Investigating human speech recognition: Reverse-engineering the machine solution with EMEG and RSA

Cai Wingfield[1]*, Li Su[2], Xunying Liu[3], Philip Woodland[3], Chao Zhang[3], Andrew Thwaites[1], Elisabeth Fonteneau[1], William Marslen-Wilson[1]
*cw417@cam.ac.uk
[1]Department of Psychology, [2]Department of Psychiatry, [3]Department of Engineering

Speech recognition in humans and machines

Spoken language comprehension depends on the capacity to decode word identities and their meanings from a stream of rapidly varying acoustic input.

Humans

- For humans, this task seems effortless and automatic.
- There is only limited neurocomputational understanding of how this is achieved.
- Some evidence that the mapping from speech input onto word representations is in terms of phonetic (articulatory) features[1,2].

Machines

- Automatic speech recognition (ASR) systems are approaching human levels of performance with word identification rates in the 90% range.
- These systems are built strictly on engineering principles, but provide a computationally specific model of how successful speech recognition can be achieved.

Overview

The questions

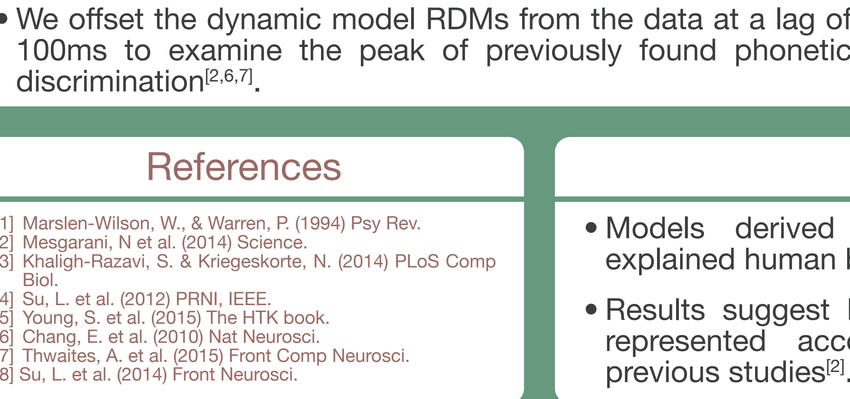
- Can we use ASR-derived computational models of speech recognition to illuminate the human solution?
- Do ASR-derived phonetic feature models significantly explain patterns of neural activation in speech-related human brain areas (bilateral superior temporal cortices)?

Research focus

- To answer these questions, we will relate dynamic brain states in the human listener to dynamic machine states in a candidate ASR system.
- Simultaneous electro- and magnetoencephalography (EMEG) provides non-invasive, high-resolution recording of neurophysiological activity in the brain.
- Representational similarity analysis (RSA) provides the abstraction necessary to compare brain and machine responses at the level of the representations they encode.

