Sensorimotor Interaction and Temporal Modulation

Louis Goldstein
University of Southern California

Sources of Evidence for Sensorimotor interaction

- Active use of sensory information in speech production:
  - Imitation and vocal learning (Meltzoff & Moore, 1977; 1997)
  - Adaptive responses to perturbed auditory feedback in speech production (e.g., Houde & Jordan, 1998)
  - Synchronous speech (e.g., Cummins, 2002)
  - Articulatory convergence (e.g., Lee et al, 2018)

Motor Engagement in Speech Perception

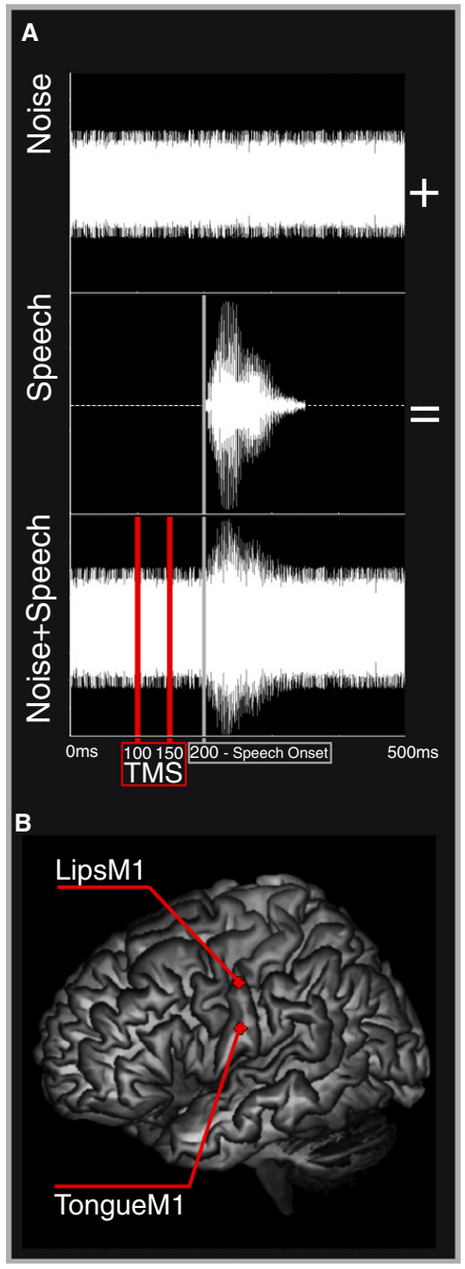
The Motor Somatotopy of Speech Perception

Alessandro D'Ausilio,[1] Friedemann Pulvermüller,[2] Paola Salmas,[3] Ilaria Bufalari,[1] Chiara Begliomini,[1] and Luciano Fadiga[1,3,*]

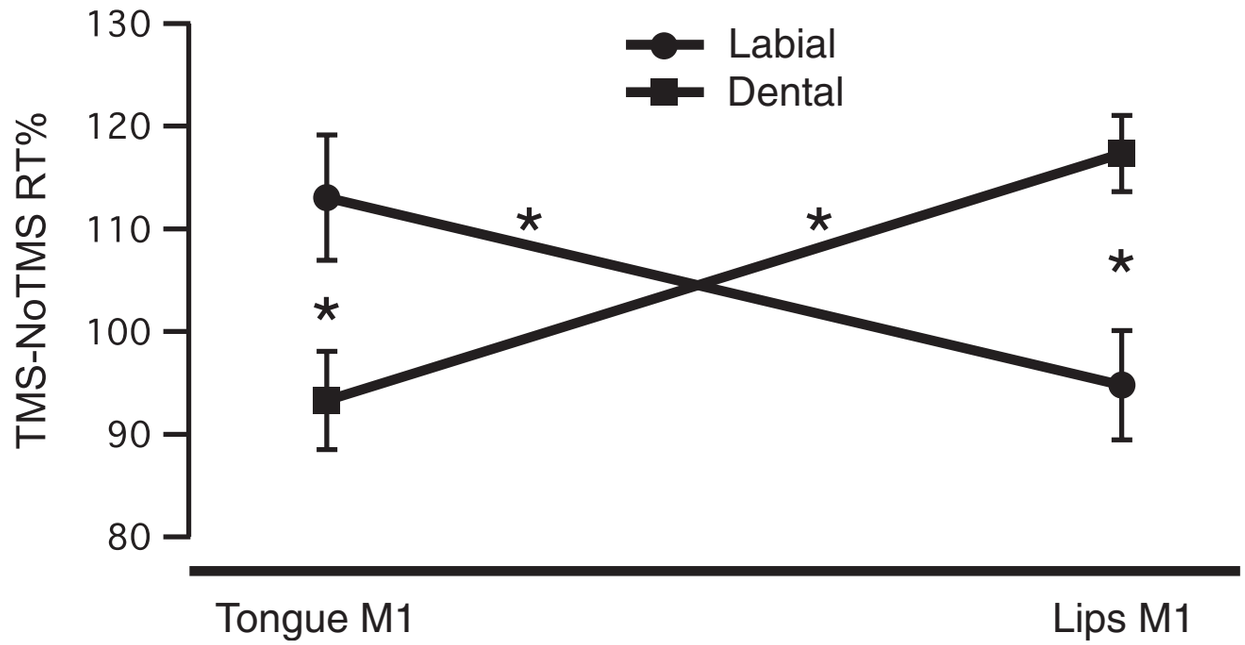
Current Biology *19*, 381–385, March 10, 2009 DOI 10.1016/j.cub.2009.01.017

- Hypothesis: If speech perception engages neural circuitry specific to production of distinct speech gestures, then pre-activation of the motor area that is compatible with particular percept (response on a perception task) should enhance its response and inhibit other responses.

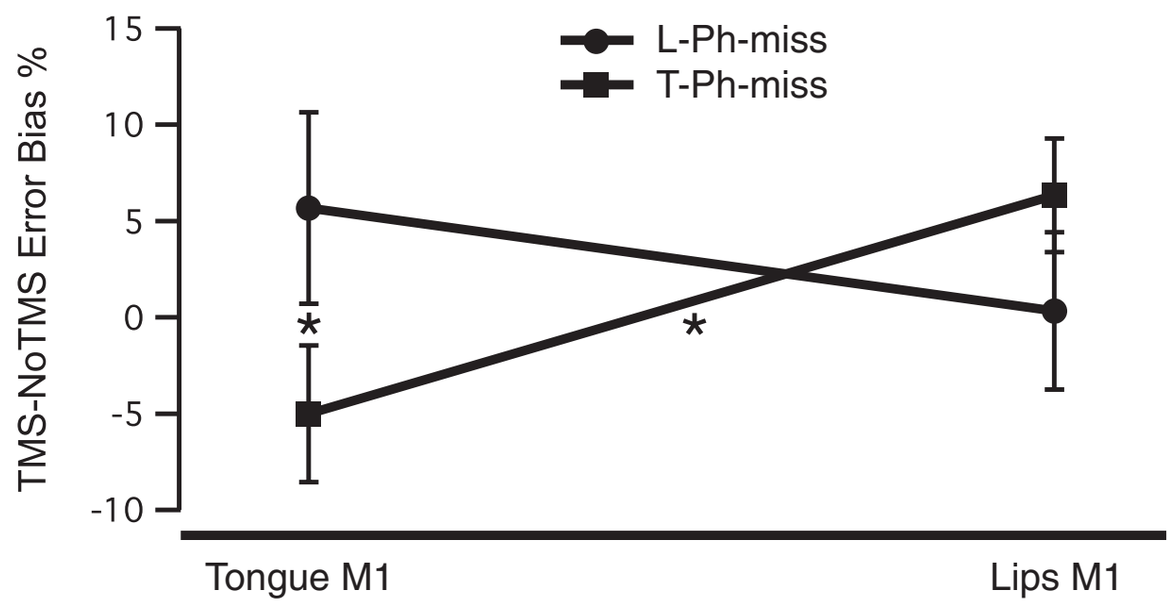
- Present /bæ/ /pæ/ /dæ/ /tæ/
- Noise
- TMS vs No-TMS



Reaction Time (RT)



Errors



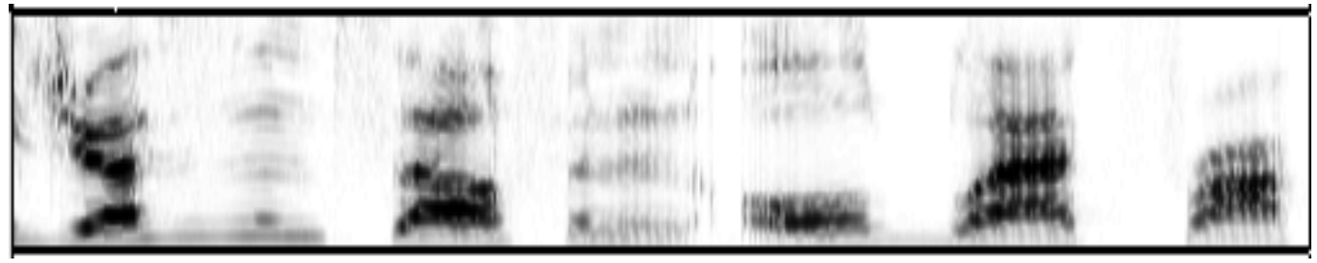
Binding Problem

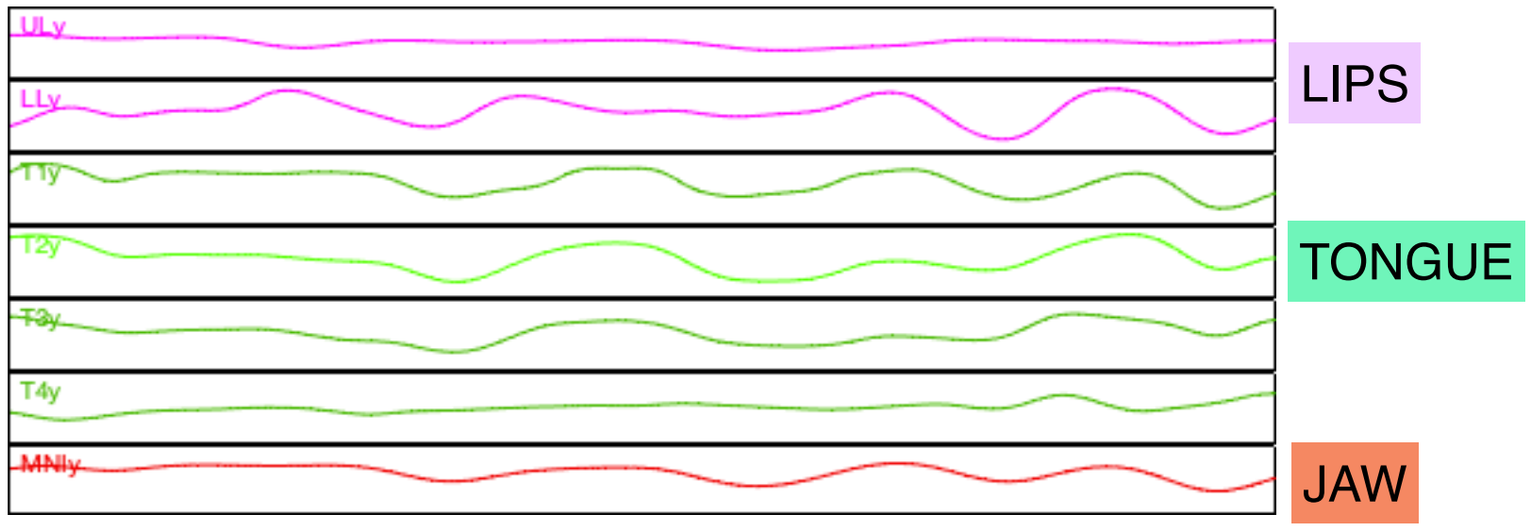
- How can auditory and articulatory neural representations can be "aligned" to one another facilitate sensorimotor interaction?

- Peripheral acoustic and articulatory signals are very distinct from one another.

- What about the neural representations of speech?
  - Auditory
    (Superior Temporal Gyrus)
  - Motor
    (Ventral Sensorimotor cortex)

"He dresses himself in an old black frock"





Possible to use deep learning to map from one to another (Anumanchipalli et al, 2019).

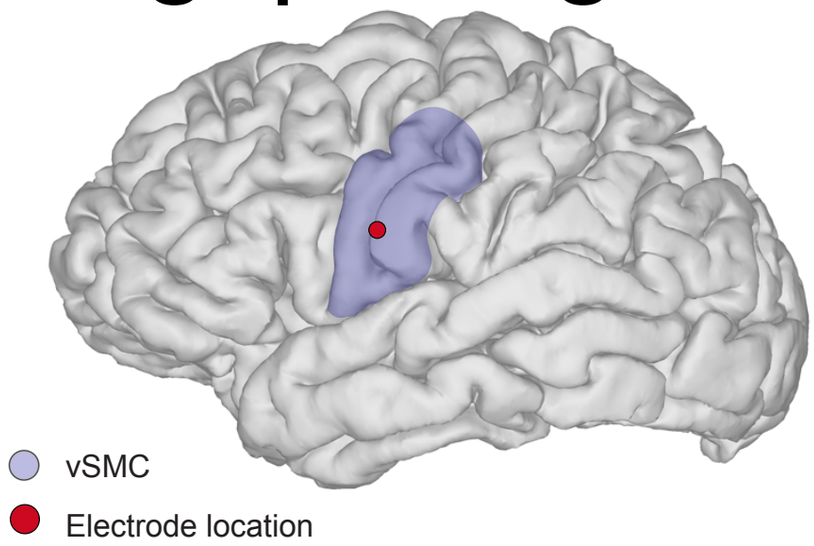
But is this how humans solve the binding problem?

Activation in motor cortex during listening

- Evidence for activation in the motor cortex during listening to speech (e.g., Wilson et al, 2004)
- Can this provide the basis for the alignment?
- Compare the structure of:
  - auditory cortex representations during listening
  - motor cortex representations during speaking
  - motor cortex representations during listening

Activation in vSMC during speaking

- Electrocorticography (ECoG) application of a mesh of tiny electrodes directly on the surface of the brain of a patient who is being prepared for brain surgery.
- Allows recording from very small populations of neurons.
- Examine multiple sites in vSMC *while patient is speaking*.
- 460 read sentences (MOCHA-TIMIT)
- 130 electrode sites (across 5 participants)
- Test which descriptions of speech best predict patterns of activation in particular electrode locations (phoneme id, formants, constriction formation, individual articulators).

