LETTER

doi:10.1038/nature11020

Selective cortical representation of attended speaker in multi-talker speech perception

Nima Mesgarani[1] & Edward F. Chang[1]

Humans possess a remarkable ability to attend to a single speaker's voice in a multi-talker background[1–3]. How the auditory system manages to extract intelligible speech under such acoustically complex and adverse listening conditions is not known, and, indeed, it is not clear how attended speech is internally represented[4,5]. Here, using multi-electrode surface recordings from the cortex of subjects engaged in a listening task with two simultaneous speakers, we demonstrate that population responses in non-primary human auditory cortex encode critical features of attended speech: speech spectrograms reconstructed based on cortical responses to the mixture of speakers reveal the salient spectral and temporal features of the attended speaker, as if subjects were listening to that speaker alone. A simple classifier trained solely on examples of single speakers can decode both attended words and speaker identity. We find that task performance is well predicted by a rapid increase in attention-modulated neural selectivity across both single-electrode and population-level cortical responses. These findings demonstrate that the cortical representation of speech does not merely reflect the external acoustic environment, but instead gives rise to the perceptual aspects relevant for the listener's intended goal.

Separating out a speaker of interest from other speakers in a noisy, crowded environment is a perceptual feat that we perform routinely. The ease with which we hear under these conditions belies the intrinsic complexity of this process, known as the cocktail party problem[1–3,6]: concurrent complex sounds, which are completely mixed upon entering the ear, are re-segregated and selected from within the auditory system. The resulting percept is that we selectively attend to the desired speaker while tuning out the others.

Although previous studies have described neural correlates of masking and selective attention to speech[4,5,7–9], fundamental questions remain unanswered regarding the precise nature of speech representation at the juncture where competing signals are resolved. In particular, when attending to a speaker within a mixture, it is unclear what key aspects (for example, spectrotemporal profile, spoken words and speaker identity) are represented in the auditory system and how they compare to representations of that speaker alone; how rapidly a selective neural representation builds up when one attends to a specific speaker; and whether breakdowns in these processes can explain distinct perceptual failures, such as the inability to hear the correct words, or follow the intended speaker.

To answer these questions, we recorded cortical activity from human subjects implanted with customized high-density multi-electrode arrays as part of their clinical work-up for epilepsy surgery[10]. Although limited to this clinical setting, these recordings provide simultaneous high spatial and temporal resolution while sampling the population neural activity from the non-primary auditory speech cortex in the posterior superior temporal lobe. We focused our analysis on high gamma (75–150 Hz) local field potentials[11], which have been found to correlate well with the tuning of multi-unit spike recordings[12]. In humans, the posterior superior temporal gyrus has been heavily implicated in speech perception[13], and is anatomically defined as the lateral parabelt auditory cortex (including Brodmann areas 41, 42 and 22)[14].

Subjects listened to speech samples from a corpus commonly used in multi-talker communication research[15,16]. A typical sentence was "ready tiger go to red two now" where "tiger" is the call sign, and "red two" is the colour–number combination. One male and one female speaker were selected, each speaking the same 12 unique combinations of two call signs (ringo or tiger), three colours (red, blue or green) and three numbers (two, five or seven). Example acoustic spectrograms from two individual speakers are shown in Fig. 1a, b. The two voices differ along several dimensions including pitch (male versus female), spectral profile (different vocal track shapes) and temporal characteristics (speaking rate). Subjects first listened to each of the speakers alone and were able to report the colour and number with 100% accuracy. Subjects then listened to a monaural, simultaneous mixture of the two speakers' phrases with different call signs, colours and numbers. The subjects were instructed to respond by indicating the colour and number spoken by the talker who uttered the target call sign. The target call sign (ringo or tiger) was fixed and shown visually on a monitor during each trial block, which contained 28 different mixture sounds. As the target speaker was changed randomly from trial to trial, the subjects were required to monitor both voices initially (divided attention) to identify the target speaker. The target call sign was switched after each block, turning the previous target speaker in each mixture into a masker. This resulted in two sets of behavioural and neural responses for each identical mixture sound, which differed only in the focus of attention. Subjects reported correct responses in 74.8% of trials.

Figure 1c illustrates the mixture spectrogram and how difficult it is to tell which sound parts belong to one speaker versus the other. The energy for both speakers is distributed broadly across the spectral and temporal domains, with overlap in some areas and isolated sound parts in others, as shown in their difference spectrogram (Fig. 1d; average spectrograms in Supplementary Fig. 1a).

