Speaker-normalized sound representations in the human auditory cortex

Matthias J. Sjerps1,2, Neal P. Fox3, Keith Johnson4 & Edward F. Chang3,5

The acoustic dimensions that distinguish speech sounds (like the vowel differences in “boot” and “boat”) also differentiate speakers’ voices. Therefore, listeners must normalize across speakers without losing linguistic information. Past behavioral work suggests an important role for auditory contrast enhancement in normalization: preceding context affects listeners’ perception of subsequent speech sounds. Here, using intracranial electrocorticography in humans, we investigate whether and how such context effects arise in auditory cortex. Participants identified speech sounds that were preceded by phrases from two different speakers whose voices differed along the same acoustic dimension as target words (the lowest resonance of the vocal tract). In every participant, target vowels evoke a speaker-dependent neural response that is consistent with the listener’s perception, and which follows from a contrast enhancement model. Auditory cortex processing thus displays a critical feature of normalization, allowing listeners to extract meaningful content from the voices of diverse speakers.
A fundamental computational challenge faced by perceptual systems is the lack of a one-to-one mapping between highly variable sensory signals and the discrete, behaviorally relevant features they reflect\(^1\). A profound example of this problem exists in human speech perception, where the main cues to speech sound identity also vary depending on speaker identity\(^2\-5\).

For example, to distinguish one speaker’s /u/ ("boot") and /o/ ("boat"), listeners rely primarily on the vowel’s first formant frequency (F1; the first vocal tract resonance, reflected as a dominant peak in the frequency spectrum) because it is lower for /u/ than for /o/\(^6\). However, people with long vocal tracts (typically tall male speakers) have overall lower resonance frequencies than those of speakers with shorter vocal tracts. Consequently, a tall person’s production of the word "boat" and a short person’s "boot" could be acoustically identical. Behavioral research has shown that preceding context allows listeners to tune-in to the acoustic properties of a particular voice and normalize subsequent speech input\(^7\-11\). One classic example of this effect is that a single acoustic token, ambiguous between /u/ and /i/ will be labelled as /i/ after a context sentence spoken by a tall-sounding person (low F1), but as /u/ after a context sentence spoken by a shorter-sounding person (high F1)\(^9\). Therefore, understanding speech involves a process that builds up a representation of the characteristics of a speaker’s voice and adjusts perception of new speech input to accommodate those characteristics.

There is considerable evidence that perceptually relevant sound representations arise within human parabelt (nonprimary) auditory cortex (AC). First, neural activity in the superior temporal gyrus (STG) is sensitive to acoustic-phonetic features, like F1, that are critical for recognizing and discriminating phonemes\(^12\-18\). For example, vowels with low F1 frequencies (e.g., /u/, /i/) can be distinguished from vowels with relatively higher F1 frequencies (e.g., /o/, /æ/) based on local activity within human STG\(^19\). Second, the STG’s encoding of speech is not a strictly linear (vertical) encoding the acoustics; rather, it reflects some properties of abstraction, including categorical perception, relative encoding of pitch, and attentional enhancement\(^13,26,21\). However, to date, it remains unknown whether speech representations in human AC also exhibit the type of context-dependence that could underlie speaker normalization.

Behavioral research in humans has previously suggested that normalization effects could partly arise from the general auditory contrast enhancement mechanisms\(^10,11,22-26\), which are known to affect neurophysiological responses throughout the auditory hierarchy in animal\(^27\-29\). Contrast enhancement models posit that adaptation to the frequency content of immediately preceding contexts—or their long-term average spectrum (see, e.g., \(^11,30\)—affects the responses to novel stimuli depending on the amount of overlap in their frequency content with that of the context. Moreover, behavioral evidence suggests that this contrast enhancement (and, hence, normalization) should arise—at least in part—centrally. For instance, contralateral (dichotic) presentation of the context (either speech or nonspeech) and target drive similar contrast enhancement effects in speech categorization as do ipsilaterally presented context and target\(^11,25,31,32\).

Taken together, past work suggests a role for contrast enhancement in speech sound normalization in AC, but this prediction has not been directly demonstrated in neurophysiological studies in humans. Models of contrast enhancement make specific predictions about the responses of feature-tuned neuronal populations in AC\(^30\). Not only should the cortical representation of the same speech target depend on context, but, more specifically, context-dependent representations should differ in a particular (contrastive) way. That is, after a low-F1 context, the encoding of an ambiguous vowel target’s F1 should more closely resemble the encoding of high F1 targets, while after a high F1 context it should more closely resemble low F1 targets. So far, contrast enhancement has been observed within frequency-tuned neurons in tonotopic primary AC in nonhuman mammals\(^27,28\), but related patterns have not yet been observed in human AC, let alone in the context of human speech perception.

To investigate the influence of speaker context on speech sound encoding in AC, we recorded neural activity from human participants implanted with subdural high-density electrode arrays that covered peri-sylvian language cortex while they listened to and identified target vowels presented in the context of sentences spoken by two different voices\(^13,33\). We found direct evidence of speaker-normalized neural representations of vowel sounds in parabelt AC, including STG. Critically, the observed normalization effects reflected the contrastive relation between the F1 range in the context sentences and F1 of the target vowels, providing direct evidence for context-dependent contrast enhancement in human speech perception. More generally, these results demonstrate the critical role of human auditory-speech cortex in compensating for variability by integrating incoming sounds with their surrounding acoustic contexts.

