Intracortical Responses in Human and Monkey Primary Auditory Cortex Support a Temporal Processing Mechanism for Encoding of the Voice Onset Time Phonetic Parameter

This study tests the hypothesis that temporal response patterns in primary auditory cortex are potentially relevant for voice onset time (VOT) encoding in two related experiments. The first experiment investigates whether temporal responses reflecting VOT are modulated in a way that can account for boundary shifts that occur with changes in first formant (F1) frequency, and by extension, consonant place of articulation. Evoked potentials recorded from Heschl’s gyrus in a patient undergoing epilepsy surgery evaluation are examined. Representation of VOT varies in a manner that reflects the spectral composition of the syllables and the underlying tonotopic organization. Activity patterns averaged across extended regions of Heschl’s gyrus parallel changes in the subject’s perceptual boundaries. The second experiment investigates whether the physiological boundary for detecting the sequence of two acoustic elements parallels the psychoacoustic result of ~20 ms. Population responses evoked by two-tone complexes with variable tone onset times (TOTs) in primary auditory cortex of the monkey are examined. Onset responses evoked by both the first and second tones are detected at a TOT separation as short as 20 ms. Overall, parallels between perceptual and physiological responses support the relevance of a population-based temporal processing mechanism for VOT encoding.

Keywords: auditory evoked potentials, Heschl’s gyrus, intracortical recording, population encoding, speech

Introduction

Understanding language encoding by the brain is predicated on clarifying neural mechanisms underlying detailed features of speech perception. One promising line of investigation focuses on the voice onset time (VOT) distinction in speech. VOT is an articulatory parameter used by most of the world’s languages, and is a measure of the interval between consonant release (onset) and the start of rhythmic vocal cord vibrations (voicing) (Lisker and Abramson, 1964). In American English, a short VOT promotes the perception of the voiced stop consonants, /b/, /d/ and /g/, while a long VOT promotes the perception of the unvoiced stop consonants, /p/, /t/ and /k/. Differential perception of these sounds is categorical, such that a small change in the duration of a short or long VOT will not significantly alter consonant perception, whereas the same small VOT change that crosses a boundary between short and long values will produce a categorical change in the perceived phoneme (e.g. Wood, 1976). This perceptual boundary in American English lies between 20 and 40 ms.

A temporal processing mechanism likely serves as the primary means by which voiced stop consonants are distinguished from unvoiced stops, despite modulation of VOT perceptual boundaries by spectral, visual and language-related lexical or linguistic manipulations (Stevens and Klatt, 1974; Lisker, 1975; Repp, 1979; Ganag, 1980; Kluender, 1995; Shannon, 1995; Borsky et al., 1998; Faulkner and Rosen, 1999; Holt et al., 2001; Lotto and Kluender, 2002; Branciano et al., 2003). This mechanism was first proposed by Pisoni (1977), who presented subjects with two-tone stimuli that varied in the relative onset timing of the two tones in a manner mimicking that of VOT (tone onset time, TOT). Subjects were measured in their ability to identify whether the tones were presented simultaneously or sequentially. Results paralleled those seen for speech; identification was categorical with a boundary at ~20 ms, and discrimination between stimuli showed a peak at the same value. These findings led Pisoni (1977) to propose that the differential perception of voiced from unvoiced stop consonants is based on whether consonant release and voicing onset are perceived as occurring simultaneously or sequentially. This speech-related example of temporal encoding was further suggested to represent a specific instance of a more general rule governing the ability to temporally order the sequence of two sounds (Hirsh, 1959).

Temporally precise speech-evoked responses in auditory cortex support the importance of temporal processing mechanisms for VOT perception. Studies in monkeys and other animals reveal a characteristic pattern of activity, wherein syllables with a short VOT evoke a single response burst time-locked to consonant release; while syllables with a longer VOT evoke response bursts time-locked to both consonant release and voicing onset (e.g. Steinschneider et al., 1994, 1995b, 2003; Eggermont, 1995a,b, 1999; McGee et al., 1996; Schreiner, 1998). Importantly, several of these studies have shown a marked increase in the response time-locked to voicing onset at VOT intervals that cross the boundary between human perception of voiced and unvoiced stop consonants. These animal model findings gain further relevance by their similarity to speech-evoked response patterns recorded directly from human auditory cortex (Liégeois-Chauvel et al., 1999; Steinschneider et al., 1999). Furthermore, this activity profile offers a plausible, physiological mechanism supporting categorical discrimination between voiced and unvoiced stop consonants. Perception of voiced stops would be facilitated when only a single response burst in auditory cortex is evoked, as by short duration VOT syllables. In contrast, perception of unvoiced stops would be promoted when two response bursts are sequentially elicited, one by consonant release and the other by voicing onset, as seen with longer duration VOTs. The border between these two response patterns would approximate the perceptual boundary.

If this cortically based temporal processing mechanism for VOT discrimination is derived from a general capacity to temporally order the sequence of two sounds through time-locked
responses, then a 20 ms physiological boundary paralleling the psychoacoustic findings of Pisoni (1977) should be present. However, studies examining responses to two-tone sequences in auditory cortex have failed to demonstrate this degree of physiological temporal acuity (Calford and Semple, 1995; Brosch and Schreiner, 1997, 2000; Horikawa et al., 1997). While methodological considerations, such as the use of anesthetized animals, may be in part responsible for this discrepancy, the fact remains that a fundamental prerequisite for this physiological temporal processing mechanism has not been met.

A second potential shortcoming of this processing scheme for VOT perception is that it does not account for significant boundary shifts that occur with changes in stop consonant place of articulation. Perceptual boundaries are shortest for the differential perception of the bilabial stop consonants /b/ and /p/ (~20 ms), intermediate for the alveolar stops /d/ and /t/ (~30 ms), and longest for the velar consonants /g/ and /k/ (~40 ms) (Lisker and Abramson, 1964). A major acoustic consequence of differences in place of articulation is that for any VOT value occurring prior to the attainment of steady-state vowel frequencies, the first formant (F1) frequency is highest for the bilabial stops, lowest for the velar consonants and intermediate for the alveolar stops (see Parker, 1988). Multiple studies have demonstrated an inverse relationship between F1 frequency and VOT boundary, and have suggested that this trading relations effect between F1 frequency and VOT is the perceptual basis for the boundary shifts observed with changes in consonant place of articulation (Lisker, 1975; Summerfield and Haggard, 1977; Summerfield, 1982; Soli, 1983; Hillenbrand, 1984).

Placed in a temporal processing framework, these findings imply that the lower F1 frequencies seen for velar stops would require a longer VOT interval for the onsets of consonant release and voicing to be perceived as sequential, and therefore as an unvoiced consonant. The ever higher F1 frequencies observed for the alveolar and bilabial stops would need progressively shorter VOT intervals to identify sequential onsets and the unvoiced character of the consonant. An auditory processing basis for the trading relations effect between spectral and temporal speech components gains additional support when considering VOT perception in animals. Animals demonstrate categorical-like perception with boundaries and boundary shifts due to changes in consonant place of articulation or F1 frequency similar to those in humans, and show heightened sensitivity to incremental changes in VOT at the boundary in a manner that mirrors human perception (Kuhl, 1986; Kluender, 1991; Kluender and Lotto, 1994; Ohlemiller et al., 1999). Since a language-specific mechanism cannot be invoked to explain perceptual phenomena in animal models, these findings indicate that at least some of the fundamental neural mechanisms responsible for VOT perception must be based on auditory system processing.

