NeuroImage 79 (2013) 201–212

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Tracing the emergence of categorical speech perception in the human auditory system

Gavin M. Bidelman [a,b,*], Sylvain Moreno [c], Claude Alain [c,d]

[a] Institute for Intelligent Systems, University of Memphis, Memphis, TN 38105, USA
[b] School of Communication Sciences & Disorders, University of Memphis, Memphis, TN 38105, USA
[c] Rotman Research Institute, Baycrest Centre for Geriatric Care, Toronto, ON M6A 2E1, Canada
[d] Department of Psychology, University of Toronto, Toronto, ON M6A 2E1, Canada

ARTICLE INFO

Article history:
Accepted 21 April 2013
Available online 3 May 2013

Keywords:
Categorical perception
Speech perception
Brainstem response
Auditory event-related potentials (ERP)
Neural computation

ABSTRACT

Speech perception requires the effortless mapping from smooth, seemingly continuous changes in sound features into discrete perceptual units, a conversion exemplified in the phenomenon of categorical perception. Explaining how/when the human brain performs this acoustic–phonetic transformation remains an elusive problem in current models and theories of speech perception. In previous attempts to decipher the neural basis of speech perception, it is often unclear whether the alleged brain correlates reflect an underlying percept or merely changes in neural activity that covary with parameters of the stimulus. Here, we recorded neuroelectric activity generated at both cortical and subcortical levels of the auditory pathway elicited by a speech vowel continuum whose percept varied categorically from /u/ to /a/. This integrative approach allows us to characterize how various auditory structures code, transform, and ultimately render the perception of speech material as well as dissociate brain responses reflecting changes in stimulus acoustics from those that index true internalized percepts. We find that activity from the brainstem mirrors properties of the speech waveform with remarkable fidelity, reflecting progressive changes in speech acoustics but not the discrete phonetic classes reported behaviorally. In comparison, patterns of late cortical evoked activity contain information reflecting distinct perceptual categories and predict the abstract phonetic speech boundaries heard by listeners. Our findings demonstrate a critical transformation in neural speech representations between brainstem and early auditory cortex analogous to an acoustic–phonetic mapping necessary to generate categorical speech percepts. Analytic modeling demonstrates that a simple nonlinearity accounts for the transformation between early (subcortical) brain activity and subsequent cortical/behavioral responses to speech (>150–200 ms) thereby describing a plausible mechanism by which the brain achieves its acoustic-to-phonetic mapping. Results provide evidence that the neurophysiological underpinnings of categorical speech are present cortically by ~175 ms after sound enters the ear.

Introduction

Sensory phenomena are typically subject to percept invariance in which a continuum of similar features is mapped onto a common identity. This many-to-one mapping is the hallmark of categorical perception (CP) which manifests in many aspects of human cognition including the perception of faces (Beale and Keil, 1995), colors (Franklin et al., 2008), and music (Klein and Zatorre, 2011). CP is particularly important in the context of speech perception whereby gradually morphed sounds along a large acoustic continuum are heard as belonging to one of only a few discrete phonetic classes (Harnad, 1987; Liberman et al., 1967; Pisoni, 1973; Pisoni and Luce, 1987). That is, listeners treat sounds within a given category as perceptually similar despite their otherwise dissimilar acoustic characteristics. Given that categorical percepts do not faithfully map from exact sensory input, they provide useful divisions of information not contained in the external world (Miller et al., 2003). Presumably, this type of "downsampling" mechanism would promote speech comprehension by generating perceptual constancy in the face of individual variation along multiple acoustic dimensions, e.g., talker variability in tempo, pitch, or timbre (Prather et al., 2009).

Categorical speech boundaries emerge early in life (Eimas et al., 1971) and are further modified based on one's native tongue (Kuhl et al., 1992) suggesting that the neural mechanisms underlying CP, while potentially innate, are also malleable to the experiential effects of learning and language experience. Indeed, the fundamental importance of this "phonetic mode" of listening (Liberman and Mattingly, 1989) to speech and language processing is evident by its integral role in speech acquisition (Eimas et al., 1971; Vihman, 1996) and the grapheme-to-phoneme mapping essential for reading and writing skills (Mody et al., 1997; Werker and Tees, 1987). Despite its importance to everyday communication,

* Corresponding author at: School of Communication Sciences & Disorders, University of Memphis, 807 Jefferson Ave., Memphis, TN 38105, USA. Fax: +1 901 525 1282.
E-mail address: g.bidelman@memphis.edu (G.M. Bidelman).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.04.093

the neural mechanisms underlying this cognitive ability remain poorly understood. The fact that speech sounds separated by equal acoustic distance may nevertheless be perceived in an unequal, categorical fashion suggests that in order to generate adequate speech percepts, the auditory system must perform a critical warping between the continuous acoustic domain of the external world and the discrete perceptual space internalized by a listener (Iverson and Kuhl, 1995; Pisoni and Luce, 1987). Understanding how and when this ubiquitous acoustic–phonetic transformation is realized within the human brain is among the many broad and widespread interests to understand how sensory features are mapped to higher order perception (Phillips, 2001; Pisoni and Luce, 1987).

Neuroimaging studies have offered a glimpse into the neural architecture underlying speech processing (Chang et al., 2010; Maiste et al., 1995; Mesgarani and Chang, 2012; Myers and Swan, 2012; Pasley et al., 2012). Cortical event-related brain potentials (ERPs), for example, have revealed systematic modulations in electrical brain activity in response to salient features of speech necessary for its comprehension including voice-onset time (Sharma and Dorman, 1999), spectrotemporal modulations (Pasley et al., 2012), voice pitch (Zhang et al., 2012), and timbre cues (Agung et al., 2006). More recently, provocative work examining human brainstem evoked responses suggests that even subcortical auditory structures, regions traditionally viewed as only passive relays, may play an active role in the early formation of speech percepts (Bidelman and Krishnan, 2010; Krizman et al., 2012; Skoe and Kraus, 2010). Indeed, the microphonic signature and remarkable clarity with which brainstem potentials preserve the properties of an acoustic signal have offered a unique window into understanding the neural transcription of a speech at initial stages along the auditory pathway. From scalp-recorded ERPs, it is thus apparent that both subcortical and cortical processes are dynamically engaged in the encoding of important speech cues. Yet, exactly how and when the nervous system translates sensory features into the abstract brain representations that subsequently render meaningful speech percepts has yet to be established. Furthermore, given the numerous reports implicating brainstem in high-level perceptual/cognitive tasks, one may posit that attributes of CP might emerge even at subcortical levels of processing. This possibility remains to be tested.

Here, we aimed to elucidate the dynamics of hierarchical processing performed by the auditory system during the perception of speech material. By recording evoked neuroelectric activity from multiple levels of the auditory system (brainstem and cortex), we probe the earliest stage and time at which neurophysiological activity maps to higher order perceptual attributes of speech. That is, we aim to reveal when along the auditory pathway neural responses begin to reflect auditory perception rather than simply acoustic properties of the stimuli. Measuring concurrent brain responses thus allows us to assess not only the output from several levels of the auditory pathway simultaneously, but also provides insight into the neurocomputations and transformations performed by the auditory *system* during speech perception. In conjunction with scalp-recorded ERPs, CP provides a unique means to dissociate neural activity that simply reflects changes in an external stimulus (exogenous response) from that which indexes a true internalized percept (endogenous response) (Pisoni and Luce, 1987). In the context of CP and speech listening, it allows us to trace the form of emergent brain representations and reveal how and when speech is converted from sensory information of the surrounding physical world into a cognitive facet of language.

Materials & methods

Participants

Fifteen English-speaking young adults (age: 23.8 ± 4.1 years; 9 females) participated in the experiment. All were right-handed and reported a collegiate level of education (17.1 ± 2.3 years). No screening was made for musical proficiency. Audiometric screening confirmed normal hearing sensitivity (i.e., ≤25 dB HL) at octave frequencies between 250 and 8000 Hz. Participants reported no previous history of neurological or psychiatric illnesses. All were paid for their time and gave informed written consent in accordance with a protocol approved by the Baycrest Centre Research Ethics Committee.

Stimuli: speech vowel continuum

Categorical speech perception requires that gradually morphed sounds along a large acoustic continuum are heard as belonging to one of only a few discrete phonetic classes (Harnad, 1987; Liberman et al., 1967; Pisoni, 1973; Pisoni and Luce, 1987). In conjunction with ERPs, employing a categorical stimulus paradigm offers clear predictions as to what each of the evoked components reflects: neuroelectric activity directly embodying speech percepts should vary with the phonetic percept itself, independent of changes in the acoustic properties of the sound stimulus. CP is typically studied using stop consonants which differ minimally in voice onset time (VOT), i.e., the initial ~40 ms containing critical formant transitions into the vowel (Chang et al., 2010; Pisoni, 1973; Sharma and Dorman, 1999). Though less salient of an effect than changes to VOT, steady-state vowels alone can be perceived categorically by simply manipulating individual formant frequencies, prominent acoustic cues that determine speech identity (Pisoni, 1973). In the present study, we employed a speech vowel continuum to probe the neural underpinnings of CP. Utilizing vowels ensured that the entire stimulus contributed to the categorical percept rather than only the initial transient onset (cf. VOT stimuli) thereby maximizing the possibility that ERPs could be used to differentiate phonetic-level information. To this end, a synthetic five-step vowel continuum was constructed such that each token differed minimally acoustically, yet was perceived categorically (Pisoni, 1973). This was accomplished by varying a single acoustic dimension across the stimuli, namely first formant frequency (F1). Each token was 100 ms including 10-ms of rise/fall time in order to reduce spectral splatter in the stimuli. Tokens contained identical voice fundamental (F0), second (F2), and third formant (F3) frequencies (F0: 100, F2: 1090 and F3: 2350 Hz, respectively) chosen to match prototypical productions from male speakers (Peterson and Barney, 1952). The critical stimulus variation was achieved by parameterizing first formant (F1) over five equal steps between 430 and 730 Hz such that the resultant stimulus set spanned a perceptual phonetic continuum from /u/ to /a/ (Fig. 1a). Stimuli were synthesized with a cascade formant synthesizer implemented in MATLAB using the techniques described by Klatt and colleagues (Klatt and Klatt, 1990).

Data acquisition and preprocessing

Stimulus presentation was controlled by a MATLAB (The MathWorks) program routed to an audiometer (GSI) via a TDT RP2 interface (Tucker-Davis Technologies) and delivered binaurally at an intensity of 83 dB SPL through insert earphones (ER-3A, Etymotic). Extended acoustic tubing and alternating polarity eliminated cochlear microphonic and the potential of electromagnetic stimulus artifact from contaminating neurophysiological responses (Aiken and Picton, 2008; Campbell et al., 2012; Skoe and Kraus, 2010). The effectiveness of these controls was confirmed by the absence of a response or artifact during a control run in which the air tubes were blocked to the ear (Aiken and Picton, 2006, 2008). Stimulus intensity was calibrated using a Larson-Davis sound pressure level (SPL) meter (Model 824, Provo, Utah) measured in a 2 cm^3, artificial ear coupler (Model AEC100l). Left and right ear channels were calibrated separately.

