Title: Neural Correlates of Individual Differences in Speech Categorization: Evidence from Subcortical, Cortical, and Behavioral Measures

Authors: Jinghua Ou*, & Alan C. L. Yu*

Department of Linguistics, University of Chicago, IL, USA

*Correspondence to:

Jinghua Ou & Alan C. L. Yu
Department of Linguistics, University of Chicago
1115 E. 58th Street, Chicago, IL, 60637
Email: jou@uchicago.edu; aclyu@uchicago.edu

Title: Neural Correlates of Individual Differences in Speech Categorization: Evidence from Subcortical, Cortical, and Behavioral Measures

Abstract

Categorization is a fundamental cognitive ability to group different objects as the same. This ability is particularly indispensable for human speech perception, yet individual differences in speech categorization are nonetheless ubiquitous. The present study investigates the neurophysiological mechanisms underlying the variability in categorization of voice-onset time (VOT). Subcortical and cortical speech-evoked responses are recorded to investigate speech representations at two functional levels of auditory processing. Individual differences in psychometric functions correlate positively with how faithfully subcortical responses encode VOT differences. Moreover, individuals also differ in how strongly the subcortical and cortical representations correlate with each other. Listeners with gradient categorization show higher correspondences between the two representations, indicating that acoustic information is relayed faithfully from the subcortical to the cortical level; listeners with discrete categorization exhibit decreased similarity between the two representations, suggesting that the subcortical acoustic encoding is transformed at the cortical level to reflect phonetic category information.

Key words: speech perception; individual differences; categorization gradient; subcortical response; cortical response.

Introduction

Categorization is an important and fundamental human cognitive ability which allows a perceiver to treat objects with different physical features as functionally equivalent classes (Goldstone & Hendrickson, 2010). A classic demonstration of this ability is the phenomenon of categorical perception of speech sounds where continuously varying acoustic stimuli are labeled by listeners as belonging to discrete sound categories with abrupt boundaries between them (Lieberman et al., 1967; Pisoni, 1973). For instance, voice-onset time (VOT), defined as the interval between the release of the stop closure and the onset of voicing in the following vowel, is a critical cue used by listeners to distinguish voiced and voiceless stop consonants in English (Lisker, 1986). While categorical information facilitates sound-to-meaning mappings, fine-grained acoustic detail can also be harnessed by listeners, allowing them to flexibly integrate multiple cues to phonetic identity (Clayards et al., 2008; Kleinschmidt & Jaeger, 2015; McMurray et al., 2002; Myers et al., 2009). Moreover, recent studies have documented extensive individual variations in how gradiently listeners categorize speech sounds among typically-developed individuals (Kapnoula et al., 2017; Kong & Edwards, 2016; Schertz et al., 2015). That is, individuals differ in how likely they consider competing response options as the category label across a sound continuum, with an increased likelihood of assigning labels to competing categories at the ambiguous region of the speech continuum resulting in a more gradient response pattern. However, the neurophysiological mechanisms underlying such individual differences remain unclear. The present study aims to examine how individual differences in neural encoding of a VOT continuum at the subcortical and cortical levels relate to individual differences in stop voicing categorization.

Neurophysiologically, categorical analysis has been suggested to be implemented at the cortical level, where the superior temporal gyrus (STG) integrates spectro-temporal features to create invariant higher order linguistic representation (Chang et al., 2010; Mesgarani et al., 2014). The acoustic-to-phonetic mapping may be subserved by a transformation in neural representation between the subcortical and cortical structures (Bidelman et al., 2013). Using a vowel continuum (from /u/ to /a/) with varying steps of first formant frequency to elicit auditory evoked responses at both subcortical and cortical levels, Bidelman et al. (2013) observed a close correspondence between the subcortical responses and the continuous acoustic changes in the vowel continuum, but found the cortical responses to reflect perceptual categories and predict listeners' perceptual boundaries between the vowel categories. These studies have revealed a general neural mechanism for speech categorization based on spectral cues. It is unclear whether a similar neurophysiological mechanism i.e. subcortical-cortical transformation, can be observed for temporal cues, such as VOT, which distinguishes voicing categories (see Fox et al., 2020 for cortical representation of a VOT continuum). Moreover, these studies have mainly focused on group-level patterning. In light of recent reports of individual variation in categorization gradience, further investigation is needed to ascertain whether the subcortical-cortical transformation observed at the group level extends to make behavioral prediction at the individual level. Kapnoula (2016) reported that differences in cortical representations may contribute to how categorically listeners perceive a VOT continuum. Specifically, listeners with steeper slope in categorization (more categorical) showed a step-like pattern of N1, a negative event-related potential (ERP) component reflecting initial formation of object representation in early auditory cortex (Näätänen & Picton, 1987), while a linear relationship between VOT steps and the amplitudes of N1 was observed among listeners with shallower slopes in categorization

(more gradient). Though this study has revealed the potential source of individual differences in categorization gradient; in measuring neural responses from a single level, this work only provides partial information about the neurophysiological processing of VOT, and the potential interactions between subcortical and cortical levels of auditory processing as well as their relationships to categorization gradient remains unknown.

The present study fills important gaps in the literature by elucidating the nature of individual variability in VOT categorization with subcortical, cortical, and behavioral measures collected from the same individuals. In contrast with previous studies where subcortical and cortical responses were elicited in a passive listening and an active labeling task respectively (Bidelman et al., 2013), the subcortical and cortical responses in this study were extracted from the same neural signal recorded during a passive listening task to minimize potential top-down influences (e.g. attention or decision). Moreover, measuring scalp-recorded electrophysiological activity from both subcortical and cortical level allows us to comprehensively assess the representations from different auditory processing stations as well as the transformations performed by the auditory system during speech perception. The subcortical structure of the auditory system encodes acoustic information and relays them to the cortex. Specifically, we used the N1 component to index the cortical representation of VOT, which is suggested to be generated in auditory cortex within Heschl's gyrus approximately 50 ms after the initial primary auditory cortex response (Pratt, 2011). For the subcortical representation of VOT, we focused on the onset component of subcortical responses. The subcortical response evoked by a speech stimulus (e.g., /ba/) consists of an onset component and a frequency-following response (FFR) component (Skoe & Kraus, 2010). The onset component reflects the encoding of transient changes in the signal (e.g., the onset of voicing), while the FFR reflects encoding of sustained

periodic information (i.e. fundamental frequency in tones and higher harmonics in vowels). Though the underlying source of FFR remains a topic of active debate, and that FFR may arise from multiple subcortical and cortical sources, not just subcortically as previously believed, the early nature of the onset response (~ 6 to 10 ms, see the Results section for more details) in the present study suggests a predominantly subcortical source, which is consistent with neural transmission time between the cochlea and inferior colliculus (IC) (see Coffey et al., 2019 for a review). The IC has often been considered as the dominant source of the subcortical responses derived from scalp-recorded EEG (Chandrasekaran & Kraus, 2010; Smith et al., 1975). In the present study, we hypothesized that individual differences in categorization might not only be influenced by differences in encoding at the subcortical and cortical stages (cf. Kapnoula, 2016), but differences in the transformation (i.e. correlation) between the two levels of auditory processing might better explain individual variation in categorization. Specifically, if categorical perception is implemented via a continuous-to-discrete transformation from the subcortical to cortical levels, individuals may differ in the degree of transformation in neural encoding and representation between these two functional levels, which might help explain individual differences in categorization gradience.