Results

Speech sound perception is dependent on context. We recorded neural activity directly from the cortical surface of five Spanish-speaking neurosurgical patients while they voluntarily participated in a speech sound identification task. They listened to Spanish sentences that ended in a (pseudo)word target, which they categorized as either “sufu” or “sofo” on each trial with a button press (Fig. 1a, b). The sentence-final targets comprised a digitally synthesized six-step continuum morphing from an unambiguous sufu to an unambiguous sofo, with four intermediate tokens (s\(\beta\)f, s\(\beta\)o, i.e., spanning a perceptually ambiguous range). On each trial, a pseudo-randomly selected target was preceded by a context sentence (A veces se halla... “At times she feels rather...”). Two versions of this sentence were synthesized, differing only in their mean F1 frequencies (Fig. 1a, c; Supplementary Fig. S1), yielding two contexts that listeners perceived as two different speakers: one with a long vocal tract (low F1; Speaker A) and one with a short vocal tract (high F1; Speaker B). Critically, F1 frequency is also the primary acoustic dimension that distinguishes between the vowels /u/ and /o/ in natural speech (in both Spanish and English) (Fig. 1a and Supplementary Fig. S1)\(^6\). Similar materials have previously been shown to induce a reliable shift in the perception of an /u-/ /o/ continuum (a normalization effect) in healthy Spanish-, English-, and Dutch-listeners\(^7\).

As expected, participants’ perception of the target continuum was affected by the F1 range of the preceding sentence context (\(\beta_{\text{context F1}} = -1.70, t = -3.51, p < 0.001;\) Fig. 1d; see Supplementary Materials for details of the mixed effects logistic regression that was used). Specifically, participants were more likely to identify tokens as sofo (the vowel category corresponding to higher F1 values) after a low-F1 voice (Speaker A) compared to the same target presented after a high-F1 voice (Speaker B). Hence, listeners’ perceptual boundary between the /u/ and /o/ vowel categories shifted to more closely reflect the F1 range of the context speaker. Past work has interpreted this classical finding in light of the contrastive perceptual effects that are ubiquitous among sensory systems\(^34\): the F1 of a speech target will sound relatively higher (i.e., sound more like an /o/) after a low-F1 context sentence than after a high-F1 context. Behaviorally, this is reflected as a shift of the category boundary to lower-F1 values.
Human AC exhibits speaker-dependent speech representations. Two of the most influential hypotheses explaining the phenomenon of speaker normalization posit that: (1) contrast enhancing processes, operating at early auditory processing levels, change the representation of the input signal in a normalizing way; or (2) alternatively, it has been suggested that normalization does not alter the perceptual representation of speech, but instead, that normalization arises as a consequence of a speaker-specific mapping of the auditory representation onto abstract linguistic units (i.e., listeners have learned to map an F1 of 400 Hz to /u/-words for speakers, or vocal tracts, that sound short, but to /o/-words for speakers that sound taller). Hence, while the contrast enhancement hypothesis predicts normalized representations in AC, the latter theory predicts that early auditory representations of speech cues remain unnormalized (i.e., independent of speaker-context).

Past neurobiological work has demonstrated that neural populations in parabelt AC are sensitive to acoustic-phonetic cues that distinguish classes of speech sounds, including vowels, and not to specific phonemes per se. Hence, the primary goal of the current study was to examine whether the neural representation of vowels in parabelt AC shifts in a contrast enhancing way relative to the acoustic characteristics of the preceding speaker, or, alternatively, whether such contrast enhancement (or normalization) is not reflected in parabelt AC processing. We first tested whether individual cortical sites that reliably differentiate between vowels (i.e., discriminate /u/ from /o/ in their neural response) exhibit normalization effects.

To this end, we examined stimulus-locked neural activity in the high-gamma band (70–150 Hz) at each temporal lobe electrode (n = 406 across patients; this number is used for all Bonferroni corrections below) during each trial. High-gamma activity is a way to exhibit normalization effects. We found that the magnitude and direction of an electrode’s context effect was predicted by the context F1 factor (Fig. 2b and bottom panel of Fig. 2c). Crucially, this strong relationship had a negative slope, such that electrodes that had high-F1 target preferences (sofo > sufu) had stronger responses to targets after low-F1 context sentences (low-F1 context > high-F1 context; $\beta_{\text{Context L-F1}} = -0.32$, $t = -5.00$, $p = 1.53 \times 10^{-8}$). Importantly, this demonstrates that the relationship between context response and target response reflects an encoding of the contrast between the formant properties of each, consistent with the normalization pattern observed in the behavioral responses (Fig. 1d) and with the predictions of a contrast enhancement model of speech normalization.

Normalization of vowel representations in all participants. Figure 2d demonstrates that local populations in AC that display tuning for specific target vowel F1 ranges exhibit normalization.
The neural response to bottom-up acoustic input is modulated by preceding context. A Target vowel preferences and locations (plotted on a standardized brain) for electrodes from all patients (three with right hemisphere [RH] and two with left hemisphere [LH] grid implants). Only those temporal lobe electrodes where the full omnibus model was significant during the context and/or the target window (F-test; p < 0.05) are displayed. Strong target F1 selectivity is relatively uncommon: electrodes with a black-and-white outline are significant at Bonferroni corrected p < 0.05 (n = 9, out of 406 temporal lobe electrodes); a single black outline indicates significance at a more liberal threshold (p < 0.05, uncorrected; n = 28). Activity from the indicated electrode (e1) is shown in b and c. b Example of normalization in a single electrode (e1); z-scored high-gamma [high-γ] response averaged across the target window (target window marked in c). c Activity from e1 across time, separating the endpoint targets (middle panel) or the contexts (bottom panel). The electrode responds more strongly to /a/ stimuli than /u/ stimuli, but also responds more strongly overall after Speaker A (low-F1). This effect is analogous to the behavioral normalization (Fig. 1d). Black bars at the bottom of the panels indicate significant time-clusters (cluster-based permutation test of significance). d Among all electrodes with significant target sound selectivity (n = 37 [9 + 28]), a relation exists between the by-electrode context effect and target preference. Both are expressed as a signed t-value, demonstrating that the size and direction of the target preferences predicts the size and direction of the context effects. e An LDA classifier was trained on the distributed neural responses elicited by the sufu and sofo endpoint stimuli using all task-related electrodes. This model was then used to predict classes for neural responses to (held-out) endpoint stimuli and for the ambiguous steps in each context condition. Proportions of neurally based “sofo” predicted trials (thick lines) display a relative shift between the two context conditions (data from one example patient). Regression curves were fitted to these data for each participant separately to estimate 50% category boundaries per condition for panel f (thin lines). f The neural classification functions display a shift in category boundaries between context conditions for all patients individually. In b and c, error bars indicate ±1 s.e.m.