During ERP recording, listeners heard 200 randomly ordered exemplars of each token and were asked to label them with a binary response as quickly as possible ("u" or "a"). The interstimulus interval (ISI) was jittered randomly between 400 and 600 ms (20-ms steps,

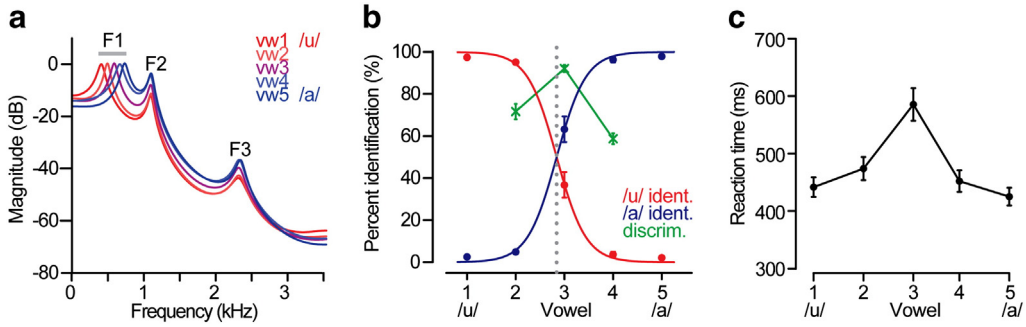

Fig. 1. Speech vowel continuum used to probe the categorical organization of speech. (a) Acoustic spectral envelopes of vowel stimuli (log scale). (b) Psychometric identification/discrimination functions and (c) reaction times for vowel identification. Despite the continuous acoustic change in stimulus first formant (F1) frequency, listeners hear a clear perceptual shift in the phonetic category (/u/ vs. /a/) around token 3 (dotted line) and are much slower at labeling stimuli near this categorical boundary as compared to within-category tokens (e.g., 1–2 or 4–5).

rectangular distribution). The choice of these parameters ensured a balance between response recording time and minimizing stimulus-specific refractory (Picton et al., 1978b) and component habituation effects (Picton et al., 1977) typical of cortical auditory ERPs. An additional 2000 trials (ISI = 150 ms) were then collected in order to detect sub-microvolt brainstem ERPs (Bidelman and Krishnan, 2010). During brainstem recordings, participants watched a self-selected muted movie with subtitles to facilitate a calm and wakeful state. Brainstem ERPs were recorded during passive listening because prior work has consistently demonstrated that they are largely unaffected by attention (Galbraith and Kane, 1993; Hillyard and Picton, 1979; Okamoto et al., 2011; Picton and Hillyard, 1974; Picton et al., 1971; Woods and Hillyard, 1978; but see Galbraith et al., 2003). Moreover, employing such a protocol allows us to make direct comparisons with the extant literature given that previous studies documenting connections between brainstem ERPs and speech perception have used identical passive listening paradigms (Bidelman and Krishnan, 2009; Krishnan et al., 2010; Krizman et al., 2012; Parbery-Clark et al., 2009; Song et al., 2010). In total, the experimental protocol took 2 h to complete.

Continuous electroencephalograms (EEGs) were recorded differentially using four Ag–AgCl scalp electrodes. Neural activity was recorded from a non-inverting (+) active electrode placed at the midline just below the hairline with reference to linked mastoids (non-inverting electrodes) (Bidelman and Krishnan, 2010; Krishnan et al., 2010). Another electrode placed on the mid forehead (~Fpz) served as the common ground. This vertical montage is optimal for recording simultaneous evoked responses of both subcortical (Bidelman and Krishnan, 2010) and cortical (Musacchia et al., 2008) origin. Inter-electrode contact impedance was maintained below 3 kΩ throughout the duration of the experiment. EEGs were digitized at 20 kHz and bandpass filtered online between 0.05 and 3500 Hz (SymAmps2, NeuroScan). Traces were then segmented (cortical ERP: −100–600 ms; brainstem ERP: −40–210 ms), baselined to the pre-stimulus interval, and subsequently averaged in the time domain to obtain ERPs for each condition (Delorme and Makeig, 2004). Trials exceeding ±50 μV were rejected prior to averaging. Grand average evoked responses were then highpass (80–2500 Hz) or lowpass (1–30 Hz) filtered to isolate brainstem and cortical evoked responses, respectively (Fig. 2a) (Musacchia et al., 2008). Filter cutoffs were chosen based on a priori knowledge of both the stimulus and ERP bandwidths. Given that brainstem ERPs are a neuro-microphonic of the acoustic waveform, they do not contain meaningful response energy below the lowest component in the acoustic stimulus (here 100 Hz). In contrast, neural activity in cortical ERPs (i.e., time-locked activity) is band-limited to about 30 Hz. Thus, filter parameters were chosen to (1) separate cortical and brainstem responses buried within the EEG, (2) maintain the entire bandwidth of the two constituent evoked responses, and (3) attenuate non-meaningful EEG activity in the recordings (e.g., myogenic, thermal noise).

Behavioral psychometric analysis

Psychometric functions were constructed according to the number of times (% of total trials) participants identified each vowel token with either the /u/ or /a/ phonetic label. Note that given the binary choice of the task, this analysis results in two complementary identification functions that are inverses of one another. Behavioral discrimination was estimated using identification scores between pairs of non-adjacent stimuli (two-step discrimination) according to Eq. (1):

$$P(C) = \frac{\left(a-a'\right)^2 + \left(b-b'\right)^2 + 2}{4} \quad (1)$$

where $P(C)$ represents the probability of correctly discriminating stimuli A and B, $a = P(a|A)$ (i.e., the probability of labeling stimulus A as phoneme a), $a' = P(a|B)$ (i.e., the probability of labeling stimulus B as phoneme a), $b = P(b|A)$, and $b' = P(b|B)$. Predicted discrimination is based on the common assumption that an observer's ability to discriminate tokens is characterized by how many sounds he/she can identify (May, 1981; Pollack and Pisoni, 1971); predicted values tend to closely match actual performance and thus provide an estimate of behavioral discrimination without measuring it explicitly. Throughout, discrimination scores are plotted along the abscissae at 2, 3, and 4 of the continuum corresponding to the discrimination between token pairs 1–3, 2–4, and 3–5, respectively (e.g., see Fig. 1b).

Brainstem and cortical ERP analysis

Brainstem first formant (F1) encoding

We quantified the salience of F1 information contained in brainstem responses. Spectral envelopes were computed from brainstem ERPs using linear predictive coding (LPC) analysis (35th-order filter) (Markel and Gray, 1976). LPCs are typically used in analyzing speech waveforms and can be used to estimate the formant resonances of a speech utterance. We adopted this common analysis but applied it to brainstem ERPs in order to visualize the spectral envelope of the responses. Neural response LPC functions were used to estimate the location and magnitude of formant-related information captured in the neural activity without a priori knowledge of the stimulus formants. F1-related encoding was quantified from each response LPC per stimulus and participant, defined as the peak magnitude in the LPC spectral envelope between 400 and 750 Hz, i.e., the expected F1 range from the input stimulus (see Fig. 1a). The location of these local maxima

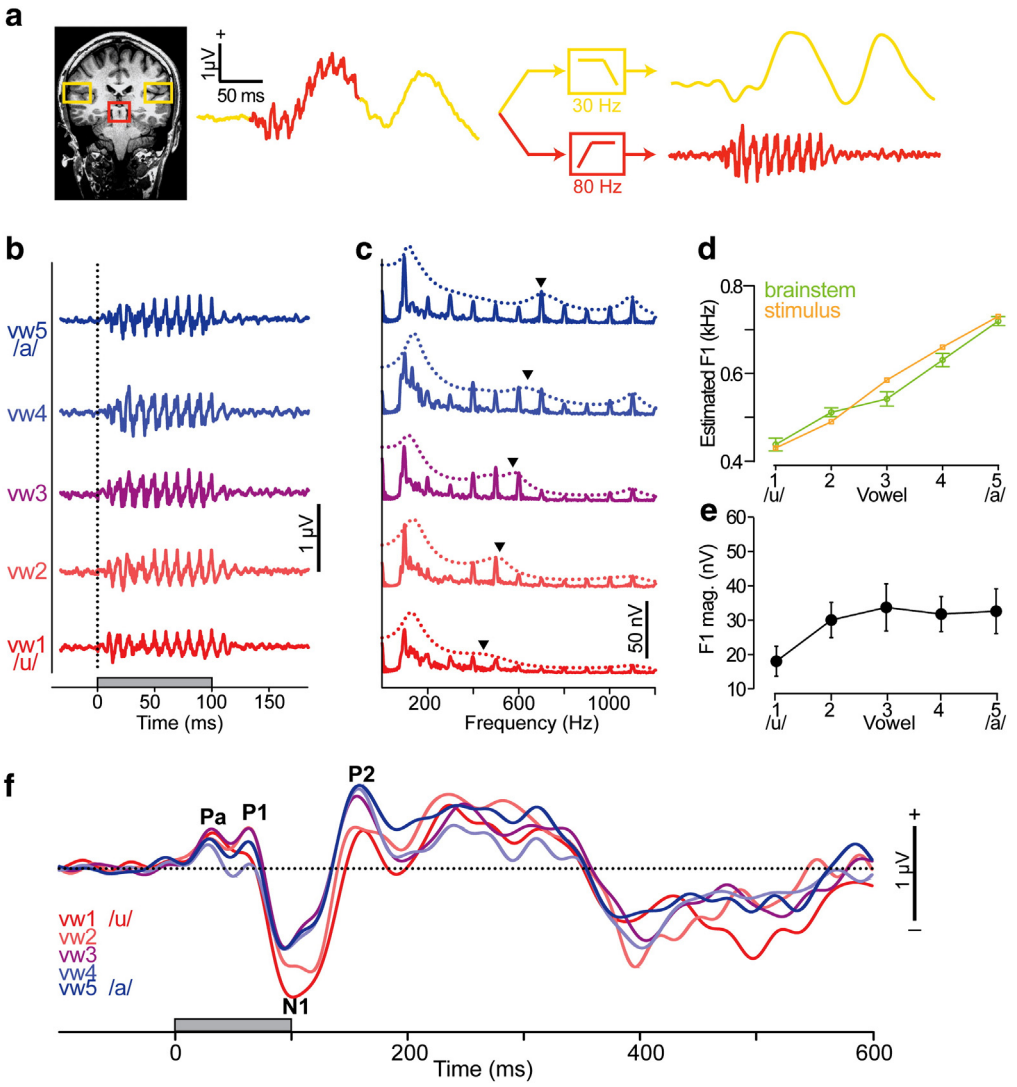

Fig. 2. Brainstem and cortical neuroelectric responses elicited during categorical speech perception. (a) Schematic derivation of brainstem (red) and cortical (yellow) responses from grand averaged evoked activity recorded at the scalp (trace time durations not to scale). The anatomy is provided only to illustrate the gross sites of the presumed ERP generators and not absolute source locations. (b) Brainstem ERPs. Subcortical neural activity mirrors the temporal waveform of the eliciting acoustic stimulus (schematized by the gray box). Stimulus onset (time = 0 ms) is indicated by the vertical dotted line. (c) Spectral encoding in brainstem responses. Solid traces show harmonic spectra of brainstem ERPs computed via Fast Fourier Transform (FFT); dotted lines, LPC spectral envelopes (linear scale). Changes in stimulus F1-energy are followed with high fidelity in the brainstem response. Arrows demarcate the location of F1-related energy encoded in brainstem responses estimated as the peak amplitude in the response LPC falling within the range of F1s from the stimulus set. (d) Correspondence between stimulus F1 and formant-related energy estimated from brainstem potentials. Subcortical responses closely follow linear changes in formant energy across the continuum. (e) F1 encoding magnitude is invariant across stimuli. (f) Group mean cortical ERPs elicited by the vowel continuum. Note the distinct modulations in middle- and long-latency components of the response (e.g., P2). Error bars here and throughout = s.e.m.

provided an estimate of the F1 frequency as represented in the brainstem ERP. One participant's F1 magnitudes were beyond two standard deviations of the group mean and hence, were excluded from subsequent statistical analyses. Fast Fourier Transforms (FFTs) of brainstem ERPs allowed for the visualization of speech harmonics encoded in subcortical responses within the limit of brainstem phase-locking (Liu et al., 2006).