Materials and Methods

Participants

Twenty-nine young adults participated in the experiment. Two participants were excluded due to technical problems during data recording, and two additional participants were excluded due to excessive artifacts in the data (electrical noise or high levels of myogenic noise). Final sample consisted of 25 subjects ($\text{age} = 22.17 \pm 0.59$; 9 males), all were university students with no reported speech, hearing, or language disorders. All of the participants were right-handed

($89.45 \pm 9.79\%$) as measured by the Edinburgh Handedness Inventory (Oldfield, 1971).

Experimental procedures were approved by the Social and Behavioral Science Institutional Review Board at the University of Chicago and were in accordance with the Declaration of Helsinki. Informed written consents were obtained from all participants.

Participants completed a survey of their language learning and musical training history (Table S1). The average musical training was 4.72 years ($SD = 4.93$), and participants with musical experience started their training between age 3 to 12 (mean = 8.20, $SD = 2.58$). Five participants had 10 or more years of musical training, but reported minimal practice for the past 24 months (i.e. no more than 1 hour/week) except for one, who practiced on average 25 hours/week. All participants identified English as their first language except for six, who identified an additional language, i.e. Hindi, Mandarin, Farsi, Albanian, Russian, and Spanish, as first language. Self-rated second-language (L2) proficiency on a 7-point Likert-type scale ranging from 1 (very poor) to 7 (native-like) for speaking and listening abilities were on average 4.65 ($SD = 1.78$) and 4.00 ($SD = 1.86$). All participants reported that English is the dominant language in daily life. To account for the potential effects of musical and language experience on speech categorization, years of musical training and self-rated L2 proficiency (averaged for speaking and listening) were entered as covariates in brain-behavior correlational analyses.

Stimuli

Auditory stimuli were synthesized tokens from a 5-step /ba-/pa/ continuum with varying VOT. The five tokens ranged from 14 ms to 70 ms VOT in steps of 14 ms (see Fig. 1A for the waveform of each token). Stimuli were constructed using the Klatt-style synthesizer by a script (<https://goo.gl/27kDuT>) in Praat (Boersma & Weenink, 2010). Tokens were designed to sound like a male talker and digitized at 44100 Hz. Stimuli began with a 4 ms burst of low-amplitude

frication. Formant transition were 50 ms in duration. The formant transition period was followed by a steady-state vowel period, during which the fundamental (F0) and all formant frequencies (F1 – F5) were identical across all tokens. The F0 rose from 90 Hz to 100 Hz during the 50 ms formant transition period and remained at 100 Hz for the rest of the steady-state portion of the vowel. The F1, F2 and F3 ramped up in frequency from 435 to 700, 1000 to 1220, 2186 to 2500 Hz respectively. The F4 and F5 were held constant at 3600 Hz and 4500 Hz. To create the VOT continuum, successive frames of voicing after the stop burst was replaced with aspiration, and the duration of the vowel was varied relative to the VOT so that the overall duration for each stimulus remained constant at 320 ms.

Behavioral Categorization Task

Prior to the electroencephalography (EEG) recording session, participants completed a forced-choice identification task in an acoustically shielded booth. The auditory stimuli were delivered binaurally through Sennheiser HD555 headphones at 80 dB SPL, and the experiment was controlled with E-Prime. Each trial began with a 500-ms fixation cross at the center of the screen followed by presentation of a stimulus. Participants were instructed to identify the sound as quickly as possible (/ba/ or /pa/) through button presses on an E-prime response box within 2 s. After their response, the experiment moved to the next trial with an inter-trial interval jittered randomly between 400 and 600 ms (50-ms steps). Each of the five tokens from the continuum was repeated 20 times in a randomized order for a total of 100 trials.

EEG Data Acquisition and Preprocessing

The /ba/-/pa/ continuum was used again to elicit electrophysiological responses. Stimulus delivery was controlled in Matlab R2018b using Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997), and delivered binaurally at an intensity of 80 dB SPL through inserted

earphones (ER-1, Etymotic Research, Elk Grove Village, IL). The audio stimulus was also routed through a StimTrak box (Brain Products GmbH, Gilching, Germany) into an auxiliary channel of the EEG amplifier (actiCHamp amplifier, Brain Products GmbH). This auxiliary recording provided accurate stimulus timing information for time locking the responses to stimulus onset during offline analysis. The stimuli were presented in alternating polarities for a total of 2000 sweeps per stimulus token (1000 per polarity) with an inter-stimulus interval (ISI) jittered among 390, 400, 410 ms. Stimuli of reversed polarity were obtained by multiplying the stimulus by -1, and this manipulation does not affect listeners' perception of the stimuli. Recording subcortical responses to both polarities and then adding the responses to the two stimulus polarities is a technique that is commonly employed to minimize stimulus artefact and cochlear microphonic (Aiken & Picton, 2008; Russo et al., 2004). The five tokens of the VOT continuum were presented in separate blocks with block order randomized across participants. The entire EEG recording lasted about 2 hours.

Electrophysiological responses were continuously recorded with Ag-AgCl scalp electrode placed at vertex (Cz, active) referenced to linked mastoids (M1/M2), with another electrode on the mid-forehead as ground. This vertical montage is shown to be optimal for recording evoked responses of both subcortical and cortical origin (Bidelman et al., 2013; Krishnan et al., 2012; Musacchia et al., 2008; Xie et al., 2018). Contact impedance was $< 5 \text{ k}\Omega$ for all electrodes. The electrodes were plugged into a bipolar amplifier that provided a gain factor of 50 (EP-preamp module, Brain Products GmbH). The bipolar amplifier interfaced with the actiCHamp amplifier through a splitter box. Responses were acquired at a sampling rate of 25 kHz using BrainVision Recorder software (Brain Products GmbH), with no online filters applied. All recordings were made passively with the participant reclining comfortably in the acoustically shielded booth. To

minimize myogenic artifacts, participants were instructed to relax and refrain from extraneous body movement, and ignore the sounds they heard.

The continuous EEG recordings were differentially bandpass filtered off-line from 1 to 30 Hz (12 dB/octave, zero phase-shift) and from 80 to 2500 Hz (12 dB/octave, zero phase-shift) to predominantly highlight cortical and subcortical auditory electrophysiological responses respectively (Bidelman et al., 2013; Krishnan et al., 2012; Musacchia et al., 2008). The recordings were then epoched into segments that were time locked to the auditory stimuli (cortical: -100 to 500 ms; subcortical: -50 to 320 ms). The timing of each stimulus onset was derived from the auxiliary channel recording. After baseline correcting each response to the mean voltage of the pre-stimulus region, trials with amplitudes exceeding a predefined range (cortical: $\pm 50 \mu\text{V}$; subcortical: $\pm 35 \mu\text{V}$) were rejected. The artifact-free trials (1683 ± 107) were averaged for each stimulus condition for each subject, and downsampled from 25 kHz to 500 Hz for cortical responses. The grand-average subcortical and cortical responses across five stimulus conditions are shown in Fig. 2A and Fig. 3A.

Behavioral Data Analysis

An identification function was constructed for each participant based on the number of times (% of total trials) participants identified each token with /pa/. To quantify the “steepness” of the identification function, that is, how strong a listener differentiates the two categories at the perceptual boundary (Xu et al., 2006), a logistic function regression model was fitted on the identification curve using the `glmfit` function in MATLAB R2018b. Slope of the logistic fit was used to index steepness of the identification curve, with steeper functions indicating stronger categorical perception.