However, only a few electrodes (n = 9, out of 406 temporal lobe electrodes) displayed very strong tuning (significance at Bonferroni-corrected p < 0.05), while the majority of F1-tuned electrodes displayed only moderate tuning and moderate context effects. Moreover, not all participants had electrodes that displayed strong target F1 tuning (see Table S2). The relative sparseness of strong tuning is not surprising given that the target vowel synthesis involved only small F1 frequency differences (~30 Hz per step), with the endpoints being separated by only 150 Hz (which is, however, a prototypical F1 distance between /iu/ and /o/). However, past work has demonstrated that even small acoustic differences among speech sounds are robustly encoded by distributed patterns of neural activity across AC13,14. In order to determine whether distributed neural representations of vowels reliably display normalization across all participants, we trained a multivariate pattern classifier model (linear discriminant analysis, LDA) on the spatiotemporal neural response patterns of each participant. Models were trained to discriminate between the endpoint stimuli (i.e., trained on the neural responses to steps 1 vs. 6, irrespective of context) using all task-related electrodes for that participant. These models were then used to predict labels for held-out neural responses to both the endpoints and the ambiguous steps in each context condition. For all participants, classification of held-out endpoint trials was significantly better than chance (Supplementary Fig. S3b). To assess the influence of target F1 and context F1 on the classifier output, a logistic generalized linear mixed model was then fit to the proportion of predicted sofo responses across all participants.

Figure 2e displays the proportion of sofo labels predicted for all stimuli by the LDA classifier based on the neural data of one example participant (thick lines). Importantly, a shift is observed in the point of crossing of the category boundary. Regression functions fitted to these data (thin lines) were used to estimate the size and direction of the context-driven neural boundary (50% crossover point) shift per participant. For each participant, the neural vowel boundaries, like the behavioral vowel boundaries, were found to be context-dependent (Fig. 2f; see Supplementary Materials and Supplementary Fig. S3 for further detail).

A combined regression analysis demonstrated that, across participants, population neural activity in the temporal lobe was modulated both by the acoustic properties of the target vowel (βTarget F1 = 0.50, t = 13.20, p < 0.001) and by the preceding
context \( (\hat{p}_{\text{Context F1}} = -0.34, t = -5.58, p < 0.001) \). Moreover, this effect was not observed for task-related electrodes outside of the temporal lobe during the target window (see Supplementary Fig. S4; nontemporal electrodes were mostly located on sensorimotor cortex and the inferior frontal gyrus).

Importantly, and consistent with participants’ perception, the neural classification functions demonstrate that the influence of the context sentences consistently affected target vowel representations in a contrastive (normalizing) direction: the neural response to an ambiguous target vowel with a given F1 is more like that of /o/ (high-F1) after a low-F1 context (Speaker A) than after a high-F1 context (Speaker B; see Supplementary Figs. S3 and S4b for more detail).

Normalization by acoustic-phonetic contrast enhancement. It has been suggested that a major organizing principle of speech encoding by human parabelt AC is its encoding of acoustic-phonetic features, which are more cross-linguistically generalizable and more physically grounded than phonemes (or other possible higher-level linguistic representations) per se. However, AC processing is diverse and may contain regions that are, in fact, selective for (more abstract) phonemes. For example, AC has also been found to display properties that are typically associated with abstract sound categories such as categorical perception. Hence, we next assessed whether the normalization effects observed here involved neural populations that display sensitivity to acoustic-phonetic features (i.e., relating to more general F1 characteristics). Because the context sentence (“A veces se halla...”) did not contain the target vowels /ul/ or /ol/, while its F1 values did cover the same general frequency range, we examined neural responses during the context window to understand individual electrode preferences.

To this end we again used the glm-based \( t \)-statistics of all temporal lobe electrodes that displayed tuning for the endpoint vowels \( (n = 37; \text{as per Fig. 2d}) \). Among these electrodes, however, we examined the relationship between their preferences for context F1 during the context window and for target F1 during the target window. Figure 3a displays context and target preferences on the cortex of a single-example patient. Among the electrodes that displayed target F1 selectivity, some also displayed selectivity for the context F1 during the context window (such dual preferences are indicated with a black-and-white outline). Figure 3b displays the activation profile of one example electrode (e2). Importantly, e2 responded more strongly to low-F1 targets during the target window (sufi preference: \( \hat{p}_{\text{Target F1}} = -0.52, t = -10.25, p = 3.9 \times 10^{-21} \)), but also to low-F1 contexts during the context window (Speaker A preference: \( \hat{p}_{\text{Target F1}} = -0.54, t = -8.05, p = 2.3 \times 10^{-14} \)). This demonstrates that this neural population responded more strongly to low-F1 acoustic stimuli in general and is not exclusively selective for a discrete phoneme category. Importantly, e2 also displayed normalization, as its activity was affected by context F1 during the target window \( (p = 2.7 \times 10^{-4}) \), and the direction of that context effect was consistent with contrastive normalization (cf. Fig. 2d).

Extending this finding to the population of electrodes, we found a significant positive relation across all target-tuned temporal lobe electrodes between an electrode’s target preference and its context preference \( (\hat{p}_{\text{Cont F1}}; t = 0.76, t = 3.59, p = 9.83 \times 10^{-4}; \text{Fig. 3c}; \text{see Supplementary Materials for more detail on this analysis}) \). Hence, a neural population’s tuning for higher or lower F1 ranges tended to be general, not vowel-specific. Moreover, when restricting the test of normalization (assessed as the relationship between target preferences and the context effect, as per Fig. 2d) to those electrodes that displayed significant tuning for both target F1 and context F1, robust normalization was again found (Supplementary Fig. S5). These findings confirm that normalization affects acoustic-phonetic (i.e., pre-phonemic) representations of speech sounds in parabelt AC.