Neural dissimilarity matrices

Motivated by the classical definition of CP (Harnad, 1987), we aimed to investigate whether evoked brain activity at different levels of auditory processing could predict the behavioral classification of speech tokens. As such, we adopted a technique originally employed by Chang et al. (2010) to classify electrophysiological responses recorded while epileptic patients performed categorical speech task. The rationale behind the algorithm is that within category speech sounds are perceived as belonging to the same class and therefore should elicit similar neural activity patterns whereas across category tokens are heard as dissimilar, and thus, should elicit more divergent neural responses (Chang et al., 2010). In speech perception, psychophysical differences are often explored using confusion matrices representing the perceptual dissimilarity/similarity between speech sounds. Analogous "neural dissimilarity" matrices (Chang et al., 2010) were computed separately within four 20-ms time windows centered at the constituent waves of the group mean cortical ERP (Pa: 30 ms; P1: 50 ms; N1: 100 ms; P2: 175 ms). This duration was chosen to maximize the window's extent while ensuring that it viewed only a single wave of the ERP at one time (component waves are separated by ~50 ms). Similar analysis windows have been successfully applied in previous studies examining the cortical correlates of CP (Chang et al., 2010). For brainstem ERPs, the analysis segment included the entire portion of the steady-state response

(15–110 ms) (Bidelman and Krishnan, 2010)[1] because we had no reason to believe that any one part of a steady-state neural response would show differential encoding when elicited by time-invariant stimuli. Within each analysis window, the standardized Euclidean distance was computed between the raw voltage waveforms of all pairwise response combinations and used to construct each dissimilarity matrix (Figs. 3a–b, top panels) (Chang et al., 2010). Each matrix cell quantifies the degree to which neuroelectric activity differs between a given pair of vowel stimuli.

Multidimensional scaling (MDS) is a well-established analysis tool often used to examine the perceptual dissimilarity and categorical perception of stimuli (Borg and Groenen, 2005; Shepard, 1980). MDS attempts to represent data observations in a low-dimensional Euclidean space such that the distances between objects reproduce the empirical matrix of dissimilarities as well as possible. Adopting this popular technique to ERPs, neural dissimilarity matrices were submitted to a metric MDS algorithm in order to visualize differences in neurophysiological responses across the stimulus set (Figs. 3a–b, bottom panels). Graphically, points in the resultant neural space represent empirical distances between brain responses akin to geographic distances between cities on a map. Within-category speech sounds are more difficult to discriminate and thus, elicit responses which are positioned closer together in MDS space; across-category tokens are easier to discriminate and hence, are well separated geometrically (Chang et al., 2010). A stress value <0.1, which represents the reconstruction's "badness of fit", was obtained with a MDS solution of only two dimensions indicating an adequate fit to the data (Borg and Groenen, 2005).

Segregating ERP data into phonetic classes

Based on visual observation of MDS maps, it was apparent that the data clustered into two distinct groupings. This segregation was quantified by applying *k*-means clustering ($k = 2$) to the MDS solution to test whether brainstem and cortical ERPs grouped in a way that parallels the psychophysical classification of the tokens. This unsupervised algorithm initially randomly partitions the data into *k* discrete sets and through an iterative process, determines class membership of each observation. That is, given a set of *n* observations ($x_1, x_2, x_3, \ldots, x_n$), the clustering aims to segregate the dataset into *k* sets ($k \le n$) $\mathbf{S} = [S_1, S_2, S_3, \ldots, S_k]$ by minimizing the within-cluster sum of squares (Eq. (2)):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{x_j \in \mathbf{S}_i} ||x_j - \boldsymbol{\mu}_i||^2 \tag{2}$$

where $\boldsymbol{\mu}_i$ represents the mean of the points in set $\mathbf{S}_i$. In minimizing the within cluster sum of squares, the between cluster variance is consequently maximized and the *k*-means algorithm converges on the final data partition. The choice of two clusters was based on the a priori knowledge that perceptually, our stimuli fall into one of two distinct phonetic categories (i.e., /u/ and /a/) (cf. Chang et al., 2010). Note that this also parallels the procedure of the behavioral task which forces listeners to make a binary decision when labeling the vowel continuum. Thus, in using *k*-mean clustering, we ask that the data naturally group itself into two classes based solely on differences between the evoked activity generated by each vowel sound (Chang et al., 2010). Objects falling within a given cluster were considered to be representatives of the same vowel identity (i.e., phonetic category). In effect, this data-driven approach implements the definition of CP (Chang et al., 2010). MDS and clustering routines were performed using built-in functions within MATLAB.

Derivation of neurometric identification and discrimination functions

Neurometric identification functions (Figs. 3c–d) were constructed for each phonetic category by computing the normalized distance between each of the individual responses (as represented in MDS space) and each of the two cluster means (representing a neural exemplar, or "template", for each vowel category) (Chang et al., 2010). Generated with respect to both cluster means, the resulting functions estimate how well the neural activity evoked by each vowel stimulus fits into one of the two discrete phonetic categories. Neural discrimination functions were estimated by computing the normalized distance between consecutive pairs of neural responses in MDS space (i.e., 2-step discrimination). Values are normalized such that the maximum distance of the sample corresponds to 100% discrimination. Again, difficult to discriminate stimuli (i.e., those crossing the categorical boundary) should elicit neural activity of closer proximity than those which are easy to discriminate (i.e., across category tokens). For each participant, we then computed a correlation coefficient between brain and behavioral functions. The significance of these correlations was examined through a single sample *t*-test where the null hypothesis was zero group mean correlation.

Note that for both brainstem and cortical ERPs, we use simple time-domain waveforms for quantifying response dissimilarity and subsequent phonetic classification. Using raw voltage traces—as opposed to say F1 magnitude for brainstem ERPs and magnitude/latency for cortical ERPs—allows us to apply the same analyses in constructing neurometric functions for both brainstem and cortical ERP data.

Statistical reliability estimates

Bootstrap resampling (Efron and Tibshirani, 1993) was used to validate the significant brain–behavior correlations observed in the group mean ERPs. Participants were randomly sampled (with replacement) N = 5000 times from the original dataset. From each bootstrap resample, the correlation between neurometric and psychometric functions was recomputed. The distribution of these statistics (i.e., probability densities) enabled us to assess the reliability of the observed correlations in the study sample.

Response based analysis of cortical ERPs

In the current stimulus set, token 3 of the continuum leads to a bistable percept. That is, listeners' perception alternated between two mutually exclusive vowel identities for a single acoustic stimulus (~50% of the time, Fig. 1b). Therefore, in an additional analysis, we divided cortical ERP trials elicited by this ambiguous token according to listener's perceptual response (/u/ vs. /a/). Thus response-dependent ERPs were composed of roughly 100 trials apiece (i.e., half of the total number of responses). This allowed us to test whether or not evoked response magnitudes are modulated depending on the perceived quality of a given speech token while holding the acoustic input constant.

Subcortical–cortical transformation function

The brainstem-to-cortical transformation was modeled with a simple sigmoidal nonlinearity (Eq. (3)) (Rosenblatt, 1958).

$$vowel(f) = \frac{1}{1 + e^{-\alpha(f - f0)}} \tag{3}$$

where *f* represents the input formant frequency and the free model parameters α and f_0 govern the width (i.e., rise time) and center of gravity (i.e., abscissa shift) of the logistic fit, respectively. Similar logistic models have been used extensively in artificial neural networks as a plausible neurobiological realization of a binary classifier

[1] In additional analyses on brainstem data, we used a more temporally restricted analysis window (~20 ms) to parallel that used on the cortical ERP waves. However, this resulted in a similar pattern of results suggesting that the duration of the analysis window did not grossly affect the measured response dissimilarity between brainstem ERPs. Thus, a window which encompassed the entire extent of the time-invariant, steady-state response was employed for the neurometric analysis of the brainstem ERPs.