EEG Data Analysis

Subcortical Responses

Based on the definition of VOT – the duration between stop release and onset of voicing, we sought to identify peaks of onset responses associated with initial noise burst (peak I), and those corresponding to voicing onset (peak II) in the subcortical responses (Zhao & Kuhl, 2018). The grand-averaged waveforms across subjects was employed as a reference to facilitate peak identification (Fig. 2A). To enhance visualization of spectrotemporal changes in the FFR, which reflect the onset of vocalic pitch, a time-frequency analysis using a continuous wavelet transform was performed on the grand average waveforms (Fig. 2B). The predefined window for each peak was then set with respect to the expected latency for a given peak based on visual inspection of the average temporal and spectral responses. While the predefined window for peak I was the same for all tokens (5 – 15 ms from stimulus onset), the window for peak II were different depending on the stimulus token (token 1: 20 – 35 ms, token 2: 35 – 50 ms, token 3: 50 – 65 ms. Token 4: 65 – 80 ms, token 5: 80 – 95 ms). Automated peak selection in predefined windows was performed using custom routines coded in MATLAB and visually confirmed by referring to the individual waveforms. The magnitude and latency values were extracted for the two peaks related to the stop consonant for each participant by identifying the maximum value and its corresponding time. The latency differences between peak I and peak II (peak II minus peak I) were computed to index the VOT values encoded in the subcortical responses. In addition to the peak analysis, we also extracted the root mean square (RMS) amplitudes of the entire response waveform to examine the strength of subcortical encoding of each VOT. The RMS measures the magnitude of neural activation by taking the mean absolute value of all sample points of the waveform.

Cortical Responses

Peak amplitude and latency were measured for the prominent deflections of the cortical ERPs (N1, P2, N2) in specific time windows. Using the grand-averaged responses across subjects as a reference to facilitate response identification, N1 was taken as the negative-going trough between 50 and 150 ms from stimulus onset, P2 as the positive-going peak between 150 and 210 ms, and N2 as the negativity between 180 and 250 ms. Prior work has shown that of the ERPs, the N1 waves are the most sensitive to acoustic changes of VOT. For instance, a one-to-one correspondence between N1 amplitude and VOT values was observed independent of phonetic categorization (Toscano et al., 2010). Evidence for the relationship between N1 and VOT changes has been mixed, however. Other studies have shown that change in N1 morphology from single to double-peak coincided with a change in perception from voiced to voiceless for a stop continuum (Sharma & Dorman, 1999; Steinschneider et al., 1999). Additionally, as compared to N1, the P2 and N2 peaks were not always readily identified for every stimuli and for every participant. Thus, we focused our analyses on N1 wave, which allowed us to assess how cortical responses changed as a function of VOT values. For each time window, the 50% fractional area latency and its associated mean amplitude were calculated using custom routines coded in MATLAB. The 50% fractional area was computed by calculating the area under the waveform within a time window and finding the time that divides the area in half. Mean amplitude was computed as the average voltage within a 50 ms window centered on the fractional area latency. The area-based measures are less susceptible to the effects of component overlap than corresponding peak measures (Luck, 2005), providing a better measure of the information indexed by each component i.e. N1, P2, and N2.

Neural Dissimilarity Matrices

As we aimed to examine whether individuals differ in the degree of transformation in neural representation between different functional levels and compare subcortical and cortical responses within each subject, we adopted a multivariate approach of pattern analysis to derive a quantitative measure, rather than extracting the peak latency/amplitude as a neural index.

We constructed neural representational dissimilarity matrices (RDMs) to comprehensively and quantitatively characterize the relationship between neural representation similarity and stimulus-derived acoustic similarity (Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013). This analysis focused on neural encoding of similarities and distinctions between tokens rather than encoding of an individual token. Specifically, neural responses to each token were extracted to calculate dissimilarity (measured as 1 minus the Pearson correlation coefficient) between each pair of tokens for constructing the RDMs. Dissimilarity was calculated for each of the token with the other four tokens, resulting in a 5 by 5 matrix, with a diagonal of zeros and off-diagonal values reflecting the degree to which neural activity differs between a given pair of tokens (see Fig. 5 for an illustration). For each subject, two neural RDMs were constructed: cortical RDM and subcortical RDM. Rather than using the entire response epoch, each RDM was computed based on segments of the pre-processed waveforms which presumably reflected differential encoding of VOT differences. For cortical RDM, the response segment ranged from 50 to 250 ms (a time window that covered the auditory evoked ERPs), and from 0 to 100 ms (a time window that covered both peak I and II across token conditions) for subcortical RDM. Additionally, we calculated the distance between each pair of tokens according to their VOT values to create an acoustic RDM. To evaluate the extent to which neural distinctions between sounds reflected acoustic differences of the stimuli, we compared neural RDMs with the

acoustic RDM using Spearman's rank correlation. The rank-based correlation is often used for comparing RDMs when a linear relationship between dissimilarities can not be assumed (e.g. when comparing brain RDMs to stimulus-based RDMs) (Kriegeskorte & Kievit, 2013). The diagonal and off-diagonal triangle of RDMs were excluded when comparing RDMs, thereby leaving only the lower or upper triangle of RDMs (Ritchie et al., 2017). The correlation coefficients were then used to indicate how strongly the two RDMs were related. In addition to the neural-acoustic correlation, we calculated a neural-neural similarity by comparing cortical and subcortical RDMs, providing the critical index of how presentational patterns of the VOT continuum at the subcortical level is related to the cortical patterns. Thus, three indices were computed for each subject: 1) subcortical-acoustic similarity, 2) cortical-acoustic similarity, and 3) subcortical-cortical similarity (Fig. 5).

Brain-behavior Relationships

To examine the whether and how subcortical and/or cortical representations predicted behavioral categorization, stepwise multiple regression analysis was performed with the behavioral index of the sharpness of categorical perception (i.e. the slope of the identification function) as the dependent variable, and the three similarity indices (i.e. subcortical-acoustic similarity, cortical-acoustic similarity and subcortical-cortical similarity) as the predictors. After the significant neural predictors were identified, a hierarchical regression model was constructed to confirm the findings and control for the potential influences of musical and L2 experiences.

Results

Behavioral Speech Categorization

In the behavioral categorization task, participants classified sounds drawn randomly from a five-step /ba/-/pa/ continuum differing minimally in VOT of the onset consonant (Fig. 1A). The

mean categorization responses for each of the five tokens are shown in Fig. 1B. Results of a mixed-effect logistic regression showed a significant effect of stimulus [$\beta = 1.52$, $z = 25.43$, $p < .001$], suggesting that listeners used VOT to categorize the stimuli, as expected, and there is a clear perceptual shift from /b/ to /p/ near token 3. To examine each participant's response pattern, a logistic function was fitted to each subject's /p/ responses across the five stimuli. The fitted results (Fig. 1C) showed that each subject's mean rate of /p/-identification response (blue dots) for the most ambiguous token (step 3) was widely dispersed between 0% and 100%, resulting in large individual differences in the shape of the fitted identification functions (red lines). Slopes of the fitted curves were then extracted to index an individual's degree of categorization gradience, with smaller slope values corresponding to more gradient categorization.

