Auditory neuroscience: sounding out the brain basis of speech perception

Article (Accepted Version)

Auditory Neuroscience: Sounding Out the Brain Basis of Speech Perception

Ediz Sohoglu
MRC Cognition and Brain Sciences Unit
University of Cambridge, Cambridge CB2 7EF, United Kingdom

Summary

What is the nature of the neural code by which the human brain represents spoken language? New research suggests that previous findings of a language-specific code in cortical responses to speech can be explained solely by simple acoustic features.

Main text

As anyone who has been repeatedly misunderstood by their phone's virtual assistant can testify, the human brain remains the best speech recognition device out there. During a rush-hour commute in any capital city, we might hear speech produced with twenty different accents (both regional and international). At the same time (at least in the case of those braving the city's metro system), we might also hear the loud screech of a train amplified by tunnel walls. In environments like these, machine speech recognition software would most probably fail. And yet most of the time, we as human listeners have no problems understanding the speech that we hear. A new study by Daube et al. [1], reported in this issue of Current Biology, provides new evidence of the neural representations by which the brain makes sense of spoken language.

To fully understand speech, the brain has to recognize words from a rapidly changing sound wave. According to one proposal [2], words are recognized by comparing the heard speech signal with an internal memory of what words should sound like (termed the 'lexicon'). A word will be recognized if there is a good match between its entry in the listener's lexicon and the heard speech signal. By this account, the acoustics of speech are mapped directly onto the lexicon without any intervening steps. However, in most other cognitive [3] and neuroscientific [4] theories, an intermediate stage is proposed whereby each spoken word is categorized into discrete units such as phonemes. The hallmark of these categorical units is that acoustic detail is stripped away from speech. Thus, despite substantial acoustic differences, a "t" phoneme is categorized as a "t" whether spoken by a woman or a man, an adult or a child, or even when whispered. To appreciate the benefit of processing speech in this way, consider encountering a speaker with an unfamiliar accent such that they pronounced the sound "t" as "p". Rather than learning the new pronunciation of every word with that sound, after hearing a few examples you might learn that "pake" is in fact "take" (and not "cake" or "lake" etc.). Mapping speech onto words via phonemes is therefore computationally efficient [5].

So what evidence is there that the brain represents categorical units such as phonemes during speech comprehension? Decades of neuroimaging and neurophysiological studies have sought to answer this question. A relatively recent development in this area has been to ask volunteers to listen to extended spoken narratives, for example story extracts, and use linear regression to test whether various speech features can predict recorded brain activity. Successful prediction would indicate that brain activity fluctuates systematically over time along with the speech features and therefore might represent those features. Previous research using this approach indicates that

cortical responses to speech, as measured using electroencephalography (EEG), track which phoneme was heard [6]. To test whether this effect was specifically related to categorical phonemes, it was important to control for the fact that phonemes differ acoustically. To do this, an acoustic description of the speech signal – the balance of sound frequencies over time or 'spectrogram' – was also regressed against the measured brain activity. The critical finding in this previous research then was that phonemes can indeed predict EEG signals, *over and above acoustics*. Similar results were obtained when replacing phonemes with articulatory features, which describe speech in terms of the underlying vocal gestures (such as the position of the tongue and lips). Like the phoneme, this description of the speech signal is categorical (e.g. either the lips come together or not e.g. when saying "b" versus "w"). These findings therefore suggest that the brain does indeed represents categorical-like units.

Despite the computational efficiency of categorical representations for spoken word recognition, whether phonemes or articulatory features are actually represented by listeners is far from universally accepted [2,7]. In their new study, Daube et al. set out to test an alternative hypothesis, one that explains cortical responses in purely acoustic terms, thereby challenging accounts based on categorical representations such as phonemes (Figure 1). They first replicated the previous finding that articulatory features can predict brain responses over and above the audio spectrogram, now with magnetoencephalography (MEG) recordings localized to auditory cortex in the superior temporal lobe. They then used the prediction success of articulatory features as a benchmark against which to compare alternative acoustic-based ones, which were all obtained by simple transformations of the spectrogram. One such transformation emphasized acoustic onsets: rapid increases in spectrogram energy that are found, for example, at the start of syllables. Previous research indicates that the cortex is especially sensitive to such acoustic onsets [8,9]. Daube et al. found that these acoustic features made very similar MEG predictions to the benchmark articulatory features. Critically, they also found that acoustic onsets explained parts of the MEG response that articulatory features could not (for further explanation, see Figure 1 and caption). Impressively, this result was again demonstrated with publically available EEG data – the same dataset that provided the original evidence for phonemes and articulatory features [6]. As well as providing another replication, this latter finding is important as EEG can pick up activity from different brain regions as compared with MEG [10]. Finally, Daube et al. show how another finding that could be interpreted as evidence for categorical representations – the ability to decode phoneme groups from EEG responses [11] – can also be attributed to acoustic onsets. Thus, cortical responses to speech were best explained in purely acoustic terms, using relatively simple computations on the spectrogram that could plausibly be realised neurally [12].