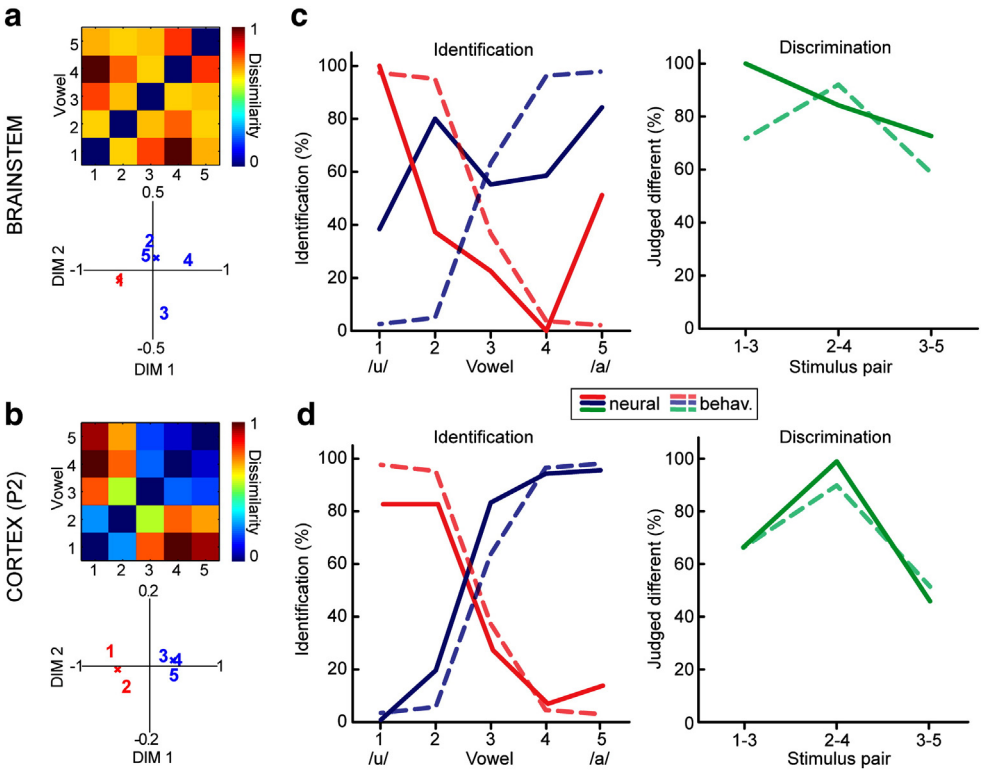

Fig. 3. ERPs reveal a transformation in the neural speech code within the ascending auditory pathway. (a, b) Dissimilarity matrices (top panels) and multidimensional scaling (bottom panels) allow for the visualization of how brain response patterns differ across vowel stimuli. ERPs segregate into two meaningful groups (red = /u/cluster; blue = /a/ cluster; 'x' = cluster centroid) mimicking the two phonemic categories of the stimulus set. (c, d) Correlations between neurometric (solid) and psychometric (dotted) categorical functions. (c) Despite the fidelity of speech cues coded at the level of the brainstem (see Fig. 2), subcortical responses do not predict the categorical boundaries as found psychophysically. (d) Cortical identification/discrimination functions derived in the time window of P2 closely match those obtained behaviorally suggesting that later cortical evoked activity provides a robust correlate of the categorical nature of speech perception.

(Rosenblatt, 1958). Similar sigmoidal models have been successfully applied in categorizing of both multiunit activity (Purushothaman and Bradley, 2005) and fMRI activation (Ley et al., 2012) in tasks requiring binary decisions. In this implementation, the particular label (*vowel*) a listener applies to a given speech sound depends on a simple threshold along the continuum's overall range (F); formant frequencies (f) straddling the internalized decision criterion ($F/2$) are heard as belonging to distinct and separate phonetic classes (Fig. 6a). Brainstem activity (see Fig. 2d) input to this transform provides a prediction for the subsequent cortical/behavioral response magnitudes generated further upstream. Given that brainstem showed invariance in magnitude across vowels (Fig. 2e), we used formant frequency estimates (Fig. 2d) as the input to the transform as it covaried with vowel stimuli. Agreement between predicted and observed functions was assessed by comparing the confidence intervals of each function's slope fitting parameter α.

Results

While recording electrical brain activity, participants labeled sounds drawn randomly from a speech continuum comprised of a set of five vowels that differed minimally only in first-formant frequency (F1) (Pisoni, 1973), a critical cue for speech understanding (Fig. 1a). Despite the small, continuous change in the tokens' F1, listeners perceived the tokens categorically, i.e., they heard a clear perceptual shift in vowel identity (/u/ vs. /a/) and were subsequently slower [one-way ANOVA, $F_{4,70} = 9.57$, $P < 0.001$] to label tokens surrounding the categorical boundary compared to those within either category ($P < 0.001$, Bonferroni corrected) (Figs. 1b–c).

Having first confirmed that our speech stimuli were indeed perceived categorically, we then examined listeners' corresponding brainstem and cortical evoked responses in order to trace how and when continuous changes in F1 were mapped into discrete vowel categories within the brain. We posited that neuronal activity following linear changes in F1 would be indicative of an acoustic coding strategy whereas activity paralleling changes in the perceptual category would reflect a neurobiological correlate of the higher-order phonetic code of speech. Selective filtering techniques (Musacchia et al., 2008) were used to extract ERPs generated at both subcortical and early cortical levels of the auditory pathway for subsequent correlation with behavioral data (Fig. 2a).

Fig. 2b shows brainstem ERPs to speech appeared as stimulus phase-locked potentials that preserved the spectrotemporal properties of the eliciting acoustic signal (Bidelman and Krishnan, 2010; Skoe and Kraus, 2010). The remarkable fidelity of the brainstem ERP was evident by the clarity in which it captured the harmonic structure and formant energy of the speech sounds (Fig. 2c). Indeed, F1-energy encoded in brainstem tightly paralleled changes in F1 along the stimulus continuum (Fig. 2d). Robust stimulus-to-response F1 correlations were observed across all participants (Pearson's correlation: $r = 0.96$; $t(14) = 1.4e3$, $P < 0.0001$, t-test against zero group mean correlation). While subcortical response properties changed in accordance with stimulus acoustics, the magnitude of formant encoding remained invariant across the continuum [ANOVA, $F_{4,65} = 1.33$, $P = 0.28$] (Fig. 2e). This implies that while the brainstem ERPs accurately track changes in the acoustic stimulus, it does not differentially encode speech sounds based on higher level phonetic characteristics. Time-locked cortical ERPs to speech were characterized by a series of early positive waves (Pa: ~30 ms, P1: ~50 ms) presumed to reflect activity in thalamus and primary auditory

cortex followed by subsequent negative and positive deflections (N1: ~100 ms, P2: ~150 ms) from putative generator(s) in or near primary and secondary auditory areas within the Sylvian fissure (Näätänen and Picton, 1987; Picton et al., 1999; Scherg et al., 1989) (Fig. 2f).

Having traced the response output from multiple levels of the auditory system, we then aimed to identify the stage of brain processing that could predict listeners' behavioral classification of speech. Using raw voltage differences between responses, we constructed analogous neural dissimilarity matrices (Chang et al., 2010) (see Materials & methods) centered around each deflection of the cortical ERPs (Pa, P1, N1, P2) in order to quantify the degree to which various components of brain activity could differentiate each vowel stimulus (Figs. 3a–b, top panels). MDS allowed for the visualization of response dissimilarities in a common Euclidean space where distances between objects quantify the magnitude of neural response dissimilarity (bottom panels). Examination of MDS "maps" showed that responses to within-category speech sounds elicited similar patterns of neural activity as they appeared with close proximity in this geometric space; across-category tokens elicited divergent activity and were consequently situated farther apart. A clustering analysis applied to the MDS solution revealed that cortical responses generated across the speech continuum could be meaningfully segregated into two distinct groupings (i.e., /u/ and /a/ clusters) mimicking the two phonetic classes heard by listeners.

To evaluate the correspondence between brain and behavioral measures, we derived neurometric identification and discrimination functions using the distance between evoked responses (as represented in MDS space) elicited by each vowel stimuli (i.e., Figs. 3a–b, bottom panels) and the two cluster means (representing a neural "template" for each vowel category) (Chang et al., 2010). The resulting functions allowed us to estimate how well the pattern of neural activity evoked by each vowel fit into one of the two discrete phonetic categories. Complementary neurometric discrimination functions were derived from pairwise distances between observations in MDS space. We reasoned that categorical speech sounds are perceived as belonging to the same class and therefore, should elicit similar neural activity patterns whereas across category tokens are heard as dissimilar, and thus, should elicit more disparate neural responses in the MDS solution (Chang et al., 2010).

Dissimilarity analyses applied to brainstem ERPs showed that subcortical neurometric functions lacked any correspondence with the perceptual phonetic categories heard by listeners (Fig. 3a; all correlations $P > 0.05$; t-test on the sample mean: /u/: $t(14) = -0.68$, $P = 0.50$; /a/: $t(14) = -1.72$, $P = 0.10$; discrim: $t(14) = -0.39$, $P = 0.70$). Thus, while we observe that brainstem ERPs clearly provide a highly detailed and faithful transcription of the acoustic signal (e.g., Figs. 2b–e), they offer primarily a continuous, stimulus-related code for speech information. Earlier deflections of the cortical ERPs (Pa, P1) similarly failed to show evidence of categorical organization (Fig. 4). Intriguingly, neurometric functions began to parallel behavioral responses (i.e., a dichotomous pattern with sharp transitions) around the time window of N1, a deflection associated with the initial formation of object representations in auditory cortex (Alain and Arnott, 2000; Näätänen and Picton, 1987). However, this brain–behavior correspondence was neither reliable ($r = 0.83$, $P = 0.08$) nor featured adequate discrimination performance ($r = 0.06$, $P = 0.96$).

Both behavioral and neuroimaging studies reveal that vowel perception is critically dependent on the relationship between first (F1) and second (F2) formant frequencies, either in terms of their ratio or difference (Monahan and Idsardi, 2010; Obleser et al., 2003; Peterson and Barney, 1952). To further test the possibility that CP may emerge in brainstem responses based on the combined encoding of F1 and F2 we compared changes in these two metrics (i.e., F1/F2 magnitude ratio and difference) across the vowel continuum. As with F1, F2 salience was extracted from brainstem ERP response spectra (Fig. 2c) in the frequency range between 1000 and 1200 Hz, the expected F2 range from the input stimulus (see Materials & methods). Only 40% of participants showed discernible encoding at F2, likely attributable to the fact that this frequency approaches the upper limit of phase-locking in subcortical nuclei (Liu et al., 2006). Of those participants which showed an identifiable F2 response, Kruskal–Wallis non-parametric ANOVAs (used given the heteroscedastic variance and limited number of observations) revealed that neither brainstem F1/F2 magnitude [$\chi^2(4) = 1.47$, $P = 0.8325$] nor the F1–F2 amplitude relationship [$\chi^2(4) = 6.17$, $P = 0.1867$] varied across the vowel continuum. We therefore infer that the absence of CP at the level of the brainstem does not depend on the specific metric used in response analysis nor is revealed when considering both first and second formant frequency information (e.g., Obleser et al., 2003).

In stark contrast to brainstem and early cortical ERP waves, neurometric identification functions derived in the time window of the cortical P2 wave (~175 ms after sound onset) were strikingly similar to their psychometric counterparts and importantly, also featured a steep boundary between phoneme categories (Fig. 3d; $r > 0.95$ for each function, $P < 0.01$). These robust brain–behavior correlations were consistent across all participants (/u/: $t(14) = 12.09$, $P < 0.001$; /a/: $t(14) = 14.48$, $P < 0.001$, t-test against zero group mean correlation). A similarly high correspondence was observed between P2 neurometric and psychometric discrimination data ($r = 0.98$, $P < 0.01$; $t(14) = 7.60$, $P < 0.001$). As with participants' perceptual decision (Fig. 1b), neural discrimination was greater for stimuli across the categorical boundary (e.g., 2–4) than for within-category tokens (e.g., 1–3). Indeed, peak neural discrimination corresponded well with the steepest parts of the neural identification functions (~token 3), thus satisfying the definition of CP (Chang et al., 2010; Harnad, 1987). Moreover, robust correlations were observed even after an extensive bootstrapping of participants, thus confirming the strong brain–behavior connection observed in our sample (Fig. S1).

