Hearing Research 305 (2013) 57–73

Contents lists available at SciVerse ScienceDirect

Hearing Research

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

Review

# Representation of speech in human auditory cortex: Is it special?

Mitchell Steinschneider[a,b,*], Kirill V. Nourski[c], Yonatan I. Fishman[a]

[a] Department of Neurology, Rose F. Kennedy Center, Albert Einstein College of Medicine, Room 322, 1300 Morris Park Avenue, Bronx, NY 10461, USA
[b] Department of Neuroscience, Rose F. Kennedy Center, Albert Einstein College of Medicine, Room 322, 1300 Morris Park Avenue, Bronx, NY 10461, USA
[c] Department of Neurosurgery, The University of Iowa, Iowa City, IA 52242, USA

ARTICLE INFO

*Article history:*
Received 30 January 2013
Received in revised form
13 May 2013
Accepted 28 May 2013
Available online 18 June 2013

ABSTRACT

Successful categorization of phonemes in speech requires that the brain analyze the acoustic signal along both spectral and temporal dimensions. Neural encoding of the stimulus amplitude envelope is critical for parsing the speech stream into syllabic units. Encoding of voice onset time (VOT) and place of articulation (POA), cues necessary for determining phonemic identity, occurs within shorter time frames. An unresolved question is whether the neural representation of speech is based on processing mechanisms that are unique to humans and shaped by learning and experience, or is based on rules governing general auditory processing that are also present in non-human animals. This question was examined by comparing the neural activity elicited by speech and other complex vocalizations in primary auditory cortex of macaques, who are limited vocal learners, with that in Heschl's gyrus, the putative location of primary auditory cortex in humans. Entrainment to the amplitude envelope is neither specific to humans nor to human speech. VOT is represented by responses time-locked to consonant release and voicing onset in both humans and monkeys. Temporal representation of VOT is observed both for isolated syllables and for syllables embedded in the more naturalistic context of running speech. The fundamental frequency of male speakers is represented by more rapid neural activity phase-locked to the glottal pulsation rate in both humans and monkeys. In both species, the differential representation of stop consonants varying in their POA can be predicted by the relationship between the frequency selectivity of neurons and the onset spectra of the speech sounds. These findings indicate that the neurophysiology of primary auditory cortex is similar in monkeys and humans despite their vastly different experience with human speech, and that Heschl's gyrus is engaged in general auditory, and not language-specific, processing.

*This article is part of a Special Issue entitled "Communication Sounds and the Brain: New Directions and Perspectives".*

## 1. Introduction

### 1.1. Complexity of phonemic perception

The ease with which speech is perceived underscores the refined operations of a neural network capable of rapidly decoding complex acoustic signals and categorizing them into meaningful phonemic sequences. A number of models have been devised to explain how phonemes are extracted from the continuous stream of speech (e.g., McClelland and Elman, 1986; Church, 1987; Pisoni and Luce, 1987; Stevens, 2002). Common to all these models is the recognition that phonemic perception is a categorization task based on sound profiles derived from a multidimensional space encompassing numerous acoustic features unfolding over time (Holt and Lotto, 2010). Features are all characterized by acoustic parameters that vary along intensity, spectral, and temporal dimensions. Increased intensity, especially

*Abbreviations:* A1, primary auditory cortex; AEP, averaged evoked potential; BF, best frequency; CSD, current source density; CV, consonant–vowel; ERBP, event-related-band-power; FRF, frequency response function; HG, Heschl's gyrus; MEG, magnetoencephalographic; MUA, multiunit activity; POA, place of articulation; SRCSD, summed rectified current source density; STG, superior temporal gyrus; tBMF, temporal best modulation frequency; VOT, voice onset time

[*] Corresponding author. Department of Neurology, Rose F. Kennedy Center, Albert Einstein College of Medicine, Room 322, 1300 Morris Park Avenue, Bronx, NY 10461, USA. Tel.: +1 718 430 4115; fax: +1 718 430 8588.

*E-mail addresses:* mitchell.steinschneider@einstein.yu.edu (M. Steinschneider), kirill-nourski@uiowa.edu (K.V. Nourski), yonatan.fishman@einstein.yu.edu (Y.I. Fishman).

0378-5955/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.heares.2013.05.013

in the low to mid-frequency ranges, helps to distinguish vowels from consonants (McClelland and Elman, 1986; Stevens, 2002). Distinct spectral (formant) patterns during these periods of increased intensity promote accurate vowel identification (Hillenbrand et al., 1995).

The temporal dimension of phonemic categorization has received increased attention in recent years. An influential proposal posits that speech perception occurs over several overlapping time scales (e.g., Poeppel et al., 2008, 2012; Giraud and Poeppel, 2012). Syllabic analyses occur within a time frame of about 150–300 m, and correlate with the amplitude envelope of speech. Speech comprehension remains high even when sentence fragments are time-reversed in 50 ms bins, and only becomes severely degraded when time-reversals occur at frequencies overlapping those of the speech envelope (Saberi and Perrott, 1999). Furthermore, temporal smearing of the speech envelope leads to significant degradation in the intelligibility of sentences only at frequencies commensurate with the speech envelope (Drullman et al., 1994).

More refined acoustic feature analyses are performed within shorter temporal windows of integration that vary between about 20 and 80 m. Segmentation of speech within this range is critical for phonetic feature encoding, especially for shorter duration consonants. Times at which rapid temporal and spectral changes occur are informationally rich landmarks in the speech waveform (Stevens, 1981, 2002). Both the spectra and formant transition trajectories occurring at these landmarks are crucial for accurate identification of true consonants such as the stops (Kewley-Port, 1983; Walley and Carrell, 1983; Alexander and Kluender, 2009). Voice onset time (VOT), the time between consonant release and the onset of rhythmic vocal cord vibrations, is a classic example of rapid temporal discontinuities that help to distinguish voiced consonants (e.g., /b/, /d/, and /g/) from their unvoiced counterparts (e.g., /p/, /t/, and /k/) (e.g., Lisker and Abramson, 1964; Faulkner and Rosen, 1999). Indeed, when semantic information is lacking, listeners of time-reversed speech have significant comprehension difficulties at the shorter temporal intervals required for phonetic feature encoding (Kiss et al., 2008).

### 1.2. Complexity of neural networks supporting phonetic processing

Early stations in the human auditory system are exquisitely tuned to encode speech-related acoustic features. Population brainstem responses accurately represent the intensity, spectrum, and temporal envelope of speech sounds (Chandrasekaran et al., 2009; Anderson and Kraus, 2010). Magnetoencephalographic (MEG) responses reflect consonant place of articulation (POA) within 50 ms after sound onset (Tavabi et al., 2007), and within 100 ms, responses differentiate intelligible versus unintelligible speech (Obleser et al., 2006). Neural responses obtained from intracranial recordings in Heschl's gyrus (HG), the putative location of primary auditory cortex in humans (Hackett et al., 2001), demonstrate categorical-like changes to syllables that vary in their VOT in a manner that parallels perception (Steinschneider et al., 1999, 2005). Spectrotemporal receptive fields derived from single unit activity in HG elicited by one portion of a movie soundtrack dialog can accurately predict response patterns elicited by a different portion of the same dialog (Bitterman et al., 2008). Finally, both MEG responses and responses obtained from invasive recordings within HG have shown that accurate tracking of the speech envelope degrades in parallel with the ability to perceive temporally compressed speech (Ahissar et al., 2001; Nourski et al., 2009; see also Peelle et al., 2013). These observations lend support to the conclusion that "acoustic–phonetic features of the speech signal such as voicing, spectral shape, formants or amplitude modulation are made accessible by the computations of the ascending auditory pathway and primary auditory cortex" (Obleser and Eisner, 2008, p. 16).

### 1.3. Plasticity of phonetic perception and neural function

An important and unresolved question is whether the representation of acoustic features of speech in the brain is based on neural processing mechanisms that are unique to humans and shaped by learning and experience with an individual's native language. The role of experience in modifying auditory cortical physiology is prominently observed during early development. The appearance of the mismatch negativity component of the event-related potential becomes restricted to native-language phonemic contrasts by 7½ months of age (Kuhl and Rivera-Gaxiola, 2008). Better native language-specific responses predict enhanced language skills at two years of age. The emergence of new event-related potentials that parallel developmental milestones in speech processing provides an additional example of neural circuitry changes derived from language experience (Friederici, 2005). In adults, both gray matter volume of primary auditory cortex and the amplitude of short-latency auditory evoked potentials generated in primary auditory cortex are larger in adult musicians than in musically-naïve subjects (Schneider et al., 2002). Recordings from animal models that are complex vocal learners such as songbirds also demonstrate pronounced modifications that occur in auditory forebrain processing of sound based on developmental exposure to species-specific vocalizations (e.g., Woolley, 2012). In sum, it remains unclear how "special" or unique in mammalian physiology human primary auditory cortex is with regard to decoding the building blocks of speech.

### 1.4. Cortical bases of speech perception: is human primary auditory cortex special?

Here, we examine this question by comparing the neural activity elicited by speech in primary auditory cortex (A1) of macaque monkeys, who are limited vocal learners, with that in HG of humans, who are obviously expert vocal learners (Petkov and Jarvis, 2012). Neural activity from human primary auditory cortex was acquired during intracranial recordings in patients undergoing surgical evaluation for medically intractable epilepsy. Measures included averaged evoked potentials (AEPs) and event-related-band-power (ERBP) in the high gamma (70–150 Hz) frequency range. Comparable population recordings were performed in the macaques. Measures included AEPs, the derived current source density (CSD), and multiunit activity (MUA). The focus of this report will be on clarifying the neural representation of acoustic features of speech that vary along both temporal and spectral dimensions. Some of the results represent a summary of previous studies from human and monkey primary auditory cortex. The remainder of the results represents new data that extend the previous findings. If perceptually-relevant features of speech are encoded similarly in humans and monkeys, then it is reasonable to conclude that human primary auditory cortex is not special.

## 2. Materials and methods

### 2.1. Monkey

#### 2.1.1. Subjects

Results presented in this report represent neurophysiological data obtained from multiple male monkeys (*Macaca fascicularis*) that have been accumulated over many years. During this time, there have been gradual changes in methodology. The reader is referred to the cited publications for methodological details (i.e., Fig. 3, Steinschneider et al., 2003; six subjects; Fig. 8, Steinschneider and Fishman, 2011; four subjects). Methods described here refer to studies involving two monkey subjects whose data are reported for the first time in this paper (Figs. 1, 4 and 6) (see Fishman and Steinschneider, 2012 for methodological details). For all monkey subjects, data were obtained from A1 in both hemispheres, and no significant laterality effects were appreciated. All experimental procedures were reviewed and approved by the AAALAC-accredited Animal Institute of Albert Einstein College of Medicine and were conducted in accordance with institutional and federal guidelines governing the experimental use of primates. Animals were housed in our AAALAC-accredited Animal Institute under daily supervision of laboratory and veterinary staff. They were routinely provided with recommended environmental enrichment protocols and regular use of expanded-size exercise units. Animals were acclimated to the recording environment and trained while sitting in custom-fitted primate chairs prior to surgery.

### 2.1.2. Surgical procedure

Under pentobarbital anesthesia and using aseptic techniques, holes were drilled bilaterally into the dorsal skull to accommodate matrices composed of 18-gauge stainless steel tubes glued together in parallel. The tubes helped guide electrodes toward A1 for repeated intracortical recordings. Matrices were stereotaxically positioned to target A1. They were oriented at a 30° anterior–posterior angle and with a slight medial–lateral tilt in order to direct electrode penetrations perpendicular to the superior surface of the superior temporal gyrus, thereby satisfying one of the major technical requirements of one-dimensional CSD analysis (Müller-Preuss and Mitzdorf, 1984; Steinschneider et al., 1992). Matrices and Plexiglas bars, used for painless head fixation during the recordings, were embedded in a pedestal of dental acrylic secured to the skull with inverted bone screws. Peri- and post-operative antibiotic and anti-inflammatory medications were always administered. Recordings began no earlier than two weeks after surgery, thus allowing for adequate post-operative recovery.

### 2.1.3. Stimuli

Stimuli used in this study included pure tones, consonant–vowel (CV) syllables varying in their VOT or POA, monkey vocalizations, and words. The pure tones were generated and delivered at a sample rate of 48.8 kHz by a PC-based system using an RX8 module (*Tucker-Davis Technologies*). Frequency response functions (FRFs) based on pure tone responses characterized the spectral tuning of the cortical sites. Pure tones used to generate the FRFs ranged from 0.15 to 18.0 kHz, were 200 ms in duration (including 10 ms linear rise/fall ramps), and were pseudo-randomly presented with a stimulus onset-to-onset interval of 658 ms. Resolution of FRFs was 0.25 octaves or finer across the 0.15–18.0 kHz frequency range tested.