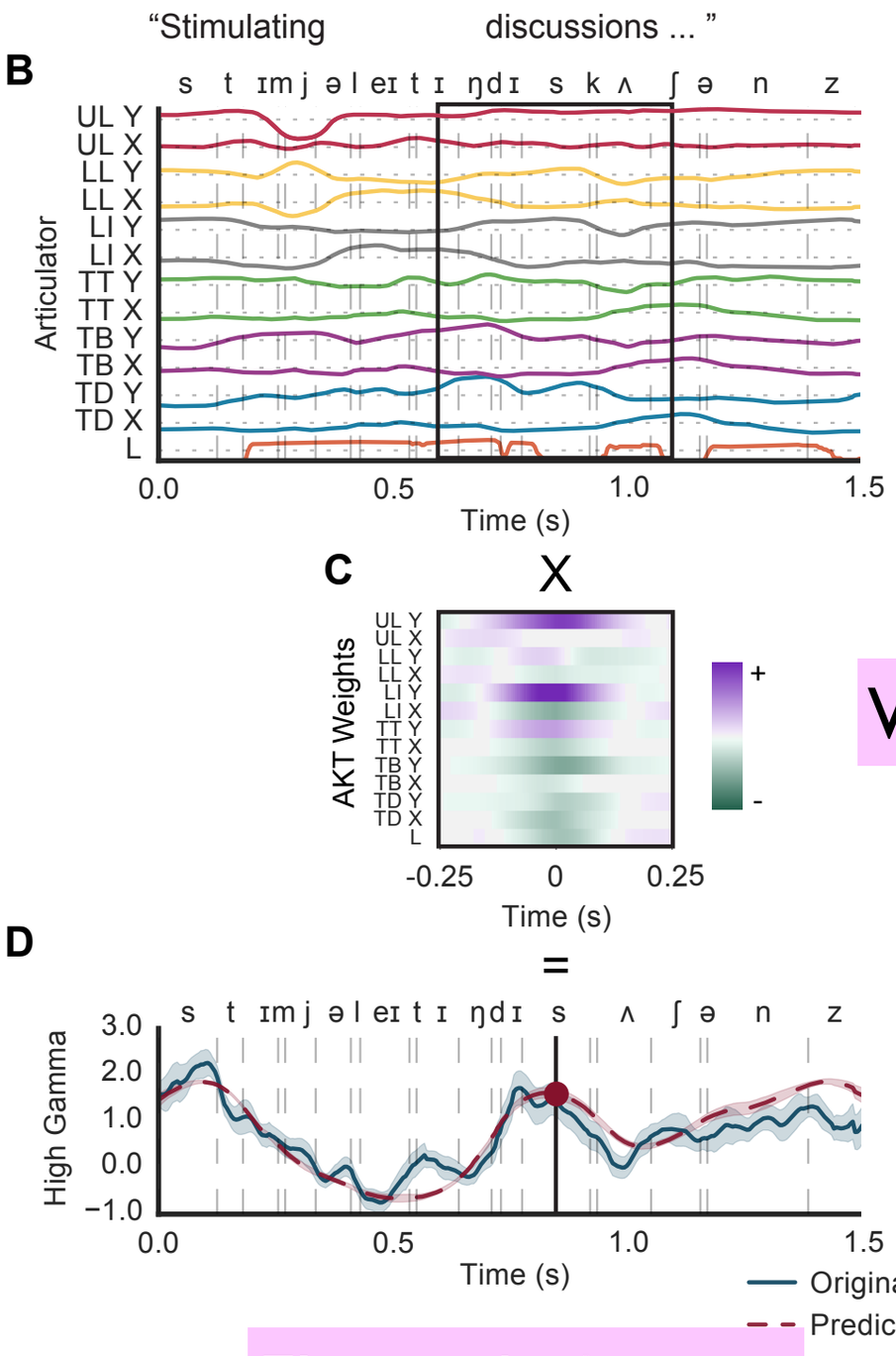


Because only acoustics are recorded, authors trained a model to infer time functions of EMA markers on Lips, Tongue, and Jaw audio (overall correlation of original and inferred EMA = ~.65 for untrained speaker).

Josh Chartier, Gopala K. Anumanchipalli, Keith Johnson, Edward F. Chang (2019). Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron*.

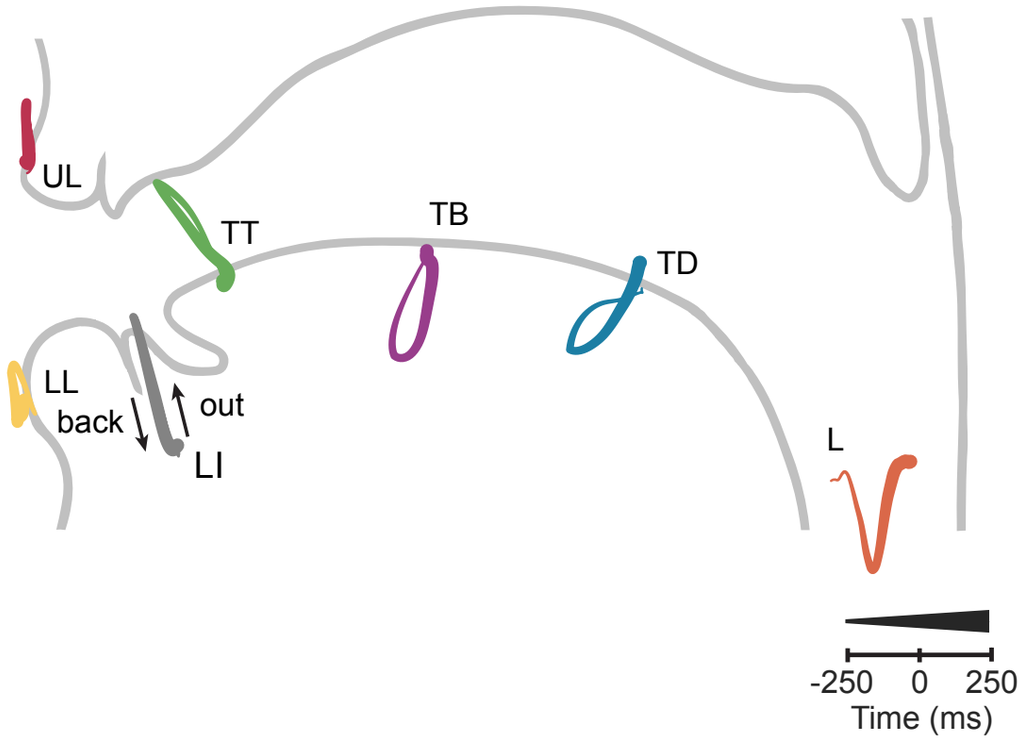
Example results

Inferred EMA



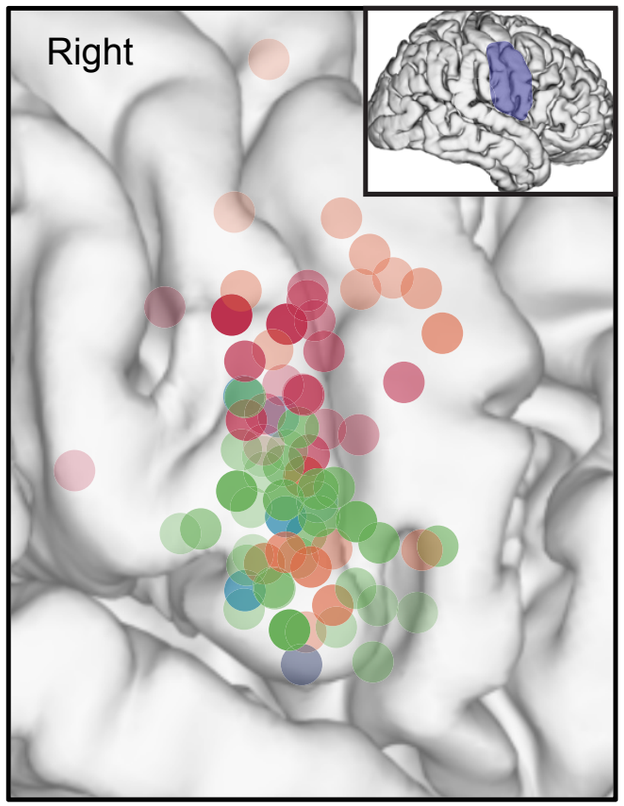
Weights

Electrode Activity



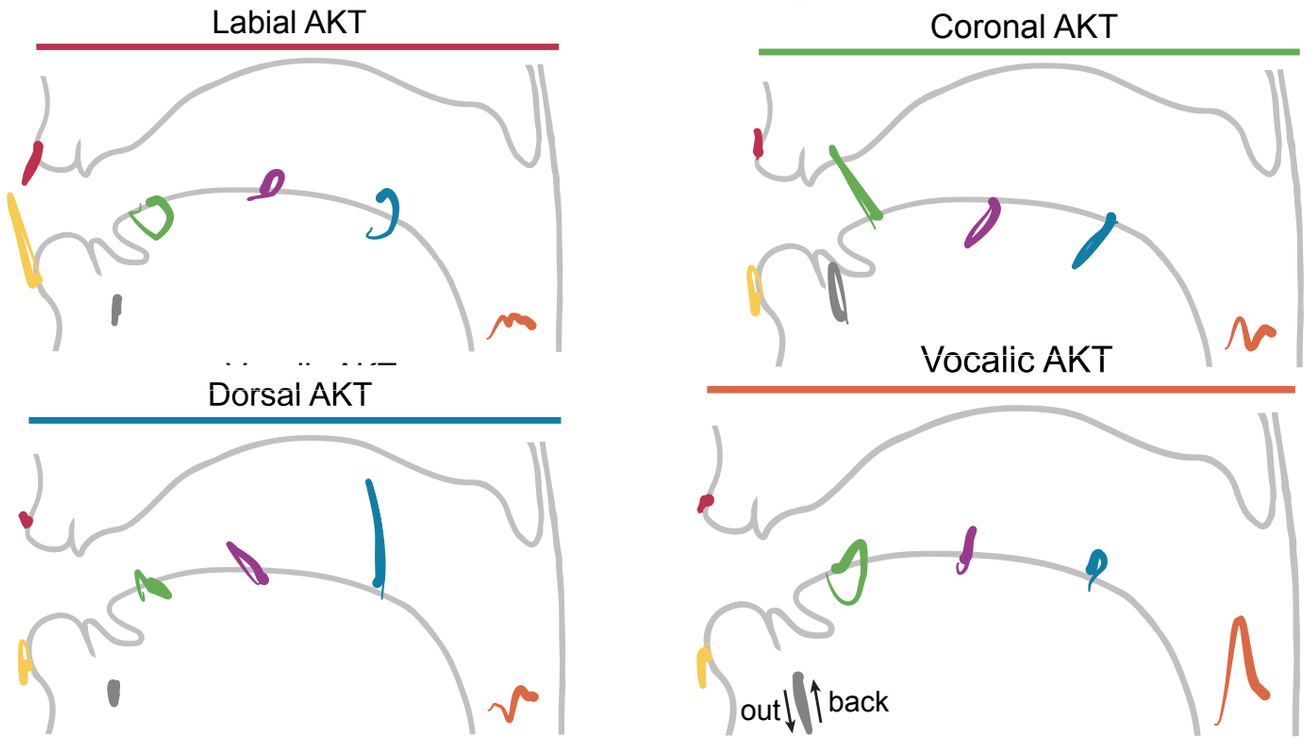
Weight pattern corresponds to coordinated articulator motion that produces and releases a coronal constriction.

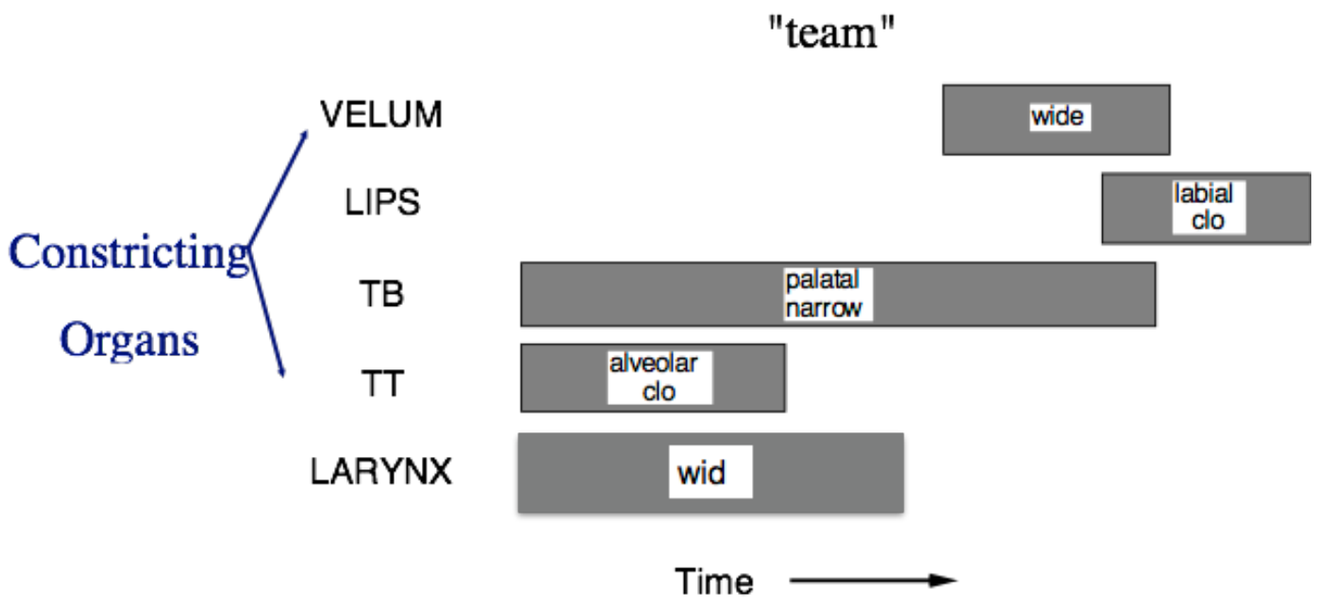
Sites code distinct constriction gestures