Thus, the goal of this study is to test the hypothesis that temporal response patterns elicited by syllables in auditory cortex are key elements for VOT perception. We test this hypothesis by examining whether perceptual boundaries are paralleled by neural patterns of activity using two related experiments. In the first experiment, we examine whether temporal responses reflecting syllable VOT are modulated by spectral components of speech in a manner that can account for the VOT boundary shifts that occur with changes in F1 frequency, and by extension, consonant place of articulation.

For this experiment, we examine auditory evoked potentials (AEP) elicited by synthetic syllables with variable F1s recorded directly from auditory cortex in a patient undergoing surgical evaluation for medically intractable epilepsy. In the second experiment, we examine whether the physiological boundary for detecting the sequence of two acoustic elements parallels the psychoacoustic result of ~20 ms. For this experiment, we examine responses evoked by two-tone complexes with variable TOTs in primary auditory cortex (A1) of the monkey.

**Materials and Methods**

**Human Electrophysiological Recordings**

One right-handed man with medically intractable epilepsy was studied. Experimental protocols were approved by the University of Iowa Human Subjects Review Board and National Institutes of Health, and informed consent was obtained from the patient prior to his study participation. The patient’s seizures often began with the perception of a ‘tuning fork sound’, and non-invasive studies suggested an epileptic focus within or near auditory cortex of the right hemisphere. Multi-contact intracranial electrodes and subdural grid electrodes were implanted for acquisition of diagnostic electroencephalographic data required to plan subsequent surgical treatment. Research recordings were performed in parallel with the diagnostical evaluation, did not disrupt acquisition of medically required information and did not add any additional health risks.

Experimental recordings were obtained from two stereotactically-placed hybrid-depth electrodes that contained evenly spaced low-impedance recording sites with higher impedance contacts interspersed along the shaft (Howard et al., 1996a,b). The first electrode was located in the anterior portion of Heschl’s gyrus, while the second was positioned at the junction of the posterior rim of Heschl’s gyrus and the planum temporale (Fig. 1). Responses elicited by musical chords from these electrodes have been previously reported (subject 1; Fishman et al., 2001b). The reference electrode was a subdural recording contact located on the ventral surface of the ipsilateral, anterior temporal lobe. Recordings were performed with the subject lying comfortably awake in a quiet room of the Epilepsy Monitoring Unit of the University of Iowa Hospitals and Clinics. The subject could abort the experimental session at any time.

AEPs were recorded at a gain of 5000 and with a band-pass of 2–500 Hz. Signals were digitized at a rate of 1000 Hz and averaged with an analysis window of 1 s, including a pre-stimulus baseline of 500 ms. Sounds were presented at an inter-trial-interval of 2 s. All single trial
epochs were examined by the lead author (board certified in clinical neurophysiology) and epochs containing epileptic spikes or high amplitude delta activity were discarded prior to generating the averages (maximum number of epochs = 50).

Stimuli were presented to the left ear (contralateral to the recording sites) via an insert earphone (Etymotic Research), and at a comfortable supratheral threshold level determined by the subject (–70 dB SPL). Stimuli were synthetic syllables, 175 ms in duration, and were generated by a parallel/cascade Klatt synthesizer (SentSyn, Sensimetrics) at a sampling rate of 10 kHz. Frequency values were chosen appropriate for the perception of /d/ and /t/. A schematic of the syllables is shown in Figure 2. Syllables contained three formants. The second formant (F2) had a starting frequency of 1600 Hz and linearly decreased to a steady-state value of 1200 Hz, while the third formant (F3) began at 3000 Hz and linearly decreased to 2500 Hz. Both formant transitions were 40 ms in duration. These formants were excited by a noise source simulating friction for the first 5 ms of the syllables. F1 was without a transition and was centered at 424, 600 or 848 Hz (1/2 octave intervals). It began after friction and after a variable period of aspiration that preceded voicing. Each syllable was presented with seven VOT values ranging from 5 to 60 ms. The subject was asked whether he heard a /d/ or a /t/ after presentation of 50 repetitions of each syllable.

**Monkey Electrophysiological Recordings**

Five male macaque monkeys (Macaca fascicularis), weighing between 2.5 and 3.5 kg, were studied following approval by the Animal Care and Use Committee of Albert Einstein College of Medicine. Experiments were conducted in accordance with institutional and federal guidelines governing the use of primates, who were housed in our AAALAC-accredited Animal Institute. Other protocols were performed in parallel with this experiment to minimize the overall number of animals used. Monkeys were trained to sit comfortably in customized primate chairs with hands restrained. Surgery was then performed using sterile techniques and general anesthesia (sodium pentobarbital). Holes were drilled into the skull to accommodate epidural matrices that allowed access to the brain. Matrices consisted of 18-gauge stainless-steel tubes glued together into a honeycomb form, and were shaped to approximate the contour of the cortical convexity. The bottom of each matrix was covered with a protective layer of sterile silastic. Matrices were stereotaxically positioned to target A1 at an angle 30° from normal to approximate the anterior–posterior tilt of the superior temporal gyrus, thus guiding electrode penetrations to be orthogonal with the surface of A1. Matrices and Flexiglas bars permitting painless head fixation were embedded in dental acrylic secured to the skull with inverted bolts keyed into the bone. Peri- and post-operative anti-inflammatory, antibiotic and analgesic agents were given. Recordings began 2 weeks after surgery.

Recordings were conducted in a sound-attenuated chamber with the animals painlessly restrained. Monkeys maintained a relaxed, but alert state, facilitated by frequent contact and delivery of juice reinforcements. Later animals were also monitored by closed-circuit television. Recordings were performed with multicontact electrodes constructed in our laboratory. They contained 14 recording contacts arranged in a linear array and evenly spaced at 150 μm intervals (<10% error), permitting simultaneous recording across multiple A1 laminae. Contacts were 25 μm stainless steel wires insulated except at the tip, and were fixed in place within the sharpened distal portion of a 30-gauge tube. Impedance of each contact was maintained at 0.1–0.4 MΩ at 1 kHz. The reference was an occipital epidural electrode. Headstage pre-amplification was followed by amplification (∼5000) with differential amplifiers (Grass P5, down 3 dB at 3 Hz and 3 kHz). Signals were digitized at a rate of 3400 Hz and averaged by Neuroscan software to generate auditory evoked potentials (AEPs). Data were also stored on a digital tape recorder (DT-1000, MicroData Instrument, Inc., sample rate 6 kHz) for 2/3 of the recording sessions. Positioning of the electrodes was performed with a microscope whose movements were guided by online inspection of AEPs and multiunit activity (MUA) evoked by 80 dB clicks. Tone bursts and two-tone complexes were presented when the recording contacts of the linear-array electrode straddled the inversion of the cortical AEP, and the largest evoked MUA was maximal in the middle electrode contacts. Response averages were generated from 50–75 stimulus presentations. One-dimensional current source density (CSD) analysis was used to define physiologically the laminar recording of location recording sites in A1. CSD was calculated from AEP laminar profiles using an algorithm that approximated the second spatial derivative of the field potentials recorded at three adjacent depths (Freeman and Nicholson, 1975). Depths of the earliest click-evoked and tone-evoked current sinks were used to locate lamina 4 and lower lamina 3 (e.g. Müller-Preuss and Mitzdorf, 1984; Steinschneider et al., 1992; Cruikshank et al., 2002). A later current sink in upper lamina 3 and a concurrent source located more superficially were almost always identified in the recordings and served as additional markers of laminar depth (e.g. Müller-Preuss and Mitzdorf, 1984; Steinschneider et al., 1992, 1994, 2003; Fishman et al., 2001b; Cruikshank et al., 2002). This physiological procedure was later checked by correlation with measured widths of A1 and its laminae at select electrode sites obtained from histological data (see below).