In the two newest subjects, stimuli were presented from a free-field speaker (Microsatellite; *Gallo*) located 60° off the midline in the field contralateral to the recorded hemisphere and 1 m away from the animal's head (*Crist Instruments*). Sound intensity was measured with a sound level meter (type 2236; *Bruel and Kjaer*) positioned at the location of the animal's ear. The frequency

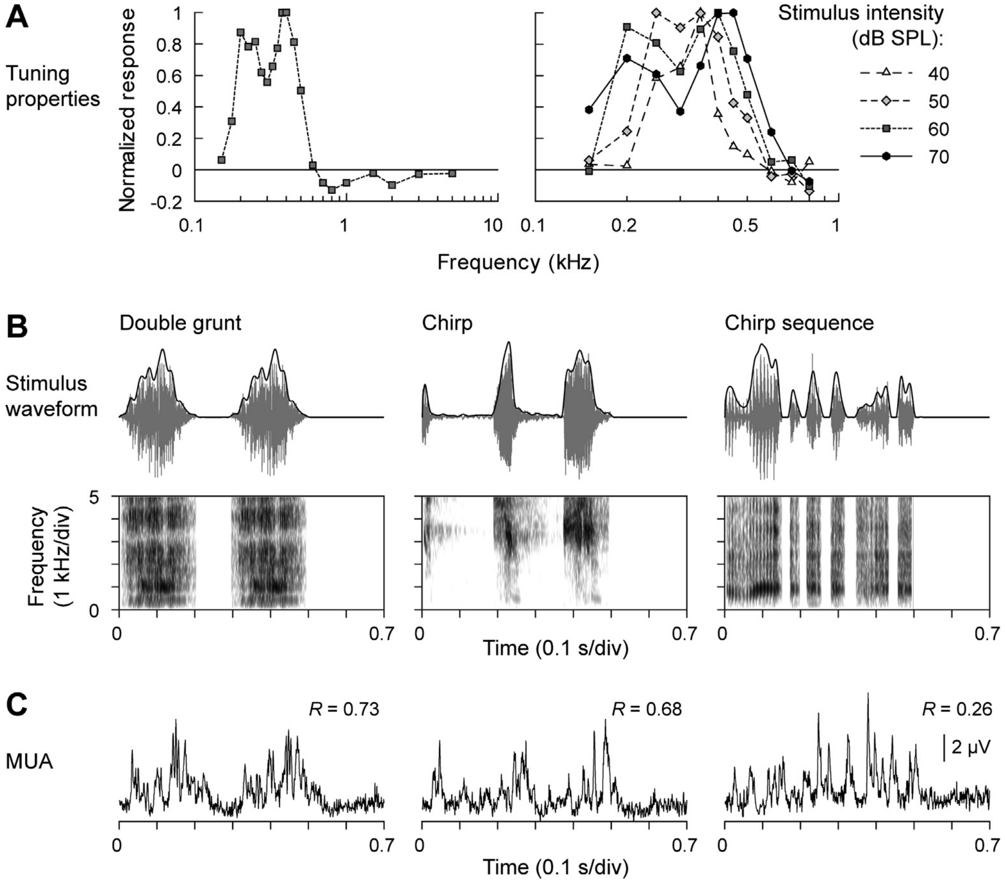

**Fig. 1.** Representation of the temporal envelope of monkey vocalizations in monkey A1. A. Frequency response functions describing the spectral selectivity of the studied multiunit cluster. Left panel: responses to 60 dB SPL tones, presented at frequencies between 0.15 and 5 kHz. Right panel: responses to tones, presented at intensities ranging from 40 to 70 dB SPL, and frequencies between 0.15 and 0.8 kHz. B. Top row: Stimulus waveforms (gray) and temporal envelopes (black) of the three vocalizations. Bottom row: Stimulus spectrograms. C. MUA recorded from lower lamina 3. Numbers indicate peak values of the cross-correlograms between vocalization envelopes and MUA.

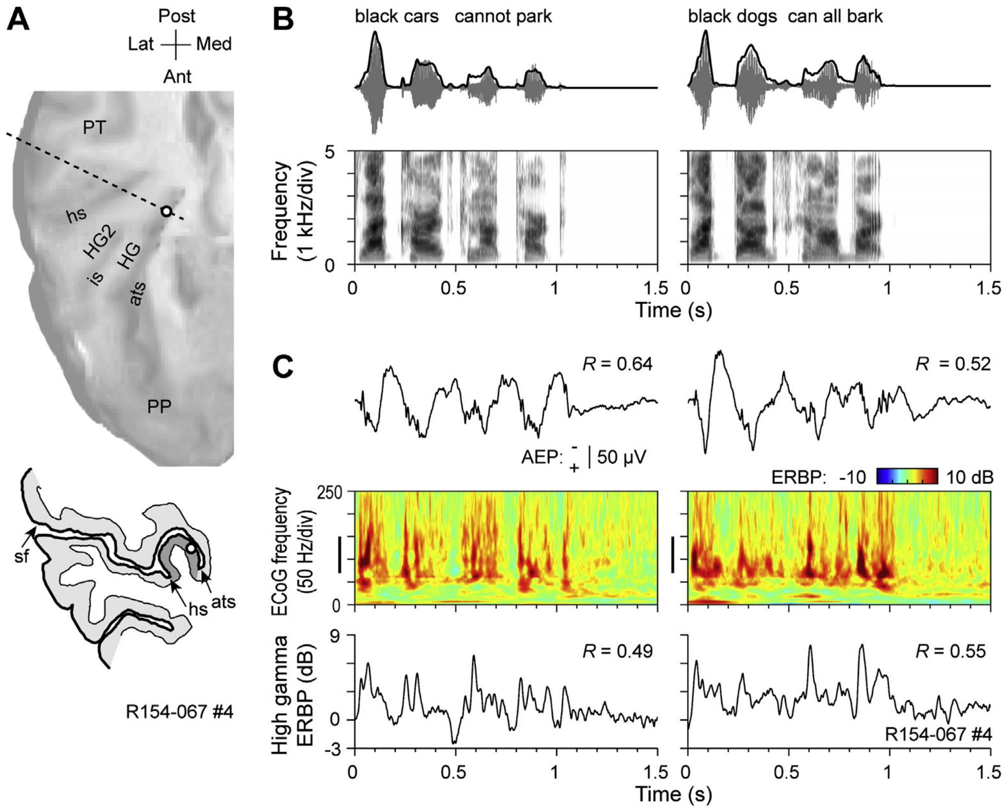

**Fig. 2.** Representation of the temporal envelope of speech in human auditory cortex. Data set from Nourski et al. (2009). A. Location of the recording contact (open circle). MRI surface rendering of the superior temporal plane and tracing of the MRI cross section (dashed line) are shown in the top and bottom panels, respectively. HG, Heschl's gyrus; HG2, second transverse gyrus, PP, planum polare; PT; planum temporale; ats, anterior temporal sulcus; is, intermediate sulcus; hs, Heschl's sulcus; sf, Sylvian fissure. B. Top row: Stimulus waveforms (gray) and temporal envelopes (black) of the two speech sentences. Bottom row: Stimulus spectrograms. C. Responses to the speech sentences. AEP waveforms, ERBP time–frequency plots and high gamma ERBP waveforms are shown in the top, middle and bottom row, respectively. Vertical bars in ERBP time–frequency plots indicate high gamma frequency range. Numbers indicate peak values of the cross-correlograms between stimulus envelopes and AEP and high gamma ERBP waveforms. Negative voltage of the AEPs is plotted upwards.

response of the speaker was essentially flat (within ±5 dB SPL) over the frequency range tested. In other subjects, sounds were presented from a dynamic headphone (MDR-7502; *Sony*) coupled to a 60-cc plastic tube that was placed against the ear contralateral to the recording site.

CV syllables (175 ms duration) varying in their VOT (0, 20, 40 and 60 ms VOT) were generated at the Haskins Laboratories (New Haven, CT). Details regarding their acoustic parameters and presentation hardware and software can be found in Steinschneider et al. (2003). Syllables varying along their POA were constructed on the parallel branch of a KLSYN88a speech synthesizer, contained 4 formants, and were also 175 ms in duration. Details can be found in Steinschneider and Fishman (2011). Macaque vocalizations were kindly provided by Dr. Yale Cohen. Words and other assorted environmental stimuli were obtained as freeware from various sites on the Internet. The monkey vocalizations, words, and other environmental sounds were edited to be 500 ms in duration, downsampled to 24,414 Hz, and presented via the *Tucker-Davis Technologies* software and hardware described above.

### 2.1.4. Neurophysiological recordings

Recordings were conducted in an electrically shielded, sound-attenuated chamber. Monkeys were monitored via closed-circuit television. The two newest subjects performed a simple auditory discrimination task to promote attention to the sounds during the recordings. The task involved release of a metal bar upon detection of a randomly presented noise burst interspersed among the test stimuli. To further maintain subjects in an alert state, an investigator entered the recording chamber and delivered preferred treats to the animals prior to the beginning of each stimulus block.

Neural population measures were examined. Recordings were performed using linear-array multi-contact electrodes comprised of 16 contacts, evenly spaced at 150 µm intervals (*U-Probe; Plexon*). Individual contacts were maintained at an impedance of about 200 kΩ. An epidural stainless-steel screw placed over the occipital cortex served as the reference electrode. Neural signals were band-pass filtered from 3 Hz to 3 kHz (roll-off 48 dB/octave), and digitized at 12.2 kHz using an RA16 PA Medusa 16-channel preamplifier connected via fiber-optic cables to an RX5 data acquisition system (*Tucker-Davis Technologies*). Local field potentials time-locked to the onset of the sounds were averaged on-line by computer to yield AEPs. CSD analyses characterized the laminar pattern of net current sources and sinks within A1 generating the AEPs. CSD was calculated using a 3-point algorithm that approximates the second spatial derivative of voltage recorded at each recording contact (Freeman and Nicholson, 1975). To derive MUA, signals were simultaneously high-pass filtered at 500 Hz (roll-off 48 dB/octave), full-wave rectified, and then low-pass filtered at 520 Hz (roll-off 48 dB/octave) prior to digitization and averaging (see Supèr and Roelfsema, 2005 for a methodological review). MUA is a measure of the envelope of summed action potential activity of neuronal ensembles within a sphere estimated to be about 100 µm in diameter (Brosch et al., 1997; Supèr and Roelfsema, 2005).

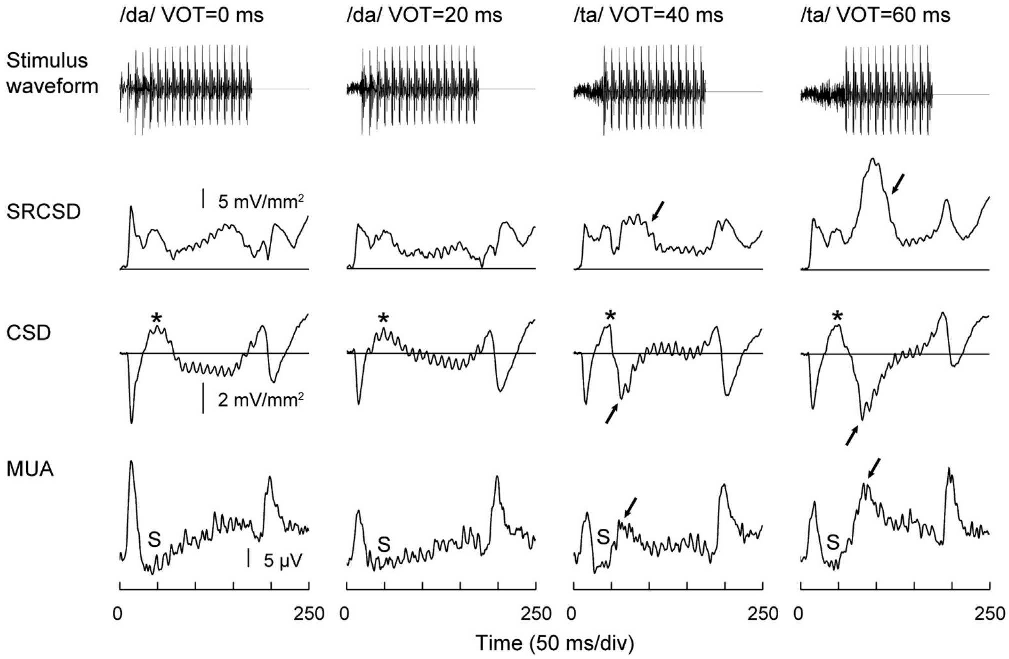

**Fig. 3.** Representation of VOT in monkey A1. Responses to 175 ms CV syllables with VOTs of 0, 20, 40 and 60 ms (left to right columns). Top-to-bottom rows: Stimulus waveforms, summed rectified CSD (SRCSD), lower lamina 3 CSD and lower lamina 3 MUA. Response components representing voicing onset are indicated by arrows. CSD sources and MUA suppression, are denoted by asterisks and 'S', respectively.

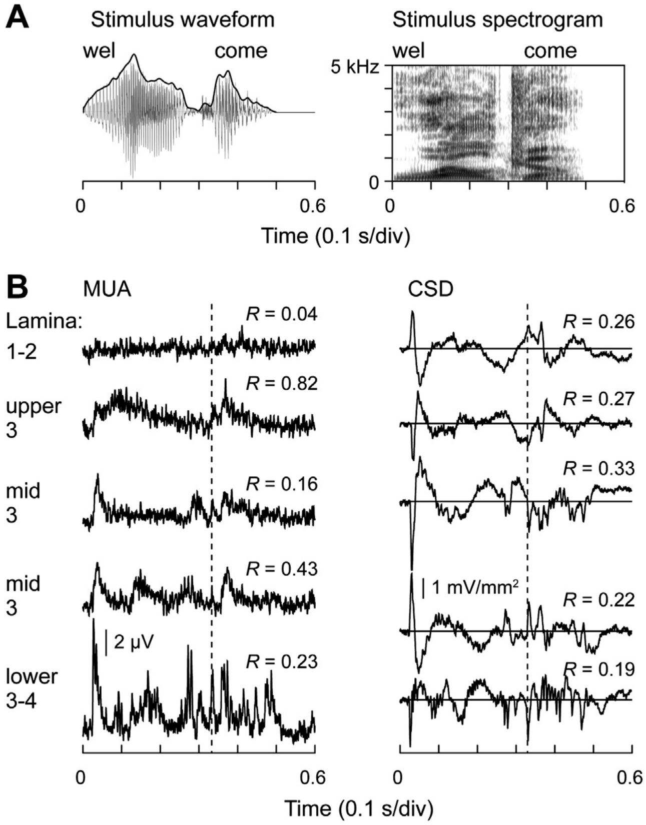

**Fig. 4.** Representation of temporal features of a spoken word ("welcome", articulated by a male speaker) in monkey A1. A. Stimulus waveform (left) and spectrogram (right). B. CSD and MUA plots (left and right column, respectively) recorded at five laminar depths, as indicated. Numbers indicate peak values of the cross-correlograms between stimulus envelopes and response waveforms. Dashed line indicates the timing of the response to the onset of the consonant /k/.

Positioning of electrodes was guided by on-line examination of click-evoked AEPs. Experimental stimuli were delivered when the electrode channels bracketed the inversion of early AEP components and when the largest MUA and initial current sink (see below) were situated in the middle channels. Evoked responses to ~40 presentations of each pure tone stimulus were averaged with an analysis time of 500 ms (including a 100-ms pre-stimulus baseline interval). The best frequency (BF) of each cortical site was defined as the pure tone frequency eliciting the maximal MUA within a time window of 10–75 m post-stimulus onset (Fishman and Steinschneider, 2009). Following determination of the BF, test stimuli were presented and averaged with an analysis time of 700 ms.