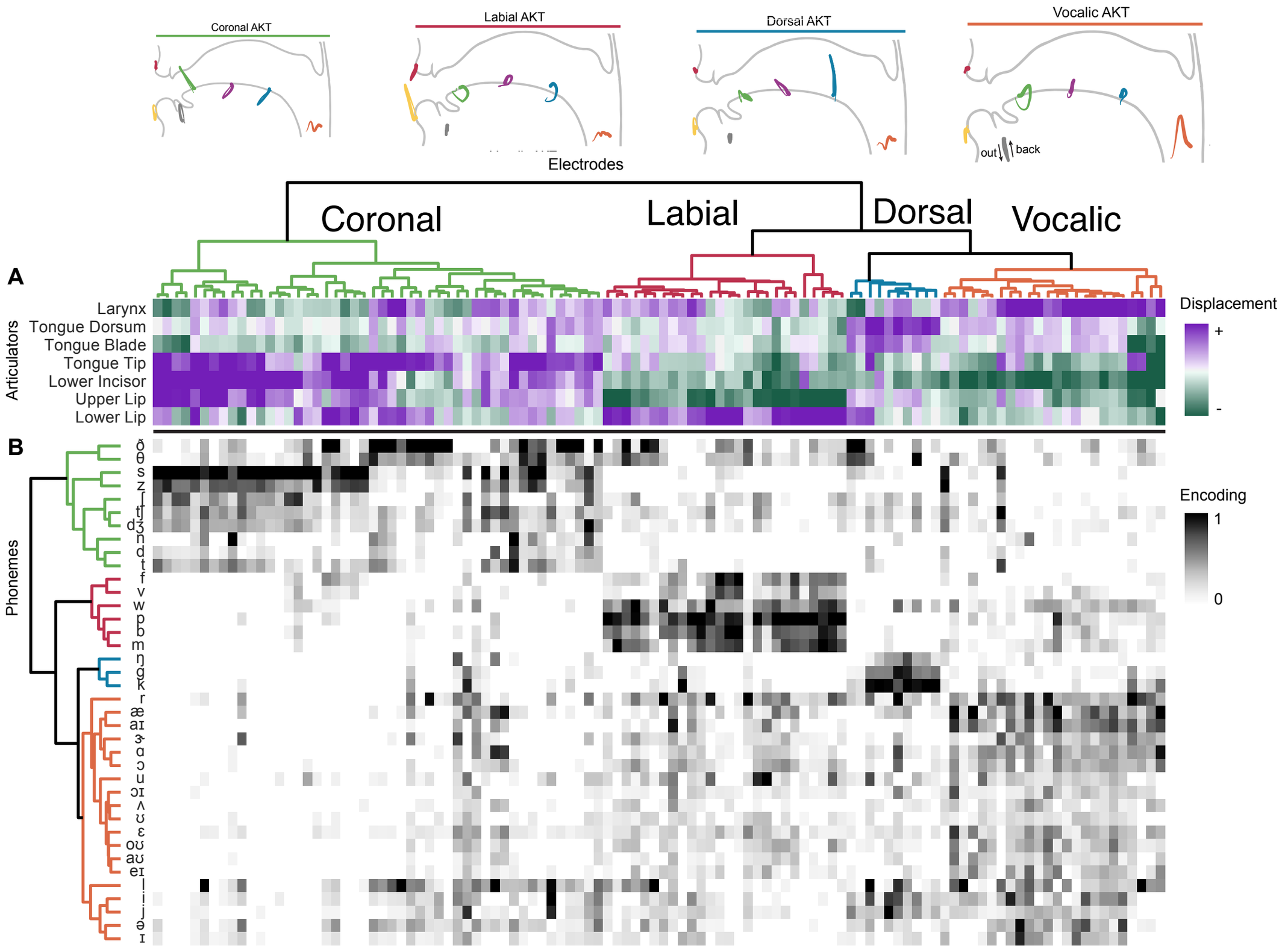
AKT Cluster
- Coronal
- Labial
- Dorsal
- Vocalic

pearson r 0.4 0.0





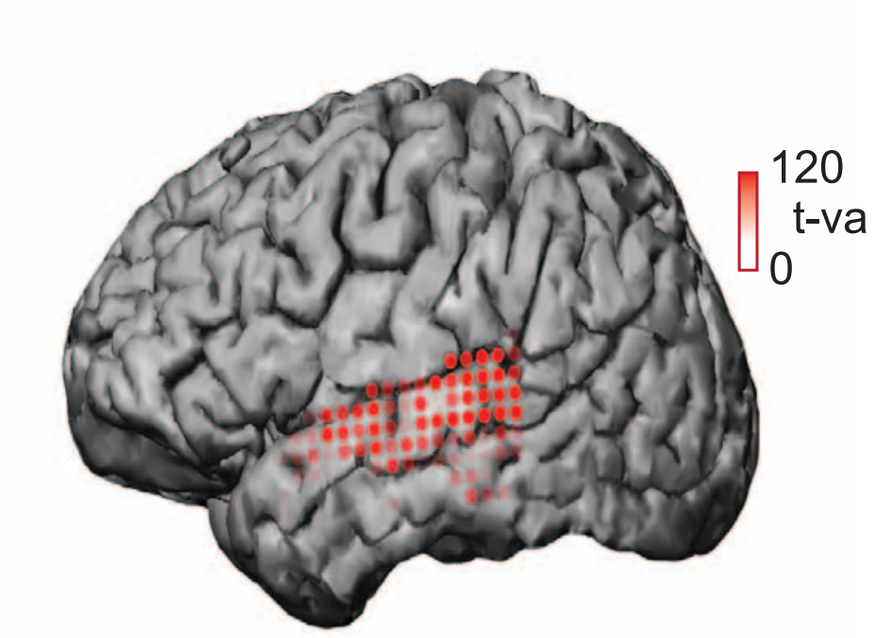
- Best predictor of electrode activity was kinematic articulatory pattern associated with a gesture: coordinated articulator activity that produces and releases a constriction.
- Organized by constriction organ ("place of articulation").



Clustering of electrodes (by articulator patterns)
Clustering of segments (by electrode patterns)

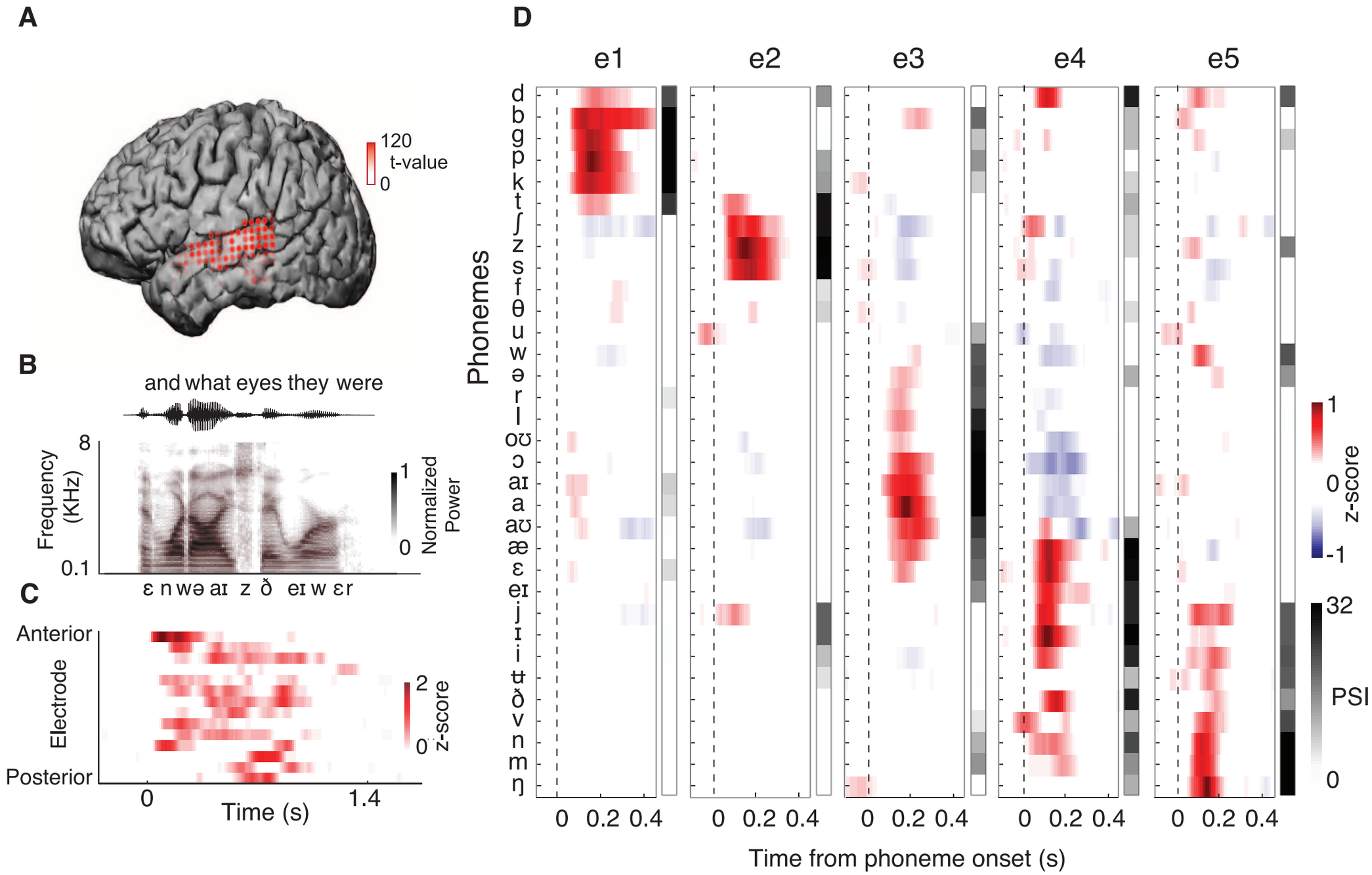
Activation in STG during listening

- Superior Temporal Gyrus:
- site of auditory representations
- Similar method as used to investigate motor cortex (vSMC) during speaking.
- 6 participants
- Listened to 500 sentences (TIMIT)
- 256 total electrodes
- How do segments cluster in their patterns of electrodes activation?
- What acoustic patterns are encoded?



Phonetic Feature Encoding in Human Superior Temporal Gyrus
Nima Mesgarani *et al.*
Science **343**, 1006 (2014);

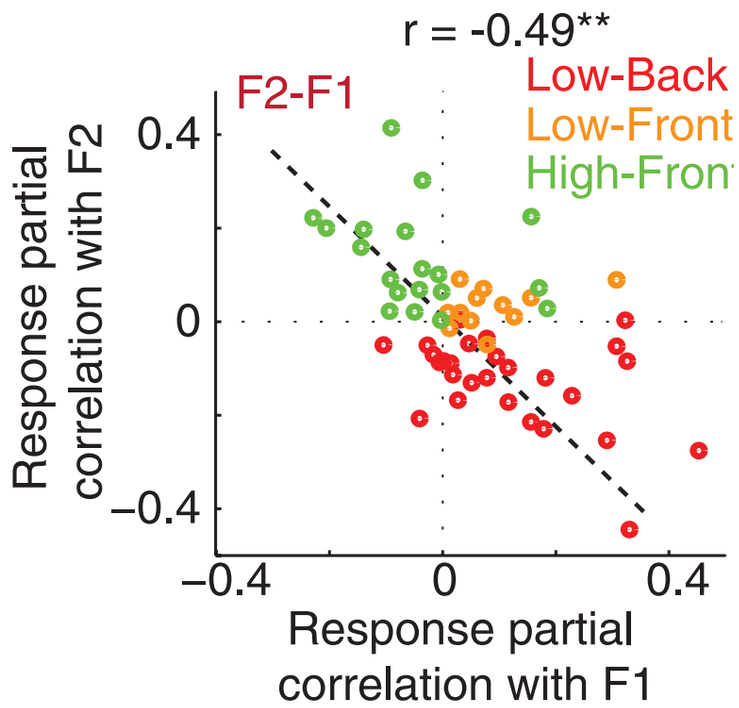
Activation in STG during listening



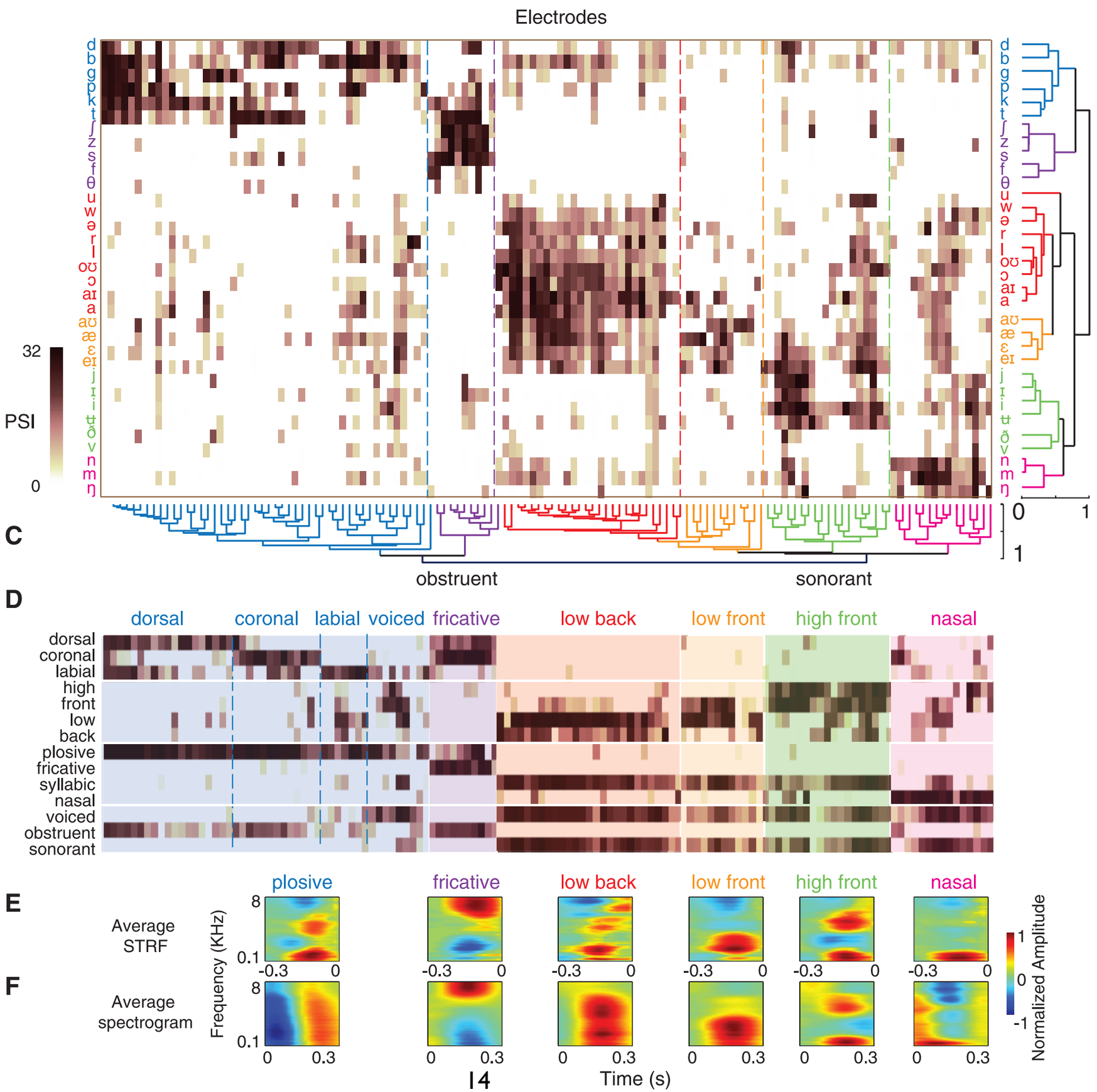
Phonetic Feature Encoding in Human Superior Temporal Gyrus
Nima Mesgarani *et al.*
Science **343**, 1006 (2014);