Subsequent analyses of single trial data further revealed that cortical response magnitudes were strongly dependent on how listeners perceived a given speech token. Indeed, despite having been elicited by an identical acoustic stimulus, we found that the ambiguous token proximal to the categorical boundary (vowel 3) evoked a unique P2 brain signature [$t(28) = 2.28$, $P = 0.03$, two-tailed t-test] depending on how listeners classified the trial (i.e., /u/ or /a/ response) (Fig. 5). N1 also appeared to vary based on listeners' perceptual responses but this effect was not significant [$t(28) = 1.01$, $P = 0.32$]. Modulations were also evident in response activity following P2; mean response amplitudes across the 250–275 ms time window were distinguishable based on listener's perceptual response [$t(28) = 2.11$, $P = 0.04$]. Consistent with recent near-field recordings (Chang et al., 2010; Mesgarani and Chang, 2012), these results demonstrate that relatively early and focal cortical activity (~150–200 ms, integrated over ~25 ms) contains adequate information which maps to higher-order perceptual attributes of human speech perception.

Taken together, our results delineate two fundamentally different coding strategies and consequently, a critical transformation between subcortical and cortical speech representations. Salient acoustic features of the speech signal are encoded with remarkable fidelity at the level of the brainstem (neuroacoustic code) (Bidelman and Krishnan, 2010; Musacchia et al., 2008). Despite such fidelity, brainstem ERPs do not contain evidence of clear categorical boundaries as found in higher cortical ERPs and psychophysical responses (phonetic code). The dissociation we observe between forms of brain activity in predicting human speech perception demonstrates that while subcortical response patterns primarily reflect the external acoustic environment, cortical representations give rise to the perceptual attributes relevant for speech listening (Mesgarani and Chang, 2012). This translation between divergent forms of the neural code represents a fundamental operation in speech perception (Liberman and Mattingly, 1989; Pisoni and Luce, 1987). We modeled a plausible neurocomputation that could realize the acoustic–phonetic transformation between low- (brainstem) and high-level (cortical/behavioral) speech representations using a

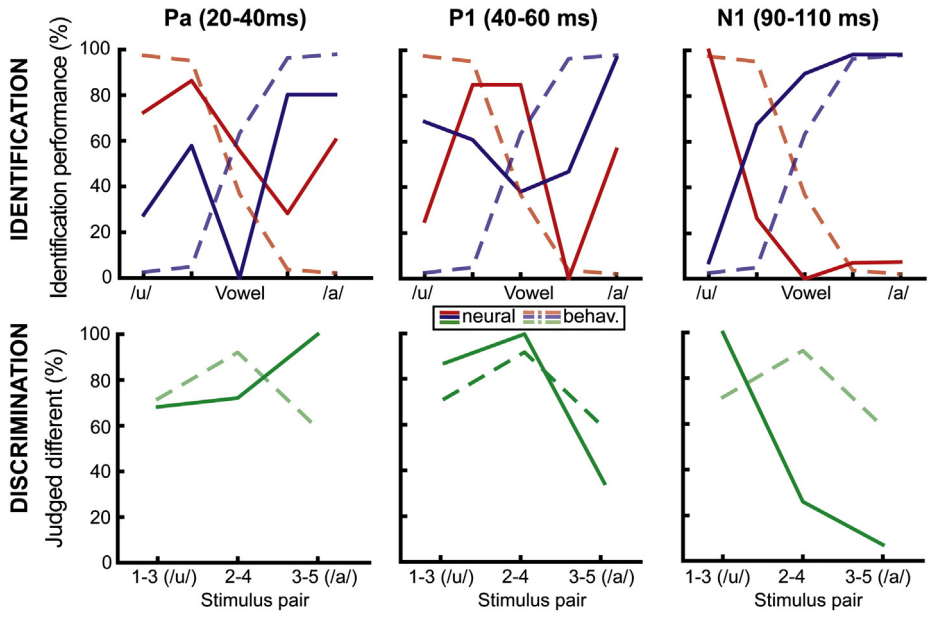

Fig. 4. Neurometric functions for early cortical response components. In contrast to the P2 component which largely reflects endogenous brain activity and shows close correspondence with perceptual speech categories (see Fig. 3d), early exogenous components (Pa, P1) of the auditory cortical field show no evidence for categorical speech processing. By N1 (~100 ms), neurometric identification loosely parallels human behavior but neural discrimination remains poor. The lack of any clear brain–behavior relationship prior to the generation of P2 (~150 ms) suggests that the neural correlates of speech likely emerge no earlier than late primary (lateral Heschl's gyrus) or secondary auditory cortex.

nonlinearity commonly employed in artificial neural networks as a Boolean classifier (Rosenblatt, 1958), in predicting monkeys' binary motor decisions based on underlying neural population activity (Purushothaman and Bradley, 2005), and the categorization of novel sounds from patterns of fMRI activation (Ley et al., 2012) (Fig. 6a). This operation takes continuous values at its input and produces a discrete, binary output. This neural transform was applied to formant frequency values estimated from brainstem responses (i.e., location of F1-energy) (see Fig. 2d) in order to predict the subsequent cortical P2 amplitudes and behavioral responses generated further upstream (Fig. 6b). As confirmed by the similarity in fitted functions, there was considerable correspondence between cortical P2 amplitudes predicted from brainstem responses and the actual P2 response magnitudes recorded at the scalp [rise-time parameter α (95% ± CI): α_{actual} = 31.1 ± 5.6, α_{pred} = 36.45.5 ± 5.5, α_{behav} = 27.7 ± 6.5]. The agreement in underlying functions implies that actual cortical ERP response properties and matching behavioral judgments to categorical speech stimuli can be viewed as a nonlinear transform of neural activity output from lower level structures (i.e., peripheral to the generator(s) of P2).

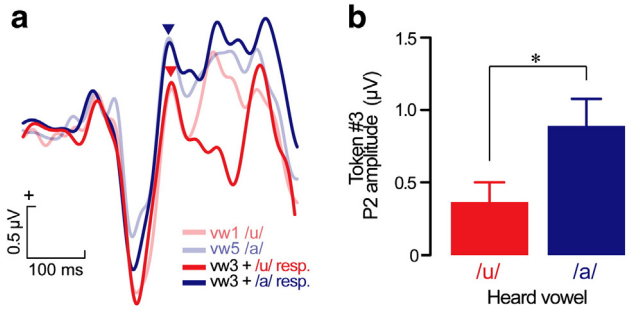

Fig. 5. Cortical P2 magnitudes reflect listeners' perceptual classification of vowel stimuli. (a) Comparison of evoked responses elicited by unambiguous tokens at the polar extremes of the stimulus continuum (light traces) versus ambiguous tokens at the categorical boundary (dark traces). Ambiguous trials are separated based on listeners' behavioral response (i.e., token 3 heard as /u/ vs. /a/). (b) Despite having been elicited by an identical stimulus, P2 amplitudes (panel a, arrows) show distinct magnitudes and are modulated based on the heard phonetic category of the stimulus. Error bars = s.e.m., $^*P < 0.05$.

Discussion

Speech perception requires that a listener maps continuous acoustic elements to an internalized discrete phonetic code in order to

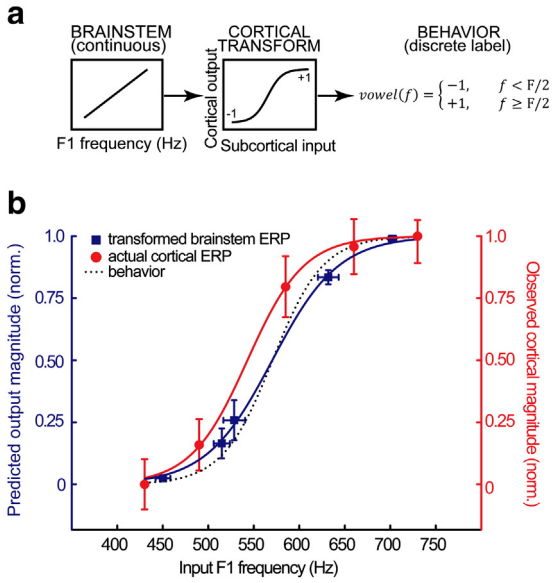

Fig. 6. Neurocomputation to account for the acoustic–phonetic transformation in neuronal speech representations. (a) Continuous acoustic parameters of speech encoded in brainstem ERPs are mapped into a discrete perceptual code upon arriving in the cortex via a nonlinear transform (Rosenblatt, 1958). (b) Formant frequencies encoded in brainstem responses were input to the classifier transform to predict cortical ERP amplitudes generated further upstream. Predicted P2 amplitudes are plotted against F1 frequencies estimated from brainstem responses (Fig. 2d); cortical and behavioral data are plotted against the stimulus F1. Transformed activity from brainstem (squares) closely mirrors both the observed cortical ERP magnitudes (circles) and behavioral responses (dotted line). This type of thresholding operation may account for how listeners convert continuous acoustic elements of speech into discrete, abstract phonetic units (Pisoni and Luce, 1987). Values are normalized to the maximum magnitude of the sample such that the resultant range is bounded between 0 and 1. Error bars = s.e.m. in either the frequency (abscissa) or amplitude (ordinate) dimension, respectively.

extract the intended message of a talker, a process exemplified in CP. By measuring scalp-recorded brain potentials to stimuli which yield categorical speech perception, we found that both cortical and subcortical processes are dynamically engaged in the encoding of important speech cues. Yet, in comparing neuroelectric activity output from these different levels of the auditory pathway, we demonstrate a dissociation in the fundamental *form of the speech code* across functional levels of the auditory brain. Our findings indicate that human brainstem responses, activity thought to receive primary contribution from the inferior colliculus (Smith et al., 1975; Sohmer et al., 1977), reflect an encoding scheme in which salient features of the physical signal (e.g., voice pitch and formant cues) are preserved in a neurophonic-like representation which closely resembles the acoustic source (Bidelman and Krishnan, 2010; Skoe and Kraus, 2010). As such, the brainstem code for speech represents a high-fidelity, primarily exogenous reflection of the input acoustic signal. We find that these representations, while providing a stable and detailed registration of signal information (Picton et al., 1978a), did not correspond with the perceptual judgments of speech as they capture only continuous (i.e., non-categorical) changes in the acoustic waveform rather than the discrete phonetic boundaries actually heard by listeners (Fig. 3d). It is worth examining these findings in light of emerging work investigating speech-evoked brainstem potentials.

On the possibility of subcortical correlates of CP

Several studies have reported correlations between response characteristics of passively recorded brainstem ERPs (e.g., magnitude, latency, spectral encoding) and behavioral pitch discrimination (Bidelman et al., 2011; Carcagno and Plack, 2011; Krishnan et al., 2010, 2012), noise-degraded listening (Parbery-Clark et al., 2009), and speech discrimination performance (Bidelman and Krishnan, 2010), where more robust and early response timing is typically associated with superior behavioral performance. These reports suggest that salient, *perceptually relevant* information contained in the acoustic signal is readily coded in subcortical auditory structures (e.g., Fig. 2). It is plausible that this information may then feed into and even influence the cortical neural generators dedicated to later perceptual processes. The output of the decision mechanism (i.e., behavioral performance) would then depend on the salience of signal representation transmitted in the neural code relative to biological noise in the system, akin to a signal detection analysis. This bottom–up, representational explanation has been used to account for the correspondence between the robustness of low-level feature extraction at the level of the brainstem and superior behavioral performance in auditory discrimination tasks (Bidelman and Krishnan, 2010; Bidelman et al., 2011). Alternatively, top–down (i.e., cortically driven) modulation via the corticofugal efferent pathway has also been invoked to explain changes in subcortical response properties during behaviorally relevant listening tasks (Suga et al., 2000; Tzounopoulos and Kraus, 2009).