At the end of the recording period, monkeys were deeply anesthetized with sodium pentobarbital and transcardially perfused with 10% buffered formalin. Tissue was sectioned in the coronal plane (80 μm thickness) and stained for Nissl substance to reconstruct the electrode tracks and to identify A1 according to previously published physiological and histological criteria (e.g., Morel et al., 1993). Based on these criteria, all electrode penetrations considered in this report were localized to A1, though the possibility that some sites situated near the lower-frequency border of A1 were located in field R cannot be excluded.

### 2.1.5. *General data analysis*

CSD profiles were used to identify the laminar locations of the recording sites (e.g., Steinschneider et al., 1992, 1994; Fishman and Steinschneider, 2009). Typically, the laminar profile would include: 1) an initial current sink in lower lamina 3/lamina 4 that is balanced by more superficial and deep sources, 2) a slightly later supragranular sink that is maximal in more superficial depths of lamina 3, and 3) a more superficial current source in laminae 1/2 that is concurrent with the supragranular sink. MUA was normalized to baseline values occurring prior to stimulus onset before analysis. Details of analyses relevant for each data set are presented in the Results.

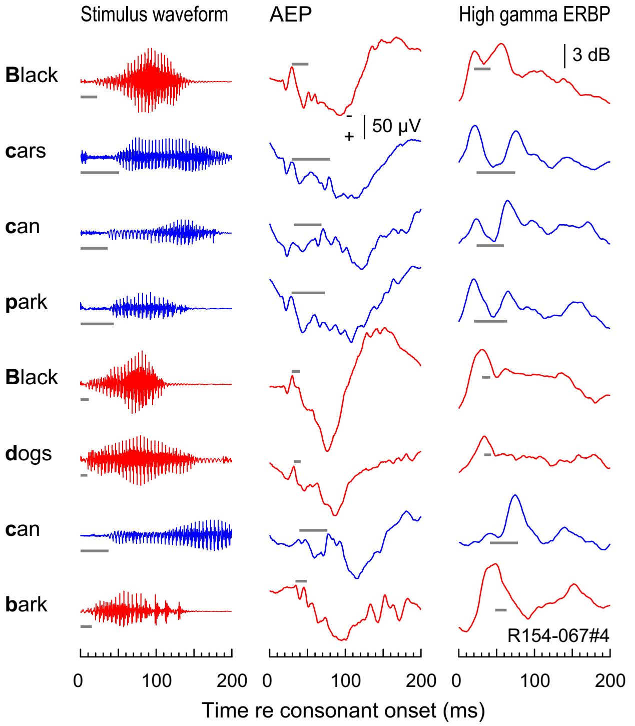

**Fig. 5.** Representation of VOT of initial stop consonants embedded in running speech in human auditory cortex. Left column: waveforms corresponding to the first 200 ms of each word (top to bottom) of the two sentences detailed in Fig. 2. Middle and right column: AEP and high gamma ERBP responses, respectively, elicited by each word. Same recording site as in Fig. 2. Red and blue plots correspond to the initial voiced and voiceless consonants, respectively. Gray horizontal lines represent the VOT aligned with the first negative peak in the AEP and the first response peak in the ERBP.

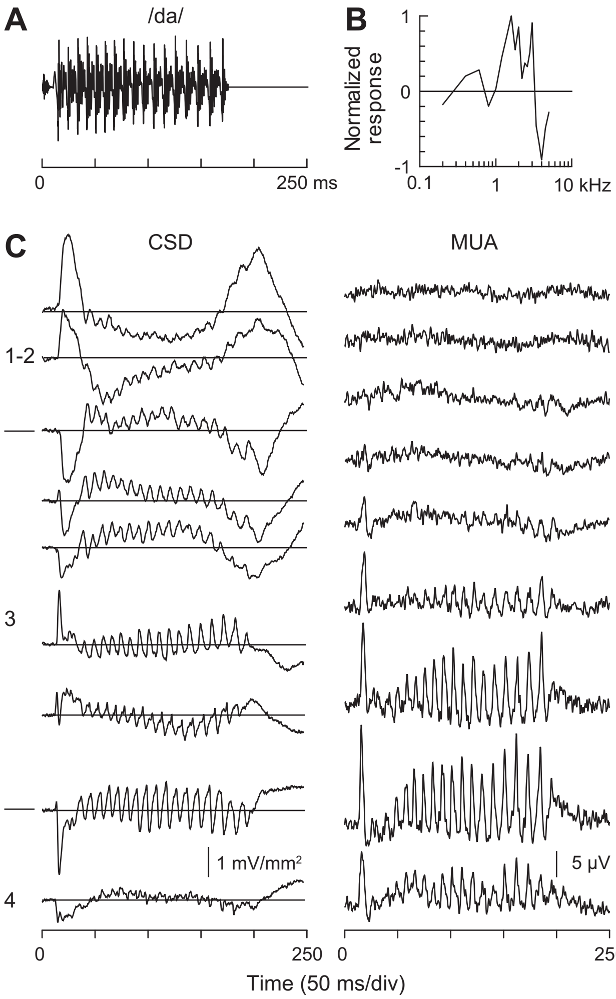

**Fig. 6.** Representation of voice $F_0$ in monkey A1. A. Stimulus waveform (syllable /da/). B. Frequency response function describing spectral selectivity of MUA recorded in lower lamina 3. Responses to pure tone stimuli (60 dB SPL) were normalized to the maximum response. C. CSD and MUA (left and right column, respectively) recorded at nine laminar depths (top to bottom plots). Cortical laminae and laminar boundaries are indicated by numbers and horizontal lines on the left, respectively. See text for details.

## 2.2. Human

### 2.2.1. Subjects

Subjects were neurosurgical patients diagnosed with medically refractory epilepsy and undergoing chronic invasive electrocorticogram (ECoG) monitoring to identify seizure foci prior to resection surgery. Newly presented data (Figs. 5, 7 and 9) were obtained from the right hemispheres of two male subjects [ages 40 (R154) and 41 (R212)]. Data illustrated in Figs. 5 and 7 were obtained from subject R154, a subject who was included in a previous study related to processing of compressed speech (Nourski et al., 2009). Data presented in Fig. 2 represents newly illustrated results from this latter manuscript. Data presented in Fig. 9 was obtained from subject R212. Placement of the electrode arrays was based on clinical considerations. Written informed consent was obtained from each subject. Research protocols were approved by The University of Iowa Institutional Review Board. All subjects underwent audiometric and neuropsychological evaluation before the study, and none were found to have hearing or cognitive deficits that might impact the findings presented in this study. All subjects were native English language speakers. Clinical analysis of intracranial recordings indicated that the auditory cortical areas on the superior temporal gyrus (STG) were not involved in the generation of epileptic activity in any of the subjects included in this study.

Each subject underwent whole-brain high-resolution T1-weighted structural magnetic resonance imaging (MRI; resolution 0.78 × 0.78 mm, slice thickness 1.0 mm, average of 2) scans before and after electrode implantation in order to determine recording contact locations relative to the pre-operative brain images. Pre-implantation MRIs and post-implantation thin-sliced volumetric computed tomography (CT) scans (in-plane resolution 0.51 × 0.51 mm, slice thickness 1.0 mm) were co-registered using a 3D linear registration algorithm (FMRIB Linear Image Registration Tool; Jenkinson et al., 2002). Coordinates for each electrode contact obtained from post-implantation CT volumes were transferred to pre-implantation MRI volumes. Results were compared to intraoperative photographs to ensure reconstruction accuracy.

Experiments were performed in a dedicated electrically shielded suite located within the Clinical Research Unit of the University of Iowa Institute for Clinical and Translational Science. The subjects were reclining awake in a hospital bed or armchair during the experiments.

### 2.2.2. Stimuli

Experimental stimuli were speech sentences “Black cars cannot park” and “Black dogs can all bark” (Ahissar et al., 2001; Nourski et al., 2009), consonant–vowel–consonant syllables /had/ (Hillenbrand et al., 1995), CV syllables /ba/, /ga/ and /da/ and 800, 1600, and 3000 Hz pure tones (Steinschneider and Fishman, 2011; Steinschneider et al., 2011). Details concerning stimulus parameters can be found in the cited papers. The experiments that used speech sentences were a part of a larger study that investigated auditory

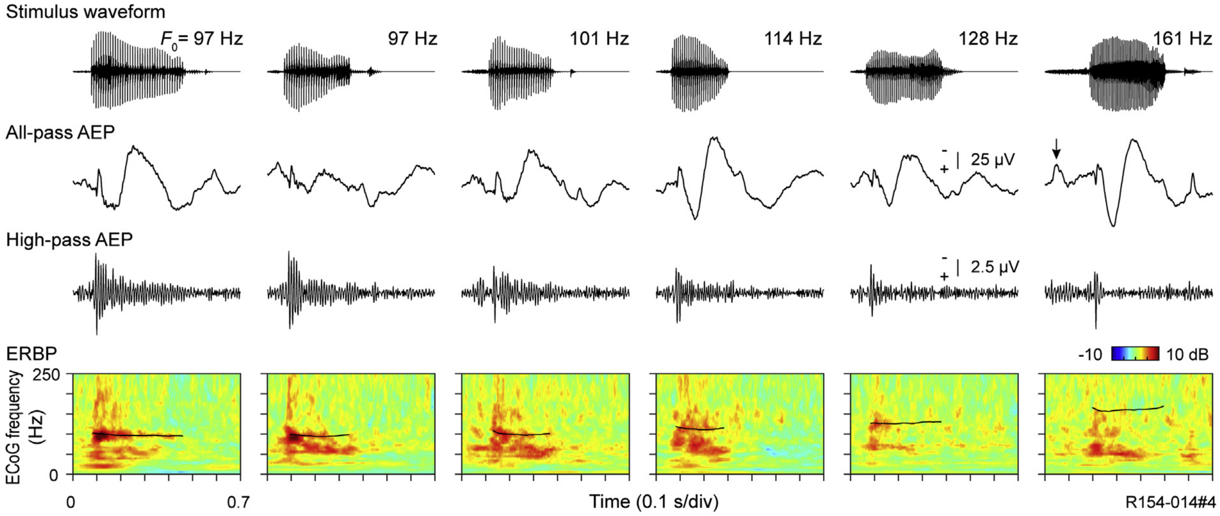

**Fig. 7.** Representation of voice $F_0$ in human auditory cortex. Same recording site as in Figs. 2 and 5. Top to bottom rows: stimulus waveforms (word "had", articulated by six male speakers, left to right); all-pass (1.6–500 Hz) AEP waveforms, high pass (>70 Hz) AEP waveforms, time–frequency ERBP plots with superimposed pitch contours (black curves).

cortical responses to time-compressed speech (Nourski et al., 2009). To that end, the sentences were time-compressed to ratios ranging from 0.75 to 0.20 of the natural speaking rate using an algorithm that preserved the spectral content of the stimuli. A stimulus set consisted of six presentations of a sentence: five time-compressed versions of the sentence "Black cars cannot park", and a sixth stimulus, "Black dogs cannot bark", which was presented with a compression ratio of 0.75 as a target in an oddball detection task to maintain an alert state in the subject. Only neural responses elicited by the sentences presented at a compression ratio of 0.75 will be discussed in this report. The subjects were instructed to press a button whenever the oddball stimulus was detected. All other sounds were presented in passive-listening paradigms without any task direction.

All stimuli were delivered to both ears via insert earphones (ER4B; *Etymotic Research*) that were integrated into custom-fit earmolds. The stimuli were presented at a comfortable level, typically around 50 dB above hearing threshold. The inter-stimulus interval was 3 s for sentences and 2 s for all other stimuli. Stimulus delivery and data acquisition were controlled by a TDT RP2.1 and RX5 or RZ2 real-time processor (*Tucker-Davis Technologies*).

### 2.2.3. Neurophysiologic recordings

Details of electrode implantation and data collection have been described previously (Nourski et al., 2009; Reddy et al., 2010). In brief, filtered (1.6–1000 Hz bandpass, 12 dB/octave rolloff) and amplified (20×) ECoG data were digitally recorded (sampling rate 12,207 Hz) from custom-designed hybrid depth electrode arrays (*AdTech*). The electrode arrays were implanted stereotactically into HG, along its anterolateral to posteromedial axis. Electrodes contained six platinum macro-contacts, spaced 10 mm apart, which were used to record clinical data. Fourteen platinum micro-contacts (diameter 40 μm, impedance 0.08–0.7 MΩ), were distributed at 2–4 mm intervals between the macro contacts and were used to record intracortical ECoG data. The reference for the micro-contacts was either a sub-galeal contact or one of the two most lateral macro-contacts near the lateral surface of the superior temporal gyrus. Reference electrodes, including those near the lateral surface of the superior temporal gyrus, were relatively inactive compared to the large amplitude activity recorded from more medial portions of HG. All recording electrodes remained in place for approximately 2–3 weeks under the direction of the patients' physicians.

### 2.2.4. General data analysis

ECoG data obtained from each recording site were analyzed as AEPs and, in the time–frequency plane, as ERBP. Data analysis was performed using custom software written in MATLAB (*MathWorks*). Pre-processing of ECoG data included downsampling to 1 kHz for computational efficiency, followed by removal of power line noise by an adaptive notch filtering procedure (Nourski et al., 2013). Additionally, single-trial (peri-stimulus) ECoG waveforms with voltage peaks or troughs greater than 2.5 standard deviations from the mean were eliminated from the data set prior to further analyses. These waveforms would include sporadic activity generated by electrical interference, epileptiform spikes, high-amplitude slow-wave activity, or movement artifacts.

Time-domain averaging of single-trial ECoG epochs yielded the AEP. Time–frequency analysis of the ECoG was performed using transforms based on complex Morlet wavelets following the approach of Oya et al. (2002) and Nourski et al. (2009). Center frequencies ranged from 10 to 250 Hz in 5 Hz increments. ERBP was calculated for each center frequency and time point on a trial-by-trial basis, log-transformed, normalized to mean baseline power, measured within a 100–200 ms window prior to stimulus onset, and averaged across trials. We focused our ERBP analysis on the high gamma frequency band ranging from 70 to 150 Hz.