STG representation during listening

- Electrode activity clusters by manner class.
  - stops
  - fricatives
  - nasals
  - back vowels & liquids
  - low front vowels
  - high front vowels

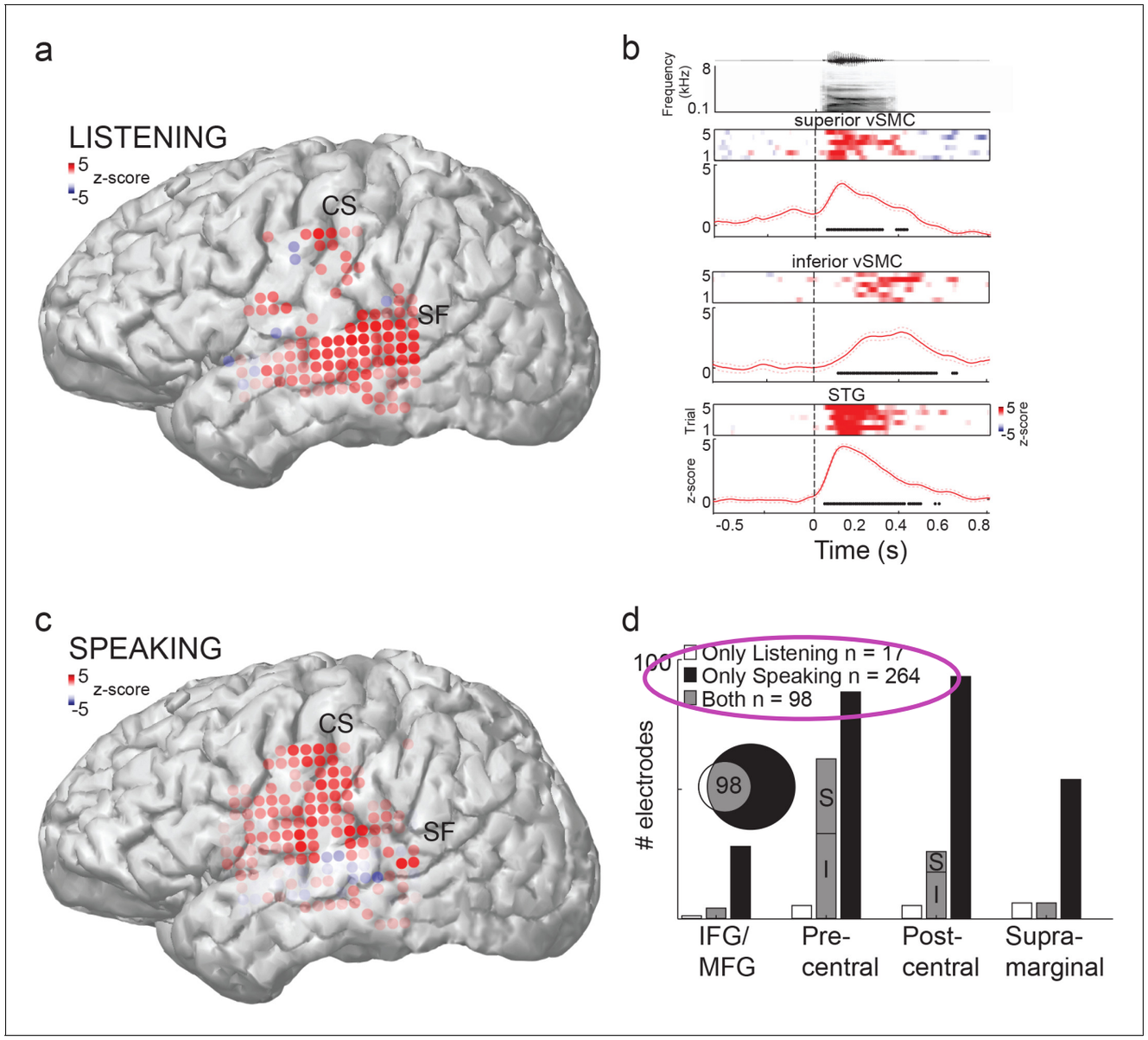


- Classes differ in gross acoustic patterns
- More fine grained representation of vowel formants.
  - Electrodes tuned to relation of F1 and F2.


Electrodes
32
PSI
0
C
obstruent
sonorant
D
dorsal
coronal
labial
high
front
low
back
plosive
fricative
syllabic
nasal
voiced
obstruent
sonorant
E
F
Average
STRF
Average
spectrogram
Frequency (KHz)
8
0.1
-0.3
0
0.3
Time (s)
Normalized Amplitude
1
0
-1

Neural activation in motor areas (vSMC) during listening

- Several studies have revealed activity in motor cortex during passive listening.
- Has been used as evidence for motor engagement during perception.
- Little is known about the structure of the neural activation during listening.
- Cheung et al (2016) measure electrode activity in both STG and vSMC during listening and speaking
- Stimuli:
  /pa ba ta da ka ga sa ʃa/
- Nine participants



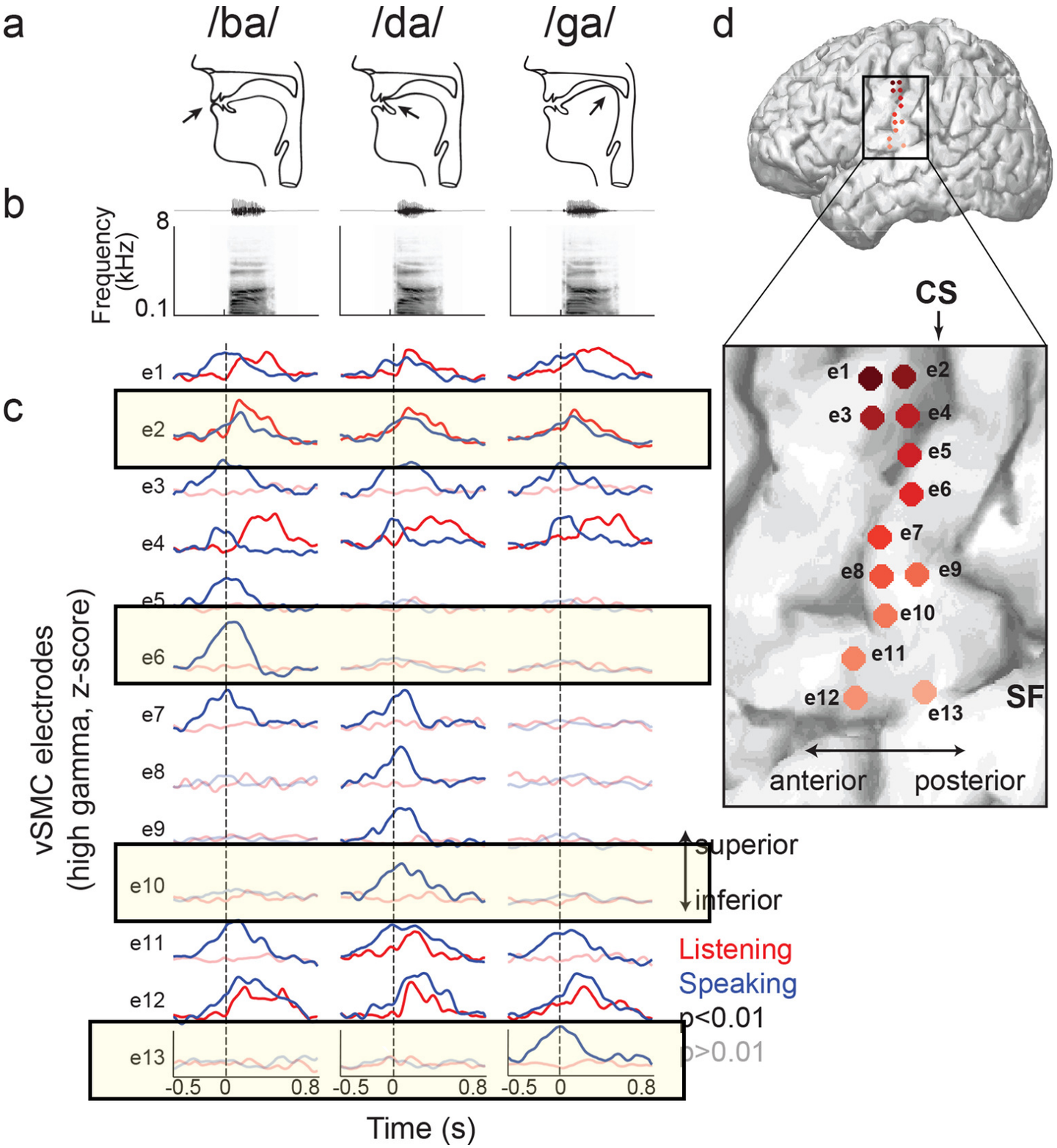
The auditory representation of speech sounds in human motor cortex

Connie Cheung[1,2,3,4†], Liberty S Hamilton[2,3,4†], Keith Johnson[5], Edward F Chang[1,2,3,4*]

Cheung *et al*. eLife 2016;5:e12577. DOI: 10.7554/eLife.12577

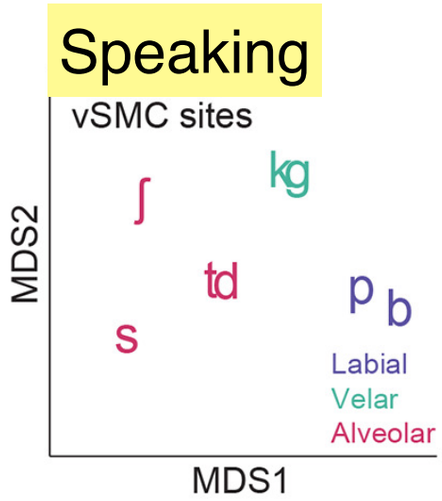
Selectivity of vSMC electrodes during speaking and listening

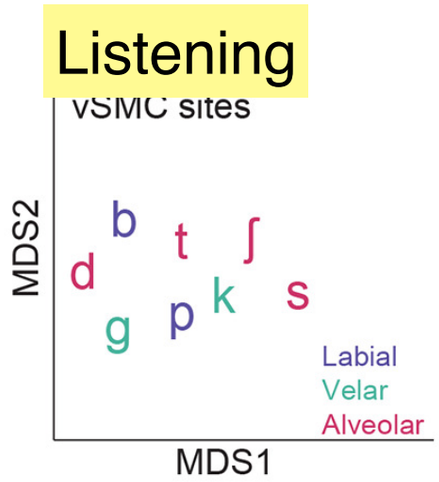
- Electrodes found that respond differentially to /b,d,g/ are typically those that are active only during speaking.
- Electrodes that are active during both listening and speaking do not exhibit clear selectivity as a function of constriction effector (labial, coronal, dorsal).

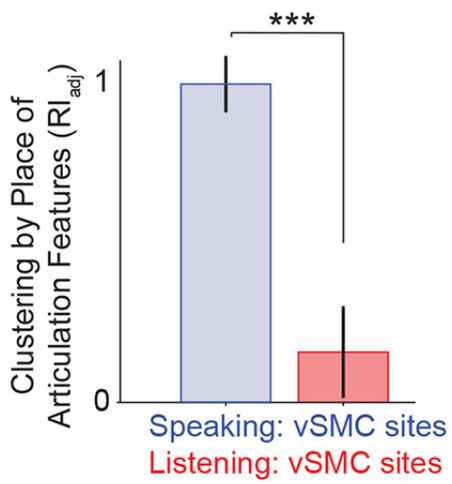


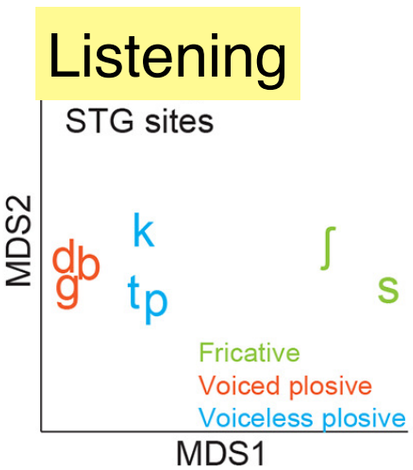
Clustering of electrode activation patterns

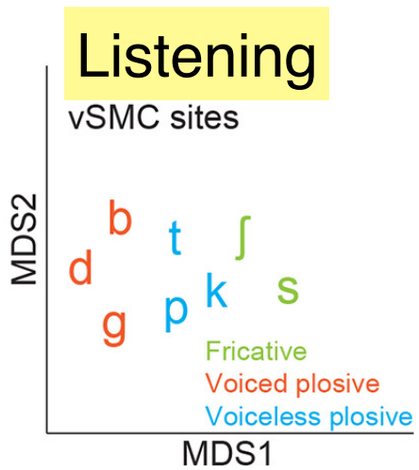
- vSMC
- Clustering by constriction type is strong during speaking.
- Clustering by constriction is weak during listening.
- Clustering by acoustic properties during listening is strong, but a bit weaker than in STG
- Clustering by Manner is stronger than clustering by constriction (place) during listening in vSMC.