To determine the spectrotemporal encoding of the attended speaker, the method of stimulus reconstruction was used[17–19] to estimate the speech spectrogram represented by the population neural responses. Reconstructed spectrograms provide an intuitive way to examine how the population neural responses encode the spectrotemporal features of speech, and more importantly, can be compared with the original acoustic spectrograms as well as across attentional conditions. We first calculated the reconstruction filters from a passive listening task using a separate continuous speech corpus (TIMIT[20]) that consisted of 499 unique short sentences spoken by 402 different speakers. The filters were then fixed and applied to a novel set of population neural responses to the single and attended mixture speech for spectrogram reconstruction.

When listening to a single speaker alone, the reconstructed spectrograms from population neural activity corresponded well to the spectrotemporal features of the original acoustic spectrograms (Fig. 1e, f compared to Fig. 1a, b, respectively), exhibiting fairly precise temporal features and spectral selectivity (for example, correspondence between the high frequency bursts of energy in "tiger" and "two", in Fig. 1a, b, e, f).

[1]Departments of Neurological Surgery and Physiology, UCSF Center for Integrative Neuroscience, University of California, San Francisco, California 94143, USA.

©2012 Macmillan Publishers Limited. All rights reserved

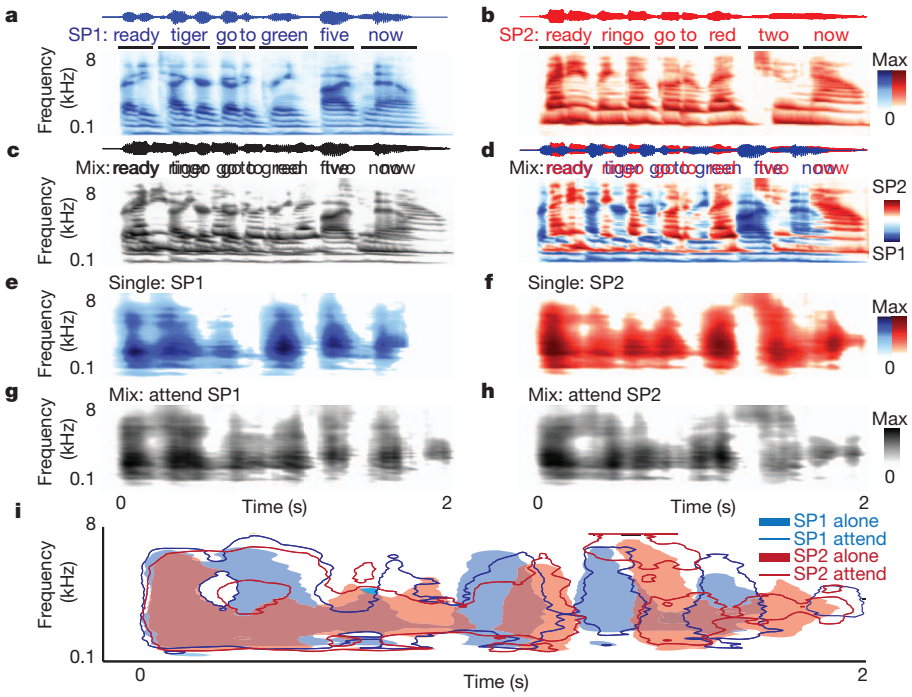

Figure 1 | Acoustic and neural reconstructed spectrograms for speech from a single speaker or a mixture of speakers. **a**, **b**, Example acoustic waveform and auditory spectrograms of speaker one (male; **a**) and speaker two (female; **b**). **c**, Waveform and spectrogram of the mixture of the two shows highly overlapping energy distributions. **d**, Difference spectrogram highlights the mixture regions where speaker one (blue) or two (red) has more acoustic energy. **e**, **f**, Neural-population-based stimulus reconstruction of speaker one (**e**) and speaker two (**f**) alone shows similar spectrotemporal features as the original spectrograms in **a** and **b**. **g**, **h**, The reconstructed spectrograms from the same mixture sound when attending to either speaker one (**g**) or two (**h**) highly resemble the single speaker reconstructions, shown in **e** and **f**, respectively. **i**, Overlay of the spectrogram contours at 50% of maximum energy from the reconstructed spectrograms in **e**, **f**, **g** and **h**.