Representation of the temporal stimulus envelope in the cortical activity was quantified in the time domain using cross-correlation analysis (Ahissar et al., 2001; Nourski et al., 2009). Peaks in cross-correlograms were found at lags between 0 and 150 ms. The representation of the VOT parameter in responses to speech sentences was characterized by fragmenting the sentences into their component words and re-plotting portions of the AEP and high gamma ERBP waveforms time-locked to these words. The frequency-following response to the voice fundamental was visualized by high-pass filtering the AEP waveforms with a cutoff frequency of 70 Hz using a 4th order Butterworth filter.

## 3. Results

### 3.1. Representation of temporal envelope

### 3.1.1. Monkey

Entrainment to the temporal envelope of vocalizations within auditory cortex is specific neither to humans nor to human speech. Fig. 1 demonstrates neural entrainment to the temporal envelope of

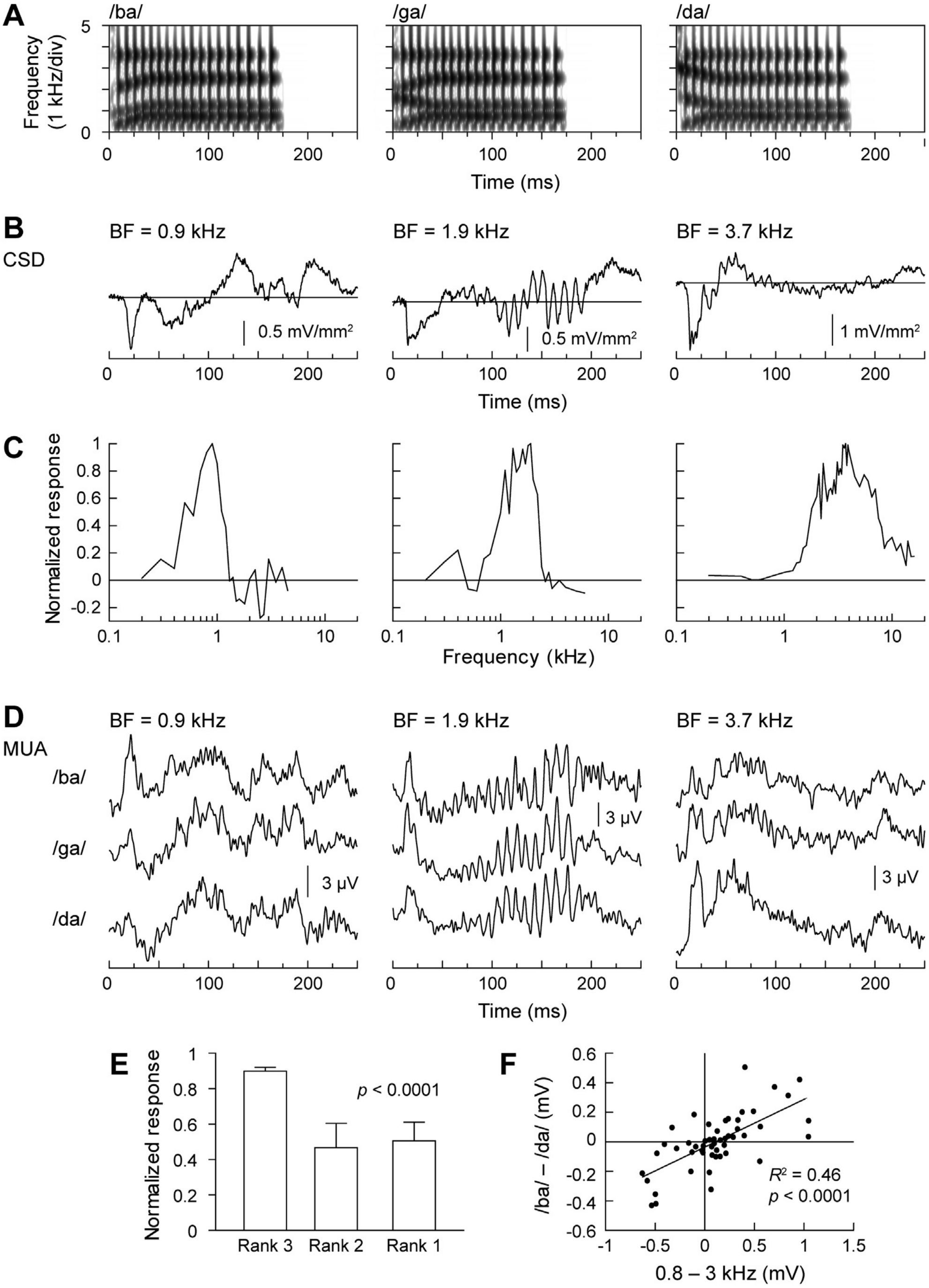

**Fig. 8.** Representation of POA in monkey A1. A. Stimulus spectrograms. B. Lower lamina 3 CSD waveforms elicited by /ba/, /ga/ and /da/, respectively. C. Frequency response functions for MUA recorded at the three sites shown in panel B. D. MUA elicited at the three sites (left to right columns) by the three syllables (top to bottom rows). E. Average (normalized) amplitude of responses to the three syllables ranked according to predictions based on responses to pure tones with frequencies corresponding to the spectral maxima of the syllables. Data set from Steinschneider and Fishman (2011). Error bars indicate standard error of the mean. F. Correlation between responses to pure-tone and speech syllable stimuli. Differences between MUA responses to 0.8 and 3.0 kHz pure tones are plotted against differences between MUA responses to syllables /ba/ and /da/. Data set from Steinschneider and Fishman (2011).

three monkey vocalizations at a low BF location within A1. The left-hand graph in Fig. 1A depicts the FRF of this site based on responses to pure tones presented at 60 dB SPL. The BF of this site is approximately 400 Hz, with a secondary peak at the 200 Hz subharmonic. FRFs based on responses to tones presented at different intensities (40–70 dB SPL) are shown in the right-hand panel of Fig. 1A. The BF at the lowest intensity presented is approximately 350 Hz. Higher intensities broaden the FRF and yield slightly higher BFs. However, in all cases, there are no significant excitatory responses above 700 Hz.

Fig. 1B depicts the stimulus waveforms (upper half) and associated spectrograms (lower half) of the monkey vocalizations. The sounds' amplitude envelopes are superimposed upon the waveforms (black line). The three sounds were randomly presented within a block that contained seven other monkey vocalizations and ten environmental sounds that included four human

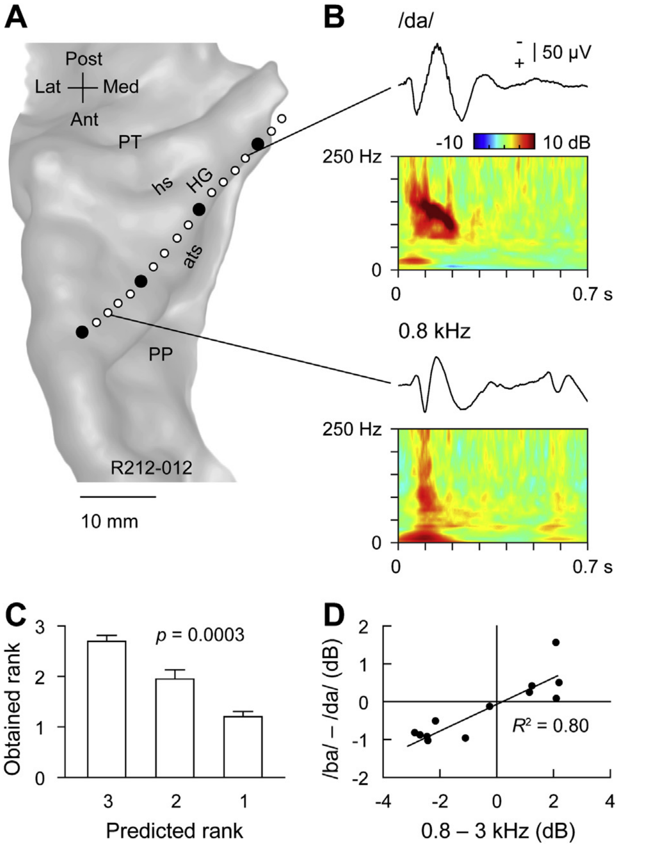

**Fig. 9.** Representation of POA in human auditory cortex. A. MRI surface rendering of the superior temporal plane showing the location of the recording array. Macro and micro contacts are represented by filled and open circles, respectively. HG, Heschl's gyrus; PP, planum polare; PT; planum temporale; ats, anterior temporal sulcus; is, intermediate sulcus; hs, Heschl's sulcus. B. Examples of pure tone- and speech sound-elicited responses recorded from Heschl's gyrus. AEP and ERBP elicited by a syllable /da/ in a medial portion of the Heschl's gyrus (top plots) and by a 0.8 kHz tone in a lateral portion of the Heschl's gyrus. C. Rank order of high gamma ERBP elicited by CV syllables compared with predicted ranks based on pure tone responses. Error bars indicate standard error of the mean. D. Correlation between response patterns elicited by pure-tone and speech syllable stimuli. Differences between high gamma responses to 0.8 and 3.0 kHz pure tones are plotted against differences between high gamma responses to syllables /ba/ and /da/. Data from 13 micro contacts are shown.

vocalizations (e.g., Fig. 4). Each monkey vocalization was characterized by a broad spectrum, with spectral centers of gravity at 1156 ± 1143 Hz, 4835 ± 1967 Hz, and 1023 ± 777 Hz, respectively. The power spectra of the temporal envelopes of the three vocalizations had modal frequencies of 5.1, 7.6, and 6.9 Hz (secondary peak at 17.2 Hz), respectively, which is comparable to the syllabic rate of intelligible human speech.

MUA elicited by the vocalizations within lower lamina 3 is shown in Fig. 1C. Despite having sound spectra whose dominant frequencies were above the BF of this recording site, each sound elicited a robust response (e.g., Wang et al., 2005). Activity at this site was phasic in nature and had an onset latency for the three vocalizations of 26, 27, and 21 ms, respectively. The vocalization depicted in the left-hand column is a 'double grunt', with the onset of each identical grunt separated by about 300 ms. The duration of each grunt is approximately 200 ms, comparable to the duration of a speech syllable. MUA elicited by each identical grunt segment is highly similar, indicating a high degree of reliability in the response pattern. MUA is clearly entrained to the vocalization's waveform envelope, as evidenced by a maximum cross-correlation coefficient (Pearson's $R = 0.73$). The three-component chirp vocalization depicted in the center column also elicits a high degree of neural entrainment to the stimulus envelope ($R = 0.68$) with the evoked activity characterized by three prominent response bursts. Even the more complex, and rapidly changing chirp sequence shown in the right-hand column elicits MUA that exhibits a weaker entrainment to the waveform envelope ($R = 0.26$).

### 3.1.2. Human

As previously reported, the envelope of running speech is represented by phase-locked activity within core auditory cortex on posteromedial HG (Nourski et al., 2009). This observation is exemplified in Fig. 2. Fig. 2A shows the location of an intracortical electrode located in HG. Both horizontal and coronal views are depicted. Fig. 2B depicts the waveforms of the two sentences presented to the subject along with associated spectrograms. Waveform envelopes are superimposed on the stimulus waveforms. In this subject, both AEPs and the high gamma ERBP calculated from the ECoG signal are phase-locked to the temporal envelope of the two sentences (Fig. 2C). Peak correlation coefficients ranged from 0.49 to 0.64.

## 3.2. Representation of VOT

Perhaps the quintessential acoustic feature of speech requiring rapid temporal analysis is the VOT of stop consonants. This feature is ubiquitous among the world's languages (Lisker and Abramson, 1964). In American English, stop consonants in syllable-initial position and with a short VOT (<20 ms) are generally perceived as voiced (i.e. /b/, /d/ or /g/), whereas those consonants with a longer VOT are generally perceived as unvoiced (i.e., /p/, /t/, and /k/). Perceptual discrimination of voiced from unvoiced stops is categorical, such that a small linear change in VOT produces a marked and non-linear change in perception (e.g., Pisoni, 1977).

### 3.2.1. Monkey

We have previously identified temporal response patterns in A1 of awake monkeys that could promote categorical perception of stop consonant VOT (Steinschneider et al., 1995b, 2003, 2005). Voiced stop CVs were found to elicit responses in A1 that contained a single "on" response time-locked to consonant release, whereas unvoiced stop CVs elicited responses that contained a "double-on" response pattern whose components were time-locked to both consonant release and voicing onset. The boundary for these dichotomous response patterns correspond to the perceptual boundary for American English and many other languages (Lisker and Abramson, 1964).

These basic findings are illustrated in Fig. 3, which depicts data obtained from an electrode penetration into a low BF region (BF = 800–1000 Hz) of monkey A1. Waveforms of the synthetically-produced syllables with VOTs ranging from 0 to 60 ms are shown above their corresponding neural responses. The topmost traces represent the rectified CSD summed across all 12 recording depths (150 µm inter-contact spacing) that spanned the laminar extent of A1 (SRCSD). SRCSD provides a global measure of across-laminar activation, and is useful when comparing temporal response patterns across stimulus conditions. The CSD waveforms recorded from lower lamina 3 are shown immediately below the SRCSD (current sinks are depicted as downward waveform deflections). Concurrently recorded MUA at the same lower lamina 3 depth is shown below the CSD waveforms.

All three response measures show categorical-like activity. MUA elicited by the syllables with 0 and 20 ms VOTs (i.e., /da/) consists of a response to stimulus onset, followed by a period of suppression ('S') below baseline levels, a steady increase in sustained activity and a subsequent response to stimulus offset. Importantly, a response elicited by voicing onset is absent in the MUA evoked by the 20 ms VOT sound. In contrast, syllables with 40 and 60 ms VOTs (i.e., /ta/) elicit an additional component time-locked to voicing onset (arrows). The concurrently recorded CSD shows similar categorical-like responses to voicing onset (arrows). The MUA suppression following the initial "on" response is paralleled by current sources in the CSD (asterisks). This combination suggests that the sources represent active regions of hyperpolarization. Finally, these categorical-like patterns are evident in the global response across cortical lamina as represented by the SRCSD.