Spatiotemporal searchlight RSA with EMEG

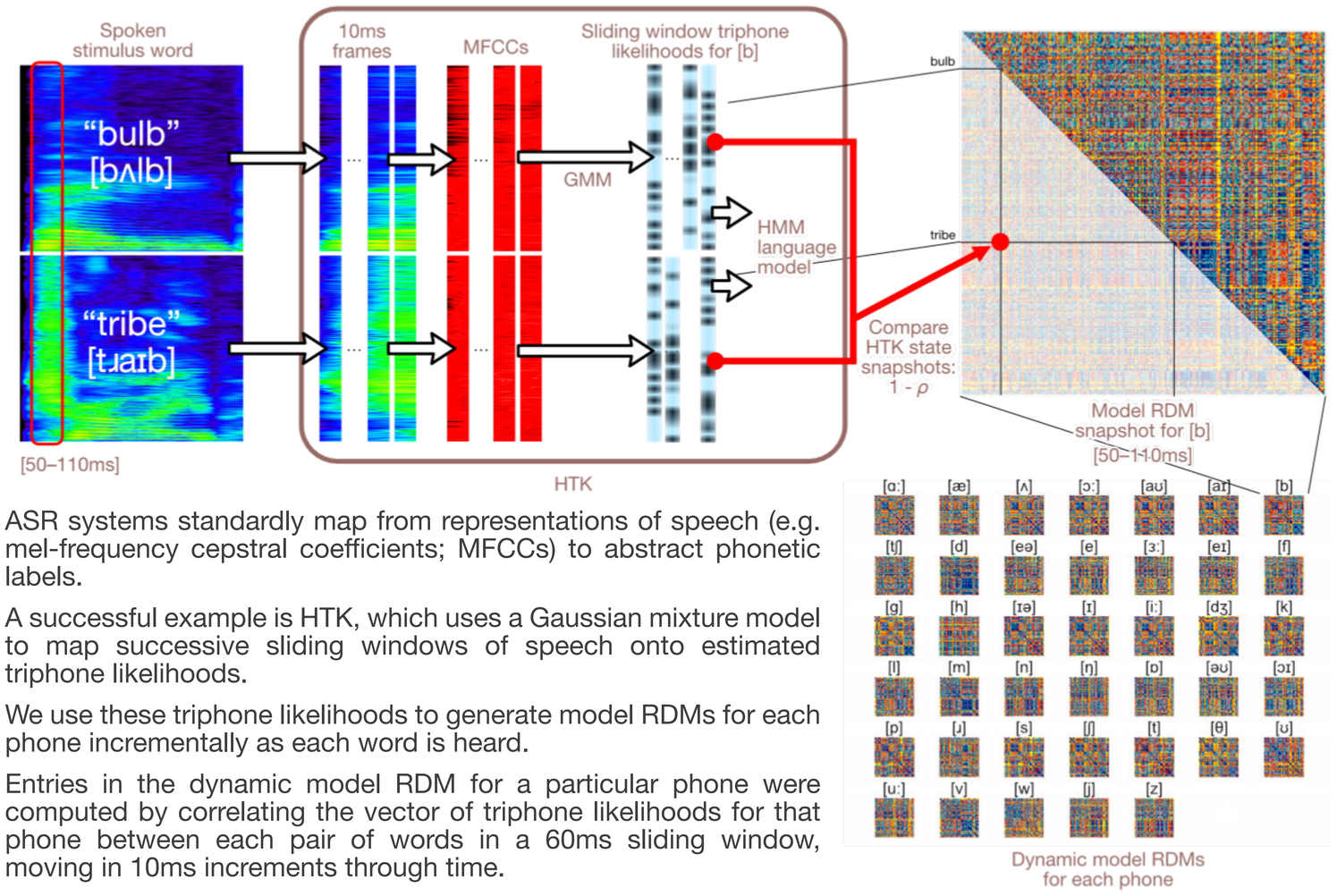
- RSA is a multivariate method which quantifies the shared similarity structures of data and model representations.
- RSA abstracts away from specific responses, allowing comparison of representations in very different formats[3].
- Spatiotemporal searchlight RSA[4] for EMEG data uses RSA to systematically map how a model fits to data throughout the brain in space and time.
- Time-varying representational dissimilarity matrices (RDMs) are computed using a sliding window of EMEG data.
- These *data* RDMs can be tested against corresponding time-varying *model* RDMs extracted from the output of an ASR system.



Materials and methods

- We recorded whole-brain EMEG data from sixteen participants while they listened to a set of 400 British English words. We used this data to produce dynamic data RDMs.
- The same 400 acoustic tokens were analysed with HTK[5], an ASR system, to produce dynamic model RDMs.
- We performed ssRSA within a language mask comprising bilateral superior temporal gyri (STG), superior temporal sulci (STS) and Heschl's gyri (HG).
- We computed brain data RDMs using a 60ms sliding window moving in 10ms increments, and a 15mm patch radius on a cortical mesh of 10,242 vertices per hemisphere. We compared the data RDMs to corresponding model RDMs which were computed using a matching temporal window.
- We offset the dynamic model RDMs from the data at a lag of 100ms to examine the peak of previously found phonetic discrimination[2,6,7].