MUA was extracted in the first four animals by high-pass filtering the raw input at 500 Hz (roll-off 24 dB/octave), further amplifying (∼8) and full-wave rectifying the derived signal, and computer averaging the resultant activity. In the last animal, rectification was followed by low-pass filtering at 600 Hz prior to digitization using newly acquired digital filters (RP2 modules, Tucker Davis Technologies). MUA measures the envelope of action potential activity generated by neuronal aggregates, weighted by neuronal location and size. MUA is similar to cluster activity, but has greater response stability (Nelken et al., 1994). We observe sharply differentiated MUA at a recording contact spacing of 75 μm (e.g. Schroeder et al., 1990), and other investigators have demonstrated a similar sphere of recording (Brosch et al., 1997). Due to limitations of the acquisition computer, sampling rates were less than the Nyquist frequency of the low-pass filter setting of the amplifiers in the first four animals. Empirical testing revealed negligible signal distortion, as almost all energy in the neural signals was <1 kHz. Samples of off-line data from the digital tape recorder were re-digitized at 6 kHz, and resultant MUA had waveforms and amplitudes nearly identical to those of data sampled at the lower rate (distortion < 1%). MUA acquired from the digitally taped data was also low-pass filtered below 800 Hz (96 dB/octave) and then averaged at a sampling rate of 2 kHz to further test the accuracy of the initial measurements. Differences between these and initial measurements were negligible (see Fishman et al., 2001b).

Peristimulus-time-histograms (PSTHs) of multiunit cluster activity were constructed from data stored on digital tape to complement MUA measures. Data were band-pass filtered between 450 and 3000 Hz (54 dB/ octave filter modules) prior to spike analysis using online inspection software and hardware (Tucker Davis Technologies, Inc.). Sample rate was 65 kHz and bin width was 1 ms. Triggers for spike acquisition were set at 2.5 times the amplitude of the high-frequency background activity.
isolated pure tones and two-tone complexes were generated and
released at a sample rate of 100 kHz by a PC-based system using R2
modules. Isolated pure tones ranged from 0.2 to 17.0 kHz and were 175
ms in duration, with linear rise/decay times of 10 ms. Two-tone
complexes of the same duration, but with 5 ms rise/decay times, were
presented with variable tone onset times (TOT) ranging from 0 to 50 ms
in 10 ms increments. The two tones ended simultaneously. All stimuli
were monaurally delivered via a dynamic headphone (MDR-7502, Sony,
Inc.) to the ear contralateral to the recorded hemisphere with a stimulus
onset asynchrony of 658 ms. Sounds were presented to the ear through a
3/4 long, 60 cc plastic tube attached to the headphone. Pure tone
intensity was 60 dB SPL measured with a Bruel and Kjær sound level
meter (type 2236) positioned at the opening of the plastic tube. Two-
tone complexes were generated through the linear addition of two
equal-amplitude 60 dB tones each beginning at 0 degree phase. The
frequency response of the headphone was flattened (±3 dB) from 0.2 to
17.0 kHz by a graphic equalizer (GE-60, Rane, Inc.).

After completion of a recording series, animals were deeply anes-
thetized with pentobarbital and perfused through the heart with
physiological saline and 10% buffered formalin. A1 was physiologically
delineated by its typically large amplitude responses and by a best
frequency (BF) map that was organized with low BFs located antero-
laterally and higher BFs postero-medially (e.g. Merzenich and Brugge,
1973; Morel et al., 1993). Electrode tracks were reconstructed from
coronal sections stained for Nissl and acetylcholinesterase, and A1 was
anatomically identified using published criteria (e.g. Morel et al., 1993).

Four adjacent channels of MUA and cell cluster activity (PSTHs)
located in lamina 4 and lower lamina 3 were averaged together for
analysis of responses to pure tones and tone pairs. BFs were defined as
the tone frequency eliciting the largest amplitude MUA within the first
20 ms after stimulus onset.

Results

Human Perceptual and Physiological Data

The subject’s perception of the syllables varied as a function of
the F1 frequency. When F1 was 600 or 848 Hz, syllables with
a VOT of 25 ms or greater were heard as /ta/, while those with
shorter VOTs were perceived as /da/. In contrast, when the F1
was lowered to 424 Hz, only the consonant with a VOT of 60 ms
was identified as /t/, while all syllables with a VOT of 40 ms or
shorter led to the perception of /d/. This effect of a later
perceptual VOT boundary for the /d/-/t/ distinction when F1 is
lowered parallels previously reported results (e.g. Lisker, 1975;
Summerfield and Haggard, 1977).

Syllable perception is associated with multiple physiological
response patterns recorded from the electrode located in
anterior Heschl’s gyrus and the more posterior electrode
located at the border with the planum temporale. The most
basic finding is that VOT is differentially represented in tempo-
rally resolved response patterns recorded within different auditory cortical
regions. This finding is illustrated in Figure 3, which depicts
AEPs averaged across the three recording sites on each elect-
rode and across the three F1 conditions. Temporal response
patterns recorded in the anterior portion of Heschl’s gyrus,
corresponding to primary auditory cortex (e.g. Hackett et al.,
2001; Wallace et al., 2002), are dramatically sensitive to the
syllable VOT. A second response component following the
initial activity is time-locked to voicing onset (arrows). This
component shows a marked decrease in amplitude at VOTs of
<30 ms, and merges with the initial response complex at
shorter values. Simultaneous recordings from the posterior
electrode, however, fail to exhibit a response to voicing onset,
despite a threefold increase in AEP amplitude relative to that
recorded from anterior Heschl’s gyrus. This finding confirms
a previous observation on differences between speech-evoked
activity recorded from anterior Heschl’s gyrus and more post-
erior areas (Steinschneider et al., 1999).

Justification for averaging the AEPs recorded from the pos-
terior electrode is illustrated in Figure 4. This figure depicts the
AEPs recorded from the three posterior electrode sites (bot-
tom) and the medial recording site on the anterior electrode
(top) in response to the syllables with the 60 ms VOT.
Responses evoked by the three F1 conditions are shown as
overlaid waveforms. The 60 ms VOT stimuli were chosen for
illustration because they evoke the largest responses to voicing
onset. Responses from the medial electrode site on the anterior
electrode are shown as this site is in close proximity to its
counterpart on the posterior electrode (see Fig. 1). Despite this
proximity, there is a marked difference in the AEPs recorded at
these medial electrode sites. AEPs recorded from the anterior
electrode contain prominent components time-locked to voic-
ing onset (arrow) and stimulus offset (asterisk). In contrast, the
AEPs recorded at the medial posterior electrode site are nearly
identical in appearance to those recorded at the other two sites
on the same electrode, and do not contain components time-
locked to voicing onset nor prominent ‘off’ responses.