Regardless of the specific mechanism, care must be exercised when drawing causation from the aforementioned correlational studies. Indeed, without exception, these reports examined auditory perception by manipulating the amount of acoustic information presented at the sensory input. Thus, modulations in the evoked response and human behavior both covaried with the acoustic properties of the signal (see Bidelman and Krishnan, 2009 for a paradigm controlling latent acoustic factors). As noted previously (Carlyon, 2004, p. 470), this aliasing restricts the interpretation that the observed ERP reflects a true correlate of auditory perceptual judgments. In the present study, CP stimuli allowed us to dissociate neural activity reflecting changes in the acoustic stimulus (exogenous response) from that which indexed a true internalized speech percept (endogenous response). The dissociation between brainstem responses and speech judgments we observe by employing this novel design thus leads us to qualify recent suggestions (e.g., Kraus and Chandrasekaran, 2010; Krishnan et al., 2010; Skoe and Kraus, 2010) that brainstem activity provides a *direct* correlate of auditory perception. The need for caution is also evident in recent studies which, through careful choice of stimuli and listening paradigm, have failed to observe a direct correspondence between brainstem representations of harmonic tones and human pitch perception (Gockel et al., 2011; but see Greenberg et al., 1987). In the context of categorical perception examined here, it is likely that the brainstem is far too low in the hierarchy of auditory processing to yield the abstract phonetic representations necessary for CP despite the presence of supporting acoustic cues contained in the response. Collectively, our findings speak to the importance to control stimulus-related factors and for using a more integrative, *systems-level* approach. Future studies investigating the neurophysiological correlates of complex percepts would benefit from examining multiple brain indices as pursued here.

The chronometry of categorical speech perception

Examining multiple levels of auditory processing allowed us to characterize the chronometry of speech listening within the brain by tracing the time-course of underlying neural activity from sound onset through perceptual recognition. To provide a correlate that maps to perception, we find that the continuous features encoded in lower-level structures must undergo a critical transformation as they propagate along the neuroaxis. This translation is the hallmark of CP and may be realized, conceivably, through a nonlinear warping of the acoustic, sensory output from brainstem and/or early cortical structures. This mapping was modeled in the present study with a relatively parsimonious sigmoidal-like operation (e.g., Fig. 6), a function often used to model categorical data from neurophysiological activity (Ley et al., 2012; Purushothaman and Bradley, 2005; Rosenblatt, 1958). We find that such an operation accounts well for the current data and thus provides a plausible biological mechanism to achieve categorical coding (at least for the single stimulus dimension examined here, i.e., changes in F1). However, this particular non-linearity is not the only candidate for producing a categorical output from a continuous input. Similar binary responses could be achieved by even simpler thresholding neurons which respond only to a select range of input along a particular dimension of speech. While a sigmoidal computation has been popular in describing the neural classification of binary decisions as examined here, it is less clear how such a mechanism might scale up. To account for the classification of multiple (>2) categories, a binary framework would have to span the extent of a particular feature space and likely operate multidimensionally, integrating different binary features of the input stimulus prior to classification.

Interestingly, the N1 deflection of the auditory cortical field approached, but failed to fully support a categorical code (Fig. 4). This is consistent with the notion that earlier activity may reflect the initial formation of object representations in early auditory cortex (Alain and Arnott, 2000; Näätänen and Picton, 1987). Upon arrival later sensory cortices (~200 ms), the speech signal is discretized, no longer reflecting mere acoustic signal properties, but rather, the learned phonemic categories (i.e., speech templates) internalized by listeners (Iverson and Kuhl, 1995). Indeed, we observe that late cortical components (P2 and later) are modulated depending on a listener's trial-by-trial perception of an identical, ambiguous speech sound (Fig. 5). This result is consistent with previous electrophysiological studies which demonstrated similar perceptual modulation in later cortical activity dependent on listeners' attention (Hillyard et al., 1973) or specific hearing of an illusory stimulus (Riecke et al., 2009) despite invariance in stimulus acoustics. Our results also converge with past work demonstrating that cortical ERPs are more robust when an identical eliciting stimulus is analyzed for phonetic versus acoustic cues, i.e., linguistic vs. nonlinguistic mode of listening (Wood, 1975; Wood et al., 1971). Taken together, these results support the notion that later cortical activity around 150–200 ms after sound onset reflects more than the external acoustic input, but instead, reflects linguistic processing that gives rise to the

abstract perceptual–phonetic attributes relevant for a listeners' intended behavior (Eulitz et al., 1995; Mesgarani and Chang, 2012).

Given that categorical percepts do not faithfully map from exact sensory input, they provide useful divisions of information not contained in the external world (Miller et al., 2003). Presumably, this type of "downsampling" mechanism would promote speech comprehension by generating perceptual constancy in the face of individual variation along multiple acoustic dimensions, e.g., talker variability in tempo, pitch, or timbre (Prather et al., 2009). More broadly, the type of reductionistic mechanism we observe here for speech processing may illustrate a generic organizational principle of neural systems where continuous, yet redundant information relayed from lower levels is pruned so as to allow for easier readout of signal identity in higher brain areas formulating a complex percept (Chechik et al., 2006). Indeed, comparing multiunit responses in brainstem and early sensory cortices in an animal model, Perez et al. (2013) have recently shown that the neural code for speech becomes more distinct as the information ascends the central auditory pathway.

Neurobiological networks engaged during speech and other sensory perception presumably involve a coordinated sequence of computations applied to neural representations at multiple stages of processing (Carandini, 2012; Hickok and Poeppel, 2004; Phillips, 2001). Understanding these neurocomputations requires comparing the output of the participating neural elements across multiple brain regions and time scales (Carandini, 2012). Taken together, the clear brain–behavior relationship we observe between late cortical (but not subcortical or early cortical) activity and human behavior leads us to infer that direct neurophysiological correlates of categorical speech percepts emerge no earlier than late primary or secondary auditory cortex. These results are broadly consistent with recent intracranial recordings obtained in epileptic patients which suggest that the neural correlates of CP arise within the superior temporal gyrus (Chang et al., 2010; Mesgarani and Chang, 2012). Using noninvasive neuroimaging methodology, our results confirm and extend these findings to neurologically intact young adults and reveal that brain correlates of categorical speech percepts emerge within the first few hundred milliseconds after sound onset.

Our findings are also convergent with results obtained from multiunit activity recorded in response to categorical stimuli in awake primates (Steinschneider et al., 2003) and humans (Chang et al., 2010; Steinschneider et al., 1999). These near-field recordings demonstrate that the acoustic cues necessary to establish a categorical percept are represented spatiotemporally within ~100–150 ms following sound onset (Chang et al., 2010). Yet, a distribution of these auditory features across the tonotopic areas of primary auditory cortex (A1) suggests that the integration of these phonetic cues, and hence the emergence of a correlate for CP, must occur in areas beyond A1, e.g., in secondary auditory cortical fields (Steinschneider et al., 2003). Indeed, recordings directly on the posterior superior temporal gyrus have shown brain responses that mimic CP to speech sounds, consistent with the notion that lateral belt areas of auditory cortex participate in the pattern recognition (cf. classification) of sound objects (Rauschecker, 1998). ERPs undoubtedly reflect the activation of brain tissue comprising a population of neural elements making it impossible to fully reconcile the neural activity we observe at the scalp with previous intracranial recordings. Nevertheless, the categorical organization in the far-field P2 observed here along with a comparable organization in the near-field activity along the lateral belt areas suggests that categorical speech percepts emerge at or near secondary auditory cortex, areas proximal to the presumed neural generator of the P2 wave (Picton et al., 1999; Scherg et al., 1989).

With regard to the origin of this phonetic organization, our data cannot speak to whether the high correspondence between late cortical activity and human perception is mediated by bottom–up or top–down influences. Intracranial recordings from human superior temporal gyrus show that early response components akin to N1 display categorical organization even during passive listening, supporting a bottom–up coding scheme for categorical speech (Chang et al., 2010). Alternatively, we cannot rule out the possibility that the task demands and goal directed action during speech identification may act to modulate early sensory components. Indeed, distal non-auditory regions including inferior frontal cortex play a role in shaping perceptual sensitivity during CP tasks (Myers and Swan, 2012). Such top–down modulation to auditory areas could result in the categorical organization of P2 activity we observe at the scalp. While either scenario is plausible, it is clear that CP materializes within the timeframe of the N1–P2 complex (cf. Figs. 3d and 4) and is firmly established within 200 ms after sound onset.

Potential limitations and directions for future work

Though our integrative systems-style approach is sufficient to reveal categorical neural organization as well as assess speech-evoked activity across time and functional levels of the auditory brain (brainstem, cortex), it is not without limitation. Indeed, only a single frontal electrode site was employed in the present study. Unfortunately, typical ERP systems limit the maximal permissible throughput of their digital-to-analog conversion. As a result, the high sample rate (>10 kHz) necessary to record brainstem potentials prohibits the use of a high number of recordable channels and hence, our ability to address questions requiring a distribution of electrodes around the scalp (e.g., topography, lateralization, source orientation). Such questions are largely irrelevant to brainstem responses anyway given the depth and concentration of their midbrain neural generator. As such, investigators examining auditory brainstem responses (Galbraith and Kane, 1993; Krishnan et al., 2012; Musacchia et al., 2008) have often employed only a single channel of recording, which is what we opted for in the present study. Nevertheless, dense, multichannel montages may be useful in future work to illuminate additional mechanisms underlying the *cortical* organization of categorical speech. Indeed, systematic changes in neural topography have been observed for speech sounds across the acoustic vowel space (Poeppel et al., 1996, 1997; Scharinger et al., 2011; Shestakova et al., 2004). It is conceivable that this spatial distribution may also vary dependent on the phonetic category of the stimulus (Chang et al., 2010).