The average and standard deviation of the correlation between reconstructed and original spectrograms over 24 sentences were 0.60 ± 0.034 (0.60 and 0.62 for the examples in Fig. 1e, f). When attending to each of the two speakers, the reconstructed spectrograms from the same speech mixture showed a marked difference depending upon which speaker was attended (Fig. 1g, h). For each pair, the key temporal and spectral features of the target speaker are enhanced relative to the masker speaker (Fig. 1g, h compared to Fig. 1e, f, respectively). To compare directly, the energy contours from these reconstructed spectrograms are overlaid in Fig. 1i. Important spectrotemporal details of the attended speaker were extracted, while the masker speech was effectively suppressed.

Attentional modulation of the neural representation was quantified, separately for correct and error trials, by measuring the correlation of the reconstructed spectrograms from the mixture in two attended conditions with original acoustic spectrograms of the speakers alone (Fig. 2a–d). During correct trials (Fig. 2a, c), we observed a significant shift of average correlation values towards the target speaker representation. During error trials, in contrast, no significant shift was observed (Fig. 2b, d). Furthermore, the correlations between the reconstructed mixture and the masker speaker were higher than the average intrinsic correlation between randomly chosen original acoustic speech phrases (Fig. 2c, d, dashed lines), revealing a weak presence of the masker speaker in mixture reconstructions, even in correct trials.

The difference in speaking rate of the two speakers, coupled with the stereotyped structure of the carrier phrases, results in specific average temporal modulation profiles for each speaker (average spectrogram for each speaker is shown in Supplementary Fig. 1a, b). To investigate encoding of the distinct spectral profile and characteristic temporal rhythm of the target compared to the masker speaker, we estimated the average difference between reconstructed spectrograms of the two speakers, when presented alone and in the attended mixture (Fig. 2e, f). The comparison between the two average difference reconstructed spectrograms reveals enhanced encoding of both temporal and spectral aspects of the attended speaker (Supplementary Fig. 1c, d). To study the time course of attention-induced modulation of reconstructed mixture spectrograms towards the attended speaker, we

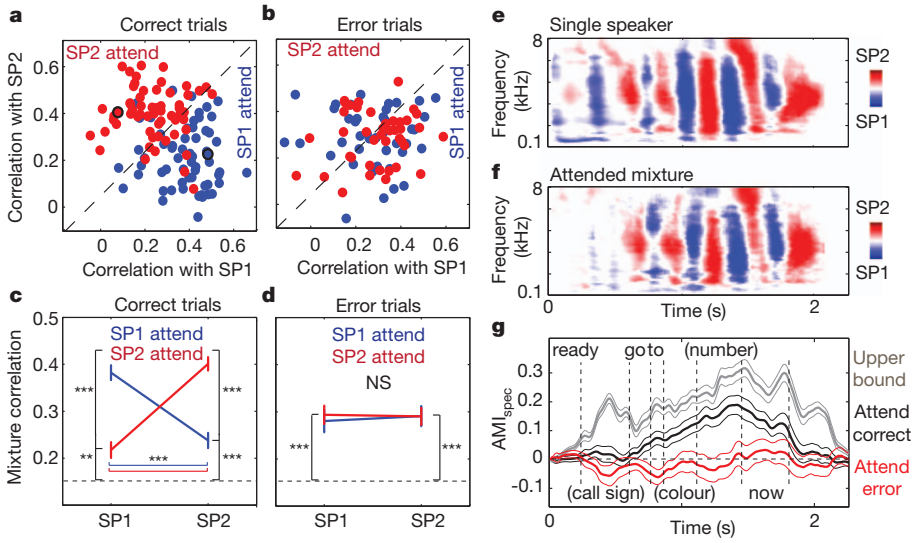

Figure 2 | Quantifying the attentional modulation of neural responses. **a**, **b**, Correlation coefficients of reconstructed mixture spectrograms under attentional control and the corresponding single speaker original spectrograms in correct and error trials (examples in Fig. 1g, h shown with black outline). **c**, **d**, Mean and standard error of correlation values for correct and error trials (28 mixtures). The dashed line corresponds to the average intrinsic correlation between randomly chosen original speech phrases. Brackets indicate pairwise statistical comparisons. NS, not significant. **e**, **f**, Average difference reconstructed spectrograms of speakers one and two from responses to single speaker (**e**) and attended mixture (**f**). **g**, Time course of average and standard error of AMI_{spec} of 28 mixtures for correct (black) and error (red) trials. Grey curve shows the upper bound of AMI_{spec}.