Discussion

A critical challenge that listeners must overcome in order to understand speech is the fact that different speakers produce the same speech sounds differently, a phenomenon that is known as the lack-of-invariance problem in speech perception. This issue is partly due to the fact that different speakers’ voices span different formant ranges. We investigated the neural underpinnings of how listeners use speaker-specific information in context to normalize phonetic processing. First, we observed behavioral normalization effects, replicating previous findings. More importantly, we observed normalized representations of vowels in parabelt AC. These normalized representations were observed broadly across parabelt AC and were observed for all participants individually. Moreover, we found that these effects follow the predictions of a general auditory contrast enhancement model of normalization, affecting speech sound representations at a level that precedes the mapping onto phonemes or higher-level
linguistic units. These findings suggest that contrast enhancement plays an important role in normalization.

Recent research has demonstrated that AC responds to acoustic cues that are critical for both recognizing and discriminating phonemes\textsuperscript{12,13,15–18,43} and different speakers\textsuperscript{46–50}. However, since cues that are critical for speaker and phoneme identification are conflated in the acoustic signal, such findings could be consistent with either context-dependent or context-independent cortical representations of acoustic properties. Our experiments demonstrate that rapid and broadly distributed normalization through contrast enhancement is a basic principle of how human AC encodes speech.

Qualitatively similar contrast enhancing operations have been widely documented in animal neurophysiological research, where it has been demonstrated to involve neural mechanisms such as adaptive gain control\textsuperscript{27–29,53} or stimulus specific adaptation\textsuperscript{28,29}. An intuitive mechanism for the implementation of contrast enhancement that follows from that work involves sensory adaptation. This could be based on neuronal fatigue: when a neuron, or neuronal population, responds strongly to a masker stimulus, its response to a subsequent probe is often attenuated when the frequency of the probe falls within the neurons’ excitatory receptive field\textsuperscript{52,53}. But in addition to such local forms of adaptation, adaptation has also been thought to arise through (inhibitory) interactions between separate populations of neurons (which may have partly non-overlapping receptive fields)\textsuperscript{27,51}. In the present study, spectral peaks in the two context sentences and those in the endpoint target vowels were partly overlapping (see Supplementary Fig. S1). These forms of adaptation may, hence, play a role in the type of normalization observed here. Indeed, we observed a number of populations for which a strong preference for one of the context sentences during the context window was associated with a decreased response during the target window (i.e., the normalization effect; Fig. 3b).

The prediction that contrast enhancement may play an important role in human speech sound normalization was previously made based on behavioral studies on contrastive context effects in speech perception\textsuperscript{10,11,23,30,34,54}. A relevant observation from that literature is that, under specific conditions, nonspeech context sounds (e.g., broadband noise and musical tones) have also been observed to affect the perception of speech sounds\textsuperscript{11,22,23,31,55}. This interpretation has been challenged, however, suggesting that speech- and nonspeech-based context effects could be based on qualitatively different processes\textsuperscript{56,57}. We did not test the influence of nonspeech contexts here, and such an investigation would provide an important next step in the study of context effects in speech perception. However, our findings most closely align with a model that assumes that normalization effects may not be speech-specific and that normalization can, at least in part, be explained by more general auditory adaptation effects\textsuperscript{23,34,58}.

An interesting additional question concerns the main locus of emergence of normalization. Broadly speaking, normalization could be inherited from primary AC or subcortical regions (from which we were unable to record; see\textsuperscript{59} for a more detailed discussion of these potential influences); it may largely emerge within parabelt AC itself; or it could be driven by top-down influences from regions outside of the AC. In our study, context and target sounds were separated in time by a 500 ms silent interval. It has been suggested that, over such relatively long latencies, adaptation effects become especially dominant at cortical levels of processing but are reduced at more peripheral levels of processing\textsuperscript{55,80,61}. Furthermore, behavioral experiments have demonstrated robust normalization effects with contralateral presentation of context and target sounds\textsuperscript{11,31,52} (i.e., a procedure that reduces precortical interaural interactions). Both observations thus suggest that normalization can also arise when the contribution of context-target interactions in the auditory periphery may be limited. With respect to the potential role of top-down modulations from regions outside of the AC, inferior frontal and sensorimotor cortex have been suggested to be involved in the resolution of perceptual ambiguities in speech perception\textsuperscript{62,63} and could, hence, have been expected to play a role in normalization, too. Here, we observed considerable activation in these regions, but, intriguingly, they did not display normalization during the processing of the target sounds (see Supplementary Fig. S4). While tentative, these combined findings highlight human AC as the most likely locus for the emergence of the context effects in speech processing observed here.

In the current experiment, we recorded neural activity from cortical sites in both the left and right hemispheres. It has previously been demonstrated that the right hemisphere is more strongly involved in the processing of voice information\textsuperscript{64,65}. Here, normalization was observed in every participant, irrespective of which hemisphere was the source of a given participant’s recordings (Fig. 2f). Importantly, however, recordings from any given participant only included measurements from a single hemisphere, so no strong conclusions regarding lateralization should be drawn based on this dataset.

