Reconstructing Speech from Human Auditory Cortex

Alex Francois-Nienaber

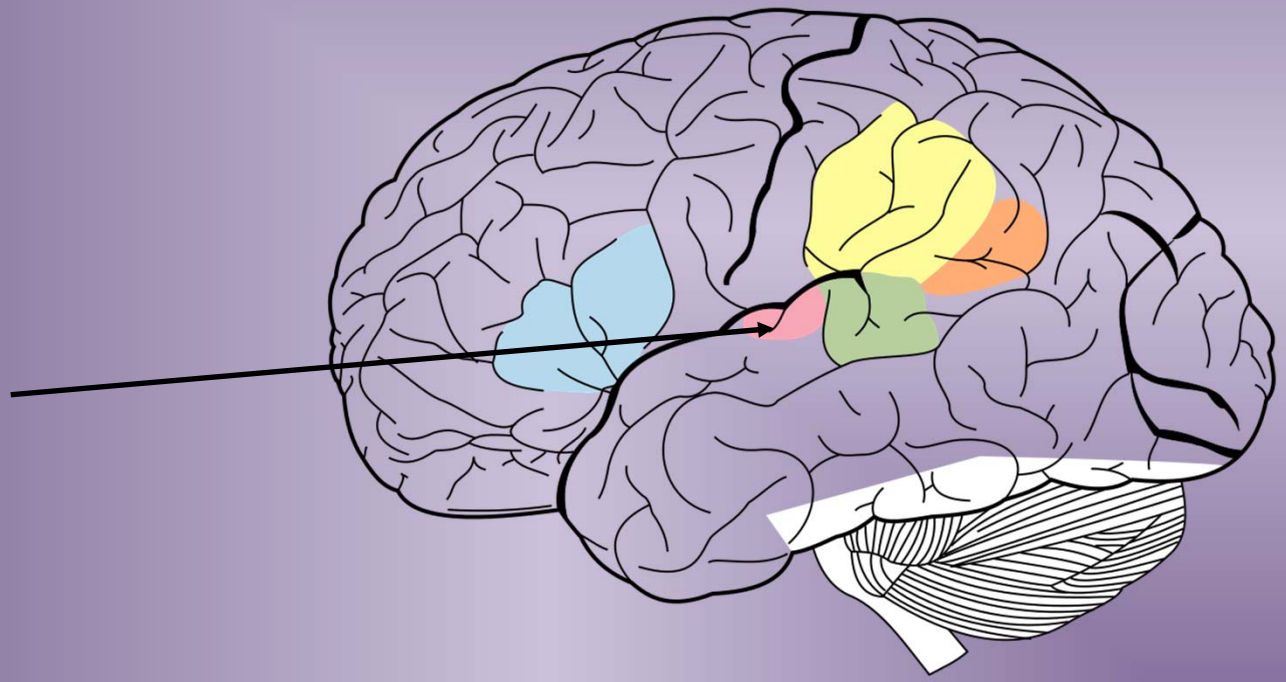
CSC2518 Fall 2014

Department of Computer Science, University of Toronto

Introduction to Mind Reading

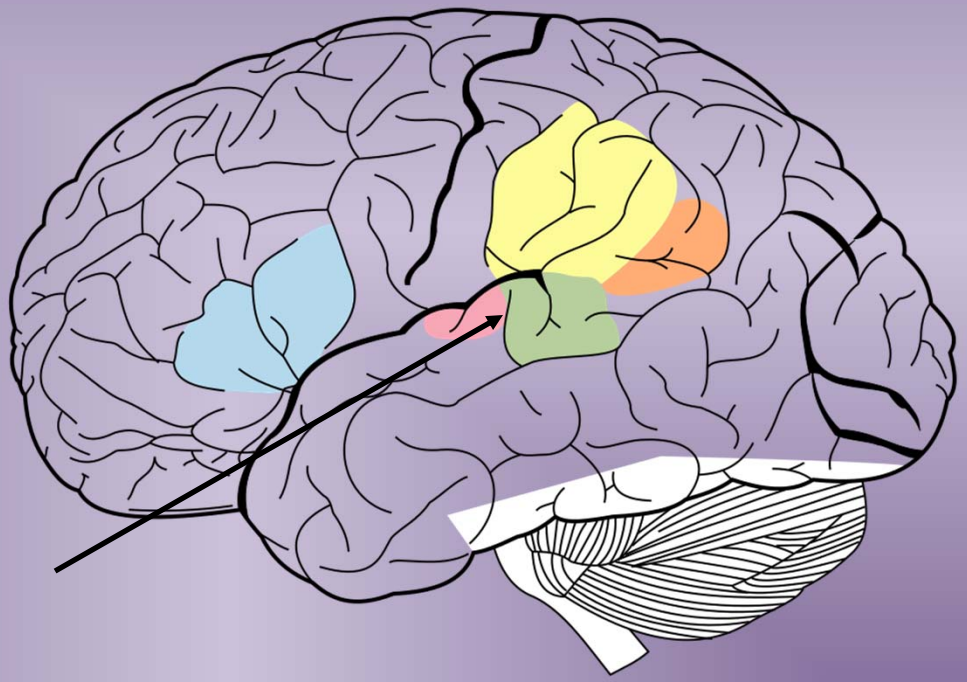
Introduction to Mind Reading

- Acoustic information from the auditory nerve is preprocessed in the Primary Auditory Cortex.



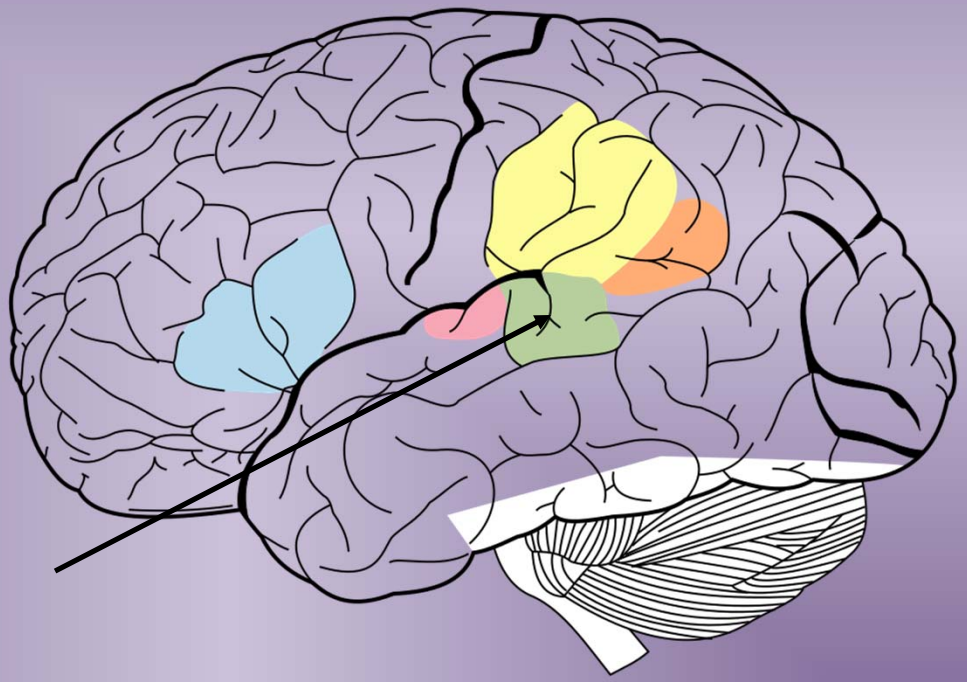
Introduction to Mind Reading

- Extracted features are relayed to the posterior Superior Temporal Gyrus (pSTG).