©2012 Macmillan Publishers Limited. All rights reserved

calculated an attentional modulation index (AMI_{spec}), using a sliding window of 250 ms throughout the trial duration:

$$AMI_{spec} = Corr(SP1_{spec}, SP1_{attend}) - Corr(SP1_{spec}, SP2_{attend}) + Corr(SP2_{spec}, SP2_{attend}) - Corr(SP2_{spec}, SP1_{attend}) \quad (1)$$

where $SP1_{spec}$ and $SP2_{spec}$ are the original acoustic spectrograms of speakers one and two, respectively, and $SP1_{attend}$ and $SP2_{attend}$ are the spectrograms reconstructed from neural responses to the mixture with attended targets, speaker one and two, respectively. Positive values of this index reflect shifts towards the target, negative values reflect shifts to the masker representation, and values around zero reflect no shift ($AMI_{spec} = 0.58$ for the example in Fig. 1). An upper bound for the AMI_{spec} was calculated by assuming that attention, at best, restores the single speaker reconstructions of the target speaker (replacing $SP1_{attend}$ and $SP2_{attend}$ in equation (1) with $SP1_{alone}$ and $SP2_{alone}$; Fig. 2g, grey line). The AMI_{spec} from the mixture was first estimated from correct trials (Fig. 2g, black line), and could resolve the time point at which the reconstructed spectrograms were modulated by attention. After the end of the call sign, which cues the speaker that should be attended, a rapid positive shift in the AMI_{spec} was observed, implying the enhanced representation of the target speaker. In error trials, this effect shows a bias towards the masker speaker, which, in contrast, occurred far earlier in the time course. The neural response shift towards the masker, which occurs as early as the call sign, suggests that listeners had prematurely attended to the wrong speaker during those error trials.

Although the reconstruction analyses showed clear attention-based spectrotemporal modulation, we wanted to determine explicitly whether the attended speech in a mixture could be decoded from a model of a single speaker. A regularized linear classifier[21] was trained on neural responses to the single speakers and then used to decode both the spoken words and speaker identity of the attended speech mixture. To keep the chance performance at 50% across all comparisons, classification results were limited only to the choices that were present in each mixture. For correct trials, the colour and number of the attended speech were decoded with high accuracy (77.2% and 80.2%, $P < 10 \times 10^{-4}$, *t*-test; Fig. 3a). However, the decoding performance during error trials was significantly below chance (30.0%, 30.1%, $P < 10 \times 10^{-4}$, *t*-test; Fig. 3b), indicating a systematic bias towards decoding the words of the masker speaker. In addition, for correct trials, the call sign was classified at chance performance (Fig. 3a). However, for incorrect trials the classifier detected the masker call sign significantly more often than the target call sign (34.1%, $P < 10 \times 10^{-4}$, *t*-test; Fig. 3b), which again shows errors due to an early selection of the masker (incorrect) speaker.

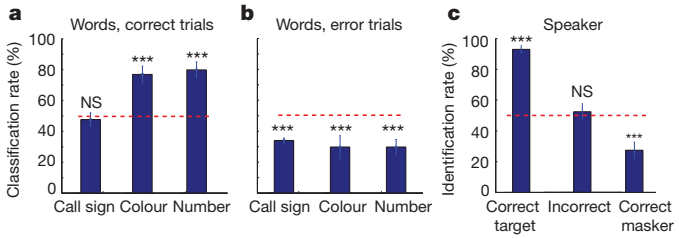

Figure 3 | Decoding spoken words and the identity of the attended speaker. **a**, Classification rate and standard deviation for spoken words (call sign, colour and number) of the attended speaker from the neural responses to the 28 mixtures. Classifiers were trained on single speaker examples only. Colour and number of the attended speech are decoded with high accuracy (77.2% and 80.2%, $P < 10 \times 10^{-4}$, *t*-test) in correct trials, but not the call sign (48.0%, not significant (NS), *t*-test). **b**, In error trials, the classifier showed a systematic bias towards the words of the masker speaker (34.1%, 30.0%, 30.1%, $P < 10 \times 10^{-4}$, *t*-test). **c**, Attended speaker identification rate and standard deviation in correct for target, incorrect (for both target and masker), and correct for masker trials.

For the speaker identification analyses, we divided the behavioural error types into two subsets. The first type occurred when the reported colour–number combination was incorrect for either speaker ('incorrect'; 16.5% of trials). The second type occurred when subjects reported the correct colour–number for the masker instead of the target speaker ('correct for masker'; 8.6% of trials).