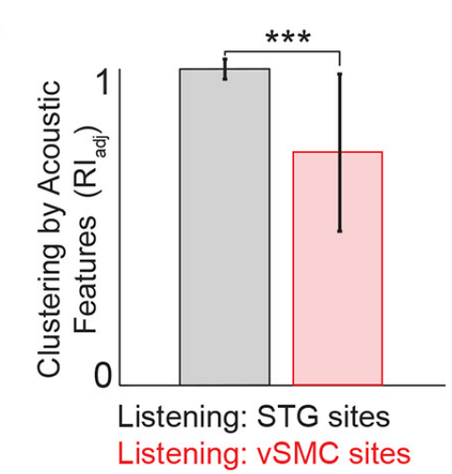


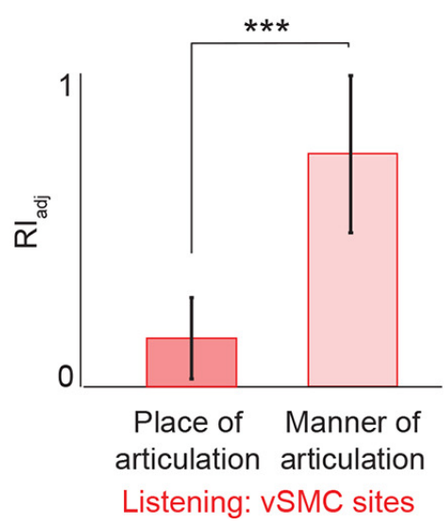










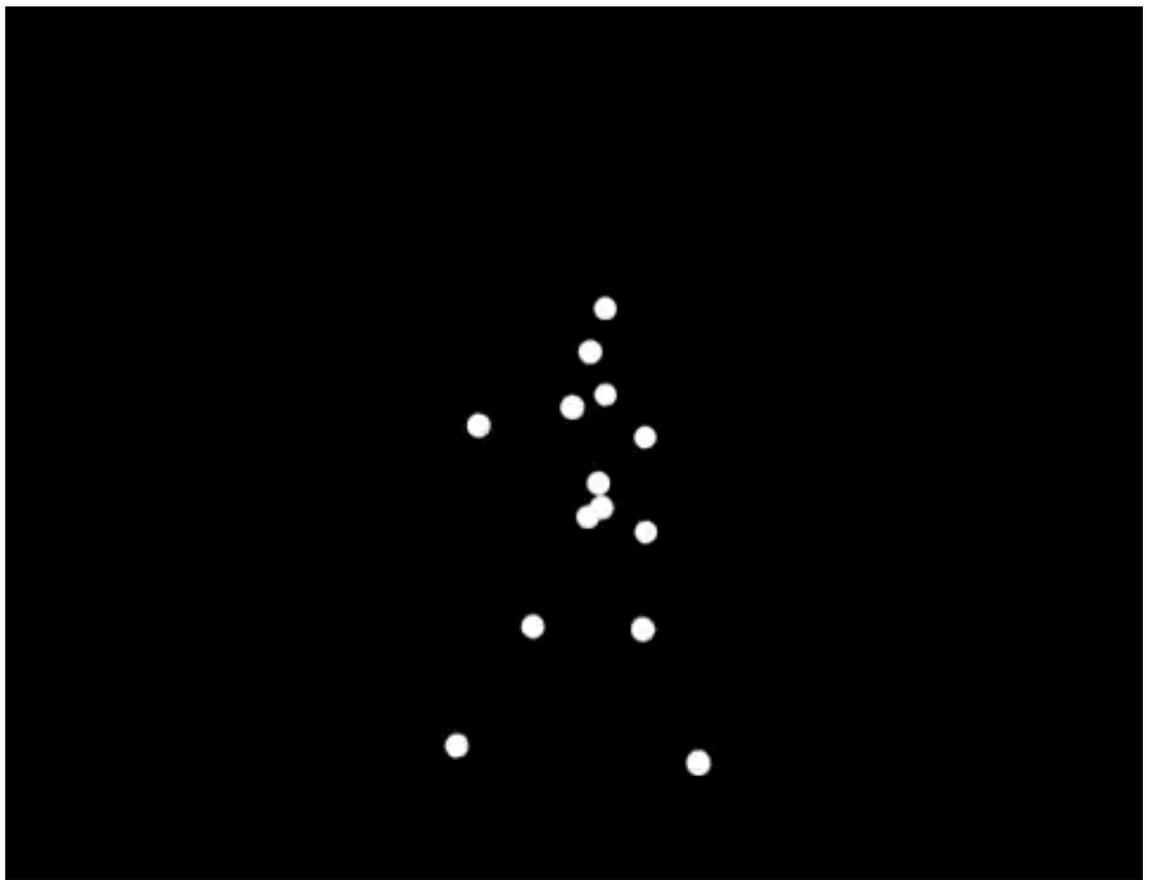


Summary: Binding Problem

- Patterns of neural activation during listening and speaking are distinct.
  - Constriction gesture responses during speaking
  - Acoustics / manner responses during listening.
- Given evidence for sensorimotor interactions, what binds them together?
- A critical aspect of speech is not considered in these studies that examine change over time.
  - The way the acoustics, auditory patterns, articulatory patterns change over time ought to be linked to one another.
  - The articulatory changes cause the acoustic changes that cause the auditory changes
  - Goal of this work is to begin an exploratory analysis of those patterns of change and their relation to syllable structure.

Why change ?

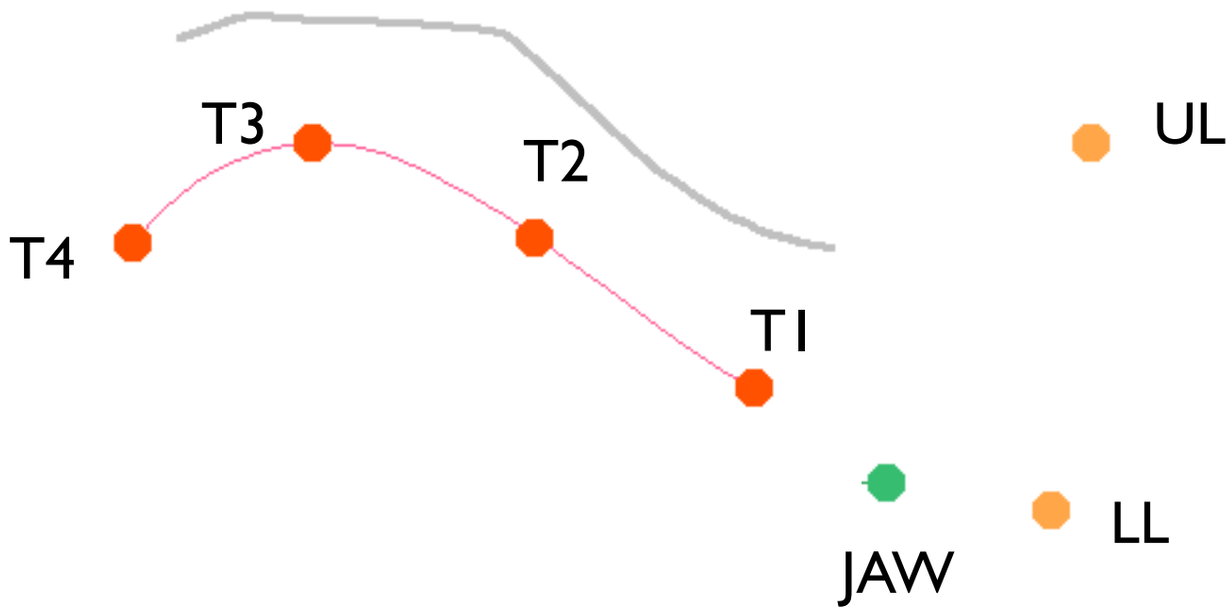
- Coherent perception of biological motion through change (only).
- Point-light displays (Johansson, 1973)
- Sine wave speech (Remez, Rubin, Pisoni & Carrell, 1981)



"Where were you a year ago?"

Articulatory Modulation functions:

- Calculate the global motion of the vocal tract at each instant in time
- X-ray microbeam (XRMB) corpus (Westbury et al., 1994)
- 7 markers on the surface of the oral articulators
- Sum the squared displacement over all markers and dimensions (x, y) at every time frame.
- Measure related to kinetic energy of vocal tract, if mass of articulators is ignored.



- No markers on soft palate or larynx
- No information relevant to nasalization or source changes

$$MBEAM(k) = \sum_{i=1}^{7} \sum_{j=1}^{2} (m(i,j,k+1) - m(i,j,k))^2$$

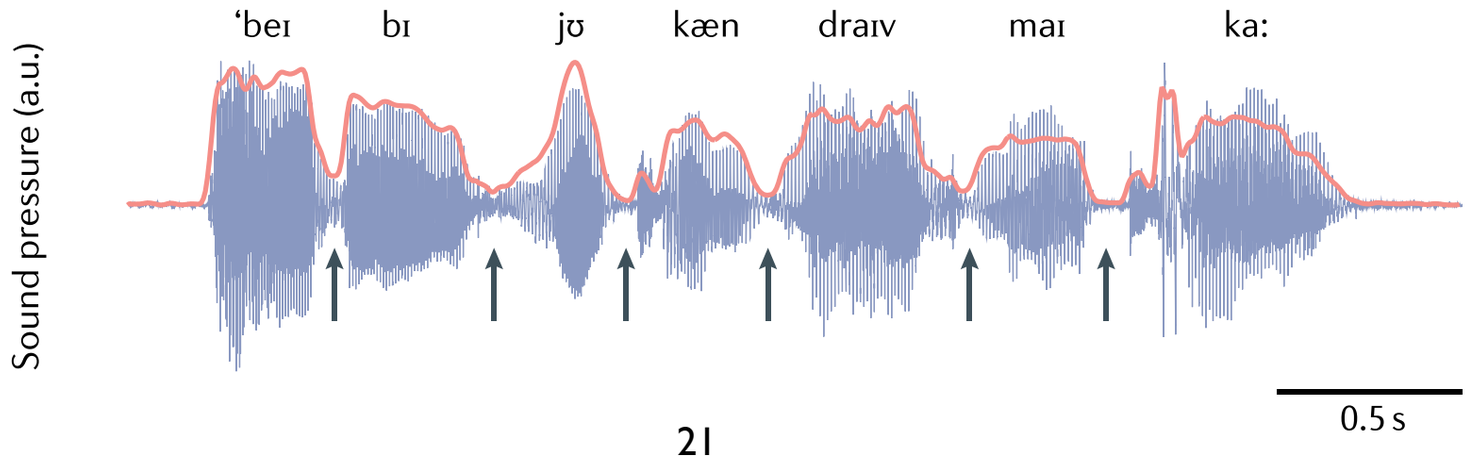
Acoustic Modulation function

- Acoustics were represented using 13 Mel-Frequency Cepstral Coefficients (MFCC).
- MFCC is a robust representation of spectral properties used in automatic speech recognition, forced text alignment, articulatory-acoustic inversion.
- Modulation was calculated in a similar way to the MBEAM modulation function: sum of the squared changes over all coefficients.

- Measure of kinetic energy of spectrum.

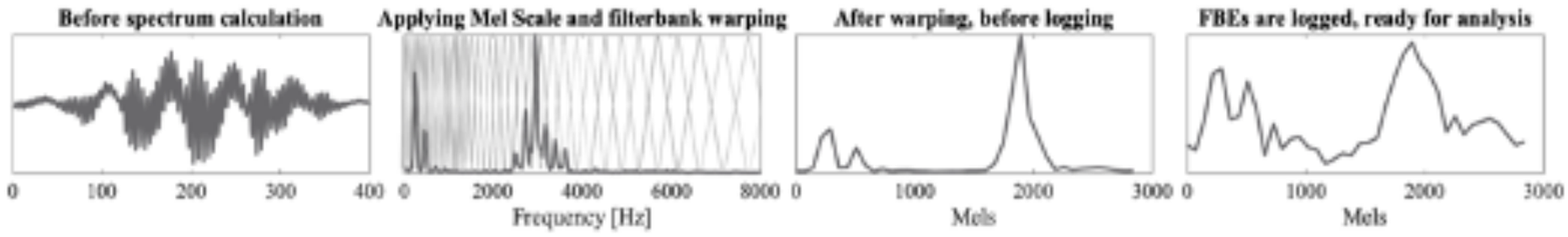
$$MFCC(k) = \sum_{i=1} (f(i, k+1) - f(i, k))^2$$

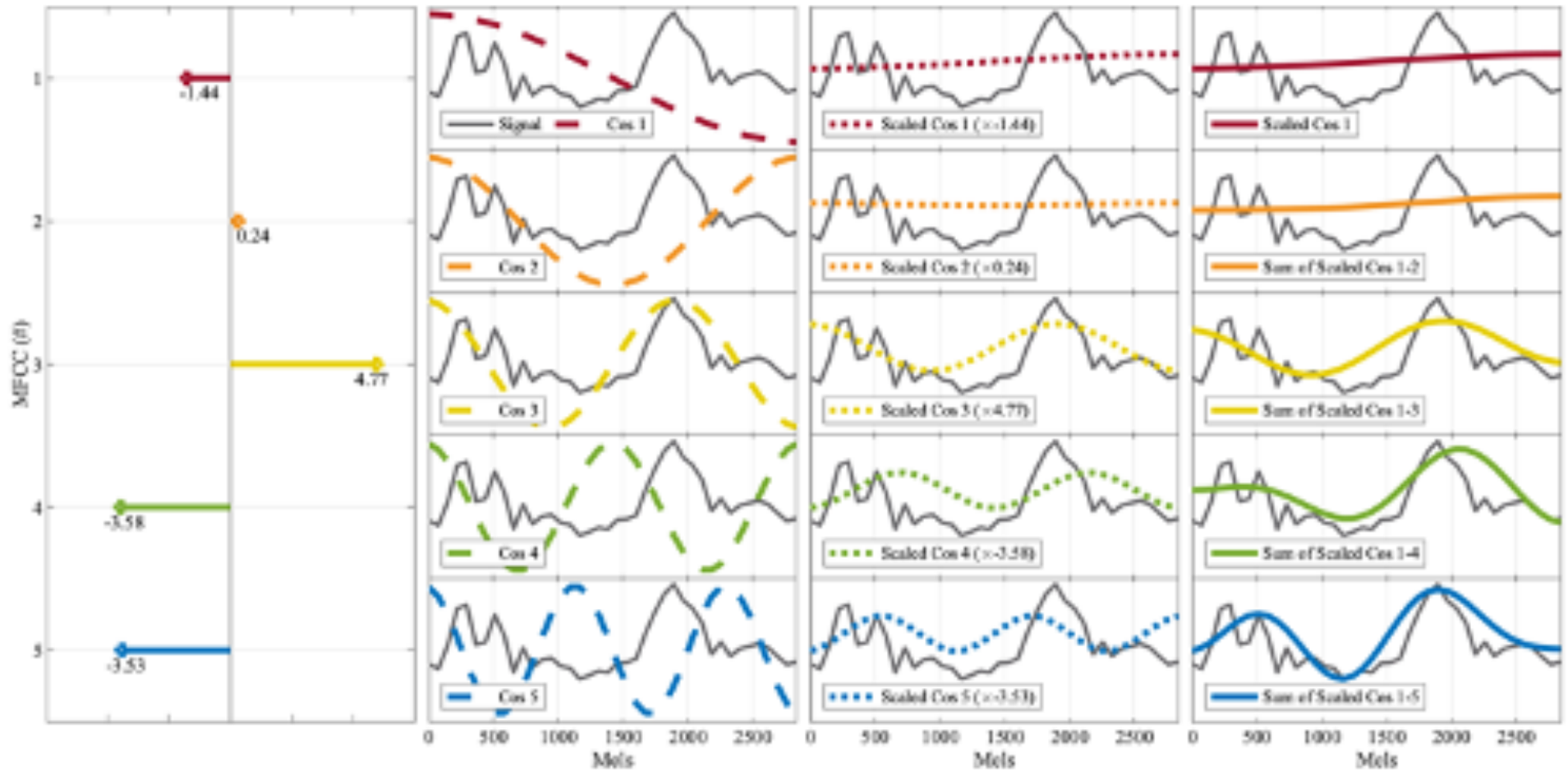
- The acoustic modulation function calculated in this way is different from broadband analysis of the amplitude envelope (Poeppel & Assaneo, 2000) and more like narrowband analysis they discuss.