These studies have not demonstrated, however, whether this response pattern would also be generated in more complex acoustic environments where an unvoiced stop consonant is embedded in a multisyllabic word. Fig. 4 illustrates temporal encoding of VOT for the unvoiced stop consonant /k/ in the word "welcome", where it occurs as the fourth phoneme articulated by a male speaker. Recordings are from the same low BF site illustrated in Fig. 1. The stimulus waveform and spectrogram are shown in Fig. 4A. In contrast to isolated, synthetic syllables, this naturally spoken word is highly coarticulated, and it is difficult to segment the word into a discrete sequence of phonemes. Laminar profiles of CSD and MUA are shown in the left and right columns of Fig. 4B, respectively. While the entire laminar profile was obtained from 16 recording contacts that spanned all laminae, the figure depicts data from only five recording depths beginning in lower lamina 3 and extending upward in 300 μm increments.

Despite the coarticulation of this rapidly changing speech sound, neural activity parses out the word into phonemically-relevant response components. At the lowest depicted depth, neural activity measured by CSD is highly phasic, and sinks representing net excitatory synaptic activity parse the word into discrete segments. Higher frequency oscillations phase-locked to the fundamental frequency of the vowels are superimposed on the slower waveform changes. At progressively higher depths, VOT continues to be represented by temporal features of the CSD waveforms (dotted line) and the response simplifies to one where prominent deflections are elicited by the onsets of the two syllables.

The MUA exhibits similar response patterns. In the thalamorecipient zone (lower 3–4), MUA is characterized by phasic response bursts (after an initial "on" response) that temporally coincide with each of the phonemes. A more sustained component is time-locked to the increase in stimulus power associated with the first vowel. The response burst associated with the onset of the /k/ is indicated by the dotted line. This burst is followed by another response component time-locked to voicing onset of the following vowel. Thus, the unvoiced CV sequence is characterized by a "double-on" response analogous to the response to the isolated CV syllable /ta/ shown in Fig. 3. At progressively more superficial laminar depths, the MUA becomes less phasic, and in upper lamina 3, exhibits two bursts time-locked to the speech envelope. This is reflected by the high maximum cross correlation ($R = 0.82$) between the MUA recorded in upper lamina 3 and the speech envelope. Despite the simplification of the response within more superficial laminae, MUA phase-locked to the VOT still persists. Thus, at this low BF frequency site, activity elicited by rapidly changing acoustic features in this two-syllable word is nested within responses elicited by the more slowly changing stimulus envelope.

### *3.2.2. Human*

Similar to responses in monkey A1, VOT of stop CV syllables is represented by categorical-like responses in posterior-medial HG (Steinschneider et al., 1999, 2005). Syllables perceived by the subjects as /da/ (VOT = 0 and 20 ms) generate a "single-on" response time-locked to consonant release in the intracranially recorded AEP, whereas syllables perceived as /ta/ (VOT = 40 and 60 ms) generate a "double-on" response with the second component time-locked to voicing onset. Further, by raising or lowering the first formant frequency, it is possible to manipulate the perception of voiced from unvoiced stop consonants. This manipulation is a form of phonemic trading relations, wherein changes in one acoustic parameter alter the perception of another parameter. In this case, lowering the first formant shifts the perceptual boundary such that syllables with a longer VOT are still identified as a voiced stop (Lisker, 1975; Summerfield and Haggard, 1977). We found that this type of manipulation would shift physiological response patterns in parallel with perceptual changes (Steinschneider et al., 2005). When the first formant was 600 or 848 Hz, the perceptual boundary between /d/ and /t/ was between 20 and 25 ms, whereas when the first formant was 424 Hz, the perceptual boundary shifted to between 40 and 60 ms. Correspondingly, the physiological boundary between a "single-on" and "double-on" response shifted from smaller VOT values to a VOT of 60 ms when the first formant was shifted from 600 Hz to 424 Hz.

It is unclear from these studies whether this pattern would also be observed in more naturally occurring running speech. To address this question, we examined the detailed temporal patterns embedded within responses entrained to the speech envelope shown in Fig. 2. The sentences were segmented into their component words, and both AEPs and the envelope of high gamma activity in the ECoG time-locked to these words were examined (Fig. 5). Gray bars in Fig. 5 indicate the duration of VOT for each word, and are aligned with the first negative peak in the AEP and with the first peak in the ERBP. While responses are not as distinct as when isolated CV syllables are tested, there is a clear tendency for words beginning with a voiced stop followed by a vowel (i.e., the words "dogs" and "bark") to elicit responses with a single peak, whereas words beginning with an unvoiced stop (i.e., the words "cars", "can", and "park") to elicit double-peaked responses. These patterns are more prominent in the high gamma activity than in the AEP. Thus, responses correlating with the perception of VOT for isolated syllables are also evident in responses to syllables embedded in running speech.

## *3.3. Representation of voice fundamental frequency ($F_0$)*

### *3.3.1. Monkey*

Neural population responses in monkey A1 have the capacity to phase-lock to components of speech faster than those required for human phonemic identification. Most notably, A1 responses are able to phase-lock to the finer temporal structure of speech sounds associated with the glottal pulsation rate in many male speakers (Hillenbrand et al., 1995). This capacity, previously reported in Steinschneider et al. (2003), is illustrated in Fig. 6, which depicts the laminar profile of CSD and MUA elicited by the syllable /da/. The syllable was presented at 60 dB SPL. Its fundamental frequency ($F_0$) began at 100 Hz after a 5 ms period of frication, rose to 120 Hz over the subsequent 70 ms, and slowly fell to 80 Hz by the end of the stimulus (Fig. 6A). The FRF is shown in Fig. 6B. The largest MUA was evoked by 1600 Hz tones, with a secondary peak in excitation at 3000 Hz. MUA was suppressed by tone frequencies of 3400 Hz and higher. Second and third formant frequencies of the syllable overlap the excitatory range of the site.

Neural responses are phase-locked to the glottal pulsation rate of the syllable and track the changing $F_0$ contour present in the stimulus (Fig. 6C). Phase-locking in the MUA is restricted to lamina 4 and lower regions of lamina 3. Phase-locking in the CSD, however, extends upward into more superficial laminae. The phase-locking of CSD responses recorded from more superficial laminae is delayed relative to the latency of lower lamina 3 responses, and several phase-reversals are evident in the sink/source profiles. This pattern is consistent with superficial phase-locked activity representing di- or polysynaptic transmission of the high-frequency responses from deeper thalamorecipient laminae.

### 3.3.2. Human

Phase-locking to the ∼100 Hz $F_0$ of synthetic CV syllables has been reported for responses in posteromedial HG (Steinschneider et al., 1999; Nourski and Brugge, 2011). Here we demonstrate the upper limits of this pattern in non-synthetic speech by illustrating the responses elicited by six male speakers articulating the word "had" (Hillenbrand et al., 1995) (Fig. 7). Waveforms of the words are arranged left to right from lowest to highest $F_0$ in the top panel. AEPs recorded from the same electrode site shown in Fig. 2 are shown immediately beneath the syllables. High-pass filtered versions of the full-pass AEPs display phase-locked activity to the glottal pulsation rate up to about 114 Hz. Phase-locking dissipates in the responses to the two syllables with higher $F_0$s. The lowest panels depict the ERBP with the $F_0$ contours superimposed on the responses (black lines). Once again, phase-locking to the $F_0$ occurs in responses to the syllables with the lower $F_0$ values, and dissipates at the higher $F_0$s.

## 3.4. Representation of place of articulation (POA)

### 3.4.1. Monkey

Responses reflecting stop consonant POA in A1 can be predicted by the spectral tuning of pure tone responses (Steinschneider et al., 1995a; Steinschneider and Fishman, 2011). This can be demonstrated by ranking responses (based on their amplitude) to pure tones with frequencies corresponding to the center frequencies of the frication bursts at the onset of speech syllables and comparing these ranks with ranked amplitudes of responses to the syllables (Fig. 8). Fig. 8A depicts the waveforms of the synthetic syllables/ba/, /ga/, and /da/. Duration of the first formant transition was 30 ms for all syllables, and 40 ms for the second and third formant transitions. Syllables were constructed such that /ba/ and /ga/ shared the same third formant transition, whereas /ga/ and /da/ shared the same second formant transition. Thus, tracking of formant transitions would not allow syllable identification. Further, all syllables shared the same second and third formant steady-state frequencies, had identical first and fourth formants, and were initiated by 5 ms of frication at sound onset. Differences among the syllables were primarily generated by increasing the amplitude of frication at 800, 1600, and 3000 Hz for /ba/, /ga/, and /da/, respectively. These manipulations produce easily discriminable stop consonants and are based on the hypothesis that lower-, medium-, and higher-frequency onset spectra determine the differential perception of stop consonants (Blumstein and Stevens, 1979, 1980).

Fig. 8B illustrates the CSD waveforms elicited by the speech sounds and used to identify lower lamina 3 at three recording sites whose BFs were 900, 1900, and 3700 Hz, respectively. Corresponding FRFs for MUA recorded at the three sites are shown in Fig. 8C. Fig. 8D depicts lamina 3 MUA elicited by /ba, /ga/, and /da/. At the site with a 900 Hz BF (left-hand column), /ba/ elicited the largest onset response. This is consistent with the spectrum of /ba/ having maximal energy near 900 Hz during frication. As predicted, /ga/ elicits the largest onset response at the site with a BF of 1900 Hz (center column), as this BF was near the spectral maximum of this syllable during frication. Finally, /da/ elicits the largest response at the site with the 3700 Hz BF (right-hand column).

Average ranks of responses from the 47 electrode penetrations in the study of Steinschneider and Fishman (2011) are shown in Fig. 8E. We first ranked the amplitude of MUA elicited by 800, 1600, and 3000 Hz tones within the 10–30 m time frame at each recording site. This period includes the onset responses elicited by both tones and syllables. Then, ranks of tone responses were used to predict the rank order of responses elicited by the syllables in the same time frame. Responses elicited by the syllables were normalized to the largest response at each recording site. Without regard to any other features of spectral tuning at each site, and using the simple metric of comparing responses evoked by syllables based on the ranking of three tone responses, one is able to reliably predict the relative amplitude of stop consonants (Friedman statistic = 19.96, $p < 0.0001$). Post hoc tests show that the response elicited by the largest predicted syllable (Rank 3) is greater than either other two ranked responses ($p < 0.001$). Thus, despite the pronounced overlap of formant transitions across syllables, neural activity is capable of discriminating the speech sounds based on the relationship between their spectral maxima and the frequency tuning of neural population responses in A1.

To further demonstrate this relationship between responses evoked by stop consonants varying in their POA and the frequency selectivity of neurons in A1, we computed the differences between MUA elicited by 800 Hz and 3000 Hz tones at each site and compared those with differences between MUA elicited by /ba/ and /da/ in the same 20 ms time interval. Results are shown in Fig. 8F. There is a significant positive correlation between differences in tone responses and differences in responses to /ba/ and /da/ ($R^2 = 0.46$, $p < 0.0001$). A similar correlation was observed for differences between responses to 800 and 1600 Hz tones, and differences between responses to /ba/ and /ga/ ($R^2 = 0.35$, $p < 0.0001$, data not shown).

### 3.4.2. Human

A similar relationship between tone-evoked and syllable-evoked responses can be demonstrated in HG (Fig. 9). Fig. 9A illustrates the location of a multi-contact intracortical electrode positioned within more anterior portions of HG. Interspersed between four low impedance ECoG contacts are fourteen electrode contacts of higher impedance. High gamma activity (70–150 Hz) elicited by randomly presented 800, 1600, and 3000 Hz tones intermixed with the three syllables /ba/, /ga/, and /da/ was examined. Amplitude of high gamma activity from three 50 ms overlapping time windows beginning at 25 ms and ending at 125 ms was computed. The rationale for choosing these time windows stems from previous findings that in non-primary auditory cortex located on the lateral surface of posterior–lateral STG, responses reflecting consonant POA of CV syllables were maximal between 100 and 150 ms (Steinschneider et al., 2011). It was therefore reasoned that responses reflecting POA in core auditory cortex should begin earlier than, and end prior to, activity on the lateral surface.

Representative AEPs and ERBP elicited by speech sounds (response to /da/ recorded from a medially located high impedance contact 3) and tones (response to 800 Hz recorded from a more laterally located high impedance contact 13) are shown in Fig. 9B. Prominent high gamma activity is elicited by both sounds. Phase-locking to the $F_0$ of /da/ is evident in the ERBP, and both 'on' and 'off' responses can be seen in the response to the tone (duration = 500 ms). Ranks of tone response amplitudes were used to predict the rank order of response amplitudes to the syllables. Results are shown in Fig. 9C. Differences in response ranks are significant (Friedman statistic = 16.00, $p = 0.0003$). Post hoc analysis shows that the largest predicted rank is larger than smallest rank ($p < 0.001$). Similarly, the correlation between differences in high gamma activity elicited by 800 Hz and 3000 Hz tones and differences in high gamma activity elicited by /ba/ and /da/ is significant (Fig. 9D, $R^2 = 0.80$, $p < 0.0001$). Significant correlations are also observed (data not shown) between (1) differences in high gamma activity elicited by 800 and 1600 Hz tones and differences in activity elicited by /ba/ and /ga/ ($R^2 = 0.64$, $p = 0.001$) and (2) differences in high gamma activity elicited by 1600 and 3000 Hz tones and differences in activity elicited by /ga/ and /da/ ($R^2 = 0.74$, $p = 0.0002$).

## 4. Discussion

### 4.1. Summary and general conclusions

The key finding of this study is that neural representations of fundamental temporal and spectral features of speech by population responses in primary auditory cortex are remarkably similar in monkeys and humans, despite their vastly different experience with human language. Thus, it appears that plasticity-induced language learning does not significantly alter response patterns elicited by the acoustical properties of speech sounds in primary auditory cortex of humans as compared with response patterns in macaques. It is still possible, however, that more subtle differences might be found when examining the activity of single neurons – though it is reasonable to expect that any dissimilarity observed would reflect quantitative rather than qualitative differences in neural processing (e.g., Bitterman et al., 2008). It is important to emphasize that the present findings only apply to responses in primary auditory cortex, and do not imply that later cortical stages associated with more complex phonological, lexical and semantic processing are not unique to humans and therefore “special”.