Figure 3. AEP response patterns evoked by the syllables differ between the two
recording electrodes located in the anterior and posterior portions of Heschl’s gyrus.
Responses evoked at the three recording sites on each electrode and for each of the
three F1 conditions have been averaged together to illustrate the effect of electrode
location on the temporal activity patterns. Activity patterns are segregated according
to the VOT of the syllables. AEPs recorded from the anterior electrode contain
responses evoked by both consonant release and by voicing onset (arrows). Responses evoked by voicing markedly decrease in amplitude at VOT values of
<30 ms and merge with the initial response component evoked by consonant release.
In contrast, AEPs recorded from the more posterior electrode fail to exhibit a response
component elicited by voicing onset at any VOT value. This absence occurs despite
a threefold increase in overall amplitude of the AEPs relative to those recorded from
the more anterior electrode.
The more detailed representation of VOT in anterior Heschl’s gyrus is also non-uniform, and varies in a systematic manner across recording sites (Fig. 5). Responses evoked by the three F1 stimulus sets are collapsed and averaged to illustrate the effect of recording site. AEPs recorded from the lateral site contain prominent components time-locked to voicing onset for all values of VOT (arrows). At the center electrode contact, located 4.2 mm away, these components are decreased in amplitude relative to the initial responses evoked by consonant release. Discrete response components evoked by voicing onset (arrows) are present at longer VOTs, and merge with the initial ‘onset’ response at shorter values. This trend continues at the most medial recording contact, located 2.5 mm away from the center recording site. Here, a discrete response component evoked by voicing onset is observed only for the 60 ms VOT syllables (arrow, overlapping waveforms at the top of Figure 4 illustrate the three AEPs that were averaged to produce the composite response).

This systematic change in the relative strengths of the responses evoked by consonant release and voicing onset across the anterior Heschl’s gyrus electrode was quantified by first dividing each AEP into four separate averages derived from every fourth recorded epoch. Maximum voltages in 5 ms intervals were computed for each subdivided average. The relative strengths of the responses evoked by consonant release and voicing onset were then determined by computing the ratios of the trough-to-peak voltage excursions evoked by voicing onset in comparison to the voltage excursions from baseline of the initial positivity evoked by consonant release. The results of this analysis are shown in Figure 6, which depicts the amplitude ratios at VOTs from 10 to 60 ms. Ratios at a VOT of 5 ms were excluded because only the response at the lateral site had a discrete response to voicing onset. Similarly, ratios were only computed at the medial site at a VOT of 60 ms because this was the only VOT where a discrete response to voicing onset occurred. The response evoked by voicing onset relative to that evoked by consonant release was larger at the lateral recording site than the other two locations at all examined VOTs except 10 ms.

Physiological responses in anterior Heschl’s gyrus are also systematically modulated by changes in F1 frequency that parallel increases in the patient’s perceptual boundary at the lowest F1. This finding is illustrated in Figure 7, which depicts the AEPs averaged across all three recording sites as a function of F1 frequency and VOT. Ovals superimposed on the waveforms indicate the subject’s perceptual boundaries between the perceptions of /d/ and /t/ for each VOT. In the case of F1 centered at 424 Hz, the subject perceived /t/ only when the VOT was 60 ms. In parallel, a discrete response time-locked to

**Figure 4.** AEP response patterns evoked by the syllables with a 60 ms VOT. Responses at the medial electrode site on the anterior Heschl’s gyrus electrode and all three sites on the posterior Heschl’s gyrus electrode are shown. AEPs evoked by the syllables with the three F1 conditions are shown as overlapping waveforms. See text for details.

**Figure 5.** AEP response patterns evoked by the syllables systematically vary with recording site along the anterior Heschl’s gyrus electrode. Responses evoked by the three F1 conditions have been averaged together to illustrate the effect of electrode location on the temporal activity patterns. AEPs recorded from the lateral electrode site contain responses time-locked to voicing onset for all VOT values (arrows), while responses at the medial site contain a time-locked burst evoked by voicing onset for only the syllables with a VOT of 60 ms. Responses recorded from the center electrode display an intermediate capacity to respond to voicing onset.
voicing onset is only observed for the syllable with a VOT of 60 ms (arrow). At shorter VOTs, this component loses its distinct appearance and merges with the initial response complex evoked by consonant release. In contrast to the AEPs elicited by the syllables with a 424 Hz F1, a discrete response peak evoked by voicing onset is present at VOTs as short as 25 and 20 ms when the F1 is 600 or 848 Hz, respectively (arrows). These values of VOT where the response component elicited by voicing onset loses its distinct morphology and merges with the initial onset response complex parallel the perceptual boundary when the F1 is 600 Hz, and approximate the perceptual boundary when the F1 is 848 Hz.

This effect of F1 was statistically analyzed by examining the mean and variability of the waveforms binned according to maximum amplitudes within 5 ms intervals. Figure 8 depicts the responses from 20 ms prior to stimulus onset to 200 ms post-stimulus delivery. A single positive component (P1) is followed by a single negative component (N1) for syllables with a short VOT. These components are labeled in the responses evoked by the 5 ms VOT syllables. In contrast, syllables with a prolonged VOT contain both these components as well as a second positive-going wave (P2) that truncates the first negativity and leads to the appearance of a second negative-going wave (N2). These components are labeled in the responses evoked by the 60 ms VOT syllables. Statistical analyses involve determining whether P2, which is evoked by voicing onset, is significantly larger than the preceding N1. Arrows mark the minimum values of N2, and the maximum values of P2 that are analyzed through a series of t-tests for the former comparisons, and a one-way ANOVA for the latter comparison. See text for details. The apparent decrease in the response ratio at the lateral site when VOT was 40 ms reflects an exceptionally large initial positivity that occurred for unclear, and probably spurious, reasons.

voicing onset is only observed for the syllable with a VOT of 60 ms (arrow). At shorter VOTs, this component loses its distinct appearance and merges with the initial response complex evoked by consonant release. In contrast to the AEPs elicited by the syllables with a 424 Hz F1, a discrete response peak evoked by voicing onset is present at VOTs as short as 25 and 20 ms when the F1 is 600 or 848 Hz, respectively (arrows). These values of VOT where the response component elicited by voicing onset loses its distinct morphology and merges with the initial onset response complex parallel the perceptual boundary when the F1 is 600 Hz, and approximate the perceptual boundary when the F1 is 848 Hz.

This effect of F1 was statistically analyzed by examining the mean and variability of the waveforms binned according to maximum amplitudes within 5 ms intervals. Figure 8 depicts the responses from 20 ms prior to stimulus onset to 200 ms post-stimulus delivery. A single positive component (P1) is followed by a single negative component (N1) for syllables with a short VOT. These components are labeled in the responses evoked by the 5 ms VOT syllables. In contrast, syllables with a prolonged VOT contain both these components as well as a second positive-going wave (P2) that truncates the first negativity and leads to the appearance of a second negative-going wave (N2). These components are labeled in the responses evoked by the 60 ms VOT syllables. Statistical analyses involve determining whether P2, which is evoked by voicing onset, is significantly larger than the preceding N1. Arrows mark the minimum values of N2, and the maximum values of P2 that are analyzed through a series of t-tests for the former comparisons, and a one-way ANOVA for the latter comparison. See text for details. The apparent decrease in the response ratio at the lateral site when VOT was 40 ms reflects an exceptionally large initial positivity that occurred for unclear, and probably spurious, reasons.

with varying F1s to evoke a response to voicing onset that has a clear trend to parallel changing perceptual boundaries.