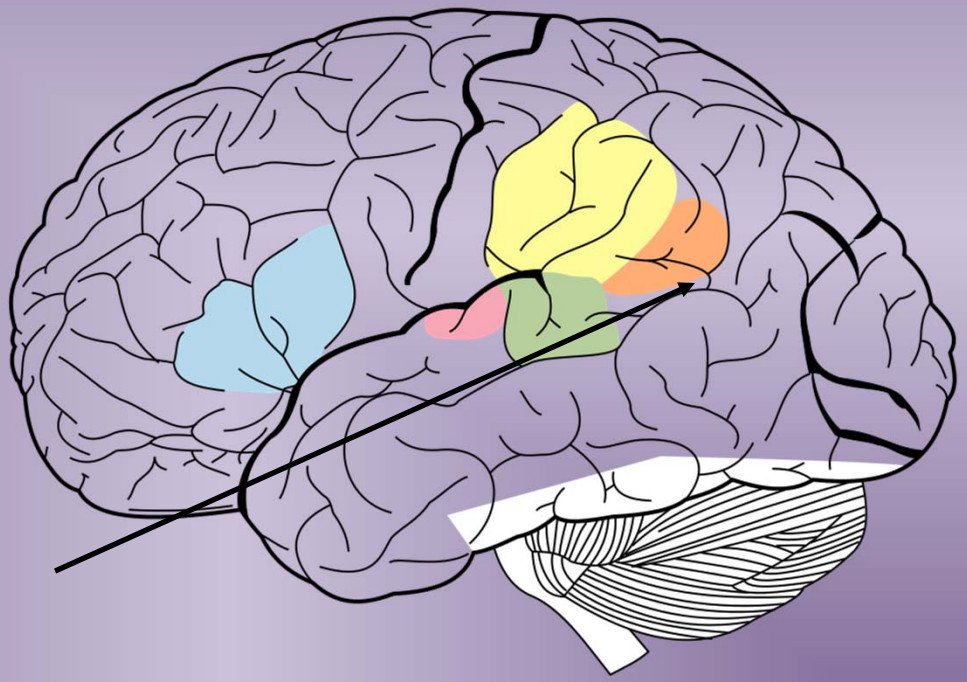
Introduction to Mind Reading

- The decoded speech features are then sent to Wernicke's area for semantic processing.



Introduction to Mind Reading

- Finally signals are sent to the TemporoParietal Junction, where they are processed with information from other modalities.



Introduction to Mind Reading

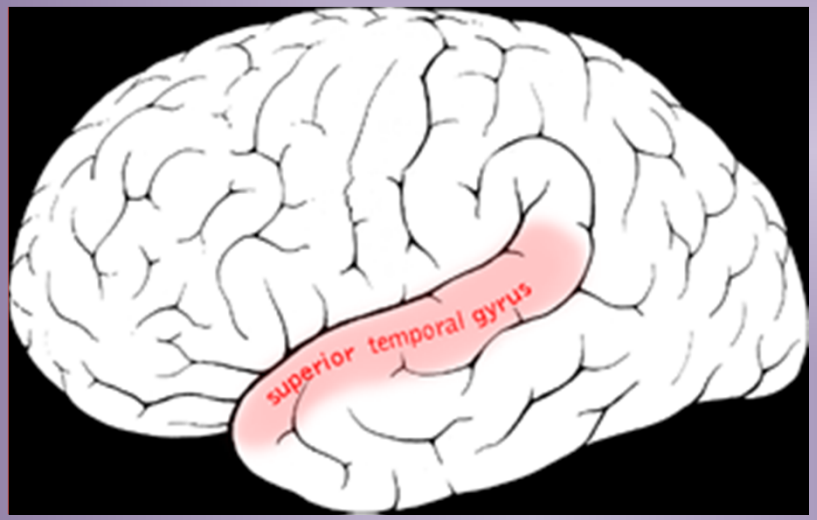
- We believe pSTG is involved in an intermediate stage of audio processing: interesting spectrotemporal features are extracted while nonessential acoustic features (i.e. noise) are filtered.
- These features are then converted to phonetic/lexical information.

That's why we would be interested in monitoring that area.

BUT how?

Electrocorticography

- Neurons are densely packed in cortical convolutions (gyri), e.g. pSTG.

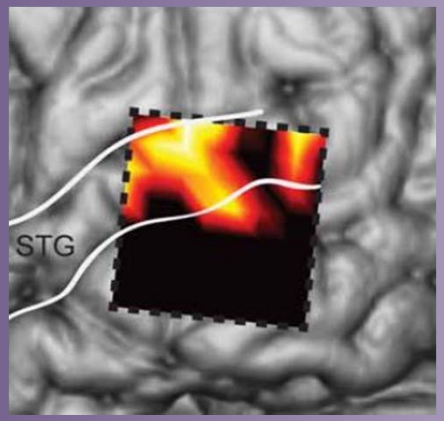


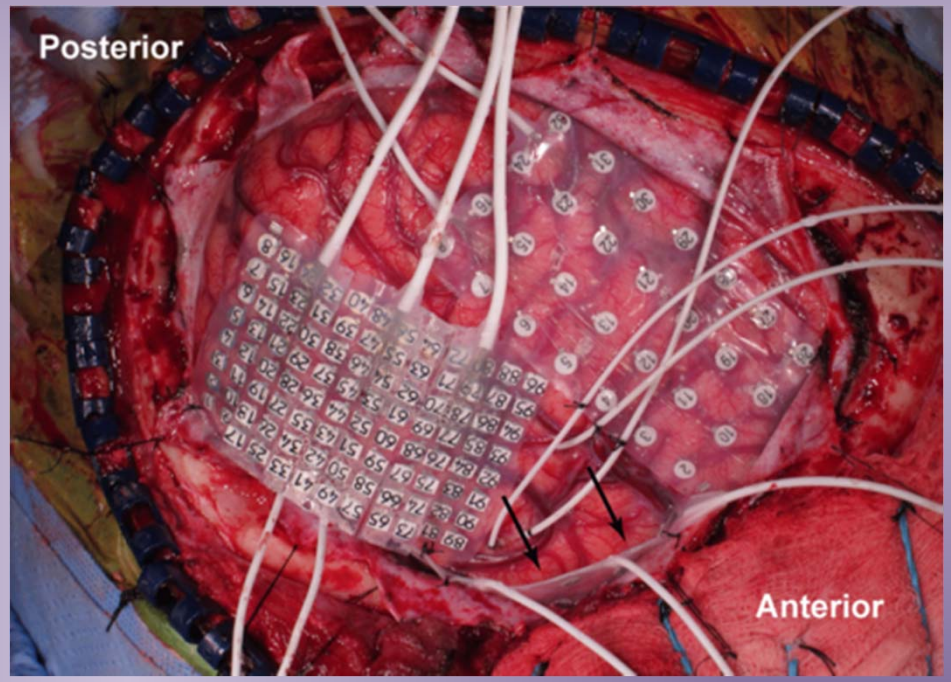
Electrocorticography

- We can record the summed-up synaptic current flowing extracellularly - the surface **field potentials** - by embedding very small electrodes directly into nerve tissue.
- By placing all the electrodes in a grid-like pattern, we can monitor an entire brain area!

Electrocorticography

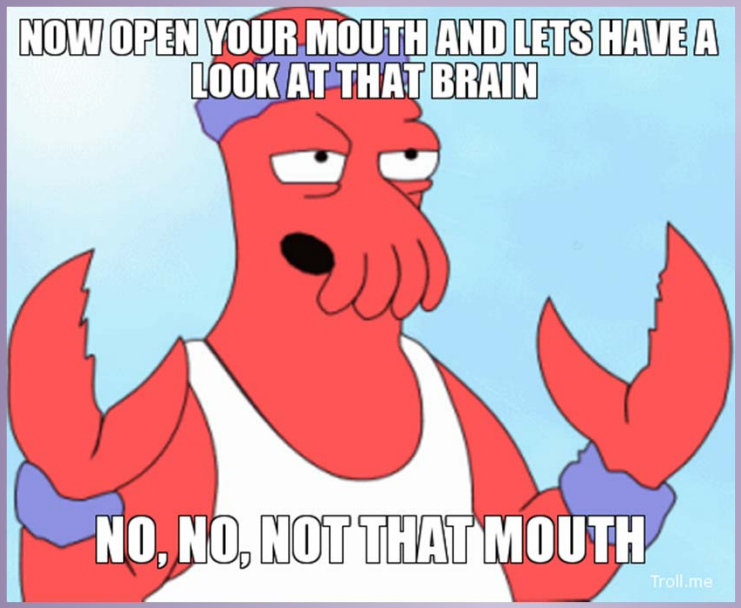
- The grid density will influence the precision of the results.





Electrocorticography

- 15 patients undergoing neurosurgery for tumors/epilepsy volunteered for this invasive experiment.