MFCCs: Mel-Frequency Cepstral Coefficents

(from Jessie Johnson)





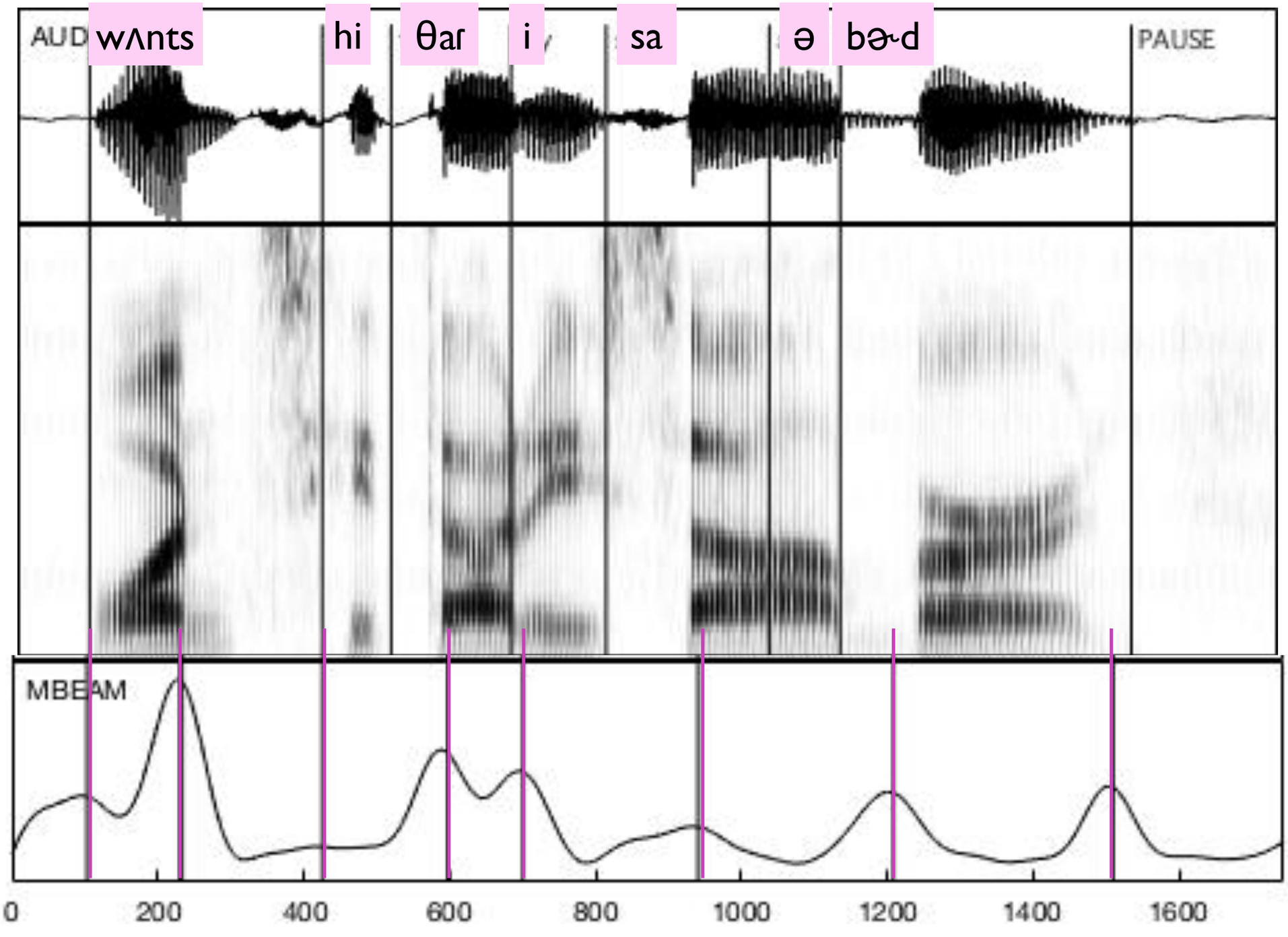
Data

- Twenty-three speakers from the XRMB corpus
- All were students at the University of Wisconsin, circa 1990
- Dialect regions:
  - 13 Wisconsin
  - 3 from Illinois,
  - 2 from Minnesota
  - 1 each from Indiana, Colorado, California, Massachusetts and New Jersey.

- One sentence from a read paragraph:
  "Once he thought he saw a bird, but it was just a large leaf that had failed to drop to the ground during the winter."

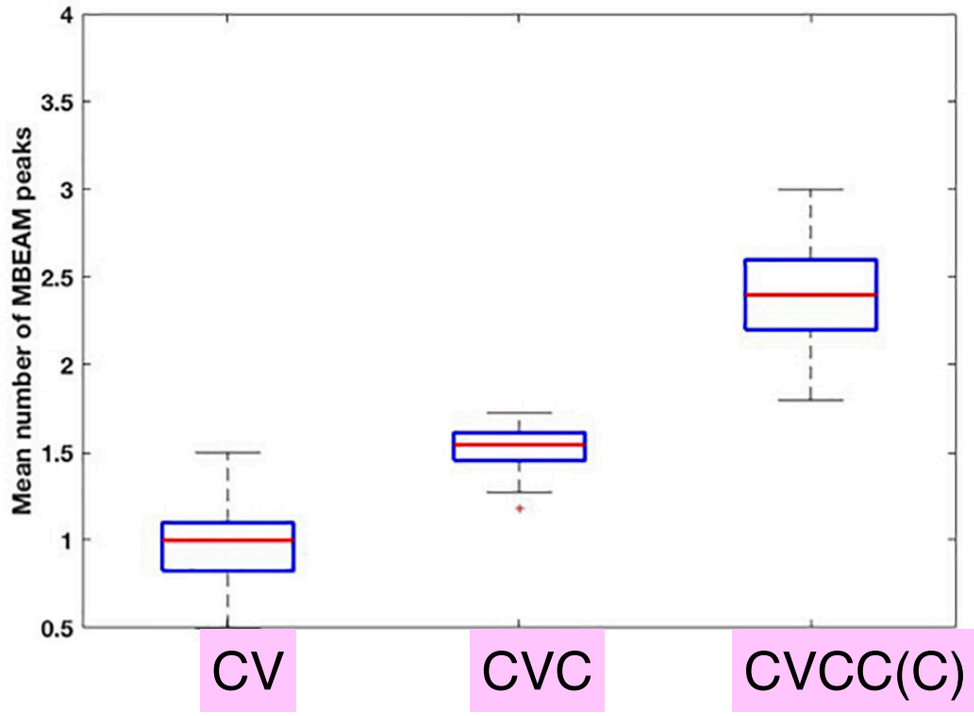
Example: Articulatory modulation

"Once he thought he saw a bird"

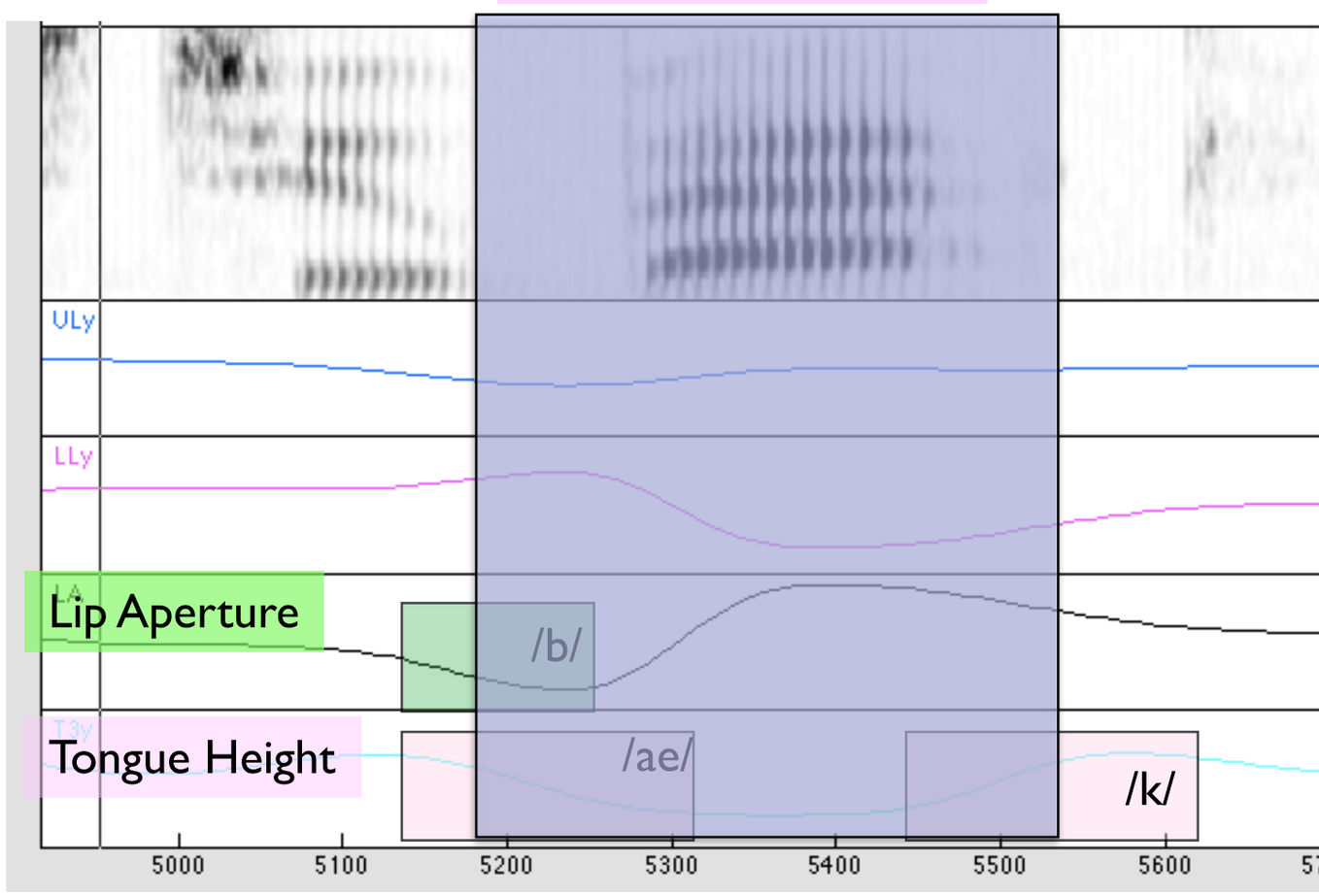


Articulatory modulation

- The articulatory modulation function has the form of a repetitive sequence of pulses, clear alternation between instants of maximal change and instants of minimal change.
- In general, there is one pulse per CV syllable, whose peak located near initial consonant or at the margins of the initial consonant and the vowel (as located in a spectrogram).
- Initial consonant gesture are co-produced with vowel gesture so there is much change going on at the beginning of the syllable, leading to the modulation peaks.
- There may be an additional pulse (or pulses) for syllables with post-vocalic consonants.



"two back"



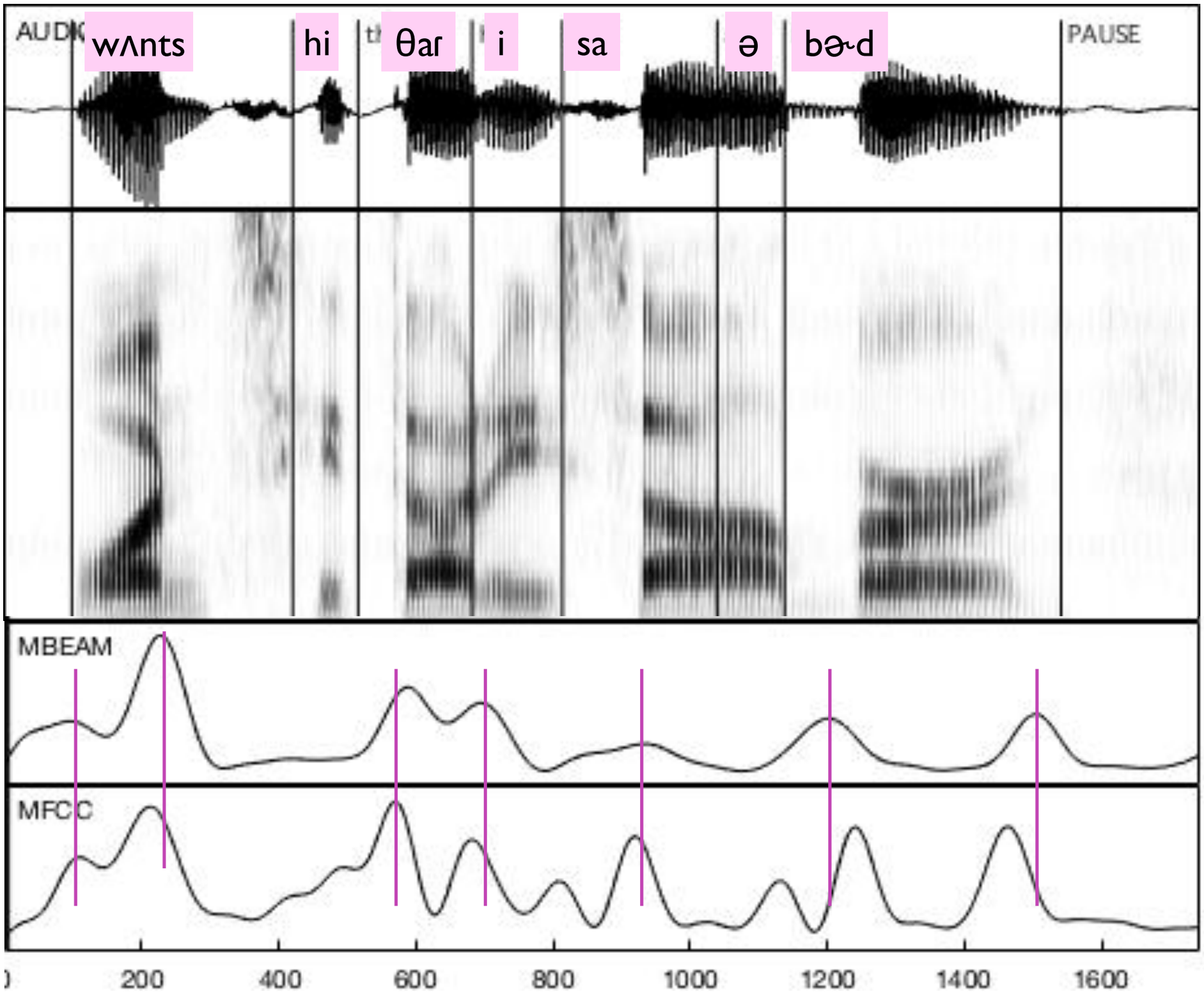
Acoustic Modulation

- The MFCC modulation function also exhibits pulses, but they slightly higher frequency (ie., there are more of them).