We further examined if there were any systematic interactions between F1 and recording sites in the amplitude response ratios using a series of ANOVA tests and post hoc analyses using the Newman–Keuls multiple comparisons test (level of significance, $P < 0.05$). The responses in the 424 Hz F1 condition were smaller at the center electrode for VOTs of 10 and 30 ms. At a VOT of 10 ms, the response ratio was smaller than the response to the 600 Hz F1 condition. However, the response to the 848 Hz F1 condition was also smaller than the same F1 condition. At a VOT of 30 ms, the response to the 424 Hz F1 condition was smaller than both other conditions. The only other statistically significant interaction was at the lateral electrode site when VOT was 25 ms. In this case, the response to the 424 Hz F1 condition was smaller than the 600 Hz F1 condition. It must be stressed that the interpretation of these latter analyses are constrained by limited power and should be viewed with caution.

The opportunity to record directly from Heschl’s gyrus is rare, necessitating studies generally based on, at most, observations obtained from only a few subjects. In this study, we report activity from only a single human subject. Without additional information, it is difficult to evaluate with confidence whether the Heschl’s gyrus responses are representative indices of auditory cortical activity, or are aberrant findings in a patient with medically intractable epilepsy. One way to support the reliability of the AEPs is to compare responses from this subject with varying F1s to evoke a response to voicing onset that has a clear trend to parallel changing perceptual boundaries.
with those obtained in other patients using identical stimuli. In
the present case, we also obtained AEPs evoked by the speech
sounds /da/ and /ta/ used in a previous study of Heschl's gyrus
activity (Steinschneider et al., 1999). VOTs varied from 0 to
80 ms in 20 ms increments. Perceptually in this patient, /ta/ was
heard when the VOT was 40–80 ms, and /da/ when the VOT

Figure 8. Responses recorded from the anterior electrode binned in 5 ms increments according to maximum values within the bins. Means and standard errors are shown for the responses from 20 ms prior to stimulus onset to 200 ms post-stimulus onset. AEPs from the three recording sites are averaged together for each F1 condition and at each VOT. See text for details.
was 0 and 20 ms. Figure 9 depicts the averaged AEP evoked by these syllables collapsed across the three anterior Heschl’s gyrus recording sites. As previously reported, discrete time-locked components elicited by voicing onset are only observed for the three stimuli heard as /ta/ (solid arrows). The dotted arrow in the AEP for the 20 ms VOT sound marks the predicted time at which the absent response evoked by voicing onset should have occurred.

In summary, data indicate that there are multiple physiological response patterns evoked by the same syllables in different regions of auditory cortex, and that activity in Heschl’s gyrus is non-uniform with respect to representation of VOT. Responses in the anterior portion of Heschl’s gyrus contain components time-locked to voicing onset, whereas more posterior regions do not. Lateral locations in anterior Heschl’s gyrus are better able to represent voicing onset relative to consonant release than more progressively medial locations. Activity averaged across extended regions of anterior Heschl’s gyrus appears to best correlate with changing perceptual boundaries as F1 is varied.

**Monkey Physiological Data**

**Sample Characteristics**

Data are based on responses obtained from 37 electrode penetrations into A1. BFs ranged from 0.4 to 12.5 kHz. The sample distribution of BFs is shown in Figure 10A. The average MUA response evoked by BF tones from these electrode penetrations has a stereotypic pattern with an onset of ~10 ms, a peak quickly reached by 15–20 ms, followed by a rapid decay that plateaus at low levels for the duration of the sound (Fig. 10B). Spectral sensitivity of onset responses to tones of moderate intensity, as defined by area measures within the first 10 ms of the responses, is fairly restricted (Fig. 10C, 3 and 6 dB down points of ~0.2 and 0.3 octaves from the BF, respectively). Similar values are obtained when peak measures are used. Computed for percentage change away from the BF, amplitude of the ‘on’ response is 3 and 6 dB down at ~10 and 20%, respectively (Fig. 10D). The rapid decrement in MUA over the first 100 ms following BF stimulus onset can be accurately modeled by a single phase exponential decay curve (Fig. 10E, $R^2 = 0.99$). This profile suggests that A1 detection of new acoustic events by synchronized onset responses will be manifested as deviations from this basic response pattern, and that goodness-of-fit (GOF) measures using a single phase exponential decay function can be a concise index to assess the magnitude of these deviations.

**Two-tone Responses**

Tone pairs with frequencies at various distances away from the BF of the recording sites were presented. Total sample responses computed from PSTHs are shown in Figure 11. Mean and standard deviations of the tone frequencies in terms of octave distance away from the BFs of the recording sites, and values for their separation in octaves, are also shown. The average response to the tone complex with a TOT of 0 ms reveals the same pattern of rapid rise in activation and subsequent exponential decay of activity as that seen for MUA evoked by isolated BF tones. PSTHs evoked by stimuli with all TOTs other than 0 ms show some evidence of response perturbation time-locked to the onset of the second tone (arrows). The degree of perturbation, however, increases nonlinearly as the TOT interval lengthens. While a small perturbation is evident when TOT = 10 ms, a clearly defined response peak to the second tone is first seen when TOT = 20 ms. Longer TOTs evoke peaks of similar amplitude. However, there is a progressive increase in activity evoked by the second tone manifested as a temporal widening of the response. This can best be appreciated in the superimposed waveforms shown at the bottom left of the figure. The enhanced response at longer TOT intervals is further revealed by the degree to which the GOF for a single phase exponential decay from the initial tone response is reduced (Fig. 11, bottom right). There is a shallow decrease in GOF with TOTs of 10–30 ms from an initial $R^2$ of 0.99 when TOT equals 0 ms. This, in turn, is followed by a more pronounced decrement at TOTs of 40 and 50 ms. These features suggest that while A1 is capable of representing new sound events by discrete time-locked responses at tone separations of between 10 and 20 ms, intervals of 40 ms or greater lead to enhanced neural differentiation.

Nearly identical patterns are observed in the MUA (Fig. 12). Frequency values for the first and second tones are represented as octave deviations from the BF of the recording sites. Average octave separations of the two tones are also displayed. Peak MUA values are binned together in 5 ms intervals, beginning at 10 ms post-stimulus onset, the approximate onset time of activity in A1. Statistical analyses are performed using the repeated measures ANOVA and Newman–Keuls multiple comparisons test for post hoc evaluations. We operationally define a response time interval as being larger than preceding activity if
that response is statistically larger than the response occurring in the time bin occurring 10 ms earlier (two bins). This definition makes allowance for the variable peak latency of the initial response evoked by stimulus onset in the interval between 10 and 20 ms post-stimulus onset. Solid black bars denote responses that are significantly larger than earlier activity ($P < 0.001$). When TOT = 0 ms, the temporal pattern of MUA displays the same exponential decay from early peak activity as seen for the PSTHs. A new peak initiated by the onset of the second tone can first be discerned with a TOT of 20 ms. Strengths of responses evoked by the second tone, however, increases at greater TOT intervals. This is reflected in both the presence of two time bins being larger than preceding activity when TOT is > 20 ms (Fig. 12), and by the change in GOF values obtained by modeling the responses as a single phase exponential decay function expected for an isolated acoustic event (Fig. 13). A shallow, progressive decrease in $R^2$ is observed as TOT intervals increase from 0 to 30 ms, with a more marked decrease occurring at higher TOT values.