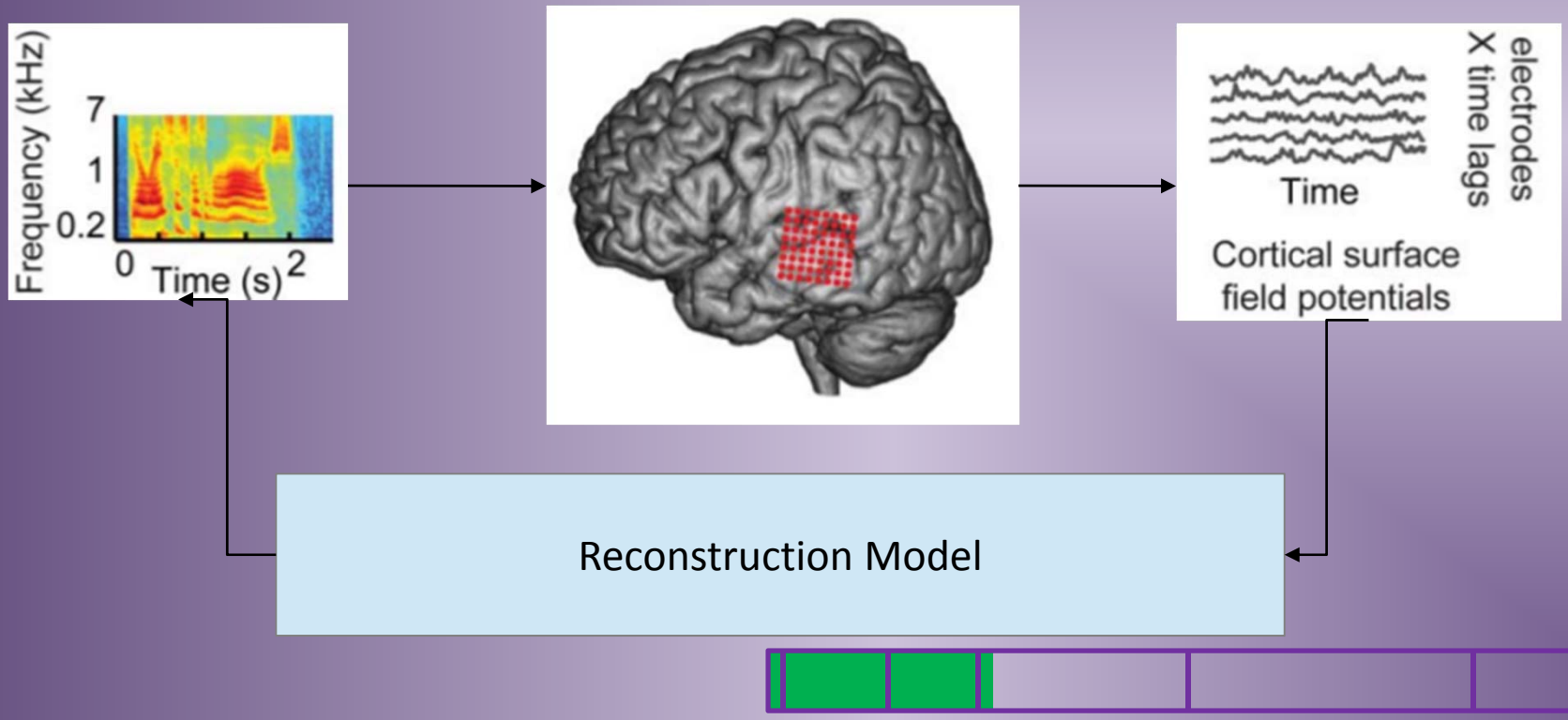
So how do we transform those cortical surface potentials into words?

So how do we transform those cortical surface potentials into words?

This will depend on how the **recorded** field potentials **represent** the acoustic information.

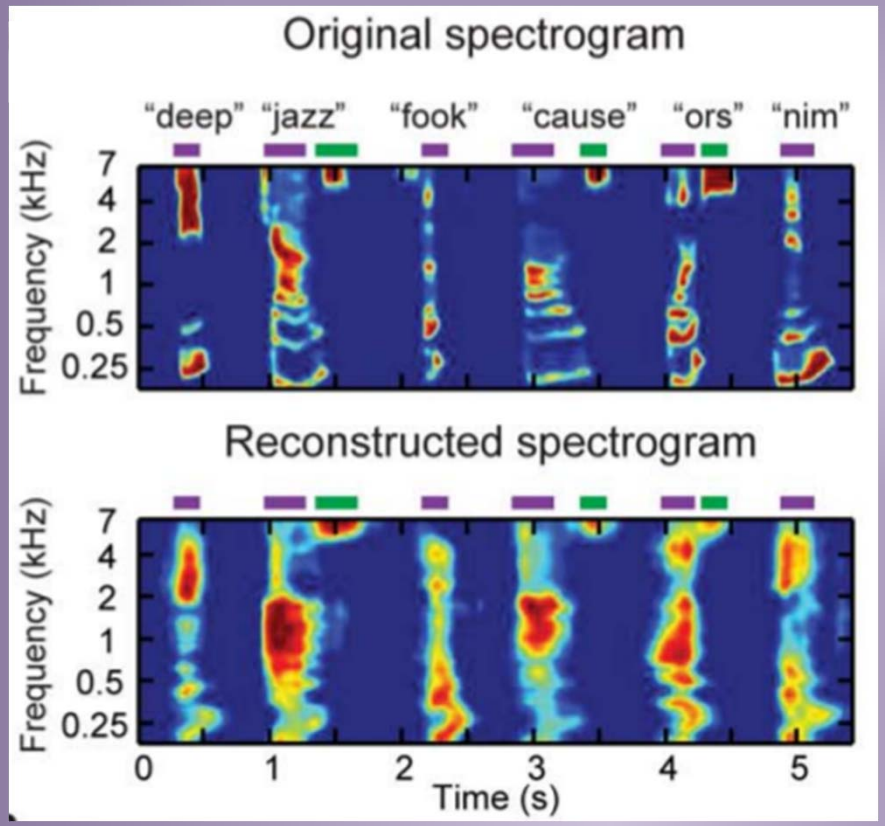
Linear Model

- An approach so far has been to assume a linear mapping between the field potentials and the stimulus spectogram.



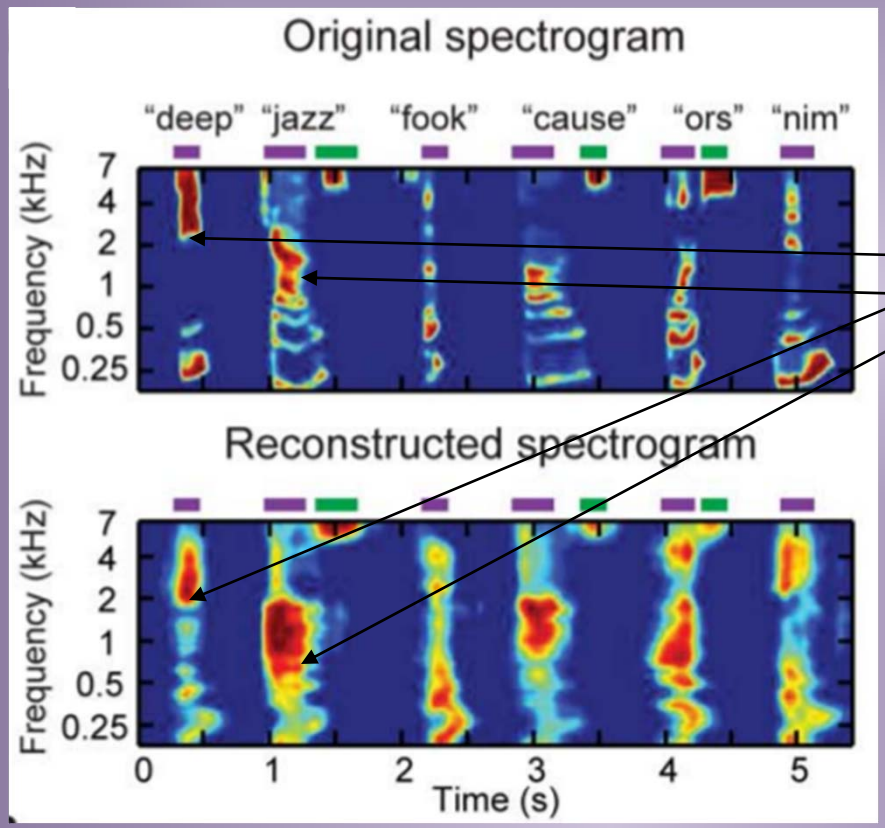
Linear Model

- This approach captures some major spectrotemporal features:



Linear Model

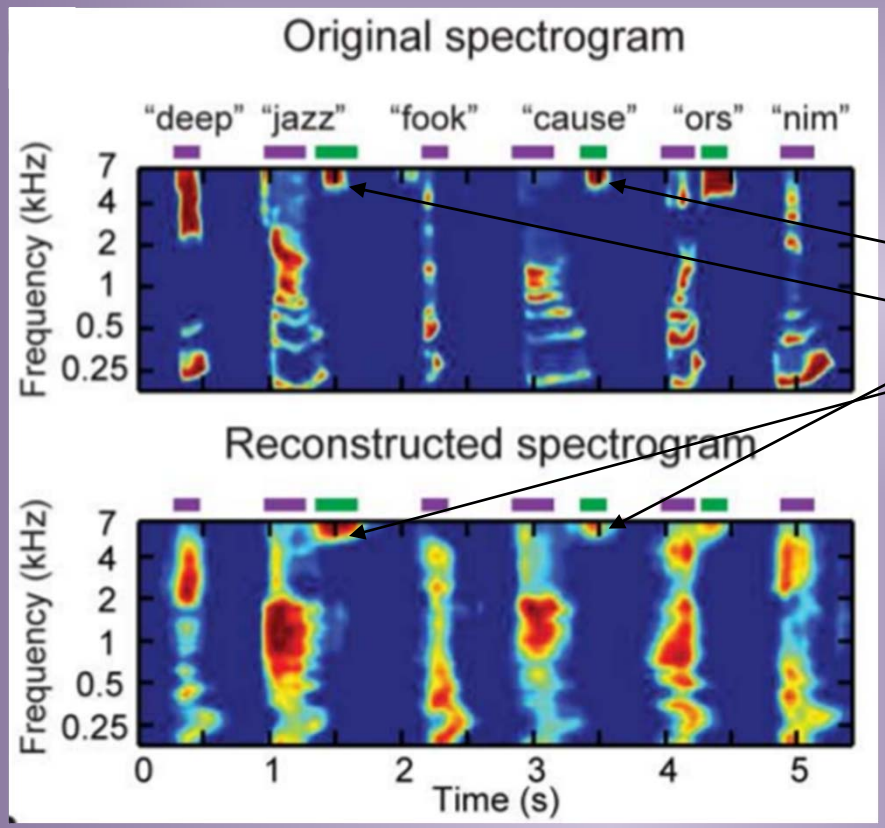
- This approach captures some major spectrotemporal features:



Vowel harmonics

Linear Model

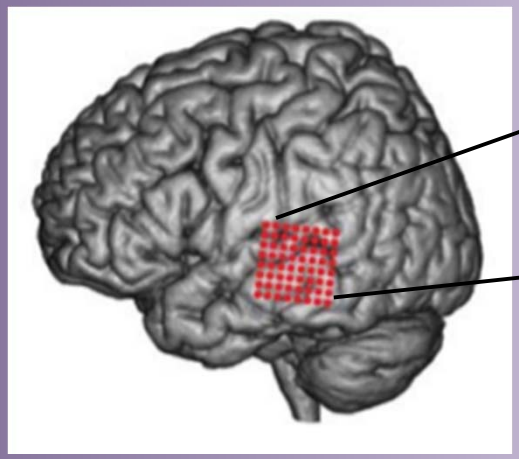
- This approach captures some major spectrotemporal features:

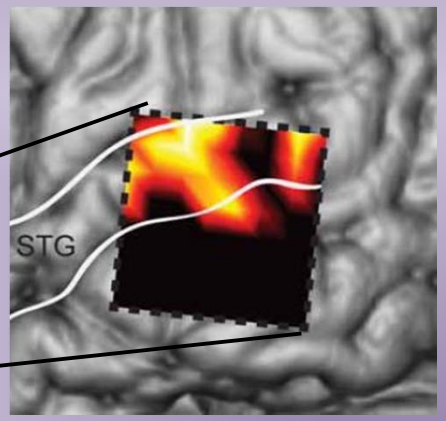


Fricative consonants

Linear Model

- The model revealed that the most informative neuronal populations were confined to pSTG.

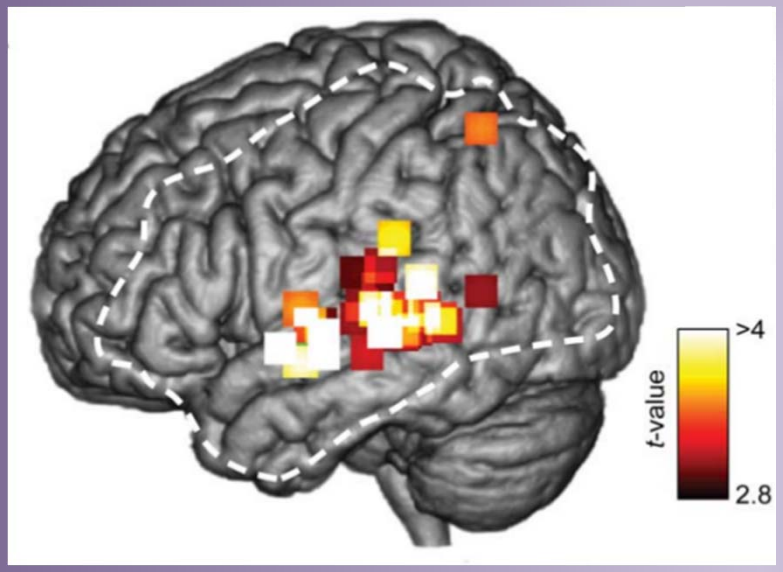




The distribution of the electrode weights in the reconstruction model

Linear Model

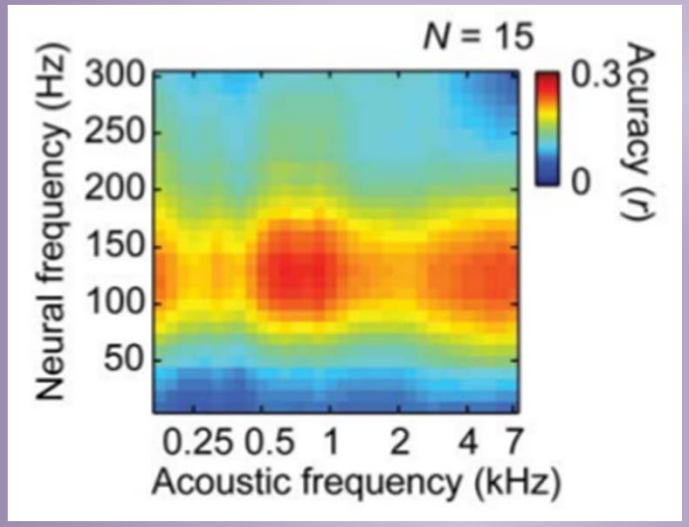
- The model revealed that the most informative neuronal populations were confined to pSTG.