Current findings support the view that temporal and spectral acoustic features important for speech perception are represented by basic auditory processing mechanisms common to multiple species. For instance, young infants are ‘language universalists’, capable of distinguishing all phonemes despite their absence from the child’s native language (Kuhl, 2010). This absence suggests that language-specific mechanisms are not responsible for infants’ capacity to discriminate phonemic contrasts that are both within and outside their language environment. The ability of animals to perceive many phonemic contrasts in a manner similar to humans provides additional support for this view (e.g., Kuhl and Miller, 1978; Sinnott and Adams, 1987; Lotto et al., 1997; Sinnott et al., 2006). The similarity in response patterns evoked by speech sounds in HG of adult humans and in monkey A1 suggests that HG is engaged in general auditory, and not language-specific, processing.

Nonetheless, this similarity in response patterns should not be interpreted to mean that primary auditory cortex does not undergo changes associated with learning and plasticity reflecting the distinct acoustic experiences of these two primate species. Neuronal plasticity in A1 of non-human animals is well-documented (e.g., de Villers-Sidani and Merzenich, 2011; Fritz et al., 2013). Likewise, studies have demonstrated plasticity in HG (e.g., Schneider et al., 2002; Gaser and Schlaug, 2003). Therefore, it can reasonably be asked why differences associated with these modifications were not readily apparent in the present study. One possibility is that experience-related changes might only occur under conditions requiring performance of complex sound processing tasks. Here, both monkey and human subjects either listened passively to the sounds or were engaged in relatively modest sound-detection tasks that were meant simply to maintain a state of alertness and to promote active listening to the stimuli. More pronounced differences reflecting language experience might have been seen if subjects were engaged in speech-related tasks or in more challenging tasks that strain processing capabilities (e.g., related to the perception of highly compressed speech). As stated above, another possibility is that the features of speech examined here tap into fundamental temporal and spectral aspects of sound processing that are shared across multiple mammalian species and which are utilized for speech sound decoding. In the following sections, the representation of the speech envelope, VOT, $F_0$, and POA will be shown to engage basic auditory cortical mechanisms that may contribute to shaping phonemic perception.

### 4.2. Representation of temporal envelope

Entrainment of neural responses in A1 to the temporal envelope of animal vocalizations has been observed in many species (e.g., Wang et al., 1995; Szymanski et al., 2011; Grimsley et al., 2012). For instance, neurons in A1 of both marmosets and ferrets respond to marmoset twitter calls with bursts time-locked to the syllable-like components of each vocalization (Wang et al., 1995; Schnupp et al., 2006). Comparable time-locked neuronal activity is observed in guinea pig A1 in response to species-specific vocalizations (Grimsley et al., 2012). Importantly, the repetition rate of call components in these vocalizations is similar to the syllabic-rate in human speech.

It is not fortuitous that the speech envelope is so well tracked by neuronal activity in primary auditory cortex. Speech syllables recur at a peak rate of 3–4 Hz with a 6 dB down point at 15–20 Hz (Drullman et al., 1994). Best temporal modulation frequencies (tBMFs) of neurons in primary auditory cortex of multiple species, including humans, are comparable to envelope modulation rates of running speech (Schreiner and Urbas, 1988; Eggermont, 1998, 2002; Oshurkova et al., 2008). For instance, repetitive frequency-modulated sweeps generally fail to elicit time-locked responses at repetition rates greater than 16 Hz in A1 of awake squirrel monkeys (Bieser, 1998). Similarly, tBMFs calculated from AEPs obtained from intracranial recordings in human auditory cortex typically range from 4 to 8 Hz (Liégeois-Chauvel et al., 2004; see also Nourski and Brugge, 2011), similar to the 8 Hz tBMF found for HG using fMRI (Giraud et al., 2000). Thus, there appears to be an optimal match between the articulatory rate of speech and the ability of primary auditory cortex to detect major modulations in the speech envelope.

Detection of modulations in the speech envelope by auditory cortex is crucial for speech perception. Typically, in running speech, each syllable embedded in larger phrases generates a discrete response (e.g., Ding and Simon, 2012). Using compressed speech, Ahissar et al. (2001) demonstrated that time-locked MEG responses to each syllable were degraded at ratios of compression where accurate perception deteriorated. This relationship between speech perception and primary auditory cortical physiology was confirmed in a study examining AEPs recorded directly from HG (Nourski et al., 2009). In contrast, high gamma activity continued to entrain to the highly compressed and incomprehensible speech sentences. This dissociation between the two response types suggests that each physiologic measure represents speech in a distinct manner. Neural processes that contribute to the AEP mark the occurrence of each new syllabic event and must be present in order to parse sentences into their constituent parts. High gamma activity, on the other hand, appears to serve as a more general marker of neuronal activation involved in the processing of the heard, but (nonetheless) incomprehensible, sound sequences.

### 4.3. Representation of VOT

Onset responses in primary auditory cortex are important not only for parsing out syllables in running speech but also for the representation of VOT in individual syllables. The pattern of “single-on” and “double-on” responses paralleling VOT has been observed in A1 of diverse animal models (e.g., Steinschneider et al., 1994, 2003; Eggermont, 1995; McGee et al., 1996; Schreiner, 1998). The relevance of these temporal response patterns for describing

VOT processing in humans is supported by similar patterns occurring in HG (Liégeois-Chauvel et al., 1999; Steinschneider et al., 1999, 2005; Trébuchon-Da Fonseca et al., 2005), and along the posterior–lateral STG (Steinschneider et al., 2011). Relevance is further supported by the ability of animals to discriminate stop CV syllables varying in their VOT in a manner that parallels human categorical perception (Kuhl and Miller, 1978; Sinnott and Adams, 1987; Sinnott et al., 2006). Further, severe degradation of spectral cues does not significantly impair discrimination of voiced from unvoiced consonants, indicating that temporal cues are sufficient (Shannon et al., 1995). Taken together, these findings suggest that the neural representation of VOT, an acoustic feature critical for speech perception, may rely on basic temporal processing mechanisms in auditory cortex that are not speech-specific.

Temporal representation of VOT could facilitate rapid and efficient differentiation of voiced from unvoiced stop consonants. Discrimination would require only that the brain distinguish between “single on” and “double on” response patterns, whereas more subtle computations would be necessary for differentiating stop consonants whose VOTs are located on the same side of a perceptual boundary (Carney et al., 1977; Pisoni et al., 1982; Kewley-Port et al., 1988). This temporal processing scheme is fully compatible with higher order lexical or semantic processing mechanisms. For instance, VOT perception was examined using synthetically produced word/nonword pairs of dash/tash and dask/task with systematic changes in the VOT of /d/ and /ta/. Only when the VOT was in an ambiguous zone was there a significant shift in the boundary toward the real word (Ganong, 1980). Otherwise, subjects heard the lexically incorrect word. Similarly, only when the VOT was in an ambiguous zone was there a significant shift in the boundary toward the word that produced a semantically correct sentence (e.g., “The dairyman hurried to milk the goat/coat”) (Borsky et al., 1998). In the preceding example, subjects heard a semantically incorrect sentence if the VOT was prolonged (i.e., /k/). These latter studies demonstrate that unambiguous VOTs, presumably represented by low-level temporal processing mechanisms, take precedence over higher-order lexical and semantic factors in the categorical perception of speech.

Representation of VOT epitomizes how basic auditory processing mechanisms (a domain-general skill) can enhance acoustic discontinuities that promote phonemic discrimination (a domain-specific skill) (Kuhl, 2004). Indeed, the VOT boundary in American English appears to correspond to a natural psychoacoustic boundary in mammalian hearing. For instance, infants with little language exposure can discriminate syllables with VOTs of +20 with +40 ms, even when this contrast is phonetically not relevant in the child’s native language (Eimas et al., 1971; Lasky et al., 1975; Eilers et al., 1979; Jusczyk et al., 1989). This natural psychoacoustic boundary for VOT appears to reflect a specific instance of the more general perceptual boundary for determining whether two sounds are perceived as occurring synchronously or separately in time (Hirsh, 1959; Pisoni, 1977). A “single-on” response would therefore suggest that the onsets of consonant release and voicing are simultaneous, whereas a “double-on” neural response would indicate the sequential occurrence of these two articulatory events.

The categorical-like responses reflecting VOT are based on three fundamental mechanisms of auditory cortical physiology (see Steinschneider et al., 2003, 2005). The first is that transient responses elicited by sound onsets occur synchronously over a wide expanse of tonotopically-organized A1 (Wang, 2007). This mechanism would explain the “on” response to consonant release, which has a predominantly high-frequency spectrum, in lower BF regions of A1. The second is that transient elements embedded in complex sounds will elicit time-locked and synchronized responses in neurons whose frequency selectivity matches the dominant components of the sound’s spectrum (Creutzfeldt et al., 1980; Wang et al., 1995). This would explain the second “time-locked” response to voicing onset in low BF regions of A1. However, this event by itself does not explain categorical-like neuronal activity, as responses evoked by voicing onset might also occur at intermediate values of VOTs within the ambiguous zone for categorical perception. Thus, the third mechanism is forward suppression, which is triggered by consonant release and which lasts sufficiently long to prevent short-latency “on” responses evoked by voicing onset. This suppression is likely due to $GABA_A$ receptor-mediated IPSPs from inhibitory interneurons (Metherate and Ashe, 1994; Metherate and Cruikshank, 1999; Cruikshank et al., 2002), and to $Ca^{2+}$- gated $K^+$ channel-mediated after hyperpolarization (Eggermont, 2000). A second response to voicing onset can only be elicited when there is sufficient decay in suppression.

### 4.4. Representation of voice fundamental frequency ($F_0$)

Voice $F_0$ typical of many male speakers is represented in A1 and posteromedial HG by responses phase-locked to the glottal pulsation rate. In the current study, phase-locked responses in HG were observed for $F_0$s less than about 120 Hz (Fig. 7). Even higher rates of phase-locked activity have been observed in population responses elicited by click trains in primary auditory cortex of monkeys (Steinschneider et al., 1998) and humans (Brugge et al., 2009; Nourski and Brugge, 2011). Generally, however, the upper limit for phase-locking has been reported to occur at much lower rates (e.g., Lu et al., 2001; Liang et al., 2002; Bendor and Wang, 2007). There are at least two possible reasons for this disparity. The first is based on differences across studies with regard to the laminae from which cells were recorded. Wang and colleagues mainly recorded from cells located in lamina 2 and upper lamina 3 in marmoset A1 (Lu et al., 2001; Wang et al., 2008). Multiunit activity recorded from the thalamorecipient zone in A1 of awake rats can phase-lock to click trains at rates up to 166 Hz (Anderson et al., 2006). A direct comparison of phase-locking across laminae has shown enhanced phase-locking (≥64 Hz, the highest rate examined) in the thalamorecipient zone relative to more superficial layers of A1 (Middlebrooks, 2008). The second reason for the disparity in the upper limit of phase-locking is the finding that neural populations phase-lock at higher rates than single cells. For instance, in the awake macaque, pooling of single cell responses in A1 of awake macaque revealed neuronal phase-locking at 120 Hz that was not otherwise observed in single cell activity (Yin et al., 2011).

Phase-locked activity at higher rates may play an important role in representing the perceived pitch of male speakers. Pitch below about 100 Hz is based on temporal mechanisms determined by the periodicity of the sound waveform, whereas pitch above 200 Hz is based on the $F_0$ (Flanagan and Guttman, 1960a, 1960b; Carlyon and Shackleton, 1994; Shackleton and Carlyon, 1994; Plack and Carlyon, 1995). Stimulus periodicities between 100 and 200 Hz lead to more ambiguous pitch determinations (see also Bendor et al., 2012). Adult male speakers typically have $F_0$s in the range of 75–175 Hz whereas adult female speakers typically have $F_0$s in the range of 175–300 Hz (Hillenbrand et al., 1995; Greenberg and Ainsworth, 2004). A more recent study examining 12 male and 12 female speakers calculated their average $F_0$s to be 112 ± 8 Hz and 205 ± 19 Hz, respectively (Sokhi et al., 2005). Gender-ambiguous $F_0$s range from 135 to 181 Hz (mean 156–160 Hz) (Sokhi et al., 2005). Thus, most non-ambiguous male speakers would have a glottal pulsation rate that would be expected to produce phase-locking in HG, from which pitch may theoretically be extracted by neuronal ‘periodicity detectors’.

### 4.5. Representation of place of articulation (POA)

Historically, the question of how stop consonant POA is represented in the brain has been a prime focus in the debate between proponents of competing hypotheses concerning the mechanisms of phonemic perception (see Diehl et al., 2004; Samuel, 2011 for reviews). Theories are wide ranging, and include at one extreme the motor theory, which posits that the objects of phonemic perception are articulatory in nature (e.g., Liberman and Mattingly, 1985; Liberman and Whalen, 2000). At the other extreme is the general approach, which posits that basic principles of auditory processing modulated by perceptual learning underlie phonemic perception (e.g., Holt and Lotto, 2010).

The general approach has been strongly supported by data indicating that stop consonant POA is partly determined by the short-term sound spectrum occurring within 20 ms of consonant release (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980; Chang and Blumstein, 1981). The perception of /d/ is promoted when the onset spectrum has a maximum at higher frequencies, whereas the perception of /b/ is promoted when the maximum is at lower frequencies. A compact spectrum maximal at mid-frequencies promotes the perception of /g/. While modifications to this scheme have been developed that incorporate contextual effects between the onset spectrum and the spectrum of the following vowel (Lahiri et al., 1984; Alexander and Kluender, 2008), and while formant transitions are also important for consonant discrimination (e.g., Walley and Carrell, 1983), mechanisms based on onset spectra are still important for the perception of stop consonants. For instance, onset spectra are sufficient for accurate perception when syllables are very short and contain minimal vowel-related information (Bertoncini et al., 1987), and are especially relevant for stop consonant discrimination in subjects with sensory neural hearing loss (Hedrick and Younger, 2007; Alexander and Kluender, 2009).