Future work may offer further insight into the hierarchical organization of speech by employing different stimuli or ERP recording techniques. In the present study, categorical representations of speech were examined using a single continuum of steady-state vowel sounds varying only in F1. CP has been traditionally studied using stop consonants whereby changes in VOT elicit a categorical shift in the perceived phoneme (e.g., /da/ vs. /ba/). A natural question that emerges then is how the present results, particularly with regard to our brainstem findings, might generalize to other speech sounds. As in the present study, VOT stimuli elicit categorical responses at a cortical level of processing (Chang et al., 2010; Pisoni, 1973; Sharma and Dorman, 1999). Brainstem ERPs to stop-consonants have been recorded in isolation (e.g., Hornickel et al., 2009) but no study has examined these responses to a VOT continuum. While subcortical correlates of CP seem improbable regardless of the specific speech sounds, single unit data does suggest that the auditory system employs different coding strategies within various functional levels of the brain for transient and sustained speech cues (e.g., consonants vs. vowel sounds) (Perez et al., 2013). Behavioral data also suggests that transient VOT stimuli are perceptually more categorical than vowel color alone (Pisoni, 1973). As such, work should investigate the possibility of categorical organization at the level of the brainstem using other, more salient CP stimuli.

Different recording techniques may also offer a more sensitive view of formant-related information at the level of the brainstem. Following previous investigators examining the brainstem response to speech (e.g., Parbery-Clark et al., 2009; Skoe and Kraus, 2010), alternating stimulus polarity was employed in the present study to

safeguard against potential stimulus artifact and ensure that the recorded responses were of true neural origin (for detailed discussions, see Aiken and Picton, 2008; Campbell et al., 2012; Skoe and Kraus, 2010).[2] However, one consequence of this presentation mode is that the brainstem response will tend to accentuate lower-frequency stimulus content (e.g., stimulus envelope) and minimize higher spectral details critical to speech perception (e.g., formant energy). While the current recording parameters clearly allow quantification of brainstem activation to formant attributes (e.g., Fig. 2, current study; Aiken and Picton, 2008; Parbery-Clark et al., 2009), we cannot rule out the possibility that alternate recording techniques, which preserve stimulus fine structure [e.g., fixed polarity presentation (Bidelman and Krishnan, 2010) or subtracting positive and negative sweeps (Krishnan, 2002)] might provide a more detailed view of the speech cues in brainstem ERPs. Future work should explore categorical organization in subcortical responses not afforded by the current recoding methodology.

Conclusions

The complexity of studying the neural correlates of human speech perception often lies in the inherent difficulty of finding a paradigm for which modulations in the corresponding ERP reflect true underlying percepts rather than simply stimulus-related activity (i.e., endogenous vs. exogenous brain response). With few exceptions (Chang et al., 2010), neuroimaging studies of CP have employed mismatch paradigms in which the neural correlates of categorical percepts are inferred by comparing brain activity elicited between two acoustic conditions (Maiste et al., 1995; Phillips et al., 2000; Sharma and Dorman, 1999). Unfortunately, under these paradigms, it is difficult to ascertain whether such derived (second-order) ERP activity truly reflects brain signatures of higher-order perceptual attributes or merely the degree of variation in stimulus acoustics (for a well-controlled paradigm to circumvent these issues, see Phillips et al., 2000). Our innovative approach in examining multiple tiers of brain activity in conjunction with CP helps not only elucidate the underlying form and dynamics of hierarchical speech representations, but also tease apart stimulus- from behavior-related neuroelectric activity generated during speech listening. Additionally, examining how multiple tiers of the human brain contribute to perception, provides a unique, systems level approach to study the connection between brain and behavior. Beyond the domain of speech, this integrative technique would be equally suitable for tracing brain representations and emergent percepts in other important aspects of human cognition including color vision and face perception.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2013.04.093.

[2] Unlike the majority of ERPs which are transient in nature (e.g., the cortical P1–N1–P2 response) sustained brainstem responses overlap in both time and frequency with the stimulus presentation. As such, care must be taken to ensure that activity recorded at the electrodes is truly neural in origin and not simply electromagnetic stimulus bleed emitted from the earphone. With alternating polarity, every other stimulus presentation is inverted 180°. The acoustic stimulus and hence any electromagnetic artifact radiated by the transducer flip with polarity; the neural response does not. Summing an even number of positive and negative phase trials therefore preserves the true neural response while effectively canceling any artifact. Alternate strategies to reduce stimulus artifact in sustained neurophysiological recordings are to encase the earphone with electromagnetic shielding (e.g., mu-metal) and/or create a sufficient physical distance between the transducer and recording electrodes. The latter can be accomplished by placing the earphones outside the testing room and using extended acoustic tubing to deliver stimuli to a participant seated inside. Detailed discussions of combating artifact and the consequence of stimulus polarity on brainstem responses can be found elsewhere (e.g., Aiken and Picton, 2008; Campbell et al., 2012; Skoe and Kraus, 2010).

Acknowledgments

We thank Yu He for her assistance in setting up the experimental protocol, Michael Weiss for his help with data collection, and Dr. Malcolm Binns for assistance with statistical analyses. We also thank Drs. Bernhard Ross and Terry Picton for comments on earlier drafts of the manuscript. Portions of this research were supported by Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Council of Canada (NSERC) (C.A.) and the GRAMMY Foundation® (G.M.B.).

Conflict of interest

The authors declare no competing financial interests.

References

Agung, K., Purdy, S.C., McMahon, C.M., Newall, P., 2006. The use of cortical auditory evoked potentials to evaluate neural encoding of speech sounds in adults. J. Am. Acad. Audiol. 17, 559–572.
Aiken, S.J., Picton, T.W., 2006. Envelope following responses to natural vowels. Audiol. Neurootol. 11, 213–232.
Aiken, S.J., Picton, T.W., 2008. Envelope and spectral frequency-following responses to vowel sounds. Hear. Res. 245, 35–47.
Alain, C., Arnott, S.R., 2000. Selectively attending to auditory objects. Front. Biosci. 202–212.
Beale, J.M., Keil, F.C., 1995. Categorical effects in the perception of faces. Cognition 57, 217–239.
Bidelman, G.M., Krishnan, A., 2009. Neural correlates of consonance, dissonance, and the hierarchy of musical pitch in the human brainstem. J. Neurosci. 29, 13165–13171.
Bidelman, G.M., Krishnan, A., 2010. Effects of reverberation on brainstem representation of speech in musicians and non-musicians. Brain Res. 1355, 112–125.
Bidelman, G.M., Gandour, J.T., Krishnan, A., 2011. Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. Brain Cogn. 77, 1–10.
Borg, I., Groenen, P.J.F., 2005. Modern Multidimensional Scaling: Theory and Applications, 2nd ed. Springer, New York.
Campbell, T., Kerlin, J.R., Bishop, C.W., Miller, L.M., 2012. Methods to eliminate stimulus transduction artifact from insert earphones during electroencephalography. Ear Hear. 33, 144–150.
Carandini, M., 2012. From circuits to behavior: a bridge too far? Nat. Neurosci. 15, 507–509.
Carcagno, S., Plack, C.J., 2011. Subcortical plasticity following perceptual learning in a pitch discrimination task. J. Assoc. Res. Otolaryngol. 12, 89–100.
Carlyon, R.P., 2004. How the brain separates sounds. Trends Cogn. Sci. 8, 465–471.
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, 1428–1432.
Chechik, G., Anderson, M.J., Bar-Yosef, O., Young, E.D., Tishby, N., Nelken, I., 2006. Reduction of information redundancy in the ascending auditory pathway. Neuron 51, 359–368.
Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. J. Neurosci. Methods 134, 9–21.
Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.
Eimas, P.D., Siqueland, E.R., Jusczyk, P., Vigorito, J., 1971. Speech perception in infants. Science 171, 303–306.
Eulitz, C., Diesch, E., Pantev, C., Hampson, S., Elbert, T., 1995. Magnetic and electric brain activity evoked by the processing of tone and vowel stimuli. J. Neurosci. 15, 2748–2755.
Franklin, A., Drivonikou, G.V., Clifford, A., Kay, P., Regier, T., Davies, I.R., 2008. Lateralization of categorical perception of color changes with color term acquisition. Proc. Natl. Acad. Sci. U. S. A. 105, 18221–18225.
Galbraith, G.C., Kane, J.M., 1993. Brainstem frequency-following responses and cortical event-related potentials during attention. Percept. Mot. Skills 76, 1231–1241.
Galbraith, G., Olfman, D.M., Huffman, T.M., 2003. Selective attention affects human brain stem frequency-following response. Neuroreport 14, 735–738.
Gockel, H.E., Carlyon, R.P., Mehta, A., Plack, C.J., 2011. The frequency following response (FFR) may reflect pitch-bearing information but is not a direct representation of pitch. J. Assoc. Res. Otolaryngol. 12, 767–782.
Greenberg, S., Marsh, J.T., Brown, W.S., Smith, J.C., 1987. Neural temporal coding of low pitch. I. Human frequency-following responses to complex tones. Hear. Res. 25, 91–114.
Harnad, S.R., 1987. Categorical Perception: The Groundwork of Cognition. Cambridge University Press, New York.
Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.
Hillyard, S.A., Picton, T.W., 1979. Event-related brain potentials and selective information processing in man. In: Desmedt, J.E. (Ed.), Progress in Clinical Neurophysiology. Karger, Basel, pp. 1–52.
Hillyard, S.A., Hink, R.F., Schwent, V.L., Picton, T.W., 1973. Electrical signs of selective attention in the human brain. Science 182, 177–180.
Hornickel, J., Skoe, E., Nicol, T., Zecker, S., Kraus, N., 2009. Subcortical differentiation of stop consonants relates to reading and speech-in-noise perception. Proc. Natl. Acad. Sci. 106, 13022–13027.