In correct trials, the classifier identified the target speaker 93.0% of the time ($P < 10 \times 10^{-4}$, *t*-test; Fig. 3c). During incorrect trials, the classifier performance was at chance. However, during correct for masker trials, the classifier identified the masker rather than the target speaker (27.3%; $P < 10 \times 10^{-4}$, *t*-test; Fig. 3c). These classification results confirm the observed restoration seen in spectrotemporal reconstruction, without necessarily assuming a linear relationship between the neural responses and the stimulus. Furthermore, they extend recent findings using similar methods to decode speech sounds presented in isolation[22] to full words and sentences under complex listening conditions.

We next asked whether the observed robust encoding of attended speech results as an emergent property of the distributed population activity or is driven by a few spatially discrete sites. The cortical regions with reliable evoked responses to speech stimuli were found using a *t*-test between neural responses during speech and silence ($P < 0.01$), and were confined to the posterior superior and middle temporal gyri (Fig. 4a). An example of the attentional response modulation at a single electrode is shown in Fig. 4b–d. The spectrotemporal receptive field (STRF, estimated using the http://www.strflab.berkeley.edu package) of this electrode in passive listening to speech (TIMIT[20]) showed a strong preference for high frequency sounds (Fig. 4b) (STRFs for all electrodes of one subject are provided in Supplementary Fig. 2b). This tuning was also evident in the increased neural response at this electrode (Fig. 4d, dashed lines) to each of the single speakers' high frequency sound components (circled in Fig. 4c, responses are delayed about 120 ms from the stimulus). However, the responses to the same speech mixture sound (Fig. 4d, solid lines) were significantly modulated by attention. The responses to high frequency components were enhanced for the attended speaker, but suppressed for similar sounds in the masker speaker (Fig. 4d, solid lines compared to dashed lines). This highly modulated yet fixed feature selectivity probably contributes to the constancy of the single speaker representation observed in our previous analyses. To quantify this effect for each individual electrode, we measured the correlation between the neural responses to the attended mixture and to those of the speakers in isolation (AMI_{elec}, equation (2) in Methods). We found a varying degree of bias towards the attended speaker distributed across the population (Supplementary Fig. 3d; $AMI_{elec} = 0.28$ for the example

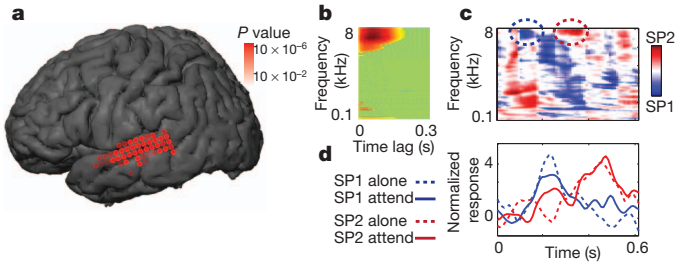

Figure 4 | Attentional modulation of individual electrode sites. **a**, Electrodes picking up a significant difference between responses to silence and speech sounds ($P < 0.01$, *t*-test). **b**, STRF of this representative electrode site shows a preference for high frequency sounds. **c**, Mixture difference spectrogram for a selected duration containing a high frequency component for each speaker (circled). **d**, The electrode shows an increased response to high frequency sounds of single speakers (dashed lines, peak neural response is delayed by about 120 ms). However, the neural response to the same mixture sound in two attention conditions (solid lines) showed an enhanced response to high frequency sounds only for the target, but with responses for similar sounds in the masker speaker suppressed.

©2012 Macmillan Publishers Limited. All rights reserved

in Fig. 4), which gradually builds up after the end of the call sign (Supplementary Fig. 3e). We did not observe any particular anatomical pattern for the attentional modulation across sites (Supplementary Fig. 3f). Rather, it appeared to be distributed over responsive sites, consistent with previous findings of higher-order sound processing[23].

In summary, we demonstrate that the human auditory system restores the representation of the attended speaker while suppressing irrelevant competing speech. Speech restoration occurs at a level where neural responses still show precise phase-locking to spectrotemporal features of speech. Population responses revealed the emergent representation of speech extracted from a mixture, including the moment-by-moment allocation of attentional focus.