These psychoacoustic findings suggest a physiologically plausible mechanism by which stop consonants varying in their POA are represented in primary auditory cortex. In this paper we demonstrate that the differential representation of stop consonants can be predicted based on the relationship between the frequency selectivity of neurons in A1 (as defined by responses to pure tones) and the maximum of the onset spectra of speech sounds (Figs. 8 and 9; Steinschneider et al., 1995a, Steinschneider and Fishman, 2011). Numerous studies have shown that the representation of vocalizations in A1 is determined by the tonotopic organization of A1 and the spectral content of the complex sounds (Creutzfeldt et al., 1980; Wang et al., 1995; Engineer et al., 2008; Bitterman et al., 2008; Mesgarani et al., 2008). Due to the tonotopic organization of sound frequency in A1, phonemes with distinct spectral characteristics will be represented by unique spatially distributed patterns of neural activity across the tonotopic map of A1 (e.g., Schreiner, 1998). Transmission of this patterned distribution of activity in A1 to non-primary auditory cortical areas may be one factor in promoting the categorical-like distribution of responses reflecting stop consonant POA observed in human posterior–lateral STG (Chang et al., 2010). These categorical responses are based, in part, on encoding the spectral composition of the speech sounds at consonant onset.

### 4.6. Concluding remarks

Results demonstrate that basic neurophysiological mechanisms govern responses to speech in primary auditory cortex of both humans and monkeys. It appears that responses to speech sounds in primary auditory cortex are neither species- nor speech-specific. It remains an open question whether neural mechanisms in primary auditory cortex have evolved to track information-bearing features common to both human speech and animal vocalizations, or (conversely) whether the latter have adapted to exploit these basic auditory cortical processing mechanisms.

In the past, it was felt that studying speech processing in experimental animals was not an appropriate line of research. In fact, studies of speech encoding in humans and animal models are synergistic: the human work brings relevance to the animal studies while the animal studies provide detailed information not typically obtainable in humans. For instance, single neuron recordings in A1 have revealed additional organizational schemes based on sensitivity to rate of spectral change in phonemes and on preference for speech sounds with broad or narrow spectral bandwidths that enhance our understanding of how the cortex categorizes phonemic elements (Mesgarani et al., 2008). The roles of learning and plasticity in shaping phonemic representation can be examined at the single cell level (e.g., Schnupp et al., 2006; Liu and Schreiner, 2007; David et al., 2012). Clarifying how representations of speech sounds are transformed across cortical regions becomes a more tractable endeavor when studies in animals and humans are conducted in parallel (e.g., Chang et al., 2010; Tsunada et al., 2011). Studies examining modulation of auditory cortical activity by attention and top-down processing in general provide further connections between auditory cortical neurophysiology in experimental animals and humans (e.g., Fritz et al., 2010; Mesgarani and Chang, 2012). Finally, theories of developmental language disorders posit abnormalities that include deficits in the physiologic representation of the amplitude envelopes of speech (e.g., Goswami, 2011; Goswami et al., 2011) and in the representations of more rapidly occurring phonetic features (e.g., Tallal, 2004; Vandermosten et al., 2010). As we have shown here, these representations are not unique to humans. Thus, physiologic studies directed at defining aberrant processing of speech in appropriate genetically engineered animal models (e.g., Newbury and Monaco, 2010) offer the potential for unraveling the neural bases of developmental language disorders.

## Contributors

All authors participated in the writing of this manuscript. Drs. Fishman and Steinschneider collected and analyzed data obtained from monkeys. Drs. Nourski and Steinschneider collected and analyzed data obtained from humans. All authors have approved the final article.

## Conflict of interest

All authors state that they have no actual or potential conflict of interest.

## Acknowledgments

The authors thank Drs. Joseph C. Arezzo, Charles E. Schroeder and David H. Reser, and Ms. Jeannie Hutagalung for their assistance in the monkey studies, and Drs. Hiroyuki Oya, Hiroto Kawasaki, Ariane Rhone, Christopher Kovach, John F. Brugge, and Matthew A. Howard for their assistance in the human studies. Primate studies supported by the NIH (DC-00657), and human studies supported by NIH (DC04290, UL1RR024979), Hearing Health Foundation, and the Hoover Fund.

## References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., Merzenich, M.M., 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc. Natl. Acad. Sci. U. S. A. 98, 13367–13372.

Alexander, J.M., Kluender, K.R., 2008. Spectral tilt change in stop consonant perception. J. Acoust. Soc. Am. 123, 386–396.

Alexander, J.M., Kluender, K.R., 2009. Spectral tilt change in stop consonant perception by listeners with hearing impairment. J. Speech Lang. Hear. Res. 52, 653–670.

Anderson, S.E., Kilgard, M.P., Sloan, A.M., Rennaker, R.L., 2006. Response to broadband repetitive stimuli in auditory cortex of the unanesthetized rat. Hear. Res. 213, 107–117.

Anderson, S., Kraus, N., 2010. Objective neural indices of speech-in-noise perception. Trends Amplif. 14, 73–83.

Bendor, D.A., Osmanski, M.S., Wang, X., 2012. Dual-pitch processing mechanisms in primate auditory cortex. J. Neurosci. 32, 16149–16161.

Bendor, D.A., Wang, X., 2007. Differential neural coding of acoustic flutter within primate auditory cortex. Nat. Neurosci. 10, 763–771.

Bertoncini, J., Bijeljac-Babic, R., Blumstein, S.E., Mehler, J., 1987. Discrimination in neonates of very short CVs. J. Acoust. Soc. Am. 82, 31–37.

Bieser, A., 1998. Processing of twitter-call fundamental frequencies in insula and auditory cortex of squirrel monkeys. Exp. Brain Res. 122, 139–148.

Bitterman, Y., Mukamel, R., Malach, R., Fried, I., Nelken, I., 2008. Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. Nature 451, 197–201.

Blumstein, S.E., Stevens, K.N., 1979. Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. J. Acoust. Soc. Am. 66, 1001–1017.

Blumstein, S.E., Stevens, K.N., 1980. Perceptual invariance and onset spectra for stop consonant vowel environments. J. Acoust. Soc. Am. 67, 648–662.

Borsky, S., Tuller, B., Shapiro, L.P., 1998. "How to milk a coat": the effects of semantic and acoustic information on phoneme categorization. J. Acoust. Soc. Am. 103, 2670–2676.

Brosch, M., Bauer, R., Eckhorn, R., 1997. Stimulus-dependent modulations of correlated high-frequency oscillations in cat visual cortex. Cereb. Cortex 7, 70–76.

Brugge, J.F., Nourski, K.V., Oya, H., Reale, R.A., Kawasaki, H., Steinschneider, M., Howard III, M.A., 2009. Coding of repetitive transients by auditory cortex on Heschl's gyrus. J. Neurophysiol. 102, 2358–2374.

Carlyon, R.P., Shackleton, T.M., 1994. Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms. J. Acoust. Soc. Am. 95, 3541–3554.

Carney, A.E., Widin, G.P., Viemeister, N.F., 1977. Noncategorical perception of stop consonants differing in VOT. J. Acoust. Soc. Am. 62, 961–970.

Chandrasekaran, B., Hornickel, J., Skoe, E., Nicol, T., Kraus, N., 2009. Context-dependent encoding in the human auditory brainstem relates to hearing speech in noise: implications for developmental dyslexia. Neuron 64, 311–319.

Chang, S., Blumstein, S.E., 1981. The role of onsets in perception of stop consonant place of articulation: effects of spectral and temporal discontinuity. J. Acoust. Soc. Am. 70, 39–44.

Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., Knight, R.T., 2010. Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. 13, 1428–1433.

Church, K.W., 1987. Phonological parsing and lexical retrieval. Cognition 25, 53–69.

Creutzfeldt, O., Hellweg, F.-C., Schreiner, C., 1980. Thalamocortical transformation of responses to complex auditory stimuli. Exp. Brain Res. 39, 87–104.

Cruikshank, S.J., Rose, H.J., Metherate, R., 2002. Auditory thalamocortical synaptic transmission in vitro. J. Neurophysiol. 87, 361–384.

David, S.V., Fritz, J.B., Shamma, S.A., 2012. Task reward structure shapes rapid receptive field plasticity in auditory cortex. Proc. Natl. Acad. Sci. U. S. A. 109, 2144–2149.

de Villers-Sidani, E., Merzenich, M.M., 2011. Lifelong plasticity in the rat auditory cortex: basic mechanisms and role of sensory experience. Prog. Brain Res. 191, 119–131.

Diehl, R.L., Lotto, A.J., Holt, L.L., 2004. Speech perception. Annu. Rev. Psychol. 55, 149–179.

Ding, N., Simon, J.Z., 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. U. S. A. 109, 11854–11859.

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Am. 95, 1053–1064.

Eggermont, J.J., 1995. Neural correlates of gap detection and auditory fusion in cat auditory cortex. Neuroreport 6, 1645–1648.

Eggermont, J.J., 1998. Representation of spectral and temporal sound features in three cortical fields of the cat. Similarities outweigh differences. J. Neurophysiol. 80, 2743–2764.

Eggermont, J.J., 2000. Neural responses in primary auditory cortex mimic psychophysical, across-frequency-channel, gap-detection thresholds. J. Neurophysiol. 84, 1453–1463.

Eggermont, J.J., 2002. Temporal modulation transfer functions in cat primary auditory cortex: separating stimulus effects from neural mechanisms. J. Neurophysiol. 87, 305–321.

Eilers, R., Gavin, W., Wilson, W., 1979. Linguistic experience and phonetic perception in infancy: a crosslinguistic study. Child Dev. 50, 14–18.

Eimas, P., Siqueland, E., Josczyk, P., Vigorito, J., 1971. Speech perception in infants. Science 171, 303–306.

Engineer, C.T., Perez, C.A., Chen, Y.H., Carraway, R.S., Reed, A.C., Shetake, J.A., Jakkamsetti, V., Chang, K.Q., Kilgard, M.P., 2008. Cortical activity patterns predict speech discrimination ability. Nat. Neurosci. 11, 603–608.

Faulkner, A., Rosen, S., 1999. Contributions of temporal encodings of voicing, voicelessness, fundamental frequency, and amplitude variation to audio-visual and auditory speech perception. J. Acoust. Soc. Am. 106, 2063–2073.

Fishman, Y.I., Steinschneider, M., 2009. Temporally dynamic frequency tuning of population responses in monkey primary auditory cortex. Hear. Res. 254, 64–76.

Fishman, Y.I., Steinschneider, M., 2012. Searching for the mismatch negativity in primary auditory cortex of the awake monkey: deviance detection or stimulus specific adaptation? J. Neurosci., 15747–15758.

Flanagan, J.L., Guttman, N., 1960a. On the pitch of periodic pulses. J. Acoust. Soc. Am. 32, 1308–1319.

Flanagan, J.L., Guttman, N., 1960b. Pitch of periodic pulses without fundamental component. J. Acoust. Soc. Am. 32, 1319–1328.

Freeman, J.A., Nicholson, C., 1975. Experimental optimization of current source density techniques for anuran cerebellum. J. Neurophysiol. 38, 369–382.

Friederici, A.D., 2005. Neurophysiological markers of early language acquisition: from syllables to sentences. Trends Cogn. Sci. 9, 481–488.

Fritz, J.B., David, S.V., Radtke-Schuller, S., Yin, P., Shamma, S.A., 2010. Adaptive, behaviorally-gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. Nat. Neurosci. 13, 1011–1019.

Fritz, J.B., David, S., Shamma, S., 2013. Attention and dynamic, task-related receptive field plasticity in adult auditory cortex. In: Cohen, Y.E., Popper, A.N., Fay, R.R. (Eds.), Springer Handbook of Auditory Research: Neural Correlates of Auditory Cognition. Springer, New York, pp. 251–292.

Ganong III, W.F., 1980. Phonetic categorization in auditory word perception. J. Exp. Psychol. Hum. Percept. Perform. 6, 110–125.

Gaser, C., Schlaug, G., 2003. Brain structures differ between musicians and non-musicians. J. Neurosci. 23, 9240–9245.

Giraud, A.-L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., Kleinschmidt, A., 2000. Representation of the temporal envelope of sounds in the human brain. J. Neurophysiol. 84, 1588–1598.

Giraud, A.-L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15, 511–517.

Goswami, U., 2011. A temporal sampling framework for developmental dyslexia. Trends Cogn. Sci. 15, 3–10.

Goswami, U., Fosker, T., Huss, M., Mead, N., Szűcs, D., 2011. Rise time and formant transition duration in the discrimination of speech sounds: the Ba–Wa distinction in developmental dyslexia. Develop. Sci. 14, 34–43.

Greenberg, S., Ainsworth, W.A., 2004. Speech processing in the auditory system: an overview. In: Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R. (Eds.), Speech Processing in the Auditory System. Springer-Verlag, New York, pp. 1–62.

Grimsley, J.M.S., Shanbhag, S.J., Palmer, A.R., Wallace, M.N., 2012. Processing of communication calls in guinea pig auditory cortex. PLoS ONE 7 (12), e51646. http://dx.doi.org/10.1371/journal.pone.0051646.

Hackett, T.A., Preuss, T.M., Kaas, J.H., 2001. Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. J. Comp. Neurol. 441, 197–222.

Hedrick, M.S., Younger, M.S., 2007. Perceptual weighing of stop consonant cues by normal and impaired listeners in reverberation versus noise. J. Speech Lang. Hear. Res. 50, 254–269.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. 97, 3099–3111.

Hirsh, I.J., 1959. Auditory perception of temporal order. J. Acoust. Soc. Am. 31, 759–767.

Holt, L.L., Lotto, A.J., 2010. Speech perception as categorization. Attent. Perc. Psychophys. 72, 1218–1227.

Jenkinson, M., Bannister, P.R., Brady, J.M., Smith, S.M., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.

Jusczyk, P.W., Rosner, B.S., Reed, M.A., Kennedy, L.J., 1989. Could temporal order differences underlie 2-month-olds. Discrimination of English voicing contrasts? J. Acoust. Soc. Am. 85, 1741–1749.

Kewley-Port, D., 1983. Time-varying features as correlates of place of articulation in stop consonants. J. Acoust. Soc. Am. 73, 322–335.

Kewley-Port, D., Watson, C.S., Foyle, D.C., 1988. Auditory temporal acuity in relation to category boundaries; speech and nonspeech stimuli. J. Acoust. Soc. Am. 83, 1133–1145.

Kiss, M., Cristescu, T., Fink, M., Wittmann, M., 2008. Auditory language comprehension of temporally reversed speech signals in native and non-native speakers. Acta Neurobiol. Exp. 68, 204–213.

