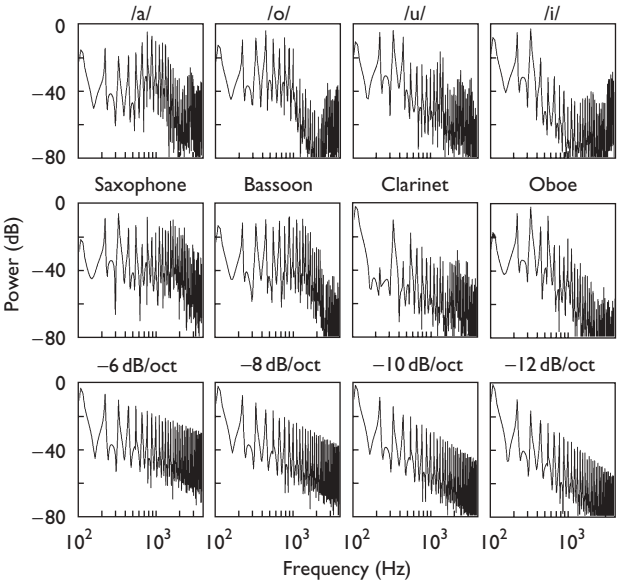

Fig. 1. Spectra of 12 stimuli with an F_0 of 110 Hz. Vocal sound category (top), from left to right: /a/, /o/, /u/, and /i/. Instrumental sound category (middle): saxophone, bassoon, clarinet, and oboe. Linear sound category (bottom): −6, −8, −10, and −12 dB/oct.

room using VectorView (Elekta Neuromag, Helsinki, Finland), which has 204 planar first-order gradiometers at 102 measurement sites on a helmet-shaped surface that covers the entire scalp. Stimulus-related epochs of 550 ms duration (including a 50-ms prestimulus baseline) were filtered with a 0.1–200-Hz band-pass filter and then recorded at a sampling rate of 600 Hz. Epochs with amplitude variations more than 3 pT/cm in any channel were excluded as artifacts. At least 100 artifact-free epochs were recorded per stimulus-type per participant, and epochs for the same sound category (i.e. at least 400 epochs for each category) were averaged together for analysis.

The averaged waveforms were low-pass filtered at 40 Hz and the 50-ms prestimulus baseline was adjusted to zero to remove DC offset. The peak latency of the main response N1m was determined for each hemisphere by the time point at which the RMS of the predefined temporal channels (Fig. 2a) reached the maximum between 70 and 140 ms after stimulus onset. The sources of N1m were modeled separately as a single equivalent current dipole (ECD) in a spherical conductor for each hemisphere. The ECDs were calculated at the peak latencies from the same temporal channels, independently for averaged data of each category in each participant. The estimated ECDs were described in a head-based coordinate system. The x-axis passes through the two preauricular points digitized before data acquisition with a positive direction to the right. The y-axis passes through the nasion and is normal to the x-axis. The z-axis points upward according to the right-handed rule and is normal to the xy-plane. Magnetic resonance imaging scans were obtained from all the participants. T1-weighted coronal, axial, and sagittal images with continuous 1.2-mm-thick slices were adopted for overlays with the ECD sources for neuroanatomical evaluation. The peak latency and ECD parameters (dipole location, orientation, and source strength) were analyzed by repeated measures analysis of variance (ANOVA) with factors of F_0 (110 and 220 Hz), stimulus category (vocal, instrumental, and linear)

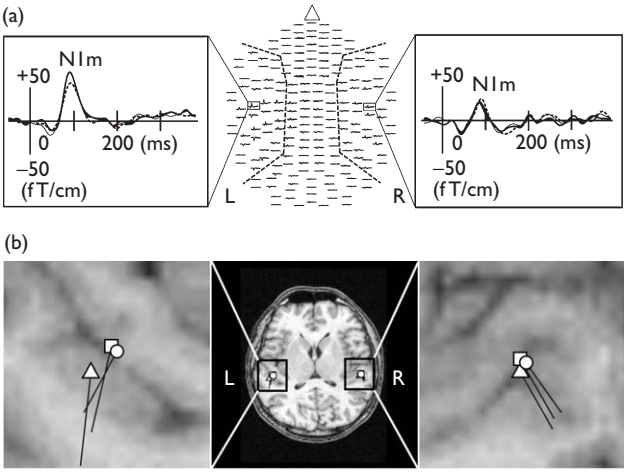

Fig. 2. (**a**) Responses to three stimulus sound categories with an F_0 of 110 Hz in a representative participant. The temporal 44 channels for analysis in each hemisphere are bordered with dashed lines (at the center). The largest responses in each hemisphere are enlarged (bilaterally). Vocal, thick line; instrumental, thin line; linear, dashed line. (**b**) N1m equivalent current dipole locations for vocal (circles), instrumental (squares) and linear (triangles) sound categories on the magnetic resonance image (at the center). Localized areas in each hemisphere are enlarged (bilaterally). Black arrows represent dipole directions.

and hemisphere (left and right). Planned comparisons were carried out using *t*-tests.