These results have implications for models of auditory scene analysis. In agreement with recent studies, the cortical representation of speech in the posterior temporal lobe does not merely reflect the acoustical properties of the stimulus, but instead relates strongly to the perceived aspects of speech[10]. Although the exact mechanisms are not fully known, multiple processes in addition to attention are likely to enable this high-order auditory processing, including grouping of predictable regularities in speech acoustics[24], feature binding[3,25] and phonemic restoration[26]. Conversely, behavioural errors seem to result from degradation of the neural representation, a direct result of inherent sensory interference such as energetic masking[16] (Supplementary Fig. 3g, h) and/or the allocation of attention[27].

In speech, the end result represented in the posterior temporal lobe appears to be unaffected by perceptually irrelevant sounds, which is ideal for subsequent linguistic and cognitive processing. Following one speaker in the presence of another can be trivial for a normal human listener, but remains a major challenge for state-of-the-art automatic speech recognition algorithms[28]. Understanding how the brain solves this problem may inspire more efficient and generalizable solutions than current engineering approaches[29]. It will also shed light on how these processes become impaired during ageing and in disorders of speech perception in real-world hearing conditions[7].

METHODS SUMMARY

Three human subjects with normal hearing underwent the placement of a subdural electrode array as part of their clinical treatment for epilepsy. We used speech samples from a publicly available database called Coordinate Response Measure (CRM[15]). One male and one female speaker were selected with two call signs (ringo and tiger), three colours (red, blue or green) and three numbers (two, five or seven). We generated 12 unique combinations of call sign, colour and number per speaker (total of 24 single speaker phrases) and 28 mixture speech samples by selecting from combinations of the 24 single speaker sentences (0 dB target-to-masker ratio). Speech sounds were presented monaurally from a loud speaker. We used stimulus reconstruction[17–19] to map the population electrocorticographic response to the spectrogram of the speech stimulus. Reconstruction filters were estimated from neural responses to a separate speech corpus (TIMIT[20]). Test speakers were not used in the estimation of filters. For word and speaker decoding analysis, a regularized linear classifier[21] was trained on neural responses of the single speakers and then used to decode the spoken words and speaker identity of the attended speech mixture.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 August 2011; accepted 5 March 2012.
Published online 18 April 2012.

1. Cherry, E. C. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* **25,** 975–979 (1953).
2. Shinn-Cunningham, B. G. Object-based auditory and visual attention. *Trends Cogn. Sci.* **12,** 182–186 (2008).
3. Bregman, A. S. *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, 1994).
4. Kerlin, J., Shahin, A. & Miller, L. Attentional gain control of ongoing cortical speech representations in a ''cocktail party''. *J. Neurosci.* **30,** 620–628 (2010).
5. Besle, J. *et al.* Tuning of the human neocortex to the temporal dynamics of attended events. *J. Neurosci.* **31,** 3176–3185 (2011).
6. Bee, M. & Micheyl, C. The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? *J. Comparative Psychol.* **122,** 235–252 (2008).
7. Shinn-Cunningham, B. G. & Best, V. Selective attention in normal and impaired hearing. *Trends Amplif.* **12,** 283–299 (2008).
8. Scott, S. K., Rosen, S., Beaman, C. P., Davis, J. P. & Wise, R. J. S. The neural processing of masked speech: evidence for different mechanisms in the left and right temporal lobes. *J. Acoust. Soc. Am.* **125,** 1737–1743 (2009).
9. Elhilali, M., Xiang, J., Shamma, S. A. & Simon, J. Z. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biol.* **7,** e1000129 (2009).
10. Chang, E. F. *et al.* Categorical speech representation in human superior temporal gyrus. *Nature Neurosci.* **13,** 1428–1432 (2010).
11. Crone, N. E., Boatman, D., Gordon, B. & Hao, L. Induced electrocorticographic gamma activity during auditory perception. *Clin. Neurophysiol.* **112,** 565–582 (2001).
12. Steinschneider, M., Fishman, Y. I. & Arezzo, J. C. Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cereb. Cortex* **18,** 610–625 (2008).
13. Scott, S. K. & Johnsrude, I. S. The neuroanatomical and functional organization of speech perception. *Trends Neurosci.* **26,** 100–107 (2003).
14. Hackett, T. A. Information flow in the auditory cortical network. *Hear. Res.* **271,** 133–146 (2011).
15. Bolia, R. S., Nelson, W. T., Ericson, M. A. & Simpson, B. D. A speech corpus for multitalker communications research. *J. Acoust. Soc. Am.* **107,** 1065–1066 (2000).
16. Brungart, D. S. Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* **109,** 1101–1109 (2001).
17. Mesgarani, N., David, S. V., Fritz, J. B. & Shamma, S. A. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *J. Neurophysiol.* **102,** 3329–3339 (2009).
18. Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R. & Warland, D. Reading a neural code. *Science* **252,** 1854–1857 (1991).
19. Pasley, B. N. *et al.* Reconstructing speech from human auditory cortex. *PLoS Biol.* **10,** e1001251 (2012).
20. Garofolo, J. S. *et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus* (Linguistic Data Consortium, 1993).
21. Rifkin, R., Yeo, G. & Poggio, T. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences* **190,** 131–154 (2003).
22. 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).
23. Staeren, N., Renvall, H., De Martino, F., Goebel, R. & Formisano, E. Sound categories are represented as distributed patterns in the human auditory cortex. *Curr. Biol.* **19,** 498–502 (2009).
24. Shamma, S. A., Elhilali, M. & Micheyl, C. Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* **34,** 114–123 (2010).
25. Darwin, C. J. Auditory grouping. *Trends Cogn. Sci.* **1,** 327–333 (1997).
26. Warren, R. M. Perceptual restoration of missing speech sounds. *Science* **167,** 392–393 (1970).
27. Kidd, G. Jr, Arbogast, T. L., Mason, C. R. & Gallun, F. J. The advantage of knowing where to listen. *J. Acoust. Soc. Am.* **118,** 3804–3815 (2005).
28. Shen, W., Olive, J. & Jones, D. Two protocols comparing human and machine phonetic discrimination performance in conversational speech. *INTERSPEECH* 1630–1633 (2008).
29. Cooke, M., Hershey, J. R. & Rennie, S. J. Monaural speech separation and recognition challenge. *Comput. Speech Lang.* **24,** 1–15 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors would like to thank A. Ren for technical help, and C. Micheyl, S. Shamma and C. Schreiner for critical discussion and reading of the manuscript. E.F.C. was funded by National Institutes of Health grants R00-NS065120, DP2-OD00862, R01-DC012379, and the Ester A. and Joseph Klingenstein Foundation.