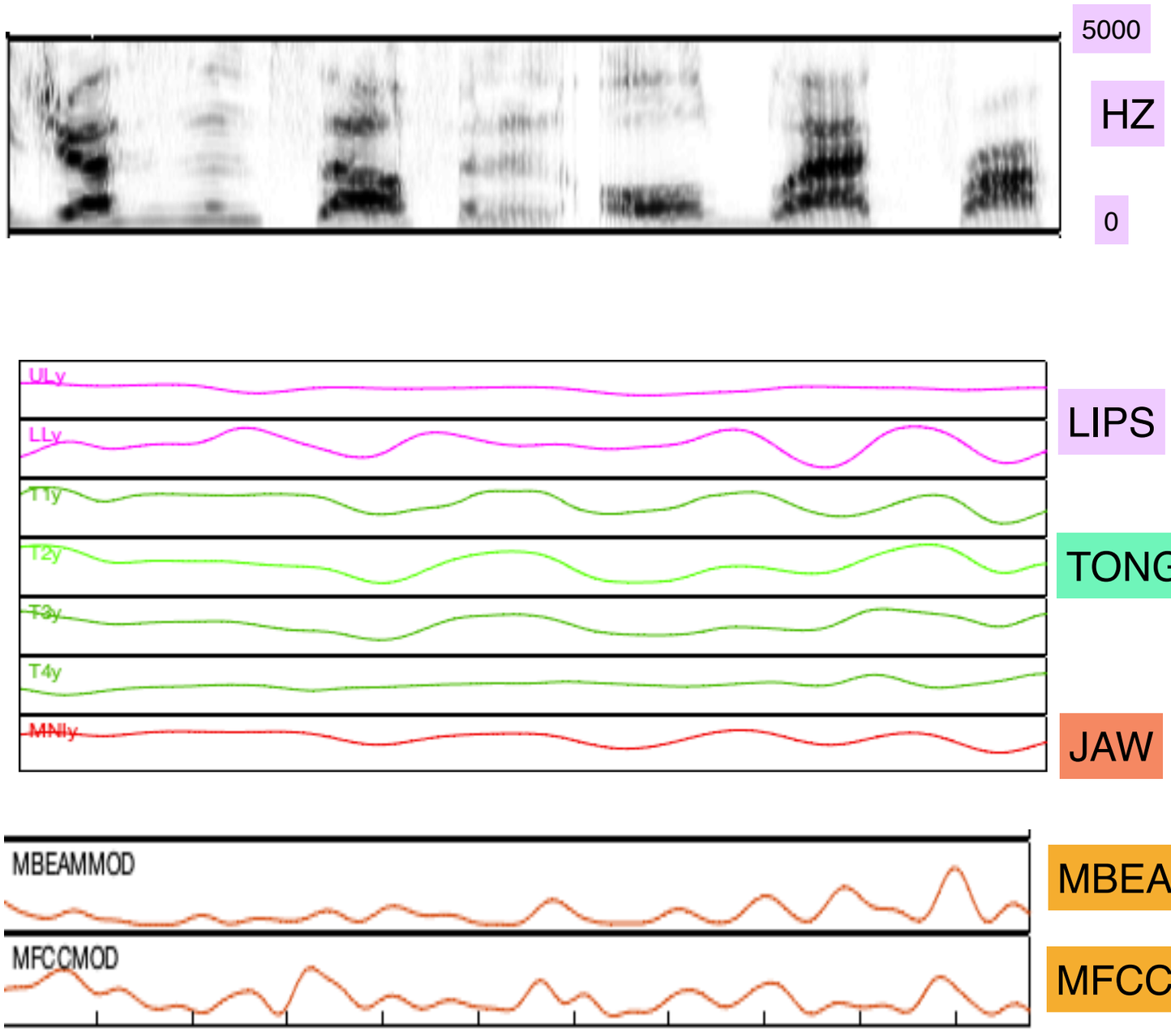
- Changes in source and nasalization may contribute to these.

- Acoustic signal is not as smooth.

- But there appear to be Acoustic pulses located in close proximity to the Articulatory pulses.

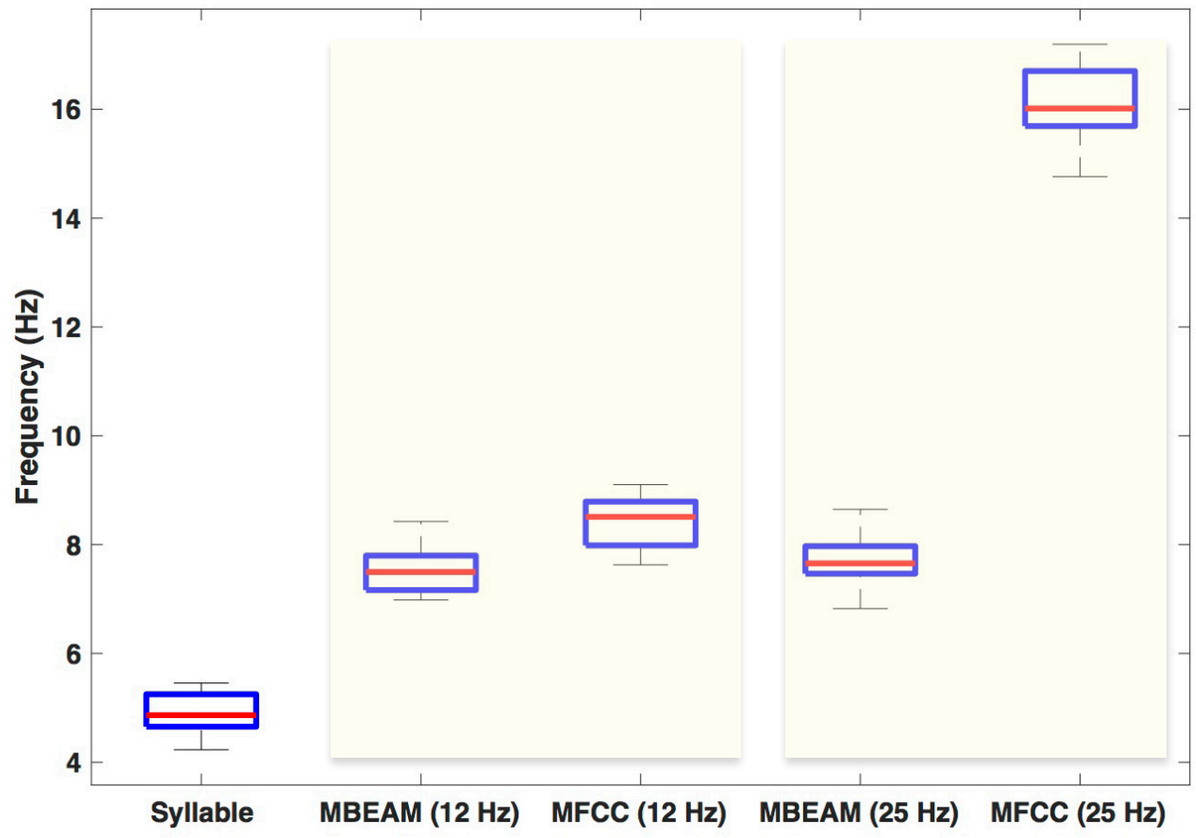

AUDIO
wʌnts
hi
θaɾ
i
sa
ə
bəˠd
PAUSE
MBEAM
MFCC
0
200
400
600
800
1000
1200
1400
1600
Time (ms)

“He dresses himself in an old black frock”



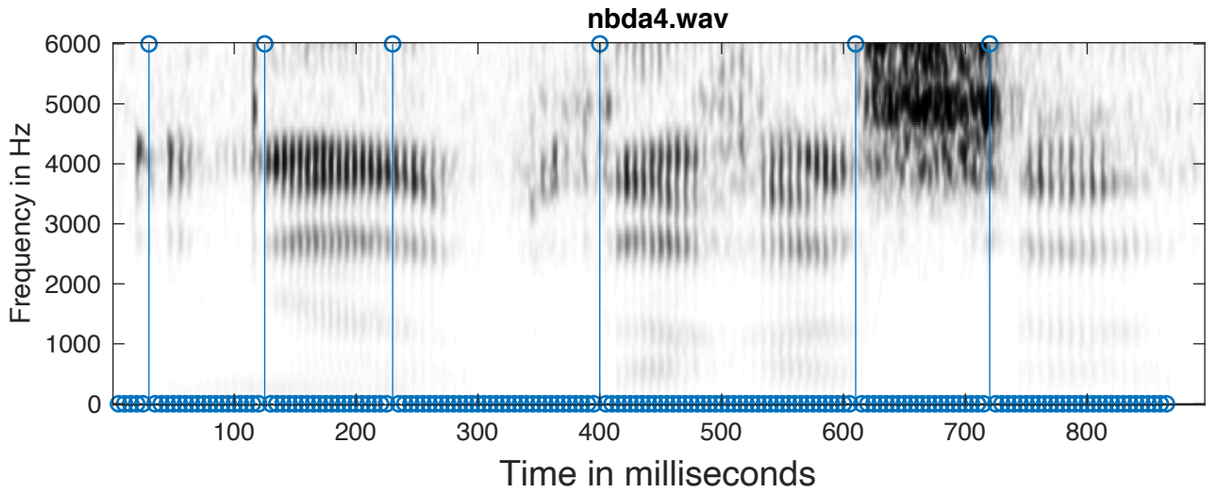
Frequency of Syllables and Pulses

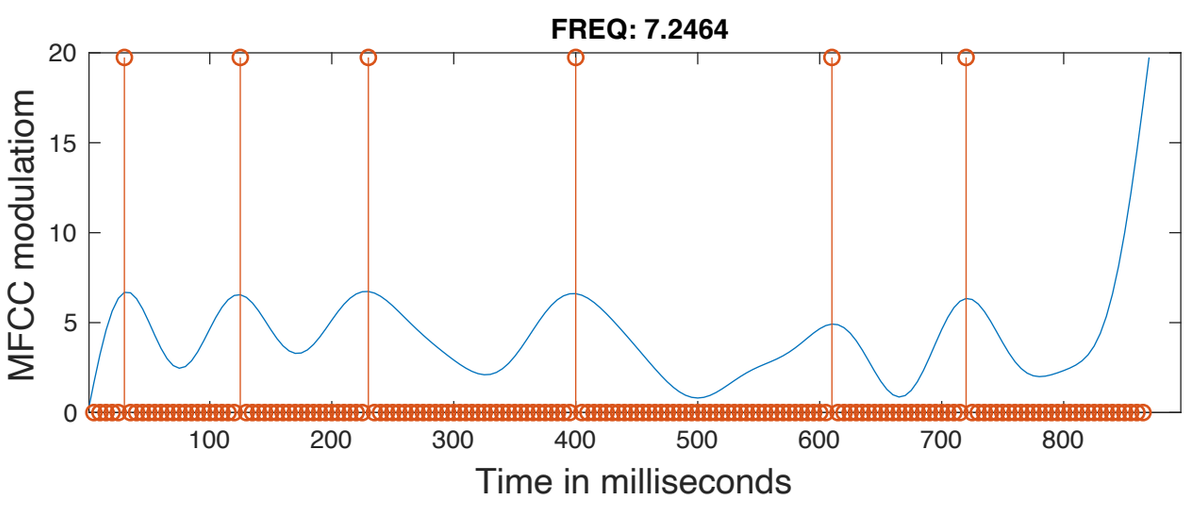
- Frequency of syllables was estimated by dividing the number of syllables by the utterance duration (minus the pause duration).
- Pulse frequency was estimated from the average time between pulses.
- Two versions of the modulation functions were compared, smoothed at 12 Hz and 25 Hz.
- Results show that the 23 speakers are reasonably consistent.
- Syllable frequencies are ~5 Hz
- For 12 Hz smoothing, MBEAM = ~7.5 Hz, MFCC = ~8Hz