RESULTS

In all participants, prominent deflections were clearly detected in the superior temporal areas in both hemispheres, in a latency range of 100 ms (N1m) (Fig. 2a). The ECDs for N1m were calculated with a goodness-of-fit larger than 82% and with the confidence volume smaller than 0.88 cm^3. In all stimulus categories, the ECDs were clustered in the superior temporal plane in both hemispheres (Fig. 2b).

The peak latencies for an F_0 of 110 Hz (mean $\pm$ SEM: 105.6 $\pm$ 1.6 ms) were more delayed [$F(1,7)=24.22$, $p<0.002$] than those for 220 Hz (101.0 $\pm$ 1.4 ms), with a significant dependence on a sound category ($p<0.01$). These delays were significant in the instrumental ($p<0.01$) and linear ($p<0.01$) sounds, but not in the vocal ($p=0.11$) sound (Fig. 3a).

The source strength for the linear sounds (30.4 $\pm$ 1.8 nAm) was smaller [$F(2,7)=5.33$, $p<0.05$] than that for the vocal (36.6 $\pm$ 2.3 nAm) and the instrumental (36.6 $\pm$ 2.6 nAm) sounds, with a significant dependence on the hemisphere ($p<0.05$). The source strength in the left hemisphere was larger for the vocal ($p<0.01$) and the instrumental ($p<0.01$) sounds, but not for the linear ($p=0.14$) sounds, than that in the right hemisphere (Fig. 3b).

The dipole locations were significantly different in only the anteroposterior direction. The ECD locations for the linear sounds (*y*-axis: 7.9 $\pm$ 1.3 mm) were posterior [$F(2,7)=7.08$, $p<0.01$] to those for the vocal (9.5 $\pm$ 1.1 mm) and the instrumental (9.2 $\pm$ 1.3 mm) sounds, with no significant dependence on the hemisphere ($p=0.48$) or the F_0 ($p=0.25$) (Fig. 3c).

DISCUSSION

The age range of the volunteers in the present study is relatively large. While the N1m localization is unaffected by age [12], both the N1 amplitude and latency increase with

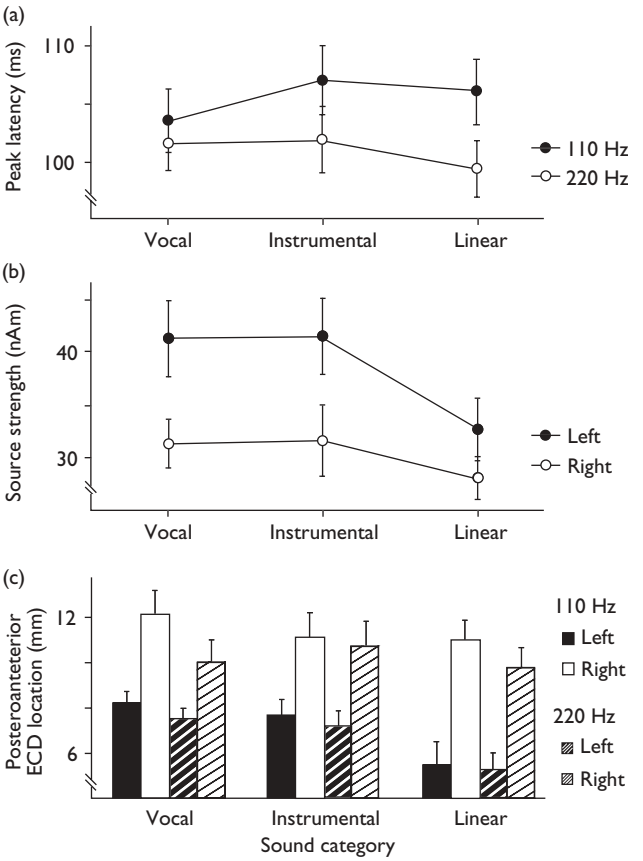

Fig. 3. (**a**) Mean ($\pm$SEM) of N1m peak latency for three sound categories with the F_0s of 110 Hz (filled circle) and 220 Hz (open circle). (**b**) Mean ($\pm$SEM) of N1m source strength for three sound categories in left (filled circle) and right (open circle) hemispheres. (**c**) Mean ($\pm$SEM) of posteroanterior dipole location (*y*-axis) of N1m for three sound categories. In the same sound category, from left to right: locations for 110 Hz sounds in left and right hemispheres, 220 Hz sounds in left and right hemispheres.