Author Contributions N.M. and E.F.C. designed the experiment, collected the data, evaluated results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.F.C. (changed@neurosurg.ucsf.edu).

©2012 Macmillan Publishers Limited. All rights reserved

METHODS

The experimental protocol was approved by the Committee for Human Research at the University of California, San Francisco.

Subjects. Three human subjects underwent the placement of a high-density subdural electrode array (4 mm pitch) over the language-dominant hemisphere as part of routine clinical treatment for epilepsy. Subjects gave their written informed consent before surgery. All subjects had self-reported normal hearing and underwent neuropsychological language testing (including the Boston naming and verbal fluency tests) and were found to be normal. The intracarotid sodium amobarbital (Wada) test was used for language dominance assessment. The electrodes in the study were located over the posterior dorsolateral temporal lobe. The location and corresponding spectrotemporal receptive fields of all the included electrodes for a subject are shown in Supplementary Fig. 2.

Data acquisition and pre-processing. The electrocorticography signal was recorded with a multichannel amplifier optically connected to a digital signal processor (TuckerDavis Technologies). Each channel time series was visually and quantitatively inspected for artefacts or excessive noise. The data were then segmented with a 100 ms pre-stimulus baseline and a 400 ms post-stimulus interval. The common mode signal was estimated using principal component analysis with channels as repetitions and was removed from each channel time series using vector projection.

Task design and behavioural testing. We used speech samples from a publicly available database called Coordinate Response Measure (CRM[15]) containing sentences in the form "ready (call sign) go to (colour) (number) now". One male and one female speaker (speakers one and five in CRM corpus) were selected with two call signs (ringo and tiger), three colours (blue (B), red (R) or green (G)) and three numbers (two, five or seven). For each of the two call signs, we generated six colour–number combinations (B2, B5, R2, R7, G5, G7), resulting in 12 different phrases. We chose the same phrases for each of the two speakers, resulting in 24 single speaker sentences. We then produced 28 unique mixture speech samples by selecting from combinations of the 24 single speaker sentences at 0 dB target-to-masker ratio. Each mixture sample was chosen such that there was no overlap between call signs, colours or the numbers of the two phrases. In addition, each speaker had the same number of call signs (ringo or tiger) in each trial block. The sounds were presented monaurally from a loudspeaker connected to a laptop, which was also used to collect subjects' responses through a customized graphical user interface. Each trial block consisted of 28 trials and the target call sign was fixed for each block. The target call sign was displayed visually before and during the trial block. Subjects first listened to each of the speakers alone and were able to report the colour and number with 100% accuracy. Subjects then listened to a monaural, simultaneous mixture of the two speakers' phrases with different call signs, colours and numbers. The subjects were instructed to respond by indicating the colour and number spoken by the talker who uttered the target call sign. The target speaker changed from trial to trial pseudo-randomly, requiring the subjects to initially monitor both speakers until they detect the target call sign. After each trial block, the target call sign was changed, switching the role of target and masker speakers in each mixture sound.