Iverson, P., Kuhl, P.K., 1995. Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. J. Acoust. Soc. Am. 97, 553–562.
Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 87, 820–857.
Klein, M.E., Zatorre, R.J., 2011. A role for the right superior temporal sulcus in categorical perception of musical chords. Neuropsychologia 49, 878–887.
Kraus, N., Chandrasekaran, B., 2010. Music training for the development of auditory skills. Nat. Rev. Neurosci. 11, 599–605.
Krishnan, A., 2002. Human frequency-following responses: representation of steady-state synthetic vowels. Hear. Res. 166, 192–201.
Krishnan, A., Bidelman, G.M., Gandour, J.T., 2010. Neural representation of pitch salience in the human brainstem revealed by psychophysical and electrophysiological indices. Hear. Res. 268, 60–66.
Krishnan, A., Bidelman, G.M., Smalt, C.J., Ananthakrishnan, S., Gandour, J.T., 2012. Relationship between brainstem, cortical and behavioral measures relevant to pitch salience in humans. Neuropsychologia 50, 2849–2859.
Krizman, J., Marian, V., Shook, A., Skoe, E., Kraus, N., 2012. Subcortical encoding of sound is enhanced in bilinguals and relates to executive function advantages. Proc. Natl. Acad. Sci. U. S. A. 109, 7877–7881.
Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N., Lindblom, B., 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. Science 255, 606–608.
Ley, A., Vroomen, J., Hausfeld, L., Valente, G., De Weerd, P., Formisano, E., 2012. Learning of new sound categories shapes neural response patterns in human auditory cortex. J. Neurosci. 32, 13273–13280.
Liberman, A.M., Mattingly, I.G., 1989. A specialization for speech perception. Science 243, 489–494.
Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M., 1967. Perception of the speech code. Psychol. Rev. 74, 431–461.
Liu, L.F., Palmer, A.R., Wallace, M.N., 2006. Phase-locked responses to pure tones in the inferior colliculus. J. Neurophysiol. 95, 1926–1935.
Maiste, A.C., Wiens, A.S., Hunt, M.J., Scherg, M., Picton, T.W., 1995. Event-related potentials and the categorical perception of speech sounds. Ear Hear. 16, 68–90.
Markel, J.E., Gray, A.H., 1976. Linear Prediction of Speech. Springer-Verlag, New York.
May, J., 1981. Acoustic factors that may contribute to categorical perception. Lang. Speech 24, 273–284.
Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236.
Miller, E.K., Nieder, A., Freedman, D.J., Wallis, J.D., 2003. Neural correlates of categories and concepts. Curr. Opin. Neurobiol. 13, 198–203.
Mody, M., Studdert-Kennedy, M., Brady, S., 1997. Speech perception deficits in poor readers: auditory processing or phonological coding? J. Exp. Child Psychol. 64, 199–231.
Monahan, P.J., Idsardi, W.J., 2010. Auditory sensitivity to formant ratios: toward an account of vowel normalization. Lang. Cognit. Process. 25, 808–839.
Musacchia, G., Strait, D., Kraus, N., 2008. Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians. Hear. Res. 241, 34–42.
Myers, E.B., Swan, K., 2012. Effects of category learning on neural sensitivity to non-native phonetic categories. J. Cogn. Neurosci. 24, 1695–1708.
Näätänen, R., Picton, T., 1987. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology 24, 375–425.
Obleser, J., Elbert, T., Lahiri, A., Eulitz, C., 2003. Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. Brain Res. Cogn. Brain Res. 15, 207–213.
Okamoto, H., Stracke, H., Bermudez, P., Pantev, C., 2011. Sound processing hierarchy within human auditory cortex. J. Cogn. Neurosci. 23, 1855–1863.
Parbery-Clark, A., Skoe, E., Kraus, N., 2009. Musical experience limits the degradative effects of background noise on the neural processing of sound. J. Neurosci. 29, 14100–14107.
Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F., 2012. Reconstructing speech from human auditory cortex. PLoS Biol. 10, e1001251.
Perez, C.A., Engineer, C.T., Jakkamsetti, V., Carraway, R.S., Perry, M.S., Kilgard, M.P., 2013. Different timescales for the neural coding of consonant and vowel sounds. Cereb. Cortex 23, 670–683.
Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of vowels. J. Acoust. Soc. Am. 24, 175–184.
Phillips, C., 2001. Levels of representation in the electrophysiology of speech perception. Cogn. Sci. 25, 711–731.
Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M., Roberts, T., 2000. Auditory cortex accesses phonological categories: an MEG mismatch study. J. Cogn. Neurosci. 12, 1038–1055.
Picton, T.W., Hillyard, S.A., 1974. Human auditory evoked potentials. II. Effects of attention. Electroencephalogr. Clin. Neurophysiol. 36, 191–199.
Picton, T.W., Hillyard, S.A., Galambos, R., Schiff, M., 1971. Human auditory attention: a central or peripheral process? Science 173, 351–353.
Picton, T.W., Woods, D.L., Baribaeu-Braun, J., Healy, T.M.G., 1977. Evoked potential audiometry. J. Otolaryngol. 6, 90–119.
Picton, T.W., Campbell, K.B., Baribeau-Braun, J., Proulx, G.B., 1978a. The neurophysiology of human attention — a tutorial review. In: Requin, J. (Ed.), Attention and Performance VII. Lawrence Erlbaum, New Jersey, pp. 429–467.
Picton, T.W., Woods, D.L., Proulx, G.B., 1978b. Human auditory sustained potentials. II. Stimulus relationships. Electroencephalogr. Clin. Neurophysiol. 45, 198–210.
Picton, T.W., Alain, C., Woods, D.L., John, M.S., Scherg, M., Valdes-Sosa, P., Bosch-Bayard, J., Trujillo, N.J., 1999. Intracerebral sources of human auditory-evoked potentials. Audiol. Neurootol. 4, 64–79.
Pisoni, D.B., 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253–260.
Pisoni, D.B., Luce, P.A., 1987. Acoustic–phonetic representations in word recognition. Cognition 25, 21–52.
Poeppel, D., Yellin, E., Phillips, C., Roberts, T.P., Rowley, H.A., Wexler, K., Marantz, A., 1996. Task-induced asymmetry of the auditory evoked M100 neuromagnetic field elicited by speech sounds. Brain Res. Cogn. Brain Res. 4, 231–242.
Poeppel, D., Phillips, C., Yellin, E., Rowley, H.A., Roberts, T.P., Marantz, A., 1997. Processing of vowels in supratemporal auditory cortex. Neurosci. Lett. 221, 145–148.
Pollack, I., Pisoni, D.B., 1971. On the comparison between identification and discrimination tests in speech perception. Psychon. Sci. 24, 299–300.
Prather, J.F., Nowicki, S., Anderson, R.C., Peters, S., Mooney, R., 2009. Neural correlates of categorical perception in learned vocal communication. Nat. Neurosci. 12, 221–228.
Purushothaman, G., Bradley, D.C., 2005. Neural population code for fine perceptual decisions in area MT. Nat. Neurosci. 8, 99–106.
Rauschecker, J.P., 1998. Parallel processing in the auditory cortex of primates. Audiol. Neurootol. 3, 86–103.
Riecke, L., Esposito, F., Bonte, M., Formisano, E., 2009. Hearing illusory sounds in noise: the timing of sensory-perceptual transformations in auditory cortex. Neuron 64, 550–561.
Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 386–408.
Scharinger, M., Idsardi, W.J., Poe, S., 2011. A comprehensive three-dimensional cortical map of vowel space. J. Cogn. Neurosci. 23, 3972–3982.
Scherg, M., Vajsar, J., Picton, T.W., 1989. A source analysis of the late human auditory evoked potentials. J. Cogn. Neurosci. 1, 336–355.
Sharma, A., Dorman, M.F., 1999. Cortical auditory evoked potential correlates of categorical perception of voice-onset time. J. Acoust. Soc. Am. 106, 1078–1083.
Shepard, R.N., 1980. Multidimensional scaling, tree-fitting and clustering. Science 210, 390–398.
Shestakova, A., Brattico, E., Soloviev, A., Klucharev, V., Huotilainen, M., 2004. Orderly cortical representation of vowel categories presented by multiple exemplars. Brain Res. Cogn. Brain Res. 21, 342–350.
Skoe, E., Kraus, N., 2010. Auditory brain stem response to complex sounds: a tutorial. Ear Hear. 31, 302–324.
Smith, J.C., Marsh, J.T., Brown, W.S., 1975. Far-field recorded frequency-following responses: evidence for the locus of brainstem sources. Electroencephalogr. Clin. Neurophysiol. 39, 465–472.
Sohmer, H., Pratt, H., Kinarti, R., 1977. Sources of frequency-following responses (FFR) in man. Electroencephalogr. Clin. Neurophysiol. 42, 656–664.
Song, J., Skoe, E., Banai, K., Kraus, N., 2010. Perception of speech in noise: neural correlates. J. Cogn. Neurosci. 23, 2268–2279.
Steinschneider, M., Volkov, I.O., Noh, M.D., Garell, P.C., Howard 3rd, 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.
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.
Suga, N., Gao, E., Zhang, Y., Ma, X., Olsen, J.F., 2000. The corticofugal system for hearing: recent progress. Proc. Natl. Acad. Sci. U. S. A. 97, 11807–11814.
Tzounopoulos, T., Kraus, N., 2009. Learning to encode timing: mechanisms of plasticity in the auditory brainstem. Neuron 62, 463–469.
Vihman, M., 1996. Phonological Development: The Origins of Language in the Child. Wiley-Blackwell, Cambridge.
Werker, J.F., Tees, R.C., 1987. Speech perception in severely disabled and average reading children. Can. J. Psychol. 41, 48–61.
Wood, C.C., 1975. Auditory and phonetic levels of processing in speech perception: neurophysiological and information-processing analyses. J. Exp. Psychol. Hum. Percept. Perform. 104, 3–20.
Wood, C.C., Goff, W.R., Day, R.S., 1971. Auditory evoked potentials during speech perception. Science 173, 1248–1251.
Woods, D.L., Hillyard, S.A., 1978. Attention at the cocktail party: brainstem evoked responses reveal no peripheral gating. In: Otto, D.A. (Ed.), Multidisciplinary Perspectives in Event-related Brain Potential Research. U.S. Government Printing Office, pp. 230–233 (EPA 600/9-77-043).
Zhang, L., Xi, J., Wu, H., Shu, H., Li, P., 2012. Electrophysiological evidence of categorical perception of Chinese lexical tones in attentive condition. Neuroreport 23, 35–39.

SUPPLEMENTARY FIGURES

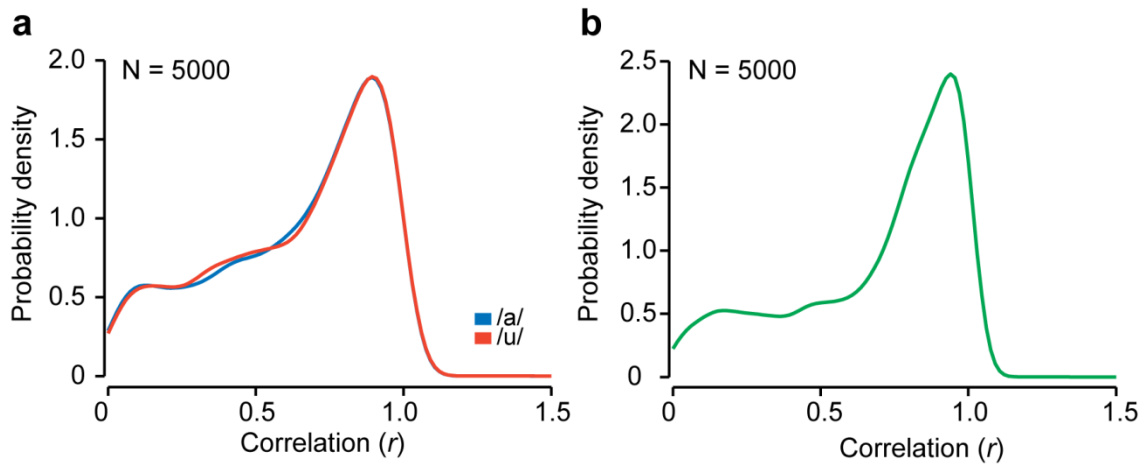

Fig. S1. Bootstrap distributions for correlations between neurometric (derived in the time window of the cortical P2 component) and psychometric functions generated from N=5000 resampled iterations of the data. (**a**) neurometric identification and (**b**) discrimination functions. The concentration of the probability densities around values of $r = 0.9$ indicates that the observed robustness in brain-behavior correlations of the sample (see Fig. 3d) are unlikely to have occurred by chance.