age [13]. The mean amplitude and latency of N1m reported in the present study may be affected by age, although the results of repeated measures ANOVA are unaffected by interindividual mean difference. Several investigators reported that the N1 latency tends to decrease as F_0 increases not only for pure tones [14,15] but also for complex sounds [16]. Recently, Mäkelä *et al.* [17] reported that the N1m latency for vocal sounds was relatively stable compared with that for pure tones independent of F_0. In our study, this stability was replicated even across the periodic complex sounds with similar spectral contents. This result suggests that the vocal feature of sounds is processed differently in a latency range of N1m and the perceptual specificity of vocal sounds is predominantly attributable to the static structure of harmonics, that is, spectral envelopes.

A significant difference was not detected between the source strength of N1m for the vocal and that for the instrumental sounds, in agreement with previous findings [8]. The source strength for the linear sounds was, however, significantly smaller than that for the other sounds. Even when a physical scale such as an averaged RMS power is balanced, the power distribution difference across stimuli may result in a difference in the sensation level [18]. Generally, the N1 latency becomes shorter and the amplitude

becomes larger as the sensation level increases. If the source amplitude difference found in this study is due to a sensation level difference, the latency for the instrumental sounds should be shorter than that for the linear sounds, which was not so in the present study. This amplitude difference may reflect the extraction and integration process of spectral information to reach the ultimate percept, considering that only the vocal and the instrumental sounds have the formant structure adopted from natural sounds. Furthermore, the amplitude in the left hemisphere was significantly larger than that in the right hemisphere only for sounds with the formant structure. Generally, the left hemisphere is dominant for speech sounds. The anterior portion of the auditory cortex in the left hemisphere is tuned to broad or multiple frequencies, while that in the right hemisphere is tuned to single frequencies [19]. Furthermore, the left auditory cortex does not tend to adapt for speech or highly complex sounds compared with pure tones [20]. Our result may include the left hemispheric persistence to the adaptation of highly complex sounds. Our results suggest that the left hemisphere plays a dominant role in the spectral integration process to reach the ultimate percept, that is, sound source identification.

The localization for the vocal and the instrumental sounds was significantly anterior to that for the linear sounds. It is well known that high-frequency tones are represented more medially than low-frequency tones in pure-tone studies [21]. Recently, another tonotopic axis was postulated by several researchers. Diesch and Luce [7] reported that N1m source location on the posteroanterior axis correlated positively with the first formant frequency of the complex sounds. Conversely, Cansino *et al.* [6] reported that the posteroanterior axis correlated negatively with the single formant frequency of the complex sounds, specifically in the right hemisphere. Tonotopic representations along the posteroanterior axis and their hemispheric laterality are not clarified yet [19]. Some researchers speculated that the auditory pathway specifically involved in sound source identification ('what' pathway) is located anterior to the primary auditory cortex [22], and the indispensable acoustic parameter for this process is spectral information [23]. Previous studies reported the existence of voice-selective areas along the upper bank of the superior temporal sulcus in humans [24], and that the anterosuperior temporal areas bilaterally respond to the spectral variation [25]. In the present study, the ECDs were not localized in the superior temporal sulcus. The localization shift observed in our study may imply that part of the identification process is already present in a latency range of N1m. However, this reasoning is preliminary and speculative, and requires substantiation by further studies using more acoustic parameters than those used in this study.

CONCLUSION

Periodic complex sounds are processed differently by the perceptual categorization of spectral envelopes in a latency range of N1m. The categorization of spectral envelopes with and without formant structures was reflected in the N1m source location and strength. The categorization of vocal and nonvocal sounds was reflected in N1m latency dependence on F_0. These findings suggest that not only vocal sounds but also natural organic sounds with formant structure are distinguished at a very early stage of cortical processing.