Relationship of Responses to BF

The previous data sets represent composites of evoked activity where the frequencies of the two tones vary widely with respect to BF. To clarify the interaction between tone frequency and temporal response patterns, data were divided into four groups, based upon the distance in octaves each tone was from the BF of the recording sites. Group 1 contains responses when both tones are less than their median distance from the BF, group 2 consists of responses when tone 1 is less than the median and tone 2 is greater than its median, group 3 is the reverse of group 2, and group 4 has both tones greater than the median. Thus, group 1 has both tones near the BF, group 2 has tone 2 farther away from the BF than tone 1, group 3 has tone 2 closer to the BF than tone 1, and group 4 has both tones at a distance from, and generally straddling, the BF.

Groups display a range of capabilities in representing both tones in a two-tone complex. Data for MUA are summarized in Table 1, which reports the statistical $P$ values of the post hoc tests for whether the response amplitudes at 10–15 and...
of the first 100 ms of the PSTHs evoked by two-tone complexes using a one phase activity occurs only when the TOT interval reaches 50 ms. For group 2, when the first tone is near and the second tone is distant from the BF, a significant increase in stimulus amplitude when the second tone is added to the PSTH data (not shown). This effect is not due to a 6 dB increase in stimulus amplitude when the second tone is added to the first, as only trivial, non-significant increases in peak amplitude are seen when both tones are near the BF of the recording sites and the TOT interval is at its most prolonged value of 50 ms (data not shown). For group 2, when the first tone is near and the second tone is distant from the BF, a significant increase in activity occurs only when the TOT interval reaches 50 ms.

More subtle differences among the responses to the four groups become apparent when data are analyzed with regard to their deviations from a single exponential decay curve typical of a response occurring when two tones are presented simultaneously (Fig. 14). GOF data for MUA and PSTHs are shown at the top and bottom of the figure, respectively. A clear ranking of GOF emerges wherein group 2 produces the least perturbation of response shape, whereas group 3 produces the greatest changes at TOT intervals as short as 20 ms. These differences reflect the fact that responses to the second tone in a complex will be largest when the second tone is nearer to the BF relative to the first tone, and will be smallest when the tone pair sequence is reversed. Group 4, where tones generally straddle the BF, produces intermediate GOF values. Interestingly, group 1, where both tones are near the BF, produces response patterns more typical of group 2 than groups 3 or 4. This finding is consistent with responses evoked by tones near the BF of a recording site acting as the most effective maskers for subsequent activity evoked by later sound stimuli.

Discussion

In this paper we test the hypothesis that VOT encoding is based, in part, on a temporal processing mechanism within auditory cortex. This physiological mechanism, in turn, represents one specific example of a more general process that facilitates the temporal ordering of acoustic events. The hypothesis, derived in part from the perceptual studies of Pisoni (1977), suggests that the presence of two response bursts evoked by consonant release and voicing onset supports the perception of two sequential acoustic events and an unvoiced stop consonant. This occurs in the setting of a prolonged VOT. At short VOTs, these two events are not resolved in time through two discrete time-locked responses, facilitating the perception of a voiced consonant. Our previous work examining population activity recorded directly from primary auditory cortex in monkeys and humans demonstrates that this hypothesis can help account for the 20–40 ms perceptual boundary between /da/ and /ta/ usually observed in American English (Steinschneider et al., 1995b, 1999, 2003).

Viability of this physiological hypothesis, however, also requires that it help account for the shorter 15–20 ms boundary that limits our ability to sequentially order two non-speech acoustic events (Hirsh, 1959; Stevens and Klatt, 1974; Miller et al., 1976; Pisoni, 1977). In this paper, we show that new onset responses evoked by both the first and second tones in a two-tone complex are reliably detected by population responses in A1 at a tone onset time separation as short as 20 ms. The minimal limit of ~20 ms is observed when both tones of the complex are near, or spectrally distant, from the BF of the recording sites, as well as when the second tone is near and the first tone is more distant from the BF. This physiological boundary parallels the perceptual data, thus supporting the relevance of a physiological processing mechanism based on synchronized onset responses for temporal order perception in audition.

The importance of synchronized, short-latency, stimulus-evoked responses within neuronal populations is a common theme in mammalian sensory cortex (e.g. Kreiter and Singer, 1996; Ehret, 1997; Phillips, 1998; Roy and Alloway, 2001; Temerencenca and Simons, 2003). Consistent with present results, it has been estimated that most stimulus-related information in primary visual and somatosensory cortices is represented by synchronized responses within 20 ms after cortical activation.
Furthermore, these synchronized responses are an especially powerful means by which A1 can effectively transmit information to secondary auditory areas for further sound processing (Eggermont, 1994; deCharms and Merzenich, 1996; see also Oram and Perrett, 1992). In addition to VOT, we have demonstrated how onset responses within A1 populations represent spectral features important for discrimination of stop consonant place of articulation, temporal pitch, musical consonance and dissonance, critical band behavior, and features of...
The relevance of synchronized onset responses in signaling temporal sound organization does not preclude the concurrent operation of other processing mechanisms. Synchronized, longer latency activity among neurons without an increase in firing rate, a property not examined in the present paper, occurs in A1 and likely plays an important role in the binding of multiple sound object attributes (deCharms and Merzenich, 1996). Neural mechanisms within A1 based on response rate instead of synchrony are an additional means by which temporal information can be physiologically encoded in cortex, especially for discrimination of rapidly changing stimuli (Lu et al., 1996). Neural mechanisms within A1 based on response rate instead of synchrony are an additional means by which temporal information can be physiologically encoded in cortex, especially for discrimination of rapidly changing stimuli (Lu et al., 1996). Neural mechanisms within A1 based on response rate instead of synchrony are an additional means by which temporal information can be physiologically encoded in cortex, especially for discrimination of rapidly changing stimuli (Lu et al., 1996). Neural mechanisms within A1 based on response rate instead of synchrony are an additional means by which temporal information can be physiologically encoded in cortex, especially for discrimination of rapidly changing stimuli (Lu et al., 1996).

**Table 1**

<table>
<thead>
<tr>
<th>Group 1</th>
<th>10</th>
<th>NS</th>
<th>NS</th>
<th>38</th>
<th>0.12 ± 0.07</th>
<th>0.12 ± 0.07</th>
<th>0.25 ± 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.15 ± 0.09</td>
<td>0.14 ± 0.08</td>
<td>0.28 ± 0.09</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>NS</td>
<td>&lt;0.01</td>
<td>50</td>
<td>0.16 ± 0.09</td>
<td>0.12 ± 0.07</td>
<td>0.27 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>NS</td>
<td>&lt;0.01</td>
<td>44</td>
<td>0.15 ± 0.09</td>
<td>0.12 ± 0.07</td>
<td>0.25 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>52</td>
<td>0.17 ± 0.08</td>
<td>0.11 ± 0.06</td>
<td>0.25 ± 0.10</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>43</td>
<td>0.11 ± 0.09</td>
<td>0.47 ± 0.21</td>
<td>0.48 ± 0.22</td>
</tr>
</tbody>
</table>

**Figure 13.** Goodness-of-fit values for curve fitting of the peak MUA responses to a one phase exponential decay function.