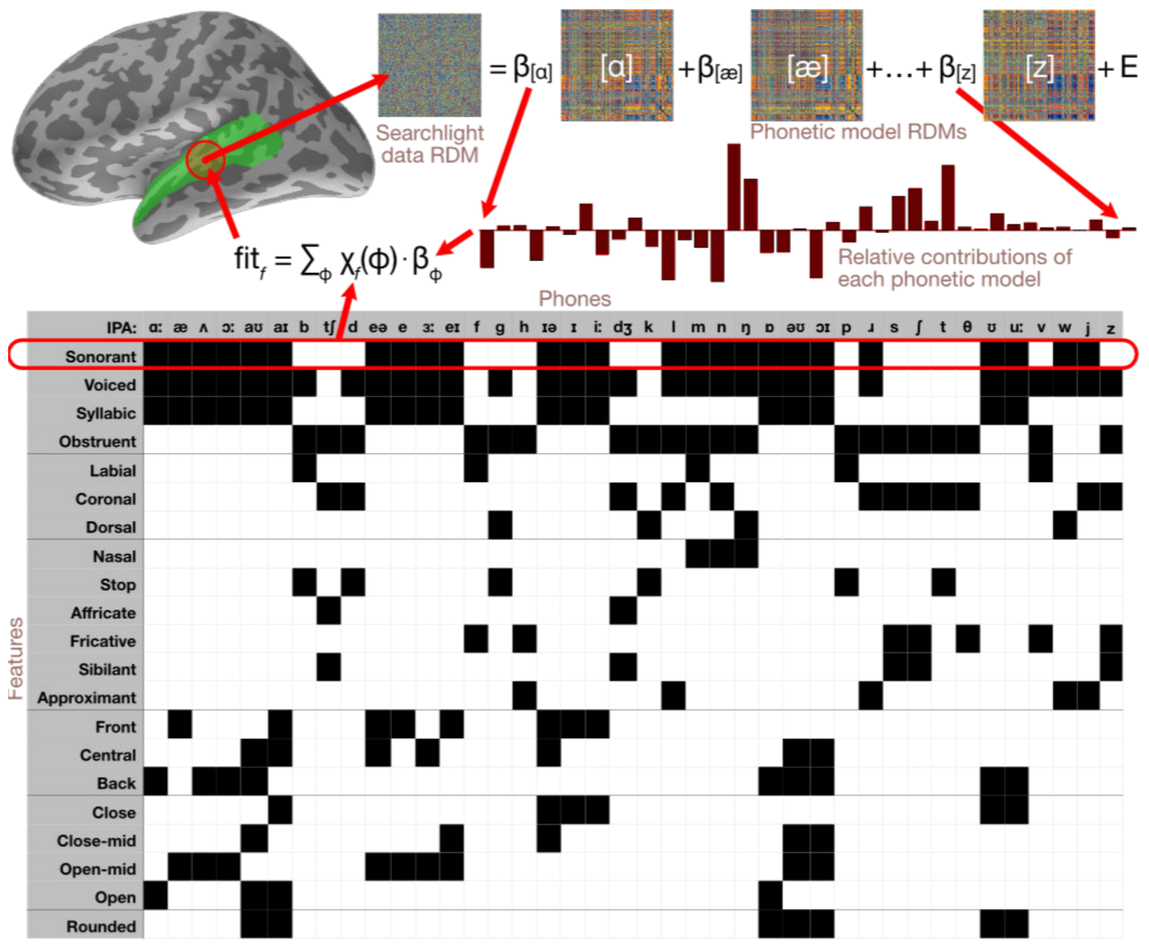
Modelling speech representations with an ASR system



- ASR systems standardly map from representations of speech (e.g. mel-frequency cepstral coefficients; MFCCs) to abstract phonetic labels.
- A successful example is HTK, which uses a Gaussian mixture model to map successive sliding windows of speech onto estimated triphone likelihoods.
- We use these triphone likelihoods to generate model RDMs for each phone incrementally as each word is heard.
- Entries in the dynamic model RDM for a particular phone were computed by correlating the vector of triphone likelihoods for that phone between each pair of words in a 60ms sliding window, moving in 10ms increments through time.
- Model RDMs are tested against corresponding data RDMs.