REFERENCES

1. Miller DC. *The Science of Musical Sounds*. New York: MacMillan; 1916.
2. Fant G. *Acoustic Theory of Speech Production*. The Hague: Mouton; 1960.
3. Liégeois-Chauvel C, de Graaf JB, Laguitton V, Chauvel P. Specialization of left auditory cortex for speech perception in man depends on temporal coding. *Cereb Cortex* 1999; **9**:484–496.
4. Helmholtz HLF. *On the Sensations of Tone*. New York: Dover; 1954.
5. von Bismarck G. Timbre of steady sounds: a factorial investigation of its verbal attributes. *Acoustica* 1974; **30**:146–159.
6. Cansino S, Ducorps A, Ragot R. Tonotopic cortical representation of periodic complex sounds. *Hum Brain Map* 2003; **20**:71–81.
7. Diesch E, Luce T. Topographic and temporal indices of vowel spectral envelope extraction in the human auditory cortex. *J Cogn Neurosci* 2000; **12**:878–893.
8. Gunji A, Koyama S, Ishii R, Levy D, Okamoto H, Kakigi R *et al*. Magnetoencephalographic study of the cortical activity elicited by human voice. *Neurosci Lett* 2003; **348**:13–16.
9. Menon V, Levitin DJ, Smith BK, Lembke A, Krasnow BD, Glazer D *et al*. Neural correlates of timbre change in harmonic sounds. *Neuroimage* 2002; **17**:1742–1754.
10. Crummer GC, Walton JP, Wayman JW, Hantz EC, Frisina RD. Neural processing of musical timbre by musicians, nonmusicians, and musicians possessing absolute pitch. *J Acoust Soc Am* 1994; **95**:2720–2727.
11. Jones SJ, Longe O, Vaz Pato M. Auditory evoked potentials to abrupt pitch and timbre change of complex tones: electrophysiological evidence of 'streaming'? *Electroencephalogr Clin Neurophysiol* 1998; **108**:131–142.
12. Pekkonen E, Huotilainen M, Virtanen J, Sinkkonen J, Rinne T, Ilmoniemi RJ *et al*. Age-related functional differences between auditory cortices: a whole-head MEG study. *Neuroreport* 1995; **6**:1803–1806.
13. Anderer P, Semlitsch HV, Saletu B. Multichannel auditory event-related brain potentials: effects of normal aging on the scalp distribution of N1, P2, N2 and P300 latencies and amplitudes. *Electroencephalogr Clin Neurophysiol* 1996; **99**:458–472.
14. Verkindt C, Bertrand O, Perrin F, Echallier JF, Pernier J. Tonotopic organization of the human auditory cortex: N100 topography and multiple dipole model analysis. *Electroencephalogr Clin Neurophysiol* 1995; **96**:143–156.
15. Woods DL, Alho K, Algazi A. Intermodal selective attention: evidence for processing in tonotopic auditory fields. *Psychophysiology* 1993; **30**: 287–295.
16. Crottaz-Herbette S, Ragot R. Perception of complex sounds: N1 latency codes pitch and topography codes spectra. *Clin Neurophysiol* 2000; **111**:1759–1766.
17. Mäkelä AM, Alku P, Mäkinen V, Valtonen J, May P, Tiitinen H. Human cortical dynamics determined by speech fundamental frequency. *Neuroimage* 2002; **17**:1300–1305.
18. Fletcher H, Munson W. Loudness, its definition, measurement, and calculation. *J Acoust Soc Am* 1933; **5**:92–108.
19. Liégeois-Chauvel C, Giraud K, Badier JM, Marquis P, Chauvel P. Intracerebral evoked potentials in pitch perception reveal a functional asymmetry of the human auditory cortex. *Ann NY Acad Sci* 2001; **930**: 117–132.
20. Teismann IK, Sörös P, Manemann E, Ross B, Pantev C, Knecht S. Responsiveness to repeated speech stimuli persists in left but not right auditory cortex. *Neuroreport* 2004; **15**:1267–1270.
21. Romani GL, Williamson SJ, Kaufman L. Tonotopic organization of the human auditory cortex. *Science* 1982; **216**:1339–1340.
22. Rauschecker JP, Tian B. Mechanisms and streams for processing of 'what' and 'where' in auditory cortex. *Proc Natl Acad Sci USA* 2000; **97**:11800–11806.
23. Rauschecker JP. Parallel processing in the auditory cortex of primates. *Audiol Neurootol* 1998; **3**:86–103.
24. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. *Nature* 2000; **403**:309–312.
25. Zatorre RJ, Belin P. Spectral and temporal processing in human auditory cortex. *Cereb Cortex* 2001; **11**:946–953.