Despite the quantitative differences between the present and cited work, there is qualitative agreement on the masking effects of the first tone in suppressing activity to the second tone. In all studies, tones at or near the BF of a recording site are the most effective masker stimuli, whereas tones at a distance from the BF are the least effective in suppressing responses to a second tone (Brosch and Schreiner, 1997). Previous studies did not examine physiological temporal acuity of A1 when both tones are distant from the BF of a recording site. We find the same 20 ms limit in the ability of synchronized neuronal activity to detect the onsets of both tones, though the strength of this activity is not as great as when the second tone is near the BF. Thus, animal model data indicate that two-tone complexes elicit multiple temporal response patterns in A1 that have varying capacity to represent both tones. Strength of response than one based on the synchronous activation of large neural populations evoked by stimulus onsets. This would explain why only under specific, low uncertainty conditions or after extensive training can subjects make certain fine-grained VOT discriminations. A temporal mechanism based on synchronized onset responses would likely dominate in the typical acoustical environment of stimulus uncertainty.
to each tone at any given site in A1 is based on the frequencies within the complex and their relationship to the BF at that site.

The human data complement findings in the monkey. Multiple temporal patterns with varying capacity to represent the onsets of consonant release and voicing occur across the three recording sites in anterior Heschl’s gyrus. The most lateral site has the greatest capacity to represent voicing onset at the shortest VOT intervals, the most medial site the least, while the central site is intermediate. Human primary auditory cortex has a tonotopic organization with lower frequencies best represented laterally and progressively higher frequencies represented medially on Heschl’s gyrus (Howard et al., 2001; Liègeois-Chauvel et al., 2001; Schönewieser et al., 2002; Formisano et al., 2003). Discussed in terms of a simplified two-tone complex, the relative strength of the response to voicing onset is greatest at the lateral site because the first tone (higher formants) is at a spectral distance from the BF and the second tone (F1) is near the BF of the recording site. In contrast, the medial site with a higher BF is a location whose first tone (higher formants) is near the BF and whose second tone (F1) is at a distance from the BF. This combination produces the least capacity for a second stimulus component to elicit a response time-locked to its onset. Responses at the center site are intermediate between these two extremes.

While each location in Heschl’s gyrus has a varying capacity to represent the onsets of consonant release and voicing, the temporal response pattern averaged across the three recording sites roughly mirrors the patient’s perceptual boundary shifts as F1 frequency is modulated. Specifically, at the lowest F1 frequency of 424 Hz, the perceptual boundary shifts from 20–25 ms to between 40 and 60 ms. In parallel, a discrete response to voicing onset is only seen at a VOT of 60 ms. Contrasting this pattern are those observed when the syllables contain higher F1 frequencies. Discrete responses evoked by voicing onset are now observed at shorter VOTs, and they maintain statistically significant increases above preceding activity to within 5 ms of the perceptual boundaries. The absence of a perfect correlation between the AEPs evoked by the higher F1 syllables and the perceptual boundaries for these stimulus sets likely reflects, in addition to the low statistical power of the single subject analysis, the fact that the presence or absence of responses to voicing onset can not be the only determinant of the voiced/voiceless distinction. For instance, the intensity and duration of aspiration noise are important cues for this perceptual discrimination (Sinnott and Adams, 1987; Lotto and Kluender et al., 2002), yet their effects upon perceptual boundaries are not evident in these AEP recordings.

Even though the observations of rough parallels between perception and temporal response patterns are limited to a single subject, we were then able to replicate parallels between physiological and perceptual boundaries using a different /da/-/ta/ series. As an additional check on whether averaged activity across auditory cortex can reflect perceptual boundaries, we reanalyzed our previously published data on VOT representation by examining averaged activity profiles across electrode sites in the human and across tonotopic regions in the monkey (Steinschneider et al., 1999, 2003). We averaged activity from subject 1 in the human study, whose three low-impedance electrode sites spanning 20 mm were amenable for analysis. In both the human and monkey data, distinct differences in response patterns were observed across the averaged responses between those evoked by /da/ with a VOT of 0 and 20 ms and those elicited by /ta/ with a VOT of 40 and 60 ms. Differences reflected a new response time-locked to voicing onset for the longer VOT stimuli (data available upon request). Thus, physiological findings support a temporal processing mechanism for VOT encoding and further suggest that the perceptual boundary is partially determined by response patterns averaged across primary auditory cortex.

Several factors likely contribute to the decreased capacity of the syllables with the lowest F1 to generate an early response to voicing onset in the averaged population responses. First, consideration of spectral tuning characteristics in A1 means that there will be a decreased contribution of the response to the 424 Hz F1 spectral component relative to the higher F1s at all but the lowest BF areas. This smaller response contribution to the average will require a longer VOT for F1 onset to be physiologically detected above the exponentially decaying activity evoked by the earlier consonant release. Compounding this effect is the diminished auditory sensitivity to the 424 Hz F1 frequency relative to F1s centered at 600 and 848 Hz (e.g. Owren et al., 1988). This diminished sensitivity translates into a functionally less intense sound component that will lead to a smaller neural response that will require a more prolonged decay of earlier activity in order for F1 onset to be identified as a new acoustic event.

Averaged population activity as a determinant for a behavioral or perceptual outcome has been repeatedly reported in both motor and sensory systems. For instance, perception of visual motion is guided by the averaged activity within area MT (Kruse

**Figure 14.** Goodness-of-fit values for curve fitting of the peak MUA responses (top graph) and peak PSTH responses (bottom graph) to a one phase exponential decay function. See text for details.
et al., 2002; Ditterich et al., 2003). Similarly, in motor and prefrontal areas, complex hand and finger movements are directed by the averaged activity of large neuronal ensembles (e.g. Georgopoulos et al., 1999; Schwartz and Moran, 1999; Averbeck et al., 2003). Generally, template-matching procedures such as population vector or maximum likelihood estimations are used to approximate the population code (Pouget et al., 2000). Ultimately, these procedures examine the overall shape and amplitude of the population activity in order to derive information regarding stimulus features or motor commands. By analogy, we propose that a physiologically plausible template in A1 for a single acoustic event is a fast rise in neuronal activity followed by a rapid exponential decay. Significant deviations from this template, as determined by the averaged activity across A1, would support a perceptual decision that more than one event has occurred in time.

Before concluding, multiple issues deserve consideration. One regards the degree to which activity in Heschl’s gyrus on the right/non-language dominant hemisphere is involved in speech perception. Both neuroimaging and behavioral studies support the importance of the right hemisphere for VOT processing (Simos et al., 1997; Laguitton et al., 2000; Jäncke et al., 2002; Papanicolaou et al., 2003). Another issue relates to whether our recordings were sufficiently extensive to adequately sample patterns of evoked activity. Electrode sites spanned 6.7 mm in the subject in this study. Anatomical maps of human A1 suggest a maximum extent of ~10–12 mm (Hackett et al., 2001; Wallace et al., 2002). The volume-conducted nature of AEPs coupled to the recording span suggests that activity patterns were reasonably approximated by our sample.