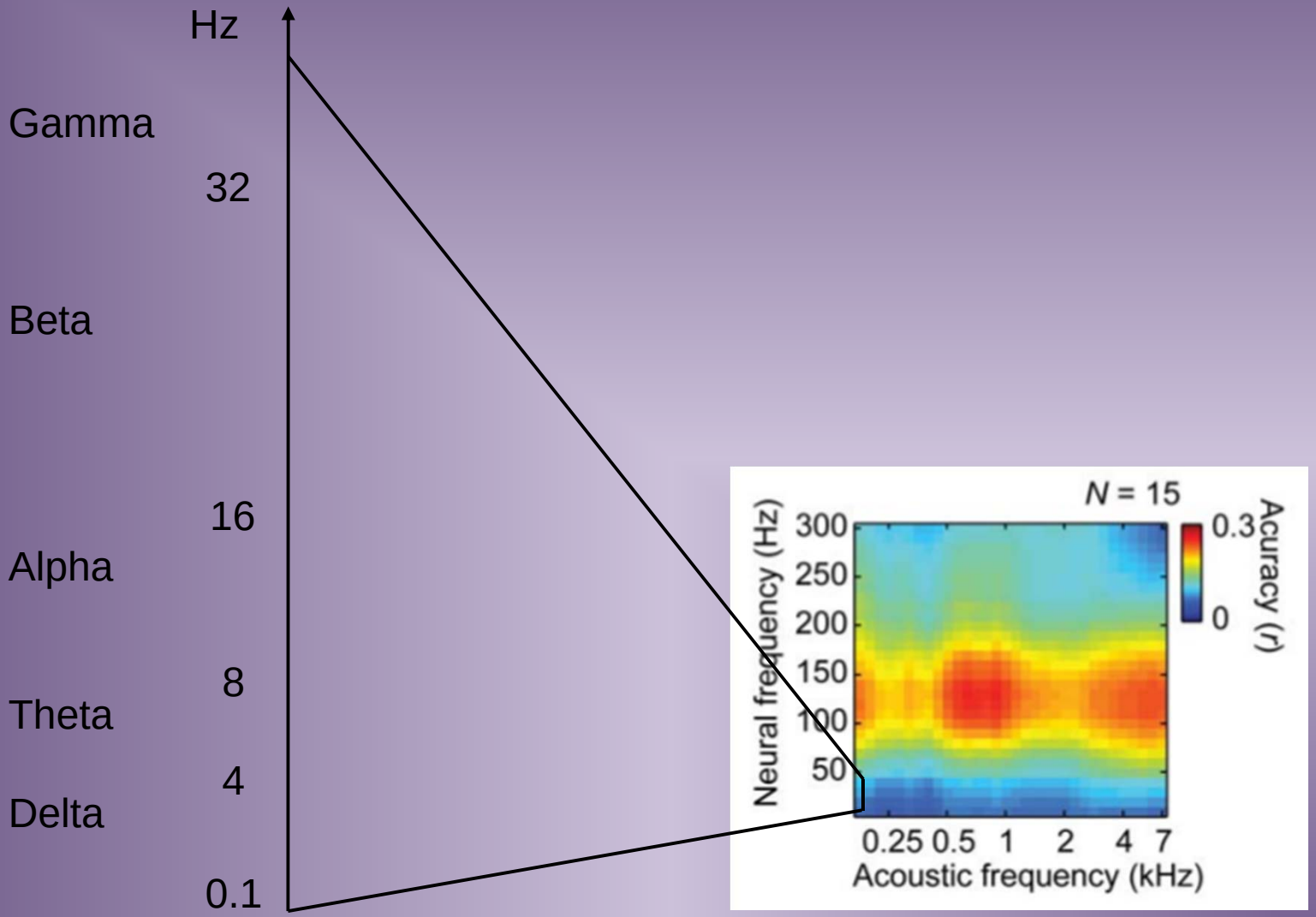
Electrode weights
in the linear model,
averaged across
all 15 participants

Linear Model

- The reconstruction model also revealed that the most useful field potential frequencies were those in the high gamma band 70-170Hz.



Linear Model



Linear Model

- Is this surprising?
- Gamma wave activity has been correlated with feature binding across modalities.
- pSTG is just anterior to the TemporoParietal Junction, a critical area of the brain responsible for integrating all modal information (among many other roles).

Linear Model

- Why does the linear model (i.e. assuming a linear mapping between stimulus spectogram and neural signals) work at all?
- The high gamma frequencies must encode at least some spectrotemporal features.

Linear Model

- Indeed, what made the mapping possible is that neurons in the pSTG behaved well:
    - They segregated stimulus frequencies: as the acoustic frequencies changed, so did the recorded field potential amplitude of certain neuronal populations.

Linear Model

- Interestingly, the full range of the acoustic speech spectrum was encoded in a distributed way across pSTG.
- This differs from the neural nets in the primary visual cortex, which are organized retinotopically.

Linear Model

- Indeed, what made the mapping possible is that neurons in the pSTG behaved well:
    - They responded relatively well to fluctuations in the stimulus spectogram. And especially well to slow temporal modulation rates (which correspond to syllable rate for instance).

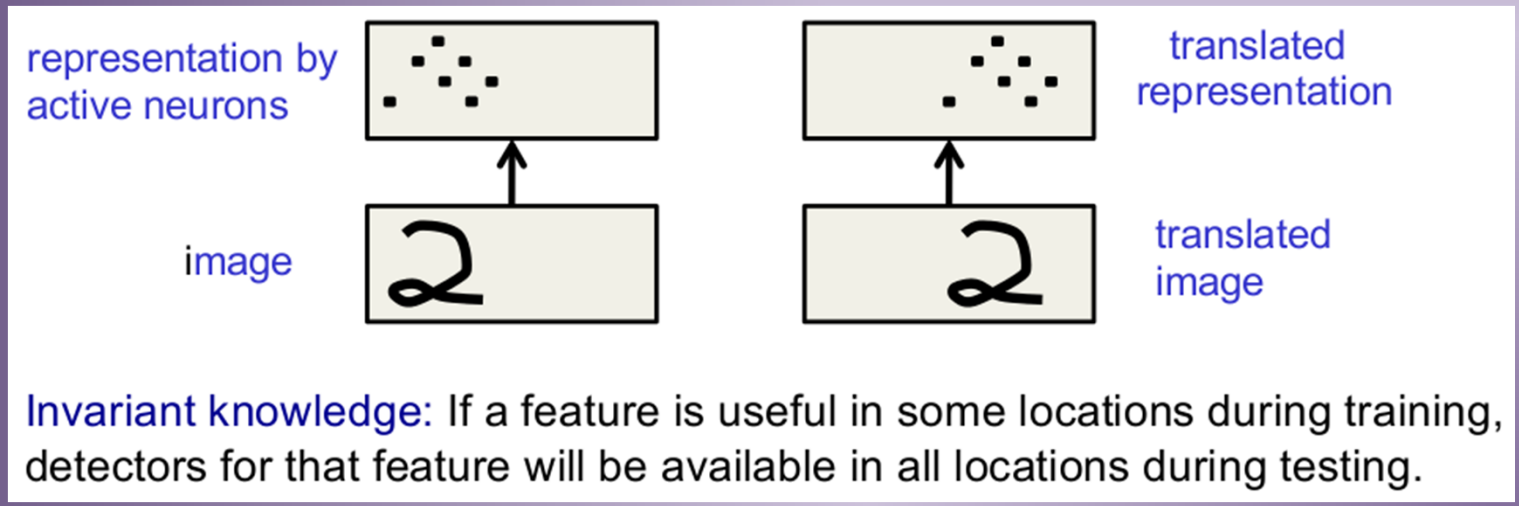
But the Linear Model failed to encode fast modulation rates (such as syllable onset)...

Energy-based Model

- The linear model was ‘time-locked’ to the stimulus spectogram, which did not permit encoding of the full complexity of its (esp. rapid) temporal modulations.
- To lift this constraint, we want a model that doesn't treat time so ‘linearly’.

Energy-based Model

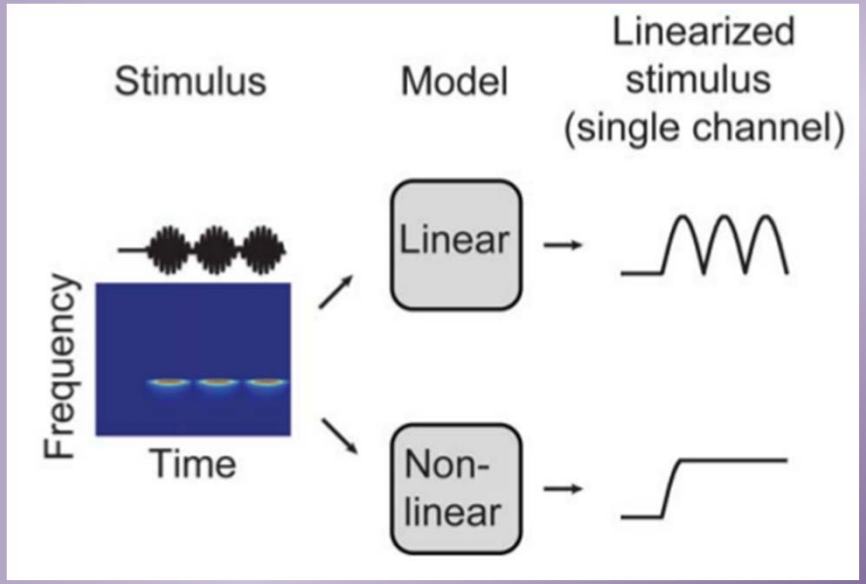
- Consider visual perception. It is well known that, even in the first stages of preprocessing (rods and cones, thalamic relay), encoded visual stimuli is robust to the point of view.