Electrode selection. The cortical sites on the superior and middle temporal gyri with reliable evoked responses to speech stimuli were selected for all the subsequent analysis. Our inclusion criteria consisted of a *t*-test between responses to randomly selected time frames during passive speech presentation (TIMIT) and in silence ($P < 0.01$, resulting in 83, 92 and 102 electrodes for subjects one to three. One example subject is shown in Supplementary Fig. 2a). Solely for visualization, we also estimated the STRFs of these selected sites from passive listening to TIMIT using normalized reverse correlation algorithm (STRFLab software package, http://www.strflab.berkeley.edu; Supplementary Fig. 2b). Correlation histogram of STRF predictions for all 275 electrode sites is shown in Supplementary Fig. 1c.

Stimulus reconstruction. We used stimulus reconstruction to map the population neural responses to the spectrogram of the speech stimulus[17–19]. Reconstruction filters were estimated from neural responses to a separate speech corpus (TIMIT[20]) containing a total of 499 unique short sentences from 402 different speakers. Filters were obtained using normalized reverse correlation to minimize the mean squared error of the reconstructed spectrograms[17] with filter time lags from −420 to 0 ms (causal filters). The filters were then fixed in all subsequent conditions and were applied to the neural responses to CRM samples. Neither of the speakers or phrases in the CRM data set was used in estimation of the filters. The output of the reconstruction algorithm was further processed with a band-pass filter applied to each frequency channel of reconstructed spectrograms to remove the baseline. All the processing steps for stimulus reconstruction were identical in all conditions (single and mixture speakers).

AMI. To quantify the change in similarity between the representation of single and attended speaker in mixture speech, we defined the AMI_{spec} in equation (1). The stereotypical format of the CRM phrases results in an intrinsic correlation between the neural responses to different sentences, particularly at the beginning ("ready") and middle of the carrier phrase ("go to"), which results in reduced possible AMI_{spec} values for these segments. To estimate an upper bound for unbiased comparison, AMI_{spec} was calculated where the representation of an attended speaker in a mixture is ideally assumed to be identical to the representation of that speaker when presented alone; therefore, replacing SP_{attend} in equation (1) with the reconstructed spectrogram of single speaker SP_{alone}. The upper bound peaks at the call sign, colour and number where different phrases are most dissimilar. The overall increase in the upper bound is due to the progressive asynchrony between the two speakers.

The same statistics can be used to estimate the AMI of an individual electrode site by calculating the correlation values between the neural response of that site to attended mixture and single speaker presentations:

$$\begin{aligned} AMI_{elec} = {} & Corr(R\text{-}SP1_{alone}, R\text{-}SP1_{attend}) - Corr(R\text{-}SP1_{alone}, R\text{-}SP2_{attend}) + \\ & Corr(R\text{-}SP2_{alone}, R\text{-}SP2_{attend}) - Corr(R\text{-}SP2_{alone}, R\text{-}SP1_{attend}) \end{aligned} \quad (2)$$

where $R\text{-}SP1_{alone}$ and $R\text{-}SP2_{alone}$ are the responses of an electrode to speakers one and two alone, respectively, and $R\text{-}SP1_{attend}$ and $R\text{-}SP2_{attend}$ are the responses of the same electrode to the mixture of the two when the attended target is speaker one and two, respectively.

Classification of spoken words and speaker identity. A linear-frame-based regularized-least-square classifier[21] was used to investigate the discriminability of the spoken words and speaker identity from electrocorticographic responses. Two binary classifiers were trained to classify the call sign and speaker identity, and two separate three-way classifiers were used for colour and for number classification. Classifiers were trained only on the neural responses of single speakers (24 sentences) and tested on the mixtures. The classifiers produced a linear weighted sum of the neural responses at each time instance and the classifier that produced the maximum average output over the duration of words was chosen as classification result. The classifier decision was limited to only the colours and numbers that occurred in each mixture, therefore resulting in same 50% chance performance in all cases.

©2012 Macmillan Publishers Limited. All rights reserved