A significant concern is whether activity profiles in a patient with epilepsy reflect normal or aberrant auditory processing. While caution must always be exercised in extrapolating normative physiologic processes from data obtained in subjects with epilepsy, there are several reasons to believe that the patterns observed in the present study represent reasonable indices of normal functions. First, temporal response patterns observed are similar to those reported by other studies examining intracranially acquired AEPs, and all conform to the known perceptual relevance of the VOT parameter (Liégeois-Chauvel et al., 1999; Steinschneider et al., 1999). Secondly, these similarities include differential sensitivities for generating responses evoked by voicing onset in anterior Heschl’s gyrus relative to more posterior regions. Thirdly, the greater amplitude of AEPs recorded from more posterior auditory cortex that is presently observed has been reported (Liégeois-Chauvel et al., 1994). Finally, while these latter studies were all performed in patients with epilepsy, they, in turn, reveal temporal patterns of activity that are also mirrored in the magnetic responses and AEPs of neurologically normal subjects (e.g. Kaukoranta et al., 1987; Joliot et al., 1994; Kuriki et al., 1995; Sharma and Dorman, 1999). In summary, the reproducibility and similarities between present results and previously reported findings enhance their potential relevance as indices of normal auditory cortical functions.

An additional issue relates to the correspondence between the intracranial data and speech-evoked activity examined using surface-recorded evoked potentials and magnetic responses. There is a decrease in the amplitude of the N1m component of the magnetic responses to speech sounds or two-tone analogs with prolonged VOTs and TOTs that correlates with perceptual boundaries (Simos et al., 1998a–c; see, however, Tremblay et al., 2003). This effect is likely based on the truncation of the N1m evoked by consonant onset by new positive-going components evoked by voicing onset (for a detailed demonstration, see Steinschneider et al., 1999). The intracranial data indicate that at short VOTs/TOTs a new response complex evoked by voicing onset is likely to be severely attenuated in amplitude, leading to an overall increase in the size of the resultant N1m component when compared against those evoked when longer VOT/TOT stimuli are presented. Thus, there is general agreement between the magnetic and intracranial responses in terms of identifying physiological activity patterns that roughly correlate with perceptual features.

Several studies, however, cast some doubt on this correlation between physiology and perception. For example, one study failed to find a parallel between the presence or absence of a single- or double-peaked N1 AEP component and shifting VOT perceptual boundaries that occurred with changes in consonant place of articulation (Sharma et al., 2000). This study needs to be carefully interpreted. The N1 component is a composite wave with multiple generators in primary and secondary auditory cortex (e.g. Wood and Wolpaw, 1982; Näätänen and Picton, 1987; Scherg et al., 1989; Liégeois-Chauvel et al., 1994; Krumholz et al., 2003). Furthermore, the dominant contributor to the scalp-recorded N1 is likely an extensive area of auditory cortex posterior to anterior Heschl’s gyrus, including the planum temporale (Liégeois-Chauvel et al., 1994). We find large differences in the capacity to respond to acoustic transients between anterior Heschl’s gyrus and cortex located more posteriorly. More subtle differences are observed within anterior Heschl’s gyrus. These distinct temporal patterns, which overlap in time, indicate that great caution must be exercised when suggesting detailed aspects of auditory cortical organization based on the modulation of a composite wave whose morphology is the result of activity in functionally disparate, yet closely spaced, auditory cortical regions.

More germane to the question of the correlation between physiology and perception are the findings of a study by Sharma and Dorman (2000). This study compared the morphology of N1 in response to bilabial consonant-vowel syllables varying in VOT from ~90 to 0 ms in Hindi and English listeners. In the former language, those syllables with prolonged pre-voicing (~<30 ms) are perceived as /ba/, while those with short VOTs are perceived as /pa/. In English, all these syllables are perceived as /ba/. Latency shifts in N1 latency correlated with VOT, but not perception, as this effect was observed in listeners from both languages. This finding indicates that the obligatory temporal response patterns we observe in primary auditory cortex help shape, but are not the ultimate determinants of, the phonetic perception. It further highlights the importance of language experience, and is in keeping with the known multiple auditory, visual, lexical and linguistic cues that all contribute to phonetic perception.

The nature of the neural elements recorded by our lower impedance electrodes in monkey A1 needs to be also addressed. Multiple lines of evidence support the conclusion that the major contributors to the MUA and PSTHs are cortical action potentials. First, the PSTHs recorded from middle laminae are derived from higher amplitude spikes, while the very small diameter of distal thalamocortical axons will generally produce lower amplitude spikes. The fact that the PSTHs and MUA have nearly identical latencies and other response characteristics supports their predominant cortical origin. Secondly, these responses are concurrent with intracortical negativities in the AEP and CSD...
sources and sinks whose dipolar spatial distribution are indicative of pyramidal cell activation (e.g. Steinschneider et al., 1994, 2003; Kisley and Gerstein, 1999; Rose and Metherate, 2001; Cruikshank et al., 2002). Thirdly, latency of the responses is in accord with other studies examining single cell activity in A1 (e.g. Phillips and Hall, 1990; Heil, 1997; Recanzone et al., 2000; Cheung et al., 2001). Finally, the earliest thalamocortical fiber volley in awake monkey A1 has an onset latency of 5–6 ms and a peak at 8–10 ms (Steinschneider et al., 1992, 2003). This activity is earlier than the responses seen in the MUA and PSTHs in the present study, indicating that cortical cells are the predominant elements recorded by our electrodes. However, it must be acknowledged that a small contribution from TC fibers to the neural responses cannot be excluded.

Finally, the relationship between the MUA/PSTH responses in the monkey and the AEP components in the human needs to be assessed. The monkey responses represent the initial activation of A1 with a peak at 15–20 ms post-stimulus onset (see Steinschneider et al., 1994, 2003; Eggermont and Ponton, 2002). In contrast, the principal AEP component examined is a positivity whose peak is ~60 ms. We and others have suggested that the homolog of this component in monkeys is a large positive wave peaking around 28 ms that is primarily generated by polysynaptic depolarizations within upper lamina 3 (Steinschneider et al., 1994; Eggermont and Ponton, 2002). The resultant current sinks are balanced by more superficial sources, leading to the positivity recorded at the scalp. While the homology between the monkey and human responses are therefore not direct, similar patterns of activity with respect to VOT encoding occur in the upper lamina 3 sinks and more superficial sources of monkey A1 (Steinschneider et al., 2003). This laminar profile suggests that initial activation in lower lamina 3 induces later synaptic events in upper lamina 3 that would be manifested in the large positive wave in the human.

In conclusion, physiological findings support a temporal processing mechanism in primary auditory cortex as important for neural encoding of VOT. Findings in the monkey bolster the hypothesis that VOT encoding represents, in part, a specific instance of a more general process governing the ability to identify the sequential order of sound events. Perceptual findings using non-speech stimuli modeling VOT indicate that a separation of ~20 ms between the onsets of two acoustic events is required for this identification. We show a nearly identical capacity in physiological response patterns of A1 populations. These A1 temporal response patterns are also systematically modulated by interactions between temporal and spectral sound components. Within primary auditory cortex of both humans and monkeys, this modulation appears to be based on the relationship between the tonotopically-organized recording location and the specific frequency components of the sounds. When viewed across the array of activated tissue, the composite temporal response patterns of large-scale neural populations in human A1 vary in a manner that supports a physiologically plausible explanation for the trading relations effect between F1 frequency and VOT boundaries. Given these positive findings, it is essential to appreciate that primary auditory cortical temporal response patterns represent just one informational component that can be used to facilitate discrimination of voiced from unvoiced phonemes. This perceptual process is ultimately decided by the activity of large-scale neural networks utilizing multiple acoustic cues, visual inputs, and higher-order lexical and linguistic constructs.

References


Fishman YI, Volkov IO, Noh MD, Garell PC, Bakken H, Arezzo JC, Howard MA, Steinschneider M (2001b) Consonance and dissonance of...