Energy-based Model

- If we can allow the model some (phase) invariance with respect to time, then we might be able to capture those fleeting rapid modulations.

We don't want to track time linearly, we want phase-invariance to capture the more subtle features of complex sounds



Energy-based Model

- Quickly: look over there without moving your head and look back.
- Did you notice that some of your neurons did not fire while others did? But seriously, those who didn't fire kept a 'still' model of space (so you could hold your head up for example).

Energy-based Model

- Why would this intuition about local *space* invariance and visual stimuli hold for local *time* invariance and acoustic stimuli?
- In other words, why would phase invariance help represent fast modulation rates better?

Energy-based Model

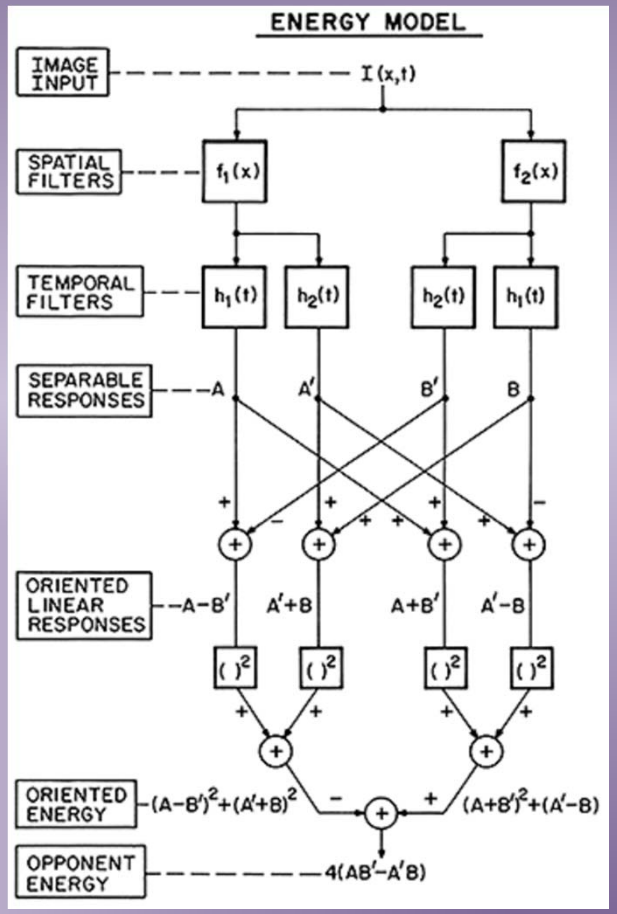
- It might be that tracking exact syllable onset is not necessary for word segregation (just as not tracking every detail of space would help segregate the motionless background from rapid visual stimuli).
- Recall that pSTG is an intermediate auditory processing area.

Energy-based Model

- So instead of a spectrotemporal stimulus representation at this intermediate stage, it could be that neuronal populations in pSTG (via the field potentials they emit) focus on encoding the '**energy**' (amplitude) of these (higher-order) modulation-based features.

Energy-based Model

- Energy-based models have been around for decades, and have been used extensively for modeling nonlinear, abstract aspects of visual perception.



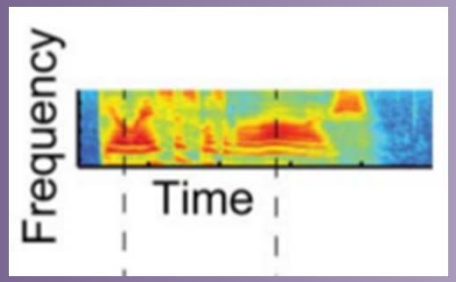
The Adelson-Bergen energy model (Adelson and Bergen 1985)

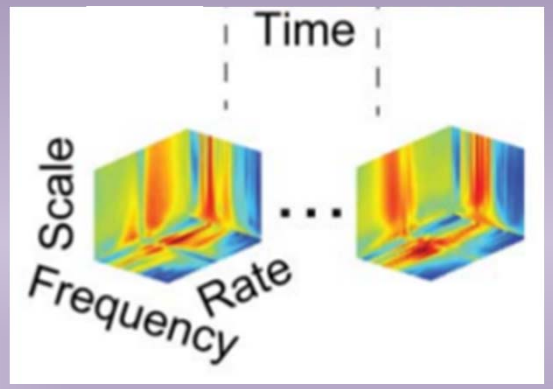
Energy-based Model

- Chi et al. 2005 proposed a model that represents modulations (temporal and spectral) explicitly as multi-resolution features.
- Their nonlinear (phase invariant) transformation of the stimulus spectogram involves complex modulation-selective filters that extract the modulation energy concentrated at different rates and scales.

Energy-based Model

- Feature extraction in the energy-based model:





The input representation is the two-dimensional spectrogram S(f,t) across frequency f and time t.

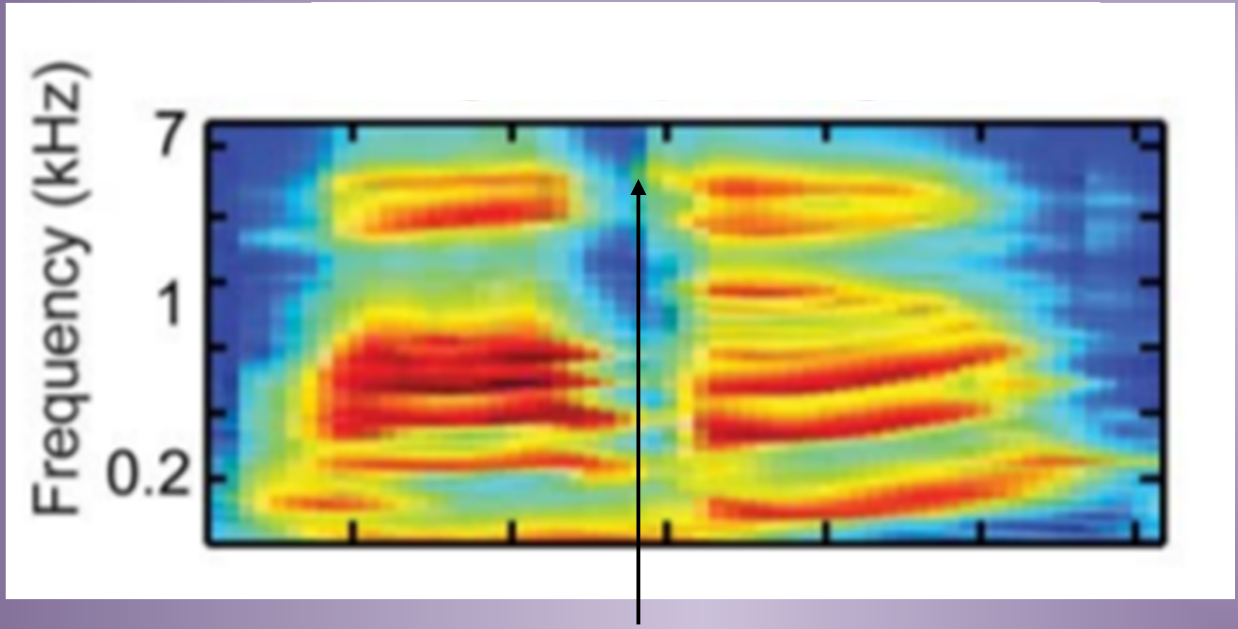
The output is the four-dimensional modulation energy representation M(s,r,f,t) across spectral modulation scale s, temporal modulation rate r, frequency f, and time t.

Energy-based Model

- The energy-based model thus achieves invariance to local fluctuations in the spectogram.
- This is in par with neural responses in the pSTG: very rapid fluctuations in the stimulus spectogram did not induce the 'big' changes the linear model was expecting.