Despite normalization of vowel representations, responses were not completely invariant to speaker differences during the context sentences (see, for example, the behavior of example electrode e2 in Fig. 3b, which displays a preference for the Low F1 sentence throughout most of the context window: i.e., it is not fully normalized). And indeed, our (and previous\textsuperscript{7,11,36}) findings show that, even for target sound processing, surrounding context rarely (if ever) results in complete normalization. That is, while the context sentence F1 differed by roughly 400 Hz between the two speakers, the normalization effect only induced a shift of ~50 Hz in the position of the category boundaries (in behavior and in neural categorization). Moreover, our target F1 values were ideally situated between the two context F1 ranges, which raises the question of whether equally large effects would have been observed for other target F1 ranges. The role of contrast enhancement in normalization should thus be seen as a mechanism that biases processing in a context-dependent direction, but not one that fully normalizes processing. Furthermore, context-based normalization is not the only means by which listeners tune-in to specific speakers: listeners categorize sound continua differently when they are merely told they are listening to a man or a woman, demonstrating the existence of normalization mechanisms that do not rely on acoustic contexts (and hence acoustic contrast) at all\textsuperscript{66}. In addition, formant frequencies are perceived in relation to other formants and pitch values in the current signal, because those features are correlated within speakers (e.g., people with long vocal tracts typically have lower pitch and lower formant frequencies, overall). These intrinsic normalization mechanisms have been shown to affect AC processing of vowels as well\textsuperscript{67–71}. Tuning-in to speakers in everyday listening must thus result from a combination of multiple mechanisms, involving at least these three distinct types of normalization\textsuperscript{8}.

To conclude, the results presented here reveal that the processing of vowels in AC becomes rapidly influenced by speaker-related acoustic properties in preceding context. These findings add to a recent literature that is ascribing a range of complex acoustic integration processes to the broader AC, suggesting that it participates in high-level encoding of speech sounds and auditory objects\textsuperscript{13,19,72–74}. Recently, it has been demonstrated that populations in parabelt AC encode speaker-invariant contours of intonation that speakers use to focus on one or the other part of a sentence\textsuperscript{20}. The current findings build on these and
demonstrate the emergence of speaker-normalized representa-
tions of acoustic-phonetic features, the most fundamental build-
ing blocks of spoken language. This context-dependence allows
AC to partly resolve the between-speaker variance present in
speech signals. These features of AC processing underscore its
critical role in our ability to understand speech in the complex
and variable situations that we are exposed to every day.

Methods
Patients. A total of five human participants (2 male; all right-handed; mean age:
30.6 years), all native Spanish-speaking (the US hospital at which participants were
recruited has a considerable Spanish-speaking patient population), were chroni-
cally implanted with high-density (256 electrodes; 4 mm pitch) multi-electrode
recording arrays subdurally on the left posterior temporal lobe. Patients
were only. All participants gave their written informed consent before the surgery, and
had self-reported normal hearing. The study protocol was approved by the UC San
Francisco Committee on Human Research. Electrode positions for reconstruction
figures were extracted from tomography scans and co-registered with the
patient’s MRI.

Stimulus synthesis. Details of the synthesis procedure for these stimuli have been
reported previously27. All synthesis was implemented in Praat software29. In brief,
using source-filter separation, the formant tracks of multiple recordings of clear
“suﬁ” and “sofo” were estimated. These estimates were used to calculate a single
average time-varying formant track for both words, now representing an average of
the formant properties over a number of instances of [su] and [so]. The height of
the formant model was increased and decreased across the whole vowel to create the new formant models for the continuum from [su] to [so]
covering the distance between endpoints in six steps. These formant tracks were
combined with a model of the glottal-pulse source to synthesize the speech sound
continuum. Synthesis parameters thus dictated that all steps were equal in pitch
contour, amplitude contour and had identical contours for the formants higher
than F1 (note that F1 and F2 values in Fig. 1a and S1 reﬂect measurements of the
resulting sounds, not synthesis parameters). The two context conditions were
created through source–filter separation of a single spoken utterance of the sentence
“a veces se halla” (“at times she feels rather...”). The first formant of the filter model
was then increased or decreased by 100 Hz and recombined with the source model
following similar procedures as for the targets.

Procedures. The participants were asked to categorize the last words of a stimulus as
either suﬁ or sofo. Listeners responded using the two buttons on a button box. The
two options suﬁ and sofo were always displayed on the computer screen. Each of
the 6 steps of the continuum was presented in both the right- and left-
ended sentence conditions. Context conditions were presented in separate mini-blocks
of 24 trials (6 steps × 4 repetitions). Participants participated in as many blocks as
they felt comfortable with (see Table S1 for trial counts).

Data acquisition and preprocessing. Cortical local ﬁeld potentials were recorded and
ampliﬁed with a multichannel ampliﬁer optically connected to a digital signal
acquisition system (Tucker-Davis Technologies) sampling at 3052 Hz. The stimuli
were presented monaurally from loudspeakers at a comfortable level. The ambient
audio (recorded with a microphone aimed at the participant) along with a direct
audio signal of stimulus presentation were simultaneously recorded with the ECoG
signals to allow for precise alignment and later inspection of the experimental
situation. Line noise (60 Hz and harmonics at 120 and 180 Hz) was removed from
the ECoG signals with notch ﬁlters. Each time series was visually inspected for
excessive noise, and trials and or channels with excessive noise or epileptiform
activity were removed from further analysis. The remaining time series were
common-average referenced across rows of the 16 × 16 electrode grid. The time-
varying analytic amplitude was extracted from eight bandpass ﬁlters (Gaussian,
with logarithmically increasing center frequencies between 70 and 150 Hz, and
semilogarithmically increasing bandwidths) with the Hilbert transform. High-
gamma power was calculated by averaging the analytic amplitude across these eight
bands. The signal was subsequently downsampled to 100 Hz. The signal was z-
scored based on the mean and standard deviation of a baseline period (from −30 to
0 ms before the onset of the context sentence) on a trial by trial basis. In the main
text, high-γ will refer to this measure.