[Figure 1 near here]

EEG Data

Group-level Patterns – Subcortical and Cortical Responses

We examined how continuous changes in VOT are represented in the subcortical and cortical responses. Based on previous studies (Bidelman et al., 2013; Toscano et al., 2010), we hypothesized that good correspondences between the neural activities and changes in VOT indicate acoustic coding while more step-like neural activities reflect a phonetic category coding of the speech signal. The group-averaged subcortical responses elicited by each of the five tokens are displayed in Fig. 2A. The magnitude and latency for peaks I and II in the response waveforms, demarcating the stop burst and the onset of voicing respectively (denoted by the black markers in Fig. 2A), were extracted. VOT is encoded in the subcortical responses in the latency of Peak II (signalling the onset of voicing) relative to Peak I (indexing the stop burst),

with its latency delayed by around 15 ms for every additional VOT step. This temporal coding pattern is also corroborated by the time-frequency representations of the subcortical responses (Fig. 2B), where the onset of vocalic pitch was delayed linearly across the five steps. The latency differences between peak I and peak II (peak II minus peak I) were computed to index the VOT values encoded in the subcortical responses. Table S2 shows the latencies of Peak I and II identified in the responses to each VOT token. The short stimulus-to-response latency of ~ 8.5 ms suggests a predominant subcortical origin of these responses. The estimated VOT in the subcortical responses tightly paralleled acoustic VOT changes along the stimulus continuum (Fig. 2C), resulting in robust stimulus-to-response VOT correlations ($r = .94$, $p < .0005$). While stimulus steps had a significant effect on the peak II latency [$\beta = 1.38$, $t = 36.3$, $p < .001$], the magnitudes of the onset peaks and RMS of the response waveform remained invariant across the continuum [peak I: $\beta = .0004$, $t = .70$, $p = .484$; peak II: $\beta = .001$, $t = 1.04$, $p = .30$; RMS: $\beta = -.19$, $t = -.42$, $p = .68$] (Fig. 2DE).

To assess the degree to which subcortical responses may reflect category information, we computed response differences (for both amplitude and latency) between tokens 1 and 2 (which were identified as /ba/) and those to tokens 4 and 5 (identified as /pa/) as within-category response differences; and response differences between tokens 2 and 3, and those to tokens 3 and 4 as between-category response differences. If the neural responses contain category information, we expect a significant effect of category condition, with within-category response differences smaller than those of between-category. No category information is observed; we did not observe a significant effect of category condition in the measures from subcortical responses [peak I amplitude: $\beta = .05$, $t = 1.43$, $p = .15$; peak II amplitude: $\beta = .43$, $t = .36$, $p = .72$; peak II latency: $\beta = .96$, $t = .57$, $p = .57$; RMS of the response waveform: $\beta = .14$, $t = .19$, $p = .84$].

[Figure 2 near here]

The group-averaged cortical responses elicited to the VOT continuum are shown in Fig. 3A. Distinct morphological changes related to categorization of VOT are seen in the N1 waveforms. A single negativity (N1) was elicited to stimulus tokens 1 to 3 while two distinct negative components (N1' and N1) were apparent for tokens 4 and 5. The time window was defined as 50 – 150 ms when one N1 component was present. When two negative components were present in the grand average waveform, the time window was defined as 50 – 100 ms for the first component N1' and 100 – 150 ms for the second component N1. The mean latency for the single N1 was 93.1 ms, while for waveforms that exhibited two negativities, the mean latency was 92.7 ms and 140.2 ms for N1' and N1 respectively. These morphological changes in N1 corresponded to the psychometric boundary observed in the behavioral categorization task (Fig. 1B). That is, stimuli that were categorized as /ba/ (tokens 1 and 2) were characterized by a single N1 component while stimuli that were identified as /pa/ (tokens 4 and 5) were characterized by two distinct N1 components (Fig. 3B). To assess the extent to which N1 reflected phonetic category information, we again computed the within-category and between-category response differences for N1 amplitude and latency, and found a significant effect of category condition on N1 latency [$\beta = -16.89$, $t = -5.58$, $p < .001$] (Fig. 3C), with within-category latency differences (between 1 & 2 and between 4 & 5; $M = 5.03$ ms) smaller than those of between-category (between 2 & 3 and between 3 & 4; $M = 21.92$ ms). For N1 amplitude, only a marginal category effect was observed [$\beta = .05$, $t = 1.43$, $p = .075$] (Fig. 3D). None of the other components such as P2 and N2 in the cortical responses showed significant effect of category condition ($ps > .185$).

To further examine whether within-category differences were reflected in the cortical responses, the stimulus effect were examined in the two sets of within-category responses

respectively. Specifically, within the /b/ category, the N1 responses to token 1 were significantly different from those to token 2 in terms of latency [$\beta = 7.22, t = 6.52, p < .001$] and amplitude [$\beta = .83, t = 4.91, p < .001$]; for the /p/ category (tokens 4 and 5), significant effects of stimuli were found for N1 latency [$\beta = 2.85, t = 4.58, p < .001$], but not N1 amplitude [$\beta = .19, t = 1.48, p = .15$], suggesting that N1 responses contain within-category differences.

To disentangle perceptual and acoustic influences on the cortical responses, a response-based analysis was performed by dividing participants into two groups (more b-percept vs. more p-percept) based on how they responded to token 3 (i.e. the most ambiguous token in the VOT continuum). Of the 25 participants, 13 of them identified token 3 as /b/ for more than 50% of the time in the behavioral categorization task, while 11 were more likely to identify the token as /p/. Two participants identified the ambiguous token equiprobably as /b/ or /p/ (at 50% of the time) and were not included in this particular analysis. The cortical responses elicited by token 3 were averaged over each type of listeners (Fig. 4A). Despite having been elicited by an identical acoustic stimulus, N1 component appeared to vary based on listeners' perceptual responses, with significantly higher N1 amplitude observed among p-percept listeners as compared to b-percept listeners [$\beta = 1.35, t = 2.22, p = .037$] (Fig. 4B), while N1 latency did not vary significantly between the two types of listeners [$\beta = 6.72, t = 1.12, p = .27$] (Fig. 4C).

Overall, these group-averaged results show that subcortical responses encode the continuous acoustic differences contained in the speech signal, while cortical N1 seems to predominantly reflect phonetic category information but also contain within-category differences. The N1 morphology (change from a single-peak to double-peak neural response) coincides with the perceptual shift from voiced to voiceless stop heard by the listeners.

[Figure 3 near here]

[*Figure 4 near here*]

Individual-level Measures – Neural Dissimilarity Matrices

An important hypothesis advanced in the present study is that individual differences in categorization gradient might be reflected in the nature of neural encoding at the subcortical and cortical levels as well as the degree of transformation in neural representations between those levels. In order to compare the representational patterns between the two functional levels, we computed neural representational dissimilarity matrices (RDMs) based on pre-processed waveforms of subcortical and cortical responses. Constructing the matrices allows us to compare representations derived from different auditory structures at the level of dissimilarities (i.e. the degree of correspondences between signals), rather than at the level of the original activity patterns. To illustrate RDMs, we used the group-averaged subcortical and cortical responses. Fig. 5 shows the subcortical RDM based on responses averaged across subjects. Each cell represents the dissimilarity between subcortical responses to a pair of VOT stimuli, with the blue cells along the diagonal marking the identical pairs. The subcortical RDM showed a high degree of dissimilarity between the five acoustic tokens as indicated by the different colors across different cells away from the diagonal. In contrast, the cortical RDM shows that within-category tokens elicited similar activity patterns (e.g. token 1 vs. token 2; token 4 vs. token 5), whereas across-category tokens (e.g. token 3 vs. token 4) elicited more disparate neural responses. A similar RDM was constructed for the dissimilarity between the target acoustic stimuli.