Kuhl, P.K., 2004. Early language acquisition: cracking the speech code. Nat. Rev. Neurosci. 5, 831–843.

Kuhl, P.K., 2010. Brain mechanisms in early language acquisition. Neuron 67, 713–727.

Kuhl, P.K., Miller, J.D., 1978. Speech perception by the chinchilla: identification functions for synthetic VOT stimuli. J. Acoust. Soc. Am. 63, 905–917.

Kuhl, P., Rivera-Gaxiola, M., 2008. Neural substrates of language acquisition. Annu. Rev. Neurosci. 31, 511–534.

Lahiri, A., Gewirth, L., Blumstein, S.E., 1984. A reconsideration of acoustic information for place of articulation in diffuse stop consonants: evidence from a cross-language study. J. Acoust. Soc. Am. 76, 391–404.

Lasky, R., Syrdal-Lasky, A., Klein, R., 1975. VOT discrimination by four to six and a half month old infants from Spanish environments. J. Exp. Child Psychol. 20, 213–225.

Liang, L., Lu, T., Wang, X., 2002. Neural representations of sinusoidal amplitude and frequency modulations in the auditory cortex of awake primates. J. Neurophysiol. 87, 2237–2261.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21, 1–36.

Liberman, A.M., Whalen, D.H., 2000. On the relation of speech to language. Trends Cogn. Sci. 4, 187–196.

Liégeois-Chauvel, C., de Graaf, J.B., Laguitton, V., Chauvel, P., 1999. Specialization of left auditory cortex for speech perception in man depends on temporal coding. Cereb. Cortex 9, 484–496.

Liégeois-Chauvel, C., Lorenzi, C., Trébuchon, A., Régis, J., Chauvel, P., 2004. Temporal envelope processing in the human left and right auditory cortices. Cereb. Cortex 14, 731–740.

Lisker, L., 1975. Is it VOT or a first-formant transition detector? J. Acoust. Soc. Am. 57, 1547–1551.

Lisker, L., Abramson, A.S., 1964. A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 384–422.

Liu, R.C., Schreiner, C.E., 2007. Auditory cortical detection and discrimination correlates with communicative significance. PLoS Biol. 5 (7), e173.

Lotto, A.J., Kluender, K.R., Holt, L.L., 1997. Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). J. Acoust. Soc. Am. 102, 1134–1140.

Lu, T., Liang, L., Wang, X., 2001. Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. Nat. Neurosci. 4, 1131–1138.

McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cognit. Psychol. 18, 1–86.

McGee, T., Kraus, N., King, C., Nicol, T., 1996. Acoustic elements of speechlike stimuli are reflected in surface recorded responses over the guinea pig temporal lobe. J. Acoust. Soc. Am. 99, 3606–3614.

Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236.

Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2008. Phoneme representation and classification in primary auditory cortex. J. Acoust. Soc. Am. 123, 899–909.

Metherate, R., Ashe, J.H., 1994. Facilitation of an NMDA receptor-mediated EPSP by paired-pulse stimulation in rat neocortex via depression of GABAergic IPSPs. J. Physiol. 481, 331–348.

Metherate, R., Cruikshank, S.J., 1999. Thalamocortical inputs trigger a propagating envelope of gamma-band activity in auditory cortex in vitro. Exp. Brain Res. 126, 160–174.

Middlebrooks, J.C., 2008. Auditory cortex phase locking to amplitude-modulated cochlear implant pulse trains. J. Neurophysiol. 100, 76–91.

Morel, A., Garraghty, P.E., Kaas, J.H., 1993. Tonotopic organization, architectonic fields, and connections of auditory cortex in macaque monkeys. J. Comp. Neurol. 335, 437–459.

Müller-Preuss, P., Mitzdorf, U., 1984. Functional anatomy of the inferior colliculus and the auditory cortex: current source density analyses of click-evoked potentials. Hear. Res. 16, 133–142.

Newbury, D.F., Monaco, A.P., 2010. Genetic advances in the study of speech and language disorders. Neuron 68, 309–320.

Nourski, K.V., Brugge, J.F., 2011. Representation of temporal sound features in the human auditory cortex. Rev. Neurosci. 22, 187–203.

Nourski, K.V., Brugge, J.F., Reale, R.A., Kovach, C.K., Oya, H., Kawasaki, H., Jenison, R.L., Howard III, M.A., 2013. Coding of repetitive transients by auditory cortex on posterolateral superior temporal gyrus in humans: an intracranial electrophysiology study. J. Neurophysiol. 109, 1283–1295.

Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Howard III, M.A., Brugge, J.F., 2009. Temporal envelope of time-compressed speech represented in the human auditory cortex. J. Neurosci. 29, 15564–15574.

Obleser, J., Eisner, F., 2008. Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn. Sci. 13, 14–19.

Obleser, J., Scott, S.K., Eulitz, C., 2006. Now you hear it, now you don't: transient traces of consonants and their nonspeech analogues in the human brain. Cereb. Cortex 16, 1069–1076.

Oshurkova, E., Scheich, H., Brosch, M., 2008. Click train encoding in primary and non-primary auditory cortex of anesthetized macaque monkeys. Neuroscience 153, 1289–1299.

Oya, H., Kawasaki, H., Howard III, M.A., Adolphs, R., 2002. Electrophysiological responses in the human amygdala discriminate emotion categories of complex visual stimuli. J. Neurosci. 22, 9502–9512.

Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-Locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cortex 23, 1378–1387.

Petkov, C.I., Jarvis, E.D., 2012. Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. Front. Evol. Neurosci. 4. article 12.

Pisoni, D.B., 1977. Identification and discrimination of the relative onset time of two component tones: implications for voicing perception in stops. J. Acoust. Soc. Am. 61, 1352–1361.

Pisoni, D.B., Aslin, R.N., Perey, A.J., Hennessy, B.L., 1982. Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. J. Exp. Psychol. Hum. Percept. Perform. 8, 297–314.

Pisoni, D.B., Luce, P.A., 1987. Acoustic–phonetic representations in word recognition. Cognition 25, 21–52.

Plack, C.J., Carlyon, R.P., 1995. Differences in frequency modulation detection and fundamental frequency discrimination between complex tones consisting of resolved and unresolved harmonics. J. Acoust. Soc. Am. 98, 1355–1364.

Poeppel, D., Emmorey, K., Hickok, G., Pylkkänen, L., 2012. Towards a new neurobiology of language. J. Neurosci. 32, 14125–14131.

Poeppel, D., Idsardi, W.J., van Wassenhove, V., 2008. Speech perception at the interface of neurobiology and linguistics. Phil. Trans. R. Soc. B 363, 1071–1086.

Reddy, C.G., Dahdaleh, N.S., Albert, G., Chen, F., Hansen, D., Nourski, K., Kawasaki, H., Oya, H., Howard III, M.A., 2010. A method for placing Heschl gyrus depth electrodes. J. Neurosurg. 112, 1301–1307.

Saberi, K., Perrott, D.R., 1999. Cognitive restoration of reversed speech. Nature 398, 760.

Samuel, A.G., 2011. Speech perception. Annu. Rev. Psychol. 69, 49–72.

Schneider, P., Scherg, M., Dosch, H.G., Specht, H.J., Gutschalk, A., Rupp, A., 2002. Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. Nat. Neurosci. 5, 688–694.

Schnupp, J.W.H., Hall, T.M., Kokelaar, R.F., Ahmed, B., 2006. Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. J. Neurosci. 26, 4785–4795.

Schreiner, C.E., 1998. Spatial distribution of responses to simple and complex sounds in the primary auditory cortex. Audiol. Neurootol. 3, 104–122.

Schreiner, C.E., Urbas, J.V., 1988. Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. Hear. Res. 32, 49–64.

Shackleton, T.M., Carlyon, R.P., 1994. The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. J. Acoust. Soc. Am. 95, 3529–3540.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.

Sinnott, J.M., Adams, F.S., 1987. Differences in human and monkey sensitivity to acoustic cues underlying voicing contrasts. J. Acoust. Soc. Am. 82, 1539–1547.

Sinnott, J.M., Powell, L.A., Camchong, J., 2006. Using monkeys to explore perceptual "loss" versus "learning" models in English and Spanish voice-onset-time perception. J. Acoust. Soc. Am. 119, 1585–1596.

Sokhi, D.S., Hunter, M.D., Wilkinson, I.D., Woodruff, P.W., 2005. Male and female voices activate distinct regions in the male brain. NeuroImage 27, 572–578.

Steinschneider, M., Fishman, Y.I., 2011. Enhanced physiologic discriminability of stop consonants with prolonged formant transitions in awake monkeys based on the tonotopic organization of primary auditory cortex. Hear. Res. 271, 103–114.

Steinschneider, M., Fishman, Y.I., Arezzo, J.C., 2003. Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey. J. Acoust. Soc. Am. 114, 307–321.

Steinschneider, M., Nourski, K.V., Kawasaki, H., Oya, H., Brugge, J.F., Howard III, M.A., 2011. Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. Cereb. Cortex 21, 2332–2347.

Steinschneider, M., Reser, D.H., Fishman, Y.I., Schroeder, C.E., Arezzo, J.C., 1998. Click train encoding in primary auditory cortex of the awake monkey: evidence for two mechanisms subserving pitch perception. J. Acoust. Soc. Am. 104, 2935–2955.

Steinschneider, M., Reser, D., Schroeder, C.E., Arezzo, J.C., 1995a. Tonotopic organization of responses reflecting stop consonant place of articulation in primary auditory cortex (A1) of the monkey. Brain Res. 674, 147–152.

Steinschneider, M., Schroeder, C., Arezzo, J.C., Vaughan Jr., H.G., 1994. Speech-evoked activity in primary cortex: effects of voice onset time. Electroencephalogr. Clin. Neurophysiol. 92, 30–43.

Steinschneider, M., Schroeder, C.E., Arezzo, J.C., Vaughan Jr., H.G., 1995b. Physiologic correlates of the voice onset time (VOT) boundary in primary auditory cortex (A1) of the awake monkey: temporal response patterns. Brain Lang. 48, 326–340.

Steinschneider, M., Tenke, C., Schroeder, C., Javitt, D., Simpson, G.V., Arezzo, J.C., Vaughan Jr., H.G., 1992. Cellular generators of the cortical auditory evoked potential initial component. Electroencephalogr. Clin. Neurophysiol. 84, 196–200.

Steinschneider, M., Volkov, I.O., Fishman, Y.I., Oya, H., Arezzo, J.C., Howard III, M.A., 2005. Intracortical responses in human and monkey primary auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter. Cereb. Cortex 15, 170–186.

Steinschneider, M., Volkov, I.O., Noh, M.D., Garell, P.C., Howard III, M.A., 1999. Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. J. Neurophysiol. 82, 2346–2357.

Stevens, K.N., 1981. Constraints imposed by the auditory system on the properties used to classify speech sounds: data from phonology, acoustics, and psychoacoustics. In: Myers, T., Laver, J., Anderson, J. (Eds.), The Cognitive Representation of Speech. North-Holland, Amsterdam, pp. 61–74.

Stevens, K.N., 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Am. 111, 1872–1891.

Stevens, K.N., Blumstein, S.E., 1978. Invariant cues for place of articulation in stop consonants. J. Acoust. Soc. Am. 64, 1358–1368.

Summerfield, Q., Haggard, M., 1977. On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. J. Acoust. Soc. Am. 62, 435–448.

Supèr, H., Roelfsema, P.R., 2005. Chronic multiunit recordings in behaving animals: advantages and limitations. In: van Pelt, J., Kamermans, M., Levelt, C.N., van Ooyen, A., Ramakers, G.J.A., Roelfsema, P.R. (Eds.), 2005. Progress in Brain Research, vol. 147. Elsevier Science, Amsterdam, pp. 263–282.

Szymanski, F.D., Rabinowitz, N.C., Magri, C., Panzeri, S., Schnupp, J.W.H., 2011. The laminar and temporal structure of stimulus information in the phase of field potentials of auditory cortex. J. Neurosci. 31, 15787–15801.

Tallal, P., 2004. Improving language and literacy is a matter of time. Nat. Rev. Neurosci. 5, 721–728.

Tavabi, K., Obleser, J., Dobel, C., Pantev, C., 2007. Auditory evoked fields differentially encode speech features: an MEG investigation of the P50m and N100m time courses during syllable processing. Eur. J. Neurosci. 25, 3155–3162.

Trébuchon-Da Fonseca, A., Giraud, K., Badier, J.-M., Chauvel, P., Liégeois-Chauvel, C., 2005. Hemispheric lateralization of voice onset time (VOT) comparison between depth and scalp EEG recordings. NeuroImage 27, 1–14.

Tsunada, J., Lee, J.H., Cohen, Y.E., 2011. Representation of speech categories in the primate auditory cortex. J. Neurophysiol. 105, 2634–2646.

Vandermosten, M., Boets, B., Luts, H., Poelmans, H., Golestani, N., Wouters, J., Ghesquiere, P., 2010. Adults with dyslexia are impaired in categorizing speech and nonspeech sounds on the basis of temporal cues. Proc. Natl. Acad. Sci. U. S. A. 177, 10389–10394.

Walley, A.C., Carrell, T.D., 1983. Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. J. Acoust. Soc. Am. 73, 1011–1022.

Wang, X., 2007. Neural coding strategies in auditory cortex. Hear. Res. 229, 81–93.

Wang, X., Lu, T., Bendor, D., Bartlett, E., 2008. Neural coding of temporal information in auditory thalamus and cortex. Neuroscience 154, 294–303.

Wang, X., Lu, T., Snider, R.K., Liang, L., 2005. Sustained firing in auditory cortex by preferred stimuli. Nature 435, 341–346.

Wang, X., Merzenich, M.M., Beitel, R., Schreiner, C.E., 1995. Representation of a species specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. J. Neurophysiol. 74, 2685–2706.

Woolley, S.M.N., 2012. Early experience shapes vocal neural coding and perception in songbirds. Dev. Psychobiol. 54, 612–631.

Yin, P., Johnson, J.S., O'Connor, K.N., Sutter, M.L., 2011. Coding of amplitude modulation in primary auditory cortex. J. Neurophysiol. 105, 582–600.