Energy-based Model

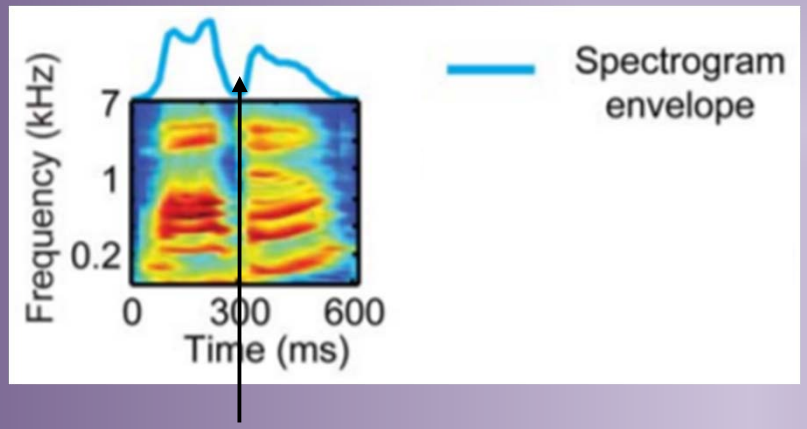
- Consider the word “WAL-DO” whose spectogram is given below:

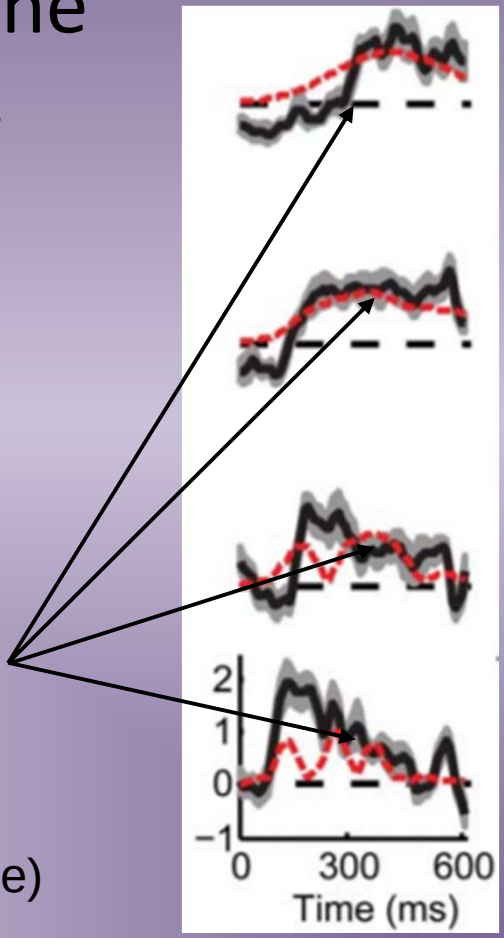


Notice the rapid fluctuation in the spectogram along this axis (300ms into the word Wal-do)

Energy-based Model

- On the right: Field Potentials (in the high gamma range) recorded at 4 electrode sites:

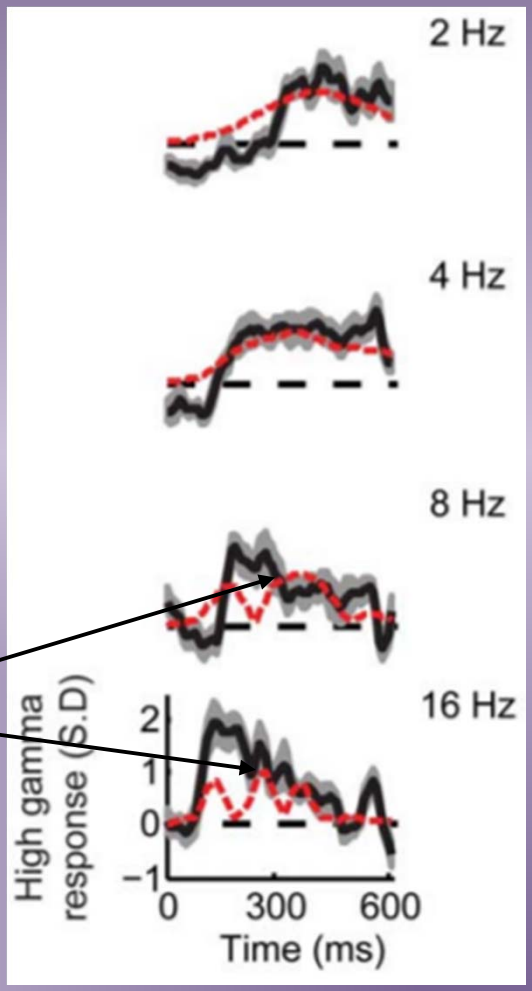




None of these rise and fall as quickly as the Wal-do spectogram does at around 300ms (actually no linear combination of them can be used to track this fast change)

Energy-based Model

- Superimposed, in red, are the temporal rate energy curves (computed from the new representation of the stimulus, for 2, 4, 8 and 16Hz temporal modulations):



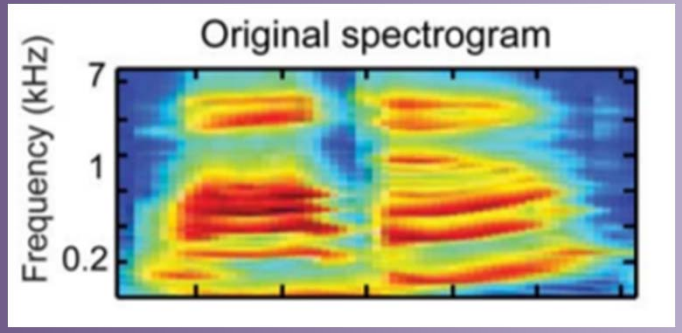
Notice that for fast temporal fluctuations (>8Hz), the red curves 'behave more informatively' at around 300ms

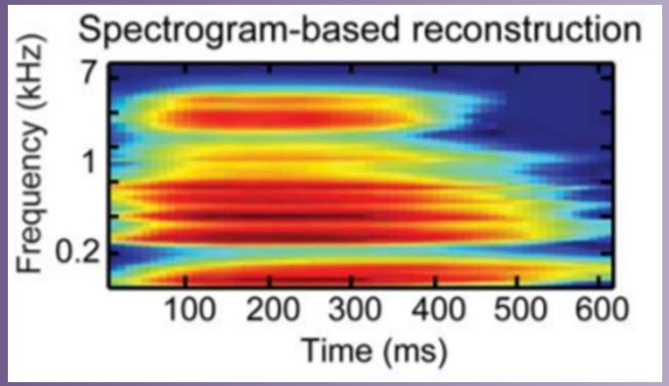
Energy-based Model

- Given the new (4D) representation of the stimulus, the model can now capture these variations in temporal energy (fast vs. slow fluctuations) from the neural field potentials more reliably.

Energy-based Model

- The linear model was too concerned with time, that it wasn't paying attention to time variation.

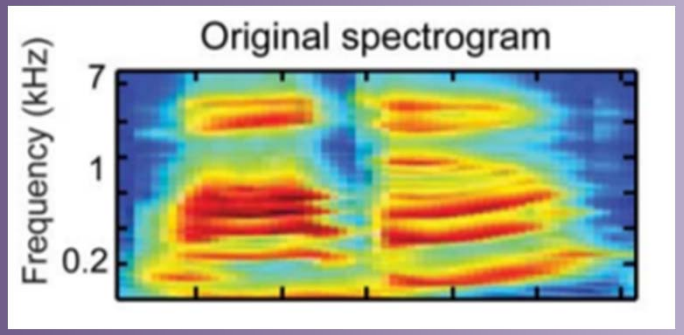


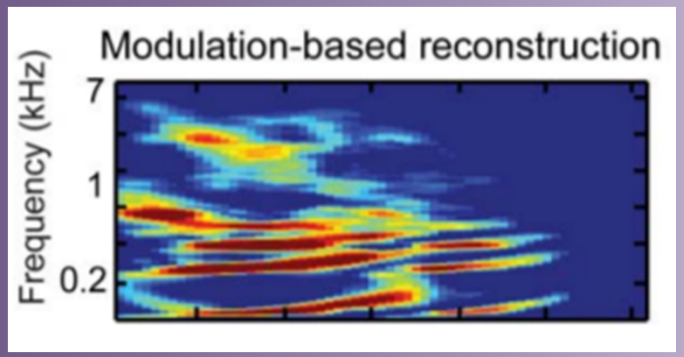


The linear model cannot segregate time variations at the scale of syllable onset

Energy-based Model

- Thanks to local temporal invariance, the energy-based model can now encode more sophisticated features.