Single-electrode encoding analysis. We used ordinary least-squares linear
regression to predict neural activity (high-γ) from our stimulus conditions (target
F1 steps, coded as −2.5, −1.5, −0.5, 0, 0.5, 1.5, 2.5; and context F1, coded as 0 and
1; as well as their interaction). These factors were used as numerical predictors to
neural responses, which were generated across the context window (from 250 to
1450 ms after context sentence onset)—a later onset was chosen to reduce the influence of large
and non-selective onset responses present in some electrodes). For each model,
R-squared (R2) provides a measure of the proportion of variance in neural activity
explained by the model. The p value associated with the omnibus F-statistic
provides a measure of significance. We set the significance threshold at
alpha = 0.05 and corrected for multiple comparisons using the Bonferroni method,
taking individual electrodes as independent samples. Supplementary Figure S2a, b
demonstrates that most of the variance in the context was explained by the factor
context F1. However, both target F1 and context F1 explained a considerable portion of the variance. The interaction term was included to accommodate a situation where the context effect is more strongly expressed on
one side of the target continuum than the other (see e.g., Fig. 2b, where the context
effect is larger toward sofo), but is not further interpreted here.

For Fig. 2d and Fig. 3c, linear mixed regression analyses were used to assess the relation between signed t-statistics of target F1 preferences and context effects
(Fig. 2d) or context preferences (Fig. 3c). Regression estimates were computed over
all signiﬁcant (9 [corrected] + 28 [uncorrected] = 37) electrodes. Linear mixed effects regression accommodates the hierarchical nature of these observations
(electrodes within patients).

Cluster-based permutation analyses. For single-electrode electrodes, a cluster-
based permutations approach was used to assess signiﬁcance of differences between
two event related high-gamma time courses (Fig. 3c and Fig. 3b, following the
method described in ref. 79). For each permutation, labels of individual trials were
randomly assigned to data (high-gama time courses), and a t test was performed for
each timepoint. Next, for each timepoint (across all 1000 permutations), a criterion
value was calculated (the highest 95% of [the absolute] t values for that timepoint). If, then, for each permutation, it was established that t was reached
above the criterion value and for how many samples it remained above criterion. A
set of subsequent timepoints above criterion is deﬁned as a cluster. Then, for each
cluster the t values were summed, and this value was assigned to that entire cluster.
For each permutation the largest (i.e., highest summed cluster value) was
stored as a single value. This resulted in a distribution of maximal 1000 cluster
values (some permutations may not result in any signiﬁcant cluster and have a
summed t value of 0). Then, using the same procedure, the size of all potential
clusters was established for the real data (correct assignment of labels), and it
was established whether the size of each cluster was larger than 95% of the
permutation-based cluster values. p < 0.001 indicates that the observed cluster
was larger than all permutation based clusters.

Stimulus classiﬁcation. Linear discriminant analysis (LDA) models were trained to
predict the stimulus from the neural population responses evoked by the stimuli.
Per participant a single model was trained on all endpoint data, which was then
used to predict labels for the ambiguous items. To predict stimulus class for the
endpoint stimuli (steps 1 and 6) a leave-one-out cross validation procedure was
used to prevent overfitting. Model features (predictors) consisted of the selected
timepoint-electrode combinations per participant. Prediction was made in a distribution of maximally 1000 cluster
values (some permutations may not result in any signiﬁcant cluster and have a
summed t value of 0). Then, using the same procedure, the size of all potential
clusters was established for the real data (correct assignment of labels), and it
was established whether the size of each cluster was larger than 95% of the
permutation-based cluster values. p < 0.001 indicates that the observed cluster
was larger than all permutation based clusters.

In all analyses, all task-related electrodes for a given participant (and region-of-
interest, see Fig. S4) were selected. Trial numbers per participant are listed in
Table S1. The analysis displayed in Fig. 2 and Supplementary Figs. S3 and S4 hence
relied on a large number of predictors (electrodes × timepoints). While a large
amount of predictors could result in overﬁtting, these parameters led to the highest
proportion of correct classiﬁcation for the endpoints: 76% correct (see
Supplementary Fig. S1b), although note that this number may be inﬂuenced because of
electrode pre-selection. This approach was chosen, however, because high
endpoint classiﬁcation performance is important to establish the presence of
normalization: a shift in a response function can only be detected if the steepness of
that function is nonzero. Importantly, speciﬁcally selecting electrodes that
distinguish the endpoints does not affect the extent of observed normalization,
because the normalization effect is orthogonal to that of target F1 (i.e.,
normalization itself was not a selection criterion). Furthermore, in all analyses,
classiﬁcation scores were only obtained from held-out data, preventing the ﬁtting of
idiomsyntactic models. In addition, averaging across time (hence decreasing the
number of predictors) led to qualitatively similar (and signiﬁcant) effects for the
important comparisons reported in this paper. Classiﬁcation analyses resulted in a
predicted class for each trial. These values were used as input for a generalized
logistic linear mixed effects model.

Linear mixed effects regression of classiﬁcation data. For the analyses that
assessed the effects of target stimulus F1 and context F1 on proportion of “sofo”
responses (both behavioral and neural-classifier-based) we employed generalized
linear mixed effects models (glm; with the dependent variable “family” set to
“binomial”). The models had Target F1 (contrast coded, with the levels −2.5; −1.5;
−0.5; 0.5; 1.5; 2.5) and Context F1 (levels = 1; 0) entered as ﬁxed effects, and
uncorrelated by-patient slopes and intercepts for these factors as random effects.
For the analysis of the behavioral data ($\beta_{\text{intercept}} = 0.52$, $t = 0.79$, $p = 0.42$), we observed more soft responses towards the soft end of the stimulus continuum ($\beta_{\text{target}} = 1.86$, $t = 4.10$, $p < 0.001$). Moreover, we observed an effect of context as items along the continuum were more often perceived as soft (the vowel category corresponding to higher-F1 values) after a low-F1 voice (Speaker A) than after a high-F1 voice (Speaker B), $\beta_{\text{context}} = -1.70$, $t = -3.31$, $p < 0.001$.