In order to compare the degree of similarity between the acoustic stimuli and their corresponding neural responses at different auditory processing stages, we conducted correlation analyses to derive three correlation coefficients to index the relationships between the subcortical, cortical and acoustic RDMs. Specifically, subcortical and cortical RDMs were

correlated with the acoustic RDM to derive two indices – the similarity between the acoustic RDM and the subcortical RDM (i.e. subcortical-acoustic similarity) and the similarity between the acoustic RDM and the cortical RDM (i.e. cortical-acoustic similarity). Crucially, we calculated a similarity index by comparing the two neural RDMs (i.e. subcortical-cortical similarity) as a way to measure the degree of information sharing between the two auditory processing stations. For example, at the group level, Spearman's rank correlation between the subcortical and acoustic RDMs ($r = .75$) was relatively high, compared to those between cortical and acoustic RDMs ($r = .57$), and between subcortical and cortical ($r = .45$), suggesting that, at the group level, subcortical encoding tracks the acoustic signal much more veridically than neural encoding at the cortical level and the weaker correspondence between subcortical and cortical RDMs suggests that the information is not faithfully relayed from the subcortical level to the cortical levels. A similar analytic approach was taken to derive three similarity correlation indices for each participant.

[Figure 5 near here]

Brain-behavior Relations

To determine the relative contribution of neural representations at the subcortical and/or cortical levels to individual differences in categorization gradience, we constructed a multiple regression model with the behavioral index of the sharpness of categorical perception (i.e. the slope of the identification function) as the dependent variable, and the three similarity indices (i.e. subcortical-acoustic similarity, cortical-acoustic similarity and subcortical-cortical similarity) as the predictors. We first estimated the correlations among the predicting variables. The subcortical-cortical similarity was not significantly correlated with either subcortical-acoustic similarity ($r = .13, p = .53$) or cortical-acoustic similarity ($r = .18, p = .38$), but the

subcortical-acoustic and cortical-acoustic similarity was significantly correlated with each other ($r = -.409, p = .009$).

Results of a stepwise regression analysis showed that only subcortical-acoustic similarity and subcortical-cortical similarity emerged as significant predictors (Table 1), accounting for 47.4% of the variance [$F(2, 22) = 11.81, p < .001$]. Specifically, the subcortical-acoustic similarity showed a positive relationship with behavioral categorization (Fig. 6A). That is, if the differences between subcortical responses were more similar to acoustic differences, the listener is more likely to show a sharper identification function. In other words, the better the subcortical responses track the differences in the acoustic signal, the more categorical the listener behaves. An opposite pattern was found for the relationship between subcortical-cortical similarity and categorization gradient (Fig. 6B), with higher similarity between subcortical and cortical representations associated with more gradient in categorization (see Fig. S1 for examples of subcortical and cortical responses from a representative categorical and gradient listener).

[Table 1 near here]

[Figure 6 near here]

To evaluate the potential effects of musical training and L2 proficiency, we constructed a hierarchical regression model with years of musical training and self-rated L2 proficiency entered in the first level, and the subcortical-acoustic and subcortical-cortical similarity added at the second level (see Table 2). Results showed that musical and L2 experience did not have a significant effect on the identification slope [$F(2, 22) = 0.61, p = .55$], and adding the neural similarity measures explained a significant portion of the identification slope variance [$F(4, 20) = 8.69, p = .0029$]. These results suggest that the subject's musical and L2 experiences had

limited influence on how categorically a listener perceive a VOT continuum as compared to neural representations in the current study.

[*Table 2 near here*]

Discussion

The present study investigated the neural mechanisms underlying individual differences in speech categorization by measuring subcortical and cortical responses elicited by a series of consonant-vowel syllables with varying VOT. There are two main sets of findings: (1) At the group level, subcortical responses encoded fine-grain acoustic differences veridically while cortical N1 signaled potential phonetic category information, suggesting an acoustic-to-phonetic transformation of speech representations across functional levels of the auditory system, echoing the previous findings (Bidelman et al., 2013; Chechik et al., 2006); (2) At the individual-level, differences in categorization gradient is positively related to subcortical representations, but negatively correlated with the degree of similarity between subcortical and cortical response patterns. The latter finding suggests that individuals who engage in representational transformation from the subcortical to the cortical levels are more likely to be categorical in their classification of speech sounds. Collectively, our findings suggest that individual differences in categorization gradient are not only driven by how speech information is encoded within each functional level *per se*, but more importantly, by how speech representation is transformed along the hierarchical structure of the auditory system.

The present findings extend prior work on categorization of spectral speech cues, i.e. vowel formants (Bidelman et al., 2013) by demonstrating that this neurophysiological mechanism generalizes to the categorization of temporal speech cues, i.e. VOT. In this study, we found a one-to-one correspondence between the acoustic landmarks associated with VOT (i.e.

noise burst and voicing onset) and the onset peaks (i.e. peaks I and II) in the subcortical responses. Additionally, we did not observe evidence for category perception at the level of subcortical, given that the between-category and within-category response differences were not significantly different. A recent study has however shown top-down modulations on subcortical representation of VOT (Zhao & Kuhl, 2018). Specifically, Spanish speakers showed a significantly longer Peak II latency in subcortical responses elicited by a stimulus of 10 ms VOT (a VOT value where Spanish and English speakers showed maximal differences in perceptual categorization) compared to English speakers, suggesting modulatory influences of linguistic categories at the subcortical level. In that study, only one VOT stimulus, rather than a continuum of VOT values, was used to elicit subcortical responses. Direct evidence for categorical perception at the subcortical level would come from different warping patterns of subcortical responses to the whole acoustic continuum between Spanish and English speakers. Findings of the present study may complement extant evidence of top-down effects by demonstrating bottom-up acoustic modulation on the subcortical representations using synthesized CV syllables as the stimuli. Future work should examine the extent to which bottom-up contributions can be modulated or overridden by top-down influences of linguistic categories using more behaviorally-meaningful stimuli such as lexical words.

Upon arrival at the cortical structures, the cortical responses, as indexed by N1, appeared to be sensitive to category distinctions reflected in the behavioral responses. Specifically, a single-peak N1 was observed in the cortical responses for syllables with a VOT less than 40 ms (tokens 1 to 3 of the continuum), while a double-peak N1 is associated with syllables with a prolonged VOT (> 40 ms; tokens 4 and 5). A VOT of 40 ms also coincided with the categorical boundary (Fig. 1B) as revealed in the speech categorization task. These results are in line with

previous physiological data demonstrating that cortical response pattern of a single versus a double onset parallels the timing of categorical perception for the voiced versus voiceless distinction of stop consonants (Sakai et al., 2009; Steinschneider et al., 2003, 2005). We also observed that N1 was modulated depending on a listener's perception of an identical, ambiguous speech sound (Fig. 4A), consistent with previous studies exhibiting perceptual modulation of cortical activity despite invariance in the physical attributes of the stimulus (Bidelman et al., 2013; Riecke et al., 2009). However, we should distinguish our results from studies that demonstrated neural correlates of categorical speech perception (Chang et al., 2010; Steinschneider et al., 1995). Given the classic definition of categorical perception (Liberman et al., 1957), neural representations are truly categorical if responses to VOT tokens of the same voicing category are indistinguishable. In this study, we found significant effects of stimuli on cortical responses elicited by tokens that are members of the same category. For instance, within the /b/ category, the N1 responses latency and amplitude to token 1 were significantly different from those to token 2. We interpreted these findings as revealing a representation of category perception, but not necessarily reflecting categorical perception, *per se*. We argue that cortical representations contain phonetic category information which is critically relevant for behavioral categorization, but the within-category differences were not lost (Fox et al., 2020; Frye et al., 2007; Toscano et al., 2010). Moreover, the present study showed that representational warping based on phonetic category can be observed without an explicit labelling task, suggesting that top-down effects may not be necessary for eliciting non-linear encoding patterns. Overall, the group-level findings provide additional support for a neurophysiological mechanism by which continuous speech representations at the subcortical level are transformed and warped at the cortical level, which is shared by both spectral and temporal speech cues.

