Multi-level representations in speech processing in brain and machine: Evidence from EMEG and RSA

✉ cw417@cam.ac.uk

Cai Wingfield[1,✉], Li Su[2], Barry Devereux[1], Xunying Liu[3,4], Chao Zhang[3], Phil Woodland[3], Elisabeth Fonteneau[1], Andrew Thwaites[1], William Marslen-Wilson[1,5]

[1]Department of Psychology, [2]Department of Psychiatry, [3]Department of Engineering; University of Cambridge
[4]Department of Systems Engineering and Engineering Management; Chinese University of Hong Kong
[5]MRC Cognition & Brain Sciences Unit; Cambridge

Introduction

Human speech recognition

- For humans, speech recognition feels effortless and automatic.
- There is only limited neurocomputational understanding of how this is achieved.
- Recent evidence suggests that responses to speech may be represented in a low-dimensional space of articulatory features[1,2].

Machine speech recognition

- Automatic speech recognition (ASR) systems using deep neural network (DNN) acoustic models approach human levels of performance, with word identif cation rates well over 90%[3].
- They provide a computationally specif c model of how successful speech recognition can be achieved.
- DNNs have been successful models for brain responses in other domains, e.g. vision[4].

The questions we ask

- Can we model human neural responses to speech with an ASR-derived DNN acoustic model?
- Can we characterise hidden-layer representations?
- Do activations in different layers of such a DNN differentially explain cortical speech responses through space and time?

- Here we use a DNN-based ASR system to model brain responses to speech using multivariate searchlight techniques[5].

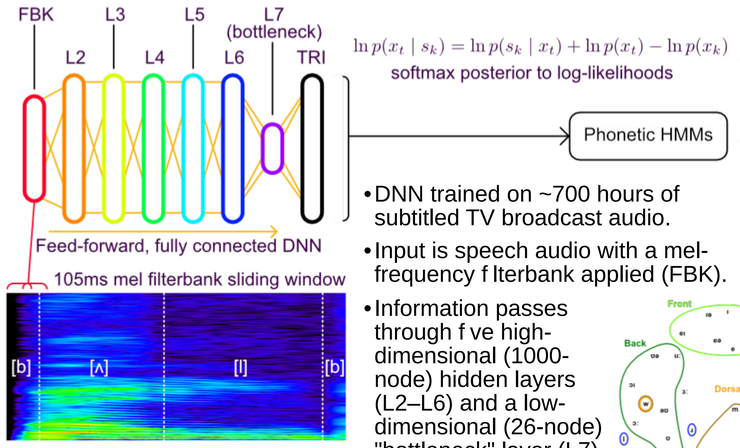
EMEG experiment

- Whole-brain simultaneous electro- and magneto-encephalography (EMEG) data was recorded from 16 participants while they listened to recordings of 400 British English words.
- EMEG data was source-localised and warped to a subject-average cortical mesh.
- The mesh was restricted to an auditory cortex (AC) mask comprising bilateral superior temporal cortices (STC) and Heschl's gyrus (HG).
- Mask matches locations previously found to exhibit phonetic feature sensitivity to speech[1,2].

References

1. Wingf eld C, et al. (submitted) *PLOS Comp Biol*.
2. Mesgarani N, et al. (2014) *Science*.
3. Young S, et al. (2015) *The HTK Book*.
4. Khaligh-Razavi S, et al. (2014) *PLOS Comp Biol*.
5. Kriegeskorte N, et al. (2006) *PNAS*.
6. Kriegeskorte N, et al. (2008) *Frot Syst Neurosci*.
7. Su L, et al. (2012) *PRNI, IEEE*.
8. Smith S, et al. (2009) *NeuroImage*.

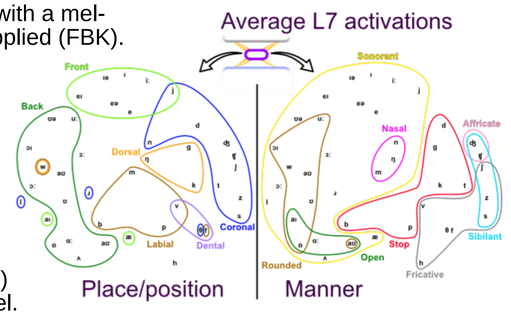
Computational modelling and analysis



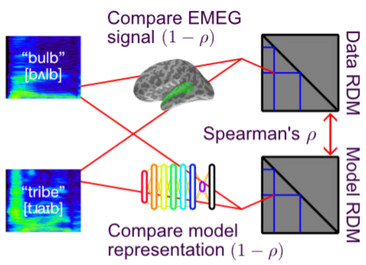
- DNN trained on ~700 hours of subtitled TV broadcast audio.
- Input is speech audio with a mel-frequency f lterbank applied (FBK).
- Information passes through f ve high-dimensional (1000-node) hidden layers (L2–L6) and a low-dimensional (26-node) "bottleneck" layer (L7).
- Output layer gives posteriors for ~6000 phonetic targets.
- These are given to a set of hidden Markov models (HMMs) associated with phonemes used in the ASR acoustic model.

Deep neural network speech recogniser

- To model the speech recognition process, we used HTK[3], an ASR system.
- HTK uses a fully-connected, feed-forward DNN is an acoustic model, mapping speech sounds to phonetic targets through time.