For the analyses of neural representations the dependent variable consisted of the classes predicted by LDA stimulus classification described above. For the overall analysis including temporal lobe electrodes, the model ($\beta_{\text{intercept}} = -1.06$, $t = -0.29$, $p = 0.27$) revealed significant classification of the continuum ($\beta_{\text{target}} = 0.50$, $t = 13.20$, $p < 0.001$) suggesting reliable neural differences between the endpoints. Furthermore, an effect was also found for the factor Context on the proportion of “soft” classifications ($\beta_{\text{context}} = -0.34$, $t = -5.58$, $p < 0.001$), reflecting the normalization effect of most interest. For the analysis focusing on the dorsal and frontal electrodes ($\beta_{\text{intercept}} = -1.6$, $t = 0.91$, $p = 0.37$) a significant effect of Step was observed; that is, significant classification of the continuum ($\beta_{\text{target}} = 0.23$, $t = 7.01$, $p < 0.001$), but no significant influence of context ($\beta_{\text{context}} = -0.02$, $t = -0.39$, $p = 0.69$) see Supplementary Fig. S4C for further detail.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

The data that support the findings of this study are publicly available through the Open Science Framework at https://osf.io/t87d2/.

**Code availability**

These results were generated using code written in Matlab. Code is publicly available through the Open Science Framework at https://osf.io/t87d2/.

Received: 19 September 2018 Accepted: 3 May 2019

Published online: 06 June 2019

**References**

1. Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. Perception of the speech code. *Psychol. Rev.* 74, 431–461 (1967).

2. Diehl, R. L., Lotto, A. J. & Holt, L. L. Speech perception. *Annu Rev. Psychol.* 55, 149–179 (2004).

3. Peterson, G. E. & Barney, H. L. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175–184 (1952).

4. Newman, R. S., Clouse, S. A. & Burnham, J. L. The perceptual consequences of within-talker variability in fricative production. *J. Acoust. Soc. Am.* 109, 1181–1196 (2001).

5. Chodroff, E. & Wilson, C. Structure in talker-specific phonetic realization: covariation of stop consonant VOT in American English. *J. Phon.* 41, 30–47 (2013).

6. Ladefoged, P. & Johnson K. A Course in Phonetics. (Cengage Learning, Stanford, 2014).

7. Sjörs, M. J. & Smiljanic, R. Compensation for vocal tract characteristics across native and non-native languages. *J. Phon.* 41, 145–155 (2013).

8. Nearey, T. M. Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* 85, 2088–2113 (1989).

9. Ladefoged, P. & Broadbent, D. E. Information conveyed by vowels. *J. Acoust. Soc. Am.* 29, 98–104 (1957).

10. Laing, E. J. C., Liu, R., Lotto, A. J. & Holt, L. L. Tuned with a tune: talker normalization via general auditory processes. *J. Exp. Psychol. Hum. Percept. Perform.* 44, 914–924 (2018).

11. Holt, L. L. & Lotto, A. J. Behavioral examinations of the level of auditory processing of speech context effects. *Hear. Res.* 167, 156–169 (2002).

12. Holt, L. L. & Kluender, K. R. General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Percept. Psychophys.* 60, 602–619 (1998).

13. Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H. & King, A. I. Contrast gain control in auditory cortex. *Neuron* 70, 1178–1191 (2011).

14. McAlpine, N., Las, L. L., Farkas, D. & Nelken, I. Multiple time scales of adaptation in auditory cortex neurons. *J. Neurosci.* 24, 10440–10453 (2004).

15. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Science* 318, 1577–1585 (2007).

16. Boatman, D., Lesser, R. P. & Gordon, B. Auditory speech perception in the left temporal lobe: an electrical interference study. *Brain Lang.* 51, 289–290 (1995).

17. Scott, S. K. & Johnsonrude, I. S. The neuroanatomical and functional organization of speech perception. *Trends Neurosci.* 26, 100–107 (2003).

18. Steinschneider, M. et al. Intraindividual study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cereb. Cortex* 21, 2332–2347 (2011).

19. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010 (2014).

20. Tang, C., Hamilton, L. S. & Chang, E. F. Intonational speech prosody encoding in the human auditory cortex. *Science* 357, 797–801 (2017).

21. Laubach, P. et al. The spectrotemporal filter mechanism of auditory selective attention. *Neuron* 77, 750–761 (2013).

22. Hickok, G. & Poeppel, D. Intraindividual study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cereb. Cortex* 21, 2332–2347 (2011).

23. Kalenick, D. E. & Broadbent, D. E. Information conveyed by vowels. *J. Acoust. Soc. Am.* 60, 808–810 (1976).

24. Newman, R. S., Clouse, S. A. & Burnham, J. L. The perceptual consequences of within-talker variability in fricative production. *J. Acoust. Soc. Am.* 109, 1181–1196 (2001).

25. Sjörs, M. J. & Smiljanic, R. Compensation for vocal tract characteristics across native and non-native languages. *J. Phon.* 41, 145–155 (2013).

26. Nearey, T. M. Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* 85, 2088–2113 (1989).

27. Ladefoged, P. & Broadbent, D. E. Information conveyed by vowels. *J. Acoust. Soc. Am.* 29, 98–104 (1957).

28. Laing, E. J. C., Liu, R., Lotto, A. J. & Holt, L. L. Tuned with a tune: talker normalization via general auditory processes. *Front Psychol.* 3, 1–9 (2012).

29. Watkins, A. J. Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *J. Acoust. Soc. Am.* 90, 2942–2955 (1991).

30. Creutzfeldt, O., Ojemann, G. A. & Lettich, E. Neuronal activity in the human lateral temporal lobe: I. Responses to speech. *Exp. Brain Res.* 77, 451–475 (1989).