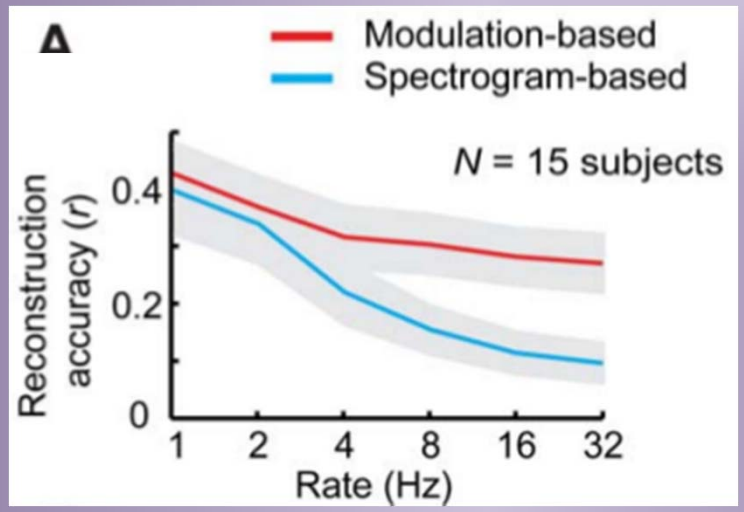




The energy-based model can decode field potentials in more detail

Energy-based Model

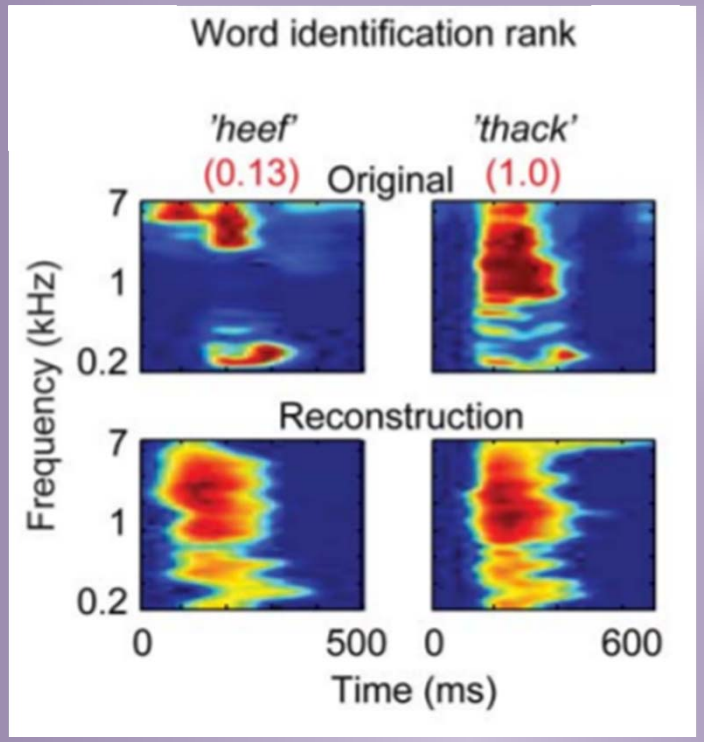
- Plotted below is the reconstruction accuracy of spectrotemporal features of the stimulus.
- Reconstruction of fast temporal energy is much better in the energy-based model.



But is this enough to let us decode words from reconstructed spectograms?

Mind reading in practice

- Pasley et al. tested the energy-based model on a set of 47 words and pseudowords (e.g. below).



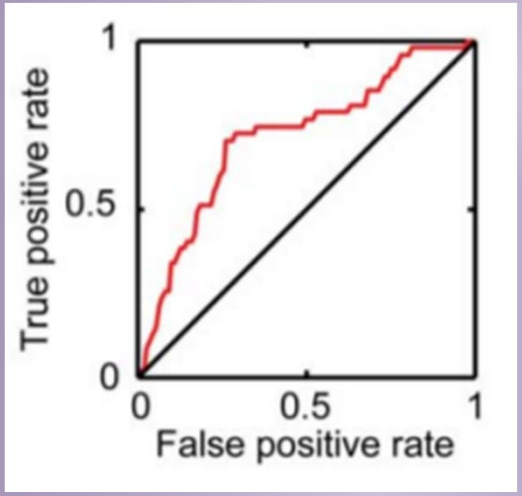
Mind reading in practice

- They used a generic speech recognition algorithm to convert reconstructed spectograms into words.
- In general, reconstruction was of poor quality.

journal.pbio.1001251.s009.wav

Mind reading in practice

- But on the 47 word set, they achieved better word recognition than would be expected by just coin flipping.



Mind reading in practice

- So it seems that we are far from being able to read minds.
- What can we do about it?

Mind reading in practice

- The results coming from Pasley et al.'s incredible study of pSTG field potential gives us hope.
- We know that those field potentials don't encode spectrotemporal features of speech information linearly. Pasley and colleagues point out a plausible dual encoding: spectogram-based for slow temporal rates, modulation-based for faster ones.

Mind reading in practice

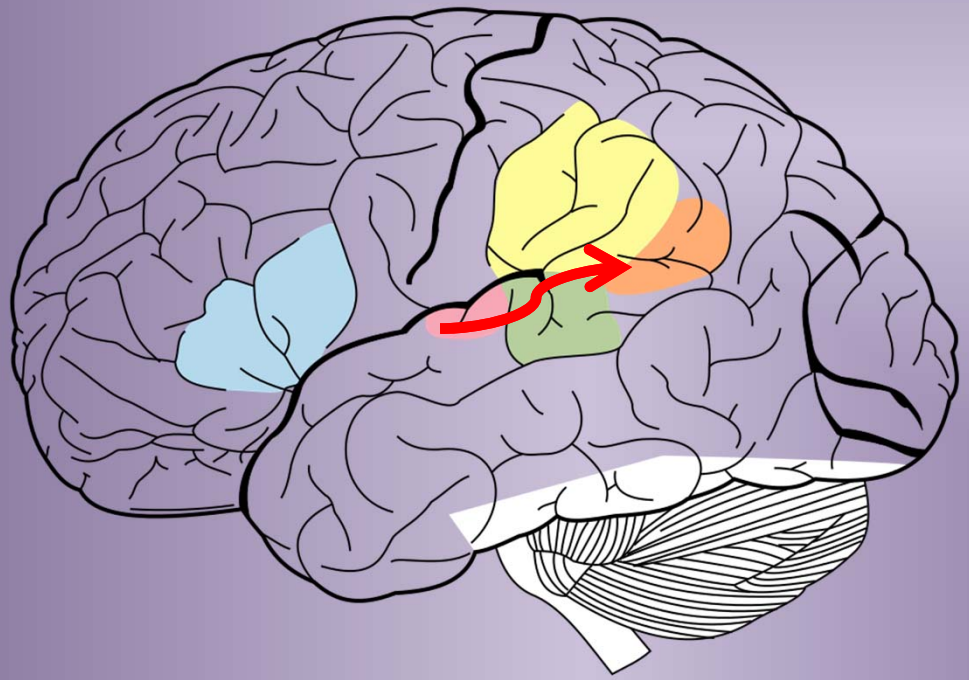
- But how would we measure cortical field potentials extracranially?
- Is it possible to expect intrusive cybernetic implants on the cortex of aphasic patients in the future?

Mind reading in practice

- Or is it more likely that we could extrapolate a (more powerful) model to convert neural signals recordable from a simple scalp implant or headset?

Mind reading in practice

- The entire ventral pathway of speech recognition could be monitored to allow for better feature detection.



Mind reading in practice

- But we still have a long way to go.

- Although…

Mind reading in practice

- I don’t need to read your minds to know that you are hungry by now…

So Thank You!

Credits

- Pasley et al. (conducted the study)
- Back to the Future (doc Brown)
- Wikipedia user Rocketooo (brain svgs)
- Futurama (Zoidberg meme)
- University of Toronto ML group (example from CSC321)
- Adelson and Bergen 1985 (energy model diagram)
- Where's Waldo (comic strips)
- Star Trek (Spock's brain)

Extra slide

- Because without this slide, this presentation would have 59 slides.