Recent methodological advances in non-invasive neuroimaging such as multivariate techniques have allowed researchers to move way from an 'activation-based' approach to one that more directly taps into content-specific representations [13,14]. With these advances it is now possible, in relatively naturalistic settings, to test a rich hypothesis space of how the human brain represents speech, including semantics [15]. The study by Daube et al. is important in highlighting the need to consider simple, acoustic-based features in these endeavours. It also raises several questions for future research. For example, Daube et al. focussed their analyses on MEG responses in the superior temporal lobe as these were the strongest and most reliable signals. Could signatures of articulatory features still be found but in higher-level regions of the cortical hierarchy? Another question is the extent to which other recent findings in the literature can be attributed to acoustic onsets, as opposed to higher-level linguistic variables? Relevant to this, it is notable that a recent study in this Journal did account for acoustic onsets when demonstrating cortical sensitivity to phonemic and lexical predictability [16]. In other work showing EEG signatures of semantic

processing of continuous speech [15], the late latency (200-600 ms) of the neural effect of interest would seem to make it unlikely a reflection of low-level acoustic onsets. It is also unclear how an acoustic onset account can explain why phonetic tracking by EEG is dependent on speech intelligibility, which might suggest a language-specific (rather than auditory-general) locus [6].

Perhaps the biggest question of all though is *why* the brain is sensitive to acoustic onsets and *how* this ultimately leads to successful speech comprehension. Daube et al. themselves offer one hypothesis: that sound onset responses are robust to background noise and thus may support listening in the noisy environments that we routinely encounter. Another perspective is based on the idea of predictive coding [17,18]. The proposal here is that rather than processing speech directly, the brain actively predicts upcoming sensory events so that only unexpected information (i.e. prediction errors) gets processed further up the cortical hierarchy. According to this view, cortical responses to acoustic onsets represent prediction errors that have the role of updating previous expectations about the short-term spectral content of sound input [19]. Whatever the interpretation, the insights provided by Daube et al. are undoubtedly a step towards a future in which we have a comprehensive account of how the brain makes sense of speech. Who knows, perhaps then we can explain why the brain is so much better at understanding you than your phone.