The picture becomes considerably more complicated once we move beyond considering aggregated group patterns and focus on individual patterns. At the behavioral level, listeners vary in the degree of categoricity in their classification responses, as indexed by the slope of the identification function at the category boundary, which may reflect a different perceptual weighting of different sources of information (e.g. between acoustic and phonological representations) (Pisoni & Tash, 1974). Specifically, a high slope value reflects a discrete mapping from cues to categories and a higher weighting of phonological representations (i.e. categorical listeners), whereas a low slope value reflects a graded mapping from cues to categories and a higher weighting of acoustic information (i.e. gradient listeners). In this study, the subcortical-cortical similarity can be taken to index the mapping between functional levels, and we observed a significant negative correlation between this neural index and the identification slope, with higher subcortical-cortical similarities linked to shallower categorization functions (Fig. 6B). A strong correspondence between subcortical and cortical representations indicates increased correspondences between functional levels. It is possible that acoustic details retained at a later processing stage may underlie the increased association between acoustic input with both categories observed among gradient listeners, which linearizes the identification slope and renders speech categorization more continuous. On the other hand, less correspondence between subcortical and cortical representations was observed among listeners with more categorical perceptual responses. These listeners may weight phonological representation more than acoustic information, and are more readily to engage the mechanism of subcortical-to-cortical transformation, resulting in less overlap in information between levels of neural representations. This critical warping between subcortical and cortical representations facilitates the extraction of behaviorally-relevant information at later processing centers, as

indicated by a negative correlation between response time and categorization slope ($r = -.41, p = .037$).

An alternative way to achieve shallower identification slopes in a 2AFC task is if listeners have a discrete mapping in their category structure but encode the acoustic signal less distinctively. The subcortical-acoustic similarity in the present study provides an index of the degree of subcortical differentiation between the five VOTs. Specifically, we observed that individual slopes were positively linked to subcortical-acoustic similarity (i.e. how faithfully the VOT differences along the continuum were encoded in subcortical responses), with shallower slopes (i.e. more gradient categorization behaviors) associated with less faithful subcortical differentiation (Fig. 6A). When the lower-level structures feed the cortical structures with a subcortical representation of reduced differentiation of the acoustic signal, the activation for both phonemes may be increased in later processes during sound identification, resulting in a less categorical identification pattern. This interpretation of the results is consistent with a bottom-up explanation for the relationship between the robustness of feature extraction at the subcortical level and individual differences in behavioral performance in speech tasks (Coffey et al., 2016; Hornickel et al., 2009; Reetzke et al., 2018). On the whole, the individual differences in speech categorization observed in the present study seems to be driven by how faithfully acoustic differences in speech are represented at the subcortical level, more importantly, how this speech information is conveyed along different processing stations (i.e. the mapping between acoustic signal and internal categories).

Potential limitations

The current approach to record subcortical and cortical responses concurrently allows us to comprehensively assess the representations from different auditory processing stations as well

as the subcortical-cortical transformation during voicing perception, it is, however, not without limitation. The simultaneous recording of subcortical and cortical response recording prohibits the use of different ISIs for subcortical and cortical activity respectively. Taking different factors (e.g. stimulus presentation, recording duration, subject fatigue) into account, an ISI jittered among 390, 400, 410 ms was adopted in the EEG recording. These ISIs may be relatively short as compared to a typical ISI (e.g. 800 to 1000 ms) for recording cortical ERPs. These short ISIs might have led to the smearing of the ERP morphology, i.e. a lack of clearly segregated N1-P2 ERPs in the current study (Fig. 3A), as the ERP response to the preceding stimulus might not have enough time to return to baseline level before the arrival of the next stimulus, resulting in overlapping between the ERPs. Alternatively, the ISI might have an opposite effect on the magnitude of the N1 and P2 waves, with short ISIs (< 400 ms) leading to increased N1 but decreased P2 amplitude (Budd & Michie, 1994; Wang et al., 2008). It may be useful in future work to record subcortical and cortical responses simultaneously or in different blocks with different ISIs in order to examine the impact of recording methods on EEG responses.

Conclusions

While the group-level results in this study provide additional support for a neurophysiological mechanism that enables the brain to achieve its acoustic-to-phonetic mapping in speech categorization (Bidelman et al., 2013; Chechik et al., 2006; Olshausen & Field, 2004; Perez et al., 2013), our individual-level results, however, reveal potential mechanistic variations subserving different categorization patterns among typically-developed adults. These findings raise questions regarding the role of rapid and discrete mapping of signal to category structures for successful communication and language development. The degree to which listeners exhibit gradience in their classification of speech sounds is thought to reflect the capacity and strength of

phonological processing, and a divergence from a step-like pattern in speech categorization has been considered a marker of atypical and non-optimal language processing (Breier et al., 2004; Stewart et al., 2018; Werker & Tees, 1987). Our findings complicate the view that categorization can be dichotomized between typically-developed and atypical individuals. The variability observed in present study not only demonstrates the extent of heterogeneity of neural representations from subcortical to cortical levels among healthy adults, but also provides a foundation for future research to elucidate the neurophysiological mechanisms underlying a much broader spectrum of variability in speech categorization, from expert/highly trained individuals (e.g. musicians) that are known to show enhanced phonological processing ability, to atypical populations (e.g., autism, dyslexia, specific language impairment etc.) that are known to exhibit phonological deficits.

Acknowledgements

This work was supported by a NSF grant (BCS1827409).

Additional Information

Competing financial interests: The authors declare no competing financial interests.