31. Chang, E. F. et al. Categorial speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432 (2010).

32. Formisano, E., De Martino, F., Bonte, M. & Goebel, R. “Who Is Saying ‘What’? Brain-based decoding of human voice and speech. *Science* 322, 970–973 (2008).

33. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402 (2007).

34. Boatman, D., Lesser, R. P. & Gordon, B. Auditory speech processing in the left temporal lobe: an electrical interference study. *Brain Lang.* 51, 289–290 (1995).

35. Scott, S. K. & Johnsonrude, I. S. The neuroanatomical and functional organization of speech perception. *Trends Neurosci.* 26, 100–107 (2003).
53. Smith, R. L. Short-term adaptation in single auditory nerve fibers: some poststimulatory effects. *J. Neurophysiol.* 40, 1988–1111 (1977).
54. Sjerps M. J., McQueen J. M. & Mitterer H. Evidence for precategorical extrinsic vowel normalization. *Atten. Percept. Psychophys.* 75, 576–587 (2013).
55. Holt, L. L. Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychol. Sci.* 16, 305–312 (2005).
56. Viswanathan, N., Magnuson, J. S. & Fowler, C. A. Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *J. Exp. Psychol. Hum. Percept. Perform.* 36, 1005–1015 (2010).
57. Viswanathan, N., Magnuson, J. S. & Fowler, C. A. Similar response patterns do not imply identical origins: an energetic masking account of nonspeech effects following reliable spectral properties in speech perception. *J. Acoust. Soc. Am.* 137, 3466–3476 (2015).
58. Stilp, C. E., Anderson, P. W. & Winn, M. B. Predicting contrast effects following reliable spectral properties in speech perception. *J. Acoust. Soc. Am.* 137, 3466–3476 (2015).
59. Stilp, C. E. & Assgari, A. A. Perceptual sensitivity to spectral properties of earlier sounds during speech categorization. *Atten. Percept. Psychophys.* 80, 1300–1310 (2018).
60. Phillips, E. A. K., Schreiner, C. E. & Hasenstaub, A. R. Cortical interneurons differentially regulate the effects of acoustic context. *Cell Rep.* 20, 771–778 (2017).
61. Fitzpatrick, D. C., Kuwada, S., Kim, D. O., Parham, K. & Batra, R. Responses of neurons to click-pairs as simulated echoes: auditory nerve to auditory cortex. *J. Acoust. Soc. Am.* 106, 3460–3472 (1999).
62. Pulvermüller, F. et al. Motor cortex maps articulatory features of speech sounds. *Proc. Natl Acad. Sci.* 103, 7865–7870 (2006).
63. Wilson, S. M. & Iacoboni, M. Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. *Neuroimage* 33, 316–325 (2006).
64. Myers, E. B. & Theodore, R. M. Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain Lang.* 165, 33–44 (2017).
65. Belin, P. & Zatorre, R. J. Adaptation to speaker’s voice in right anterior temporal lobe. *Neuroreport* 14, 2105–2109 (2003).
66. Johnson, K., Strand, E. A. & D’Imperio, M. Auditory-visual integration of talker gender in vowel perception. *J. Phon.* 27, 359–384 (1999).
67. Edmonds, B. A. et al. Evidence for early specialized processing of speech formant information in anterior and posterior human auditory cortex. *Eur. J. Neurosci.* 32, 684–692 (2010).
68. Andermann, M., Patterson, R. D., Vogt, C., Winterstetter, L. & Rupp, A. Neuromagnetic correlates of voice pitch, vowel type, and speaker size in auditory cortex. *Neuroimage* 158, 79–89 (2017).
69. Monahan, P. J. & Idsardi, W. J. Auditory sensitivity to formant ratios: toward an account of vowel normalisation. *Lang. Cogn. Process* 25, 808–839 (2010).
70. Kreiterhoff, J., Gaudrain, E. & von Kriegstein, K. A neural mechanism for recognizing speech spoken by different speakers. *Neuroimage* 91, 375–385 (2014).
71. Kriegstein, K., von, Smith, D. R. R., Patterson, R. D., Kiebel, S. J. & Griffiths, T. D. How the human brain recognizes speech in the context of changing speakers. *J. Neurosci.* 30, 629–638 (2010).
72. Engineer, C. T. et al. Cortical activity patterns predict speech discrimination ability. *Nat. Neurosci.* 11, 603–608 (2008).
73. Leonard, M. K., Baud, M. O., Sjerps, M. J. & Chang, E. F. Perceptual restoration of masked speech in human cortex. *Nat. Commun.* 7, 13619 (2016).
74. Bizley, J. K., Walker, K. M. M., Nodal, F. R., King, A. J. & Schnupp, J. W. H. Auditory cortex represents both pitch judgments and the corresponding acoustic cues. *Curr. Biol.* 23, 620–625 (2013).
75. Boersma P., Weenink D. Praat: Doing Phonetics by Computer (Version 5.1). 2009.
76. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190 (2007).

**Acknowledgements**

We are grateful to Matthew K. Leonard for commenting on an earlier version of this manuscript and to all members of the Chang Lab for helpful comments throughout this work. This work was supported by European Commission grant FP7-623072 (M.J.S.) and NIH grants R01-DC012379 (E.F.C.) and P32-DC015966 (N.P.F.). E.F.C. is a New York Stem Cell Foundation- Robertson Investigator. This research was also supported by The William K. Bowes Foundation, the Howard Hughes Medical Institute, The New York Stem Cell Foundation and The Shurl and Kay Curci Foundation.

**Author contributions**

M.J.S. and K.J. conceived the study. M.J.S. designed the experiments, generated the stimuli, and analyzed the data. M.J.S., N.P.F. and E.F.C. collected the data. M.J.S. and N.P.F. interpreted the data and wrote the paper. M.J.S., N.P.F., K.J. and E.F.C. edited the paper.

**Additional information**

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-10365-z.

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

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**Journal peer review information:** Nature Communications thanks Jonas Oblerse and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2019