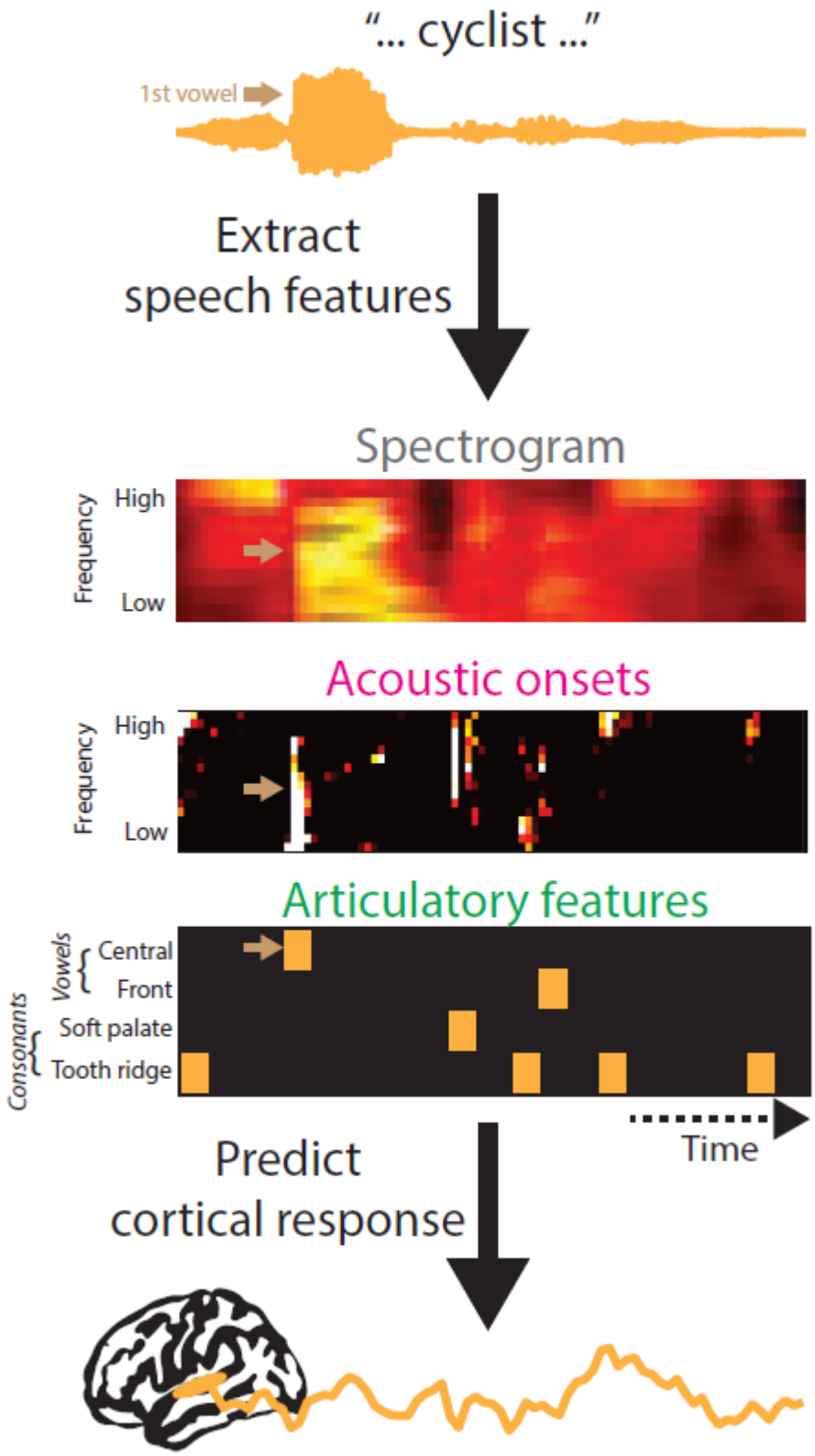

Figure 1. Daube et al. [1] tested which of several speech features was most prominently represented in MEG recordings of brain activity, shown here for the example word "cyclist". The location of the first vowel in all feature representations is marked by a brown arrow to facilitate comparison. The first vowel is signalled by a transient increase in energy for both acoustic onsets and articulatory

features. Thus, these features make the similar prediction that brain activity should change soon after hearing the start of the first vowel. Because of this similarity, combining acoustic onsets with articulatory features did not improve prediction success over acoustic onsets alone (bottom bar graph; compare magenta bars with and without green outline). Despite making similar predictions, acoustic onsets were sufficiently different from articulatory features that brain responses were better predicted by acoustic onsets (compare left magenta and green bars). This suggests that speech is neurally represented as acoustic onsets more than as articulatory features.

References

1. Daube, C., Ince, R.A.A., and Gross, J. (2019). Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr. Biol.
2. Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychol. Rev. *105*, 251–79.
3. McClelland, J.L., and Elman, J.L. (1986). The TRACE model of speech perception. Cogn. Psychol. *18*, 1–86.
4. Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402.
5. Scharenborg, O., Norris, D., Bosch, L., and McQueen, J.M. (2005). How should a speech recognizer work? Cogn. Sci. *29*, 867–918.
6. Di Liberto, G.M., O'Sullivan, J.A., and Lalor, E.C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr. Biol. *25*, 2457–2465.
7. Kazanina, N., Bowers, J.S., and Idsardi, W. (2018). Phonemes: Lexical access and beyond. Psychon. Bull. Rev. *25*, 560–585.
8. Hamilton, L.S., Edwards, E., and Chang, E.F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. Curr. Biol. *(in press)*, 1–12.
9. Martin, B.A., and Boothroyd, A. (2000). Cortical, auditory, evoked potentials in response to changes of spectrum and amplitude. J. Acoust. Soc. Am. *107*, 2155–61.
10. Molins, A., Stufflebeam, S.M., Brown, E.N., and Hämäläinen, M.S. (2008). Quantification of the benefit from integrating MEG and EEG data in minimum l2-norm estimation. Neuroimage *42*, 1069–77.
11. Khalighinejad, B., Cruzatto da Silva, G., and Mesgarani, N. (2017). Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. J. Neurosci. *37*, 2176–2185.
12. Silver, R.A. (2010). Neuronal arithmetic. Nat. Rev. Neurosci. *11*, 474–489.
13. Crosse, M.J., Di Liberto, G.M., Bednar, A., Lalor, E.C., Liberto, G.M. Di, Bednar, A., and Lalor, E.C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. Front. Hum. Neurosci. *10*, 1–14.
14. Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. *17*, 401–412.
15. Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., and Lalor, E.C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. Curr. Biol. *28*, 803–809.e3.

16. Brodbeck, C., Hong, L.E., and Simon, J.Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Curr. Biol. *28*, 3976–3983.e5.

17. Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.

18. Gagnepain, P., Henson, R.N., and Davis, M.H. (2012). Temporal Predictive Codes for Spoken Words in Auditory Cortex. Curr. Biol. *22*, 1–7.

19. Sohoglu, E., and Chait, M. (2016). Detecting and representing predictable structure during auditory scene analysis. Elife *5*.