- Average L7 response to different phones can be visualised by Sammon nonlinear multidimensional scaling.
- Bottleneck responses clustered according to place and manner of articulation of consonants, position of vowels, and broad category distinctions (e.g. sonorant–obstruent).



Spatiotemporal searchlight representational similarity analysis

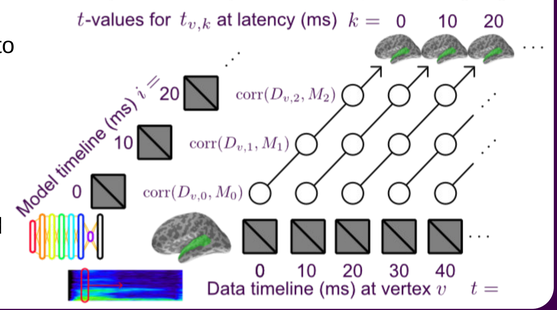
- Representational similarity analysis (RSA[6]) compares condition space representations using a representational distance matrix (RDM) with pairwise condition distances as entries.
- Data RDMs are computed from brain responses; model RDMs are computed from model representations or predicted distances.
- RSA abstracts away from specif c responses, allowing comparison of representations in very different formats[4,6].
- Spatiotemporal searchlight RSA (ssRSA[7]) computes data RDMs from a continuously moving regular spatiotemporal searchlight patch.

Computational mapping

- Model RDMs were produced from activations in each layer of the DNN as HTK listened to the same recorded words as the participants.
- We f t the model RDM time course to the data RDM time course systematically at different processing lags (0–250ms).
- This gave time-resolved maps of the f t of each model to the brain data in individual subjects at each lag, which we converted to *t*-maps across subjects.
- Threshold-free cluster enhancement (TFCE) was performed on the resultant *t*-maps[8].

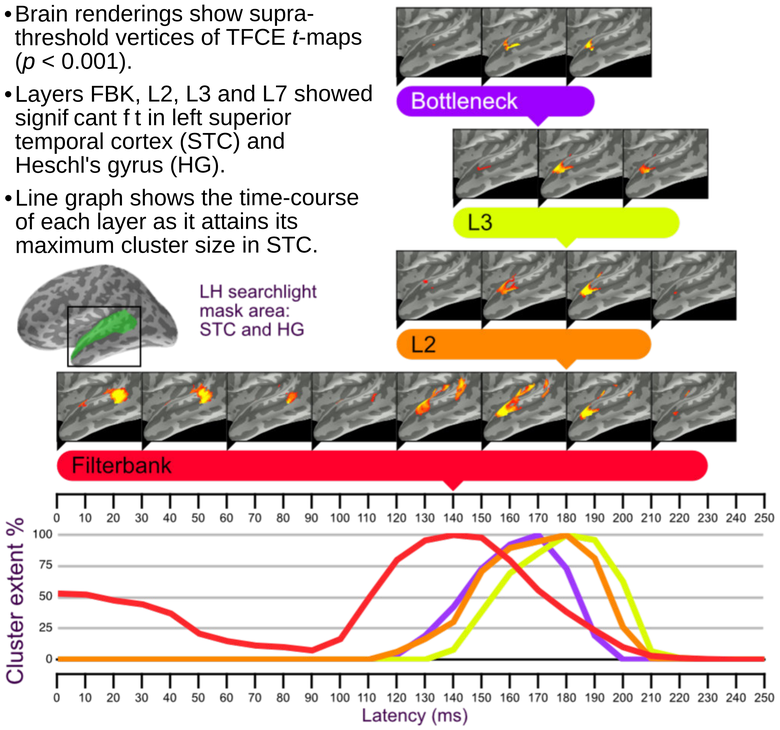
$$\mathrm{TFCE}(t_{v,k}) = \int_{h=0}^{t_{v,k}} \sqrt{e(h)} \cdot h^2 \, \mathrm{d}h$$

- Random-effects permutation across subjects was used to determine the signif cance of the TFCE values.
- Resultant maps were thresholded at $p < 0.001$.



Results

- Brain renderings show supra-threshold vertices of TFCE *t*-maps ($p < 0.001$).
- Layers FBK, L2, L3 and L7 showed signif cant f t in left superior temporal cortex (STC) and Heschl's gyrus (HG).
- Line graph shows the time-course of each layer as it attains its maximum cluster size in STC.



- Input layer FBK peaked early in posterior STC (0–90ms). Later peak in anterior STC and HG (100–230ms).
- Layers L2 and L3 peak later (120–210ms and 140–220ms) in anterior STC.
- L7 (bottleneck layer) peaked in anterior STC (120–190ms).
- Layers L4–L6 did not survive threshold ($p < 0.001$), but *t*-maps peaked in anterior STC. Clusters for L4 and L6 survived $p < 0.01$ bilaterally (not shown).
- Compared to the acoustic input-layer model, hidden-layer models from the ASR DNN f t the data at increased latencies and more anterior regions.

Conclusions

- We related different layers of a DNN-based computational model of speech recognition to EMEG data recorded from humans listening to speech.
- For FBK, L2 and L3, we saw evidence of spatial and temporal gradients of f t for higher layers of the ASR DNN.
- Higher hidden layers failed to f t brain data as well, perhaps indicating a progressive divergence of human and machine representations.
- However, the compressed code in L7, where activations cluster by articulatory features, again explained the data representations.
- Posterior–anterior STC changes in model f t over lags and layers may ref ect progressively more specif c representations of speech, with anterior STC representations relating to phonetic rather than acoustic features.