Mapping from phones to features

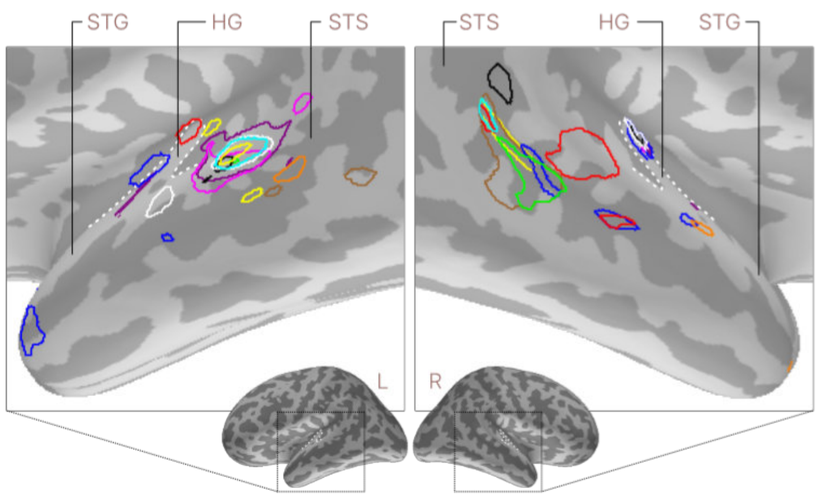
- Previous studies suggest that early representations of speech are organised according to phonetic features[1], and that these are represented in the brain in bilateral superior temporal regions[2].
- We use a generalised linear model to determine the contributions of each phonetic model RDM to explaining the data RDM[8].
- Contributions of phonetic models exhibiting each feature are summed to produce the contributions of that feature.
- This contribution is mapped back to the searchlight location.
- Resulting maps are thresholded using a permutation statistic[8].



Results

- Thresholded maps show spatially-coherent significant patches of feature sensitivity in bilateral HG, STG and STS.
- Most features showed significant fit.
- Broad phonetic category distinctions fit best in the right hemisphere. *Obstruent* phones are those where airflow is turbulent, in contrast with *sonorant* phones where airflow is non-turbulent. *Voicing* indicates whether the vocal chords vibrate.

Left	Feature	Right
*	Sonorant	***
	Voiced	**
⚙	Syllabic	⚙
	Obstruent	*
**	Labial	
	Coronal	*
**	Dorsal	
**	Nasal	⚙
	Stop	*
	Approximant	*
**	Front	
***	Central	*
	Back	*
**	Close	
*	Close-mid	
	Open-mid	*
**	Open	
**	Rounded	



- Within-category distinctions fit best in left hemisphere. *Labial*, *coronal* and *dorsal* describe place of articulation. *Nasal*, *stop* and *approximant* describe manner of articulation. The *front–back* and *close–open* dimension describe articulation of vowels, and *rounding* indicates the shape of the lips for vowels.

Conclusions

- Models derived from machine representations significantly explained human brain representations in bilateral language areas.
- Results suggest human cortical representations of speech are represented according to phonetic features, corroborating previous studies[2].
- Lateral asymmetry may be suggestive of specialisation.
- The pattern information in human brain responses can be successfully modelled using machine representations, potentially opening up many new directions of research.

References

[1] Marslen-Wilson, W., & Warren, P. (1994) Psy Rev.
[2] Mesgarani, N et al. (2014) Science.
[3] Khaligh-Razavi, S. & Kriegeskorte, N. (2014) PLoS Comp Biol.
[4] Su, L. et al. (2012) PRNI, IEEE.
[5] Young, S. et al. (2015) The HTK book.
[6] Chang, E. et al. (2010) Nat Neurosci.
[7] Thwaites, A. et al. (2015) Front Comp Neurosci.
[8] Su, L. et al. (2014) Front Neurosci.