References

- Aiken, S. J., & Picton, T. W. (2008). Envelope and spectral frequency-following responses to vowel sounds. *Hearing Research*, 245(1–2), 35–47.
- Bidelman, G. M., Moreno, S., & Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. *Neuroimage*, 79, 201–212. <https://doi.org/10.1016/j.neuroimage.2013.04.093>
- Boersma, P., & Weenink, D. (2010). Praat (Version 5.1. 35). *Amsterdam, The Netherlands: Phonetic Sciences, University of Amsterdam*.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4), 433–436.
- Breier, J. I., Fletcher, J. M., Denton, C., & Gray, L. C. (2004). Categorical perception of speech stimuli in children at risk for reading difficulty. *Journal of Experimental Child Psychology*, 88(2), 152–170. <https://doi.org/10.1016/j.jecp.2004.03.003>
- Budd, T. W., & Michie, P. T. (1994). Facilitation of the N1 peak of the auditory ERP at short stimulus intervals. *Neuroreport: An International Journal for the Rapid Communication of Research in Neuroscience*.
- Chandrasekaran, B., & Kraus, N. (2010). The scalp-recorded brainstem response to speech: Neural origins and plasticity. *Psychophysiology*, 47(2), 236–246. <https://doi.org/10.1111/j.1469-8986.2009.00928.x>
- 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. *Nature Neuroscience*, 13(11), 1428. <https://doi.org/10.1038/nn.2641>
- 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(3), 359–368. <https://doi.org/10.1016/j.neuron.2006.06.030>
- Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3), 804–809.
- Coffey, E. B., Colagrosso, E. M., Lehmann, A., Schönwiesner, M., & Zatorre, R. J. (2016). Individual differences in the frequency-following response: Relation to pitch perception. *PLoS One*, 11(3). <https://doi.org/10.1371/journal.pone.0152374>
- Coffey, E. B., Nicol, T., White-Schwoch, T., Chandrasekaran, B., Krizman, J., Skoe, E., Zatorre, R. J., & Kraus, N. (2019). Evolving perspectives on the sources of the frequency-following response. *Nature Communications*, 10(1), 1–10.
- Fox, N. P., Leonard, M., Sjerps, M. J., & Chang, E. F. (2020). Transformation of a temporal speech cue to a spatial neural code in human auditory cortex. *Elife*, 9, e53051.
- Frye, R. E., Fisher, J. M., Coty, A., Zarella, M., Liederman, J., & Halgren, E. (2007). Linear coding of voice onset time. *Journal of Cognitive Neuroscience*, 19(9), 1476–1487.
- Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1), 69–78. <https://doi.org/10.1002/wcs.26>
- Hornickel, J., Skoe, E., Nicol, T., Zecker, S., & Kraus, N. (2009). Subcortical differentiation of stop consonants relates to reading and speech-in-noise perception. *Proceedings of the*

- National Academy of Sciences*, 106(31), 13022–13027.
<https://doi.org/10.1073/pnas.0901123106>
- Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, 43(9), 1594. <https://doi.org/10.1037/xhp0000410>
- Kapnoula, E. E. (2016). *Individual differences in speech perception: Sources, functions, and consequences of phoneme categorization gradiency* (Doctoral dissertation). University of Iowa.
- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148.
- Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, 59, 40–57.
<https://doi.org/10.1016/j.wocn.2016.08.006>
- Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8), 401–412.
<https://doi.org/10.1016/j.tics.2013.06.007>
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4. <https://doi.org/10.3389/neuro.06.004.2008>
- 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(12), 2849–2859.
<https://doi.org/10.1016/j.neuropsychologia.2012.08.013>
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431. <https://doi.org/10.1037/h0020279>
- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5), 358. <https://doi.org/10.1037/h0044417>
- Lisker, L. (1986). “Voicing” in English: A catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and Speech*, 29(1), 3–11.
- Luck, S. J. (2005). Ten simple rules for designing ERP experiments. In T. C. Handy (Ed.), *Event-Related Potentials: A Methods Handbook* (pp. 17-32). Cambridge, MA: MIT Press.
- McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86(2), B33–B42.
[https://doi.org/10.1016/S0010-0277\(02\)00157-9](https://doi.org/10.1016/S0010-0277(02)00157-9)
- Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174), 1006–1010.

- Musacchia, G., Strait, D., & Kraus, N. (2008). Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians. *Hearing Research*, 241(1–2), 34–42. <https://doi.org/10.1016/j.heares.2008.04.013>
- Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science*, 20(7), 895–903.
- 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(4), 375–425. <https://doi-org/10.1111/j.1469-8986.1987.tb00311.x>
- Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, 9(1), 97–113.
- Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4), 481–487. <https://doi.org/10.1016/j.conb.2004.07.007>
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4), 437–442.
- 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. *Cerebral Cortex*, 23(3), 670–683. <https://doi.org/10.1093/cercor/bhs045>
- Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, 13(2), 253–260.
- Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 15(2), 285–290. <https://doi.org/10.3758/BF03213946>
- Pratt, H. (2011). Sensory ERP components. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potential Components* (pp. 89–114). Oxford, UK: Oxford University Press.
- Reetzke, R., Xie, Z., Llanos, F., & Chandrasekaran, B. (2018). Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. *Current Biology*, 28(9), 1419–1427. <https://doi.org/10.1016/j.cub.2018.03.026>
- 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(4), 550–561.
- Ritchie, J. B., Bracci, S., & de Beeck, H. O. (2017). Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *Neuroimage*, 148, 197–200.
- Russo, N., Nicol, T., Musacchia, G., & Kraus, N. (2004). Brainstem responses to speech syllables. *Clinical Neurophysiology*, 115(9), 2021–2030.
- Sakai, M., Chimoto, S., Qin, L., & Sato, Y. (2009). Neural mechanisms of interstimulus interval-dependent responses in the primary auditory cortex of awake cats. *BMC Neuroscience*, 10(1), 10. <https://doi.org/10.1186/1471-2202-10-10>

- Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, 52, 183–204. <https://doi.org/10.1016/j.wocn.2015.07.003>
- Sharma, A., & Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *The Journal of the Acoustical Society of America*, 106(2), 1078–1083. <https://doi.org/10.1121/1.428048>
- Skoe, E., & Kraus, N. (2010). Auditory brainstem response to complex sounds: A tutorial. *Ear and Hearing*, 31(3), 302. <https://doi.org/10.1097/AUD.0b013e3181cdb272>
- Smith, J. C., Marsh, J. T., & Brown, W. S. (1975). Far-field recorded frequency-following responses: Evidence for the locus of brainstem sources. *Electroencephalography and Clinical Neurophysiology*, 39(5), 465–472.
- 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. *The Journal of the Acoustical Society of America*, 114(1), 307–321. <https://doi.org/10.1121/1.1582449>
- Steinschneider, M., Schroeder, C. E., Arezzo, J. C., & Vaughan, H. G. (1995). Physiologic correlates of the voice onset time boundary in primary auditory cortex (A1) of the awake monkey: Temporal response patterns. *Brain and Language*, 48(3), 326–340. <https://doi.org/10.1006/brln.1995.1015>
- 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. *Cerebral Cortex*, 15(2), 170–186. <https://doi.org/10.1093/cercor/bhh120>
- 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. *Journal of Neurophysiology*, 82(5), 2346–2357. <https://doi.org/10.1152/jn.1999.82.5.2346>
- Stewart, M. E., Petrou, A. M., & Ota, M. (2018). Categorical speech perception in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders*, 48(1), 72–82. <https://doi.org/10.1007/s10803-017-3284-0>
- Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, 21(10), 1532–1540. <https://doi.org/10.1177/0956797610384142>
- Wang, A. L., Mouraux, A., Liang, M., & Iannetti, G. D. (2008). The enhancement of the N1 wave elicited by sensory stimuli presented at very short inter-stimulus intervals is a general feature across sensory systems. *PloS One*, 3(12), e3929.
- Werker, J. F., & Tees, R. C. (1987). Speech perception in severely disabled and average reading children. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 41(1), 48.

- Xie, Z., Reetzke, R., & Chandrasekaran, B. (2018). Taking attention away from the auditory modality: Context-dependent effects on early sensory encoding of speech. *Neuroscience*, 384, 64–75. <https://doi.org/10.1016/j.neuroscience.2018.05.023>
- Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America*, 120(2), 1063–1074. <https://doi.org/10.1121/1.2213572>
- Zhao, T. C., & Kuhl, P. K. (2018). Linguistic effect on speech perception observed at the brainstem. *Proceedings of the National Academy of Sciences*, 115(35), 8716–8721. <https://doi.org/10.1073/pnas.1800186115>

Supplementary Materials

Table S1. Musical and Language backgrounds for individuals in the final sample.

subject	Years of Musical training	Onset of musical training (age)	First language(s)	Other Languages and self-reported proficiencies (1-7, 1 as ‘not proficient’, 7 as ‘very proficient’) averaged for speaking and listening
S35	1	9	English	N/A
S42	7	6	English	Spanish (3)
S34	9	8	English	Spanish (4)
S29	16	3	English	N/A
S50	5	5	English	Korean (3)
S37	3	9	English; Russian	Russian (6.5); French (4.5)
S18	3	8	English	Italian (6); French (4.5)
S14	1	8	English	Spanish (5.5)
S28	0	N/A	English; Farsi	Farsi (6); Arabic (5); French (2.5)
S33	13	9	English	N/A
S25	0	N/A	English	N/A
S43	5	9	English	Spanish (3.5)
S21	1	5	English	Spanish (3.5); French (3.5)
S45	1	11	English; Spanish	Spanish (6.5); Catalan (4.5); French (3.5)
S30	11	5	English	Spanish (4.5)
S17	3	9	English; Hindi	Hindi (7); French (2.5)
S26	3	10	English; Mandarin	Mandarin (4.5)
S40	3	12	English	Spanish (5); Mandarin (2)
S22	0	N/A	English	Spanish (4)
S32	0	N/A	English; Albanian	Albanian (7); Greek (4); Japanese (4)
S36	5	10	English	N/A
S31	13	5	English	French (3.5); Croatian (2.5)
S27	0	N/A	English	N/A
S16	2	11	English	Vietnamese (3); Spanish (4.5)
S13	13	12	English	N/A

Note. This table is organized by individuals’ identification slope, from high (more categorical listeners) to low (more gradient listeners) slope value.

Table S2. Means (SD) of Peak I (associated with initial noise burst), Peak II (corresponding to voicing onset) identified in subcortical responses, and the estimated VOT (computed by Peak II minus Peak I).

Stimulus of the VOT continuum	Peak I latency (ms)	Peak II latency (ms)	Estimated VOT (Peak II – Peak I)
14 ms VOT	9.2 (4.3)	28.5 (4.9)	19.3 (6.5)
28 ms VOT	8.4 (3.5)	44.5 (5.4)	36.1 (6.2)
42 ms VOT	8.8 (3.4)	58.1 (4.8)	49.2 (5.2)
56 ms VOT	8.2 (2.6)	73.5 (3.9)	65.2 (4.8)
70 ms VOT	8.1 (2.5)	85.4 (5.4)	77.2 (5.9)

Figure S1. Subcortical (left), cortical (middle) and behavioral (right) from a representative gradient (upper) and categorical (lower) listener.

Tables:

Table 1. Summary of stepwise regression model predicting identification slope in categorization.

Predictor	β	SE	t	p
<i>Step1</i>				
Intercept	4.18	0.53	7.81	< .001***
Subcortical-cortical sim.	-3.73	1.34	-2.78	.011*
<i>Step2</i>				
Intercept	2.54	0.63	3.98	< .001***
Subcortical-cortical sim.	-4.17	1.09	-3.79	.001***
Subcortical-acoustic sim.	4.29	1.21	3.55	.002**

Note. Adjusted $R^2 = .215$ for Step 1 ($p = .011$); $\Delta R^2 = .259$ ($p < .001$). Significance code: * .05, ** .01, *** .001. sim. = similarity.

Table 2. Summary of Hierarchical regression model predicting identification slope in categorization.

Predictor	β	SE	t	p
<i>Step1</i>				
Intercept	2.71	1.02	2.64	.015*
Musical training (yrs.)	0.08	0.09	0.96	.344
L2 proficiency	-0.01	0.17	-0.09	.926
<i>Step2</i>				
Intercept	1.35	0.89	1.52	.145
Subcortical-cortical sim.	-5.49	1.18	-4.62	< .001***
Subcortical-acoustic sim.	5.36	1.38	3.88	< .001***

Note. Adjusted $R^2 = .034$ for Step 1 ($p = .55$); $\Delta R^2 = .528$ ($p < .001$). Significance code: * .05,

** .01, *** .001. yrs. = years; sim. = similarity.

Figures:

Fig. 1.

Fig. 2.

Fig. 3.

Fig. 4.

Fig. 5.

Fig. 6.

Figure captions

Fig. 1. (A) The acoustic waveform of each acoustic stimulus along the five-step /ba/-/pa/ VOT continuum. (B) Mean identification response of /b/ and /p/ for each token averaged across participants. Error bars represent ± 1 s.e.m. (C) The logistic function (in red) fitted to each participant's identification responses of /p/ (the blue dots) along the VOT continuum.

Fig. 2. Group-averaged subcortical responses to each token from the VOT continuum. (A) Group-averaged subcortical responses. Stimulus onset (time = 0 ms) is indicated by the dotted vertical line. Onset peaks corresponding to stop consonant (Peak I for initial noise burst and Peak II for voice onset) are denoted by black markers. (B) Time-frequency representations of the group-averaged subcortical responses. (C) Correspondences between stimulus VOT and acoustic VOT values estimated from subcortical responses (computed by the latency differences between peak I and peak II, i.e. peak II – peak I). (D) Group averages for peak I and peak II amplitudes. (E) Group averages for root mean square (RMS) amplitudes. Error bars = s.e.m..

Fig. 3. Group-averaged cortical responses to each stimulus from the VOT continuum. (A) Group-averaged cortical responses. Stimulus onset (time = 0 ms) is indicated by the dotted vertical line. (B) Cortical responses averaged over /b/- (tokens 1 and 2) and /p/-type (tokens 4 and 5) of stimuli. Prominent waves (e.g., N1, P2, N2) are labelled. Shaded regions indicate standard error of the mean. (C) Group averages for N1 peak latency. (D) Group-averaged N1 amplitude. Error bars = s.e.m..

Fig. 4. Cortical responses reflect perceptual categorization of VOT stimuli. **(A)** Cortical responses to the ambiguous token 3 averaged over listeners ($N = 13$) who identified the token as /b/ versus those ($N = 11$) who identified the token as /p/. Shaded regions indicate standard error of the mean. **(B)** Averaged N1 amplitude based on the heard voicing stop. **(C)** Averaged N1 peak latency based on the heard voicing stop. Error bars = s.e.m..

Fig. 5. Illustrations of neural representational dissimilarity matrices (RDMs). Three RDMs were constructed based on group-averaged subcortical response (top left), group-averaged cortical responses (top right), and stimulus acoustic waveforms (bottom), respectively. RDMs provide a way to visualize how patterns of brain responses differ across VOT steps, with diagonal entries indicating comparison of the same stimuli (dissimilarity score equals 0) and the off-diagonal values reflecting the degree of dissimilarity between two different signals. Hotter color (i.e. the redder the color) corresponds to greater dissimilarity.

Fig. 6. Brain-behavior correlations. Scatterplots between **(A)** subcortical-acoustic similarity and speech categorization slope; **(B)** subcortical-cortical similarity and speech categorization slope. Dotted lines indicate linear functions between the two measures with shaded regions showing 95% confidence interval. The subject number is shown on each data point.