Acoustic Modulation in Tashlhiyt

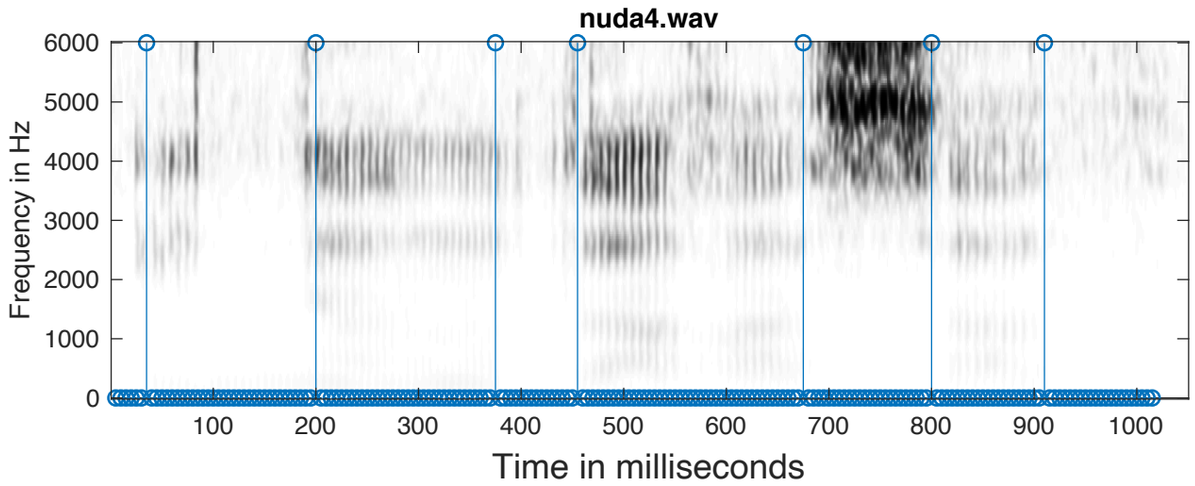
C-nucleus

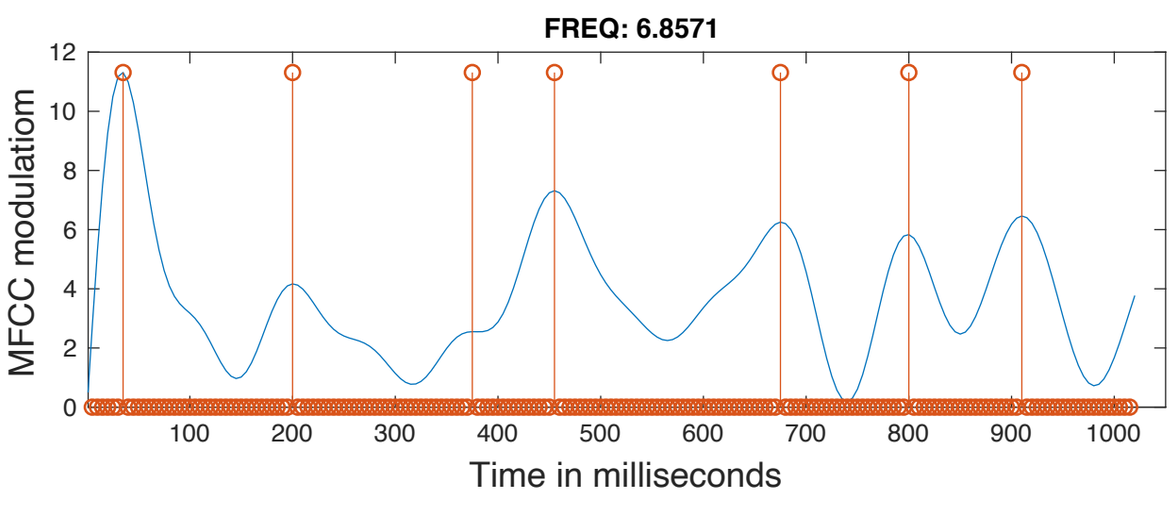




Inna nbda Rassad

V-nucleus

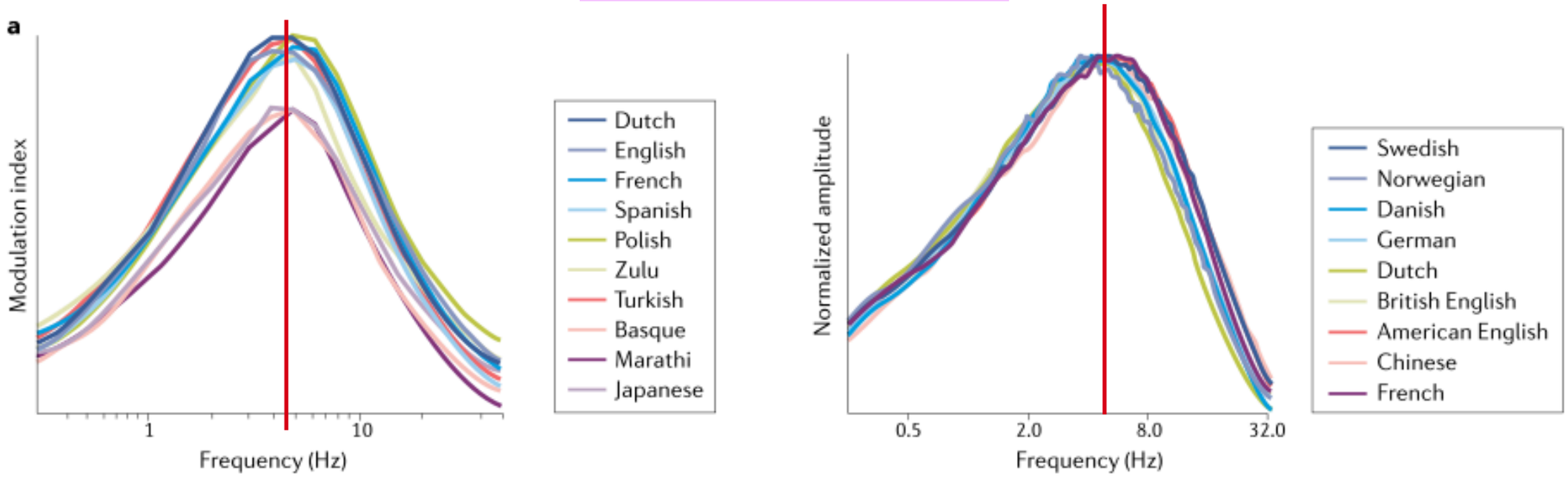


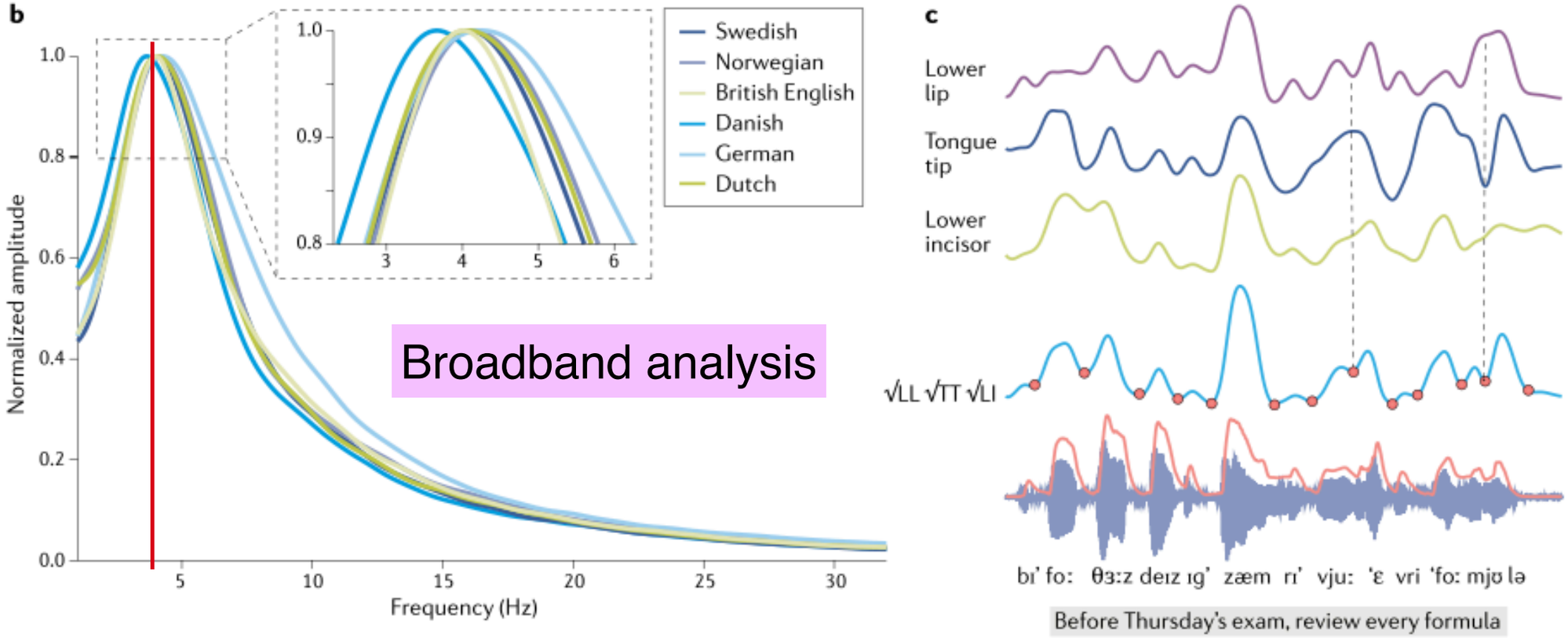


Inna nuda Rassad

Speech Rhythmicity Poeppel & Assaneo

Filter-band analysis





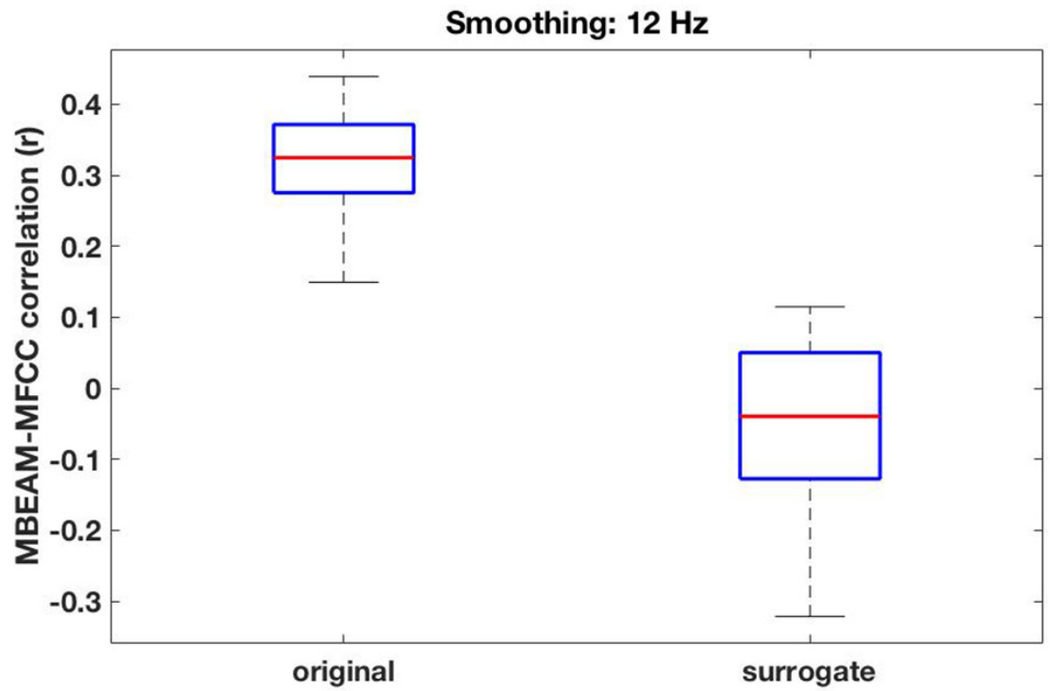
Cross-correlation of modulation functions

- Poeppel & Assaneo suggest that the rhythmic properties of acoustics, articulation and syllables are generally correlated, but don't test whether the specific modulation patterns associated with the same utterance are correlated.

- Here this was tested by correlating the articulatory and acoustic modulation functions and comparing those to the correlations in surrogate data that have the same rhythmic properties but are not causally connected.

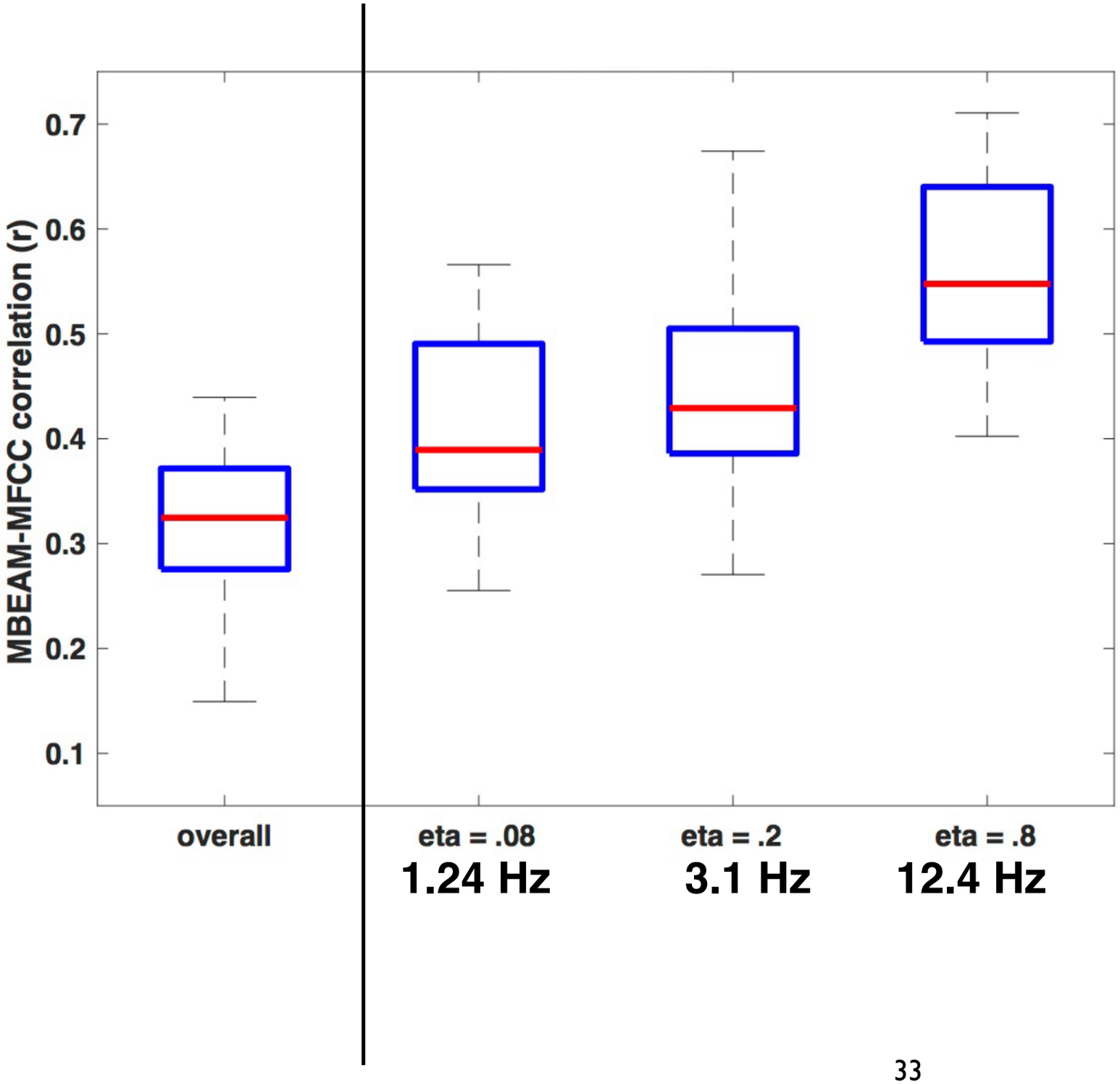
- Very weak test.

- Imagine imagine using MFCC modulation functions to identify which of a set of utterances was spoken, based on their MBEAM modulation functions. What is the probability of correct identification?



Surrogate data: first half of the MBEAM function was paired with the second half of the MFCC function, and second half of the MBEAM function was paired with the half of the MFCC function

Overall correlation and median correlation (over temporal windows) for each speaker

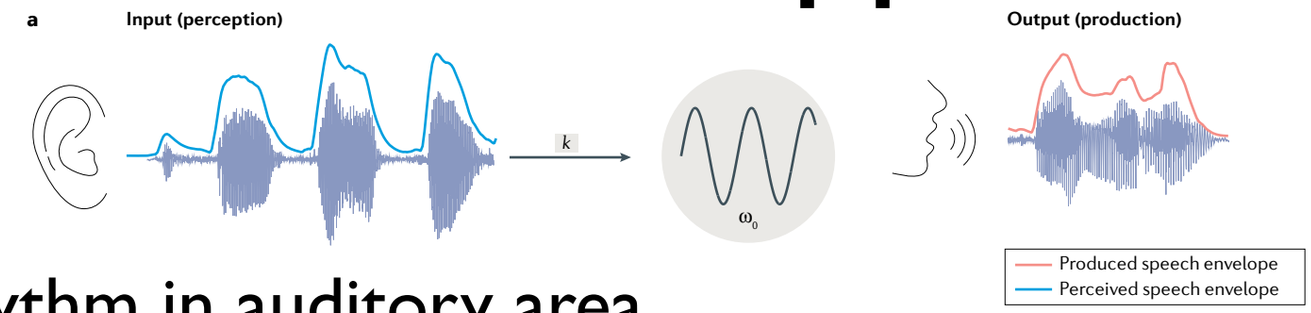


- Overall correlation is significant for all 23 speakers ($p<.001$)
- As window becomes more narrow, median correlation gets higher
- All differences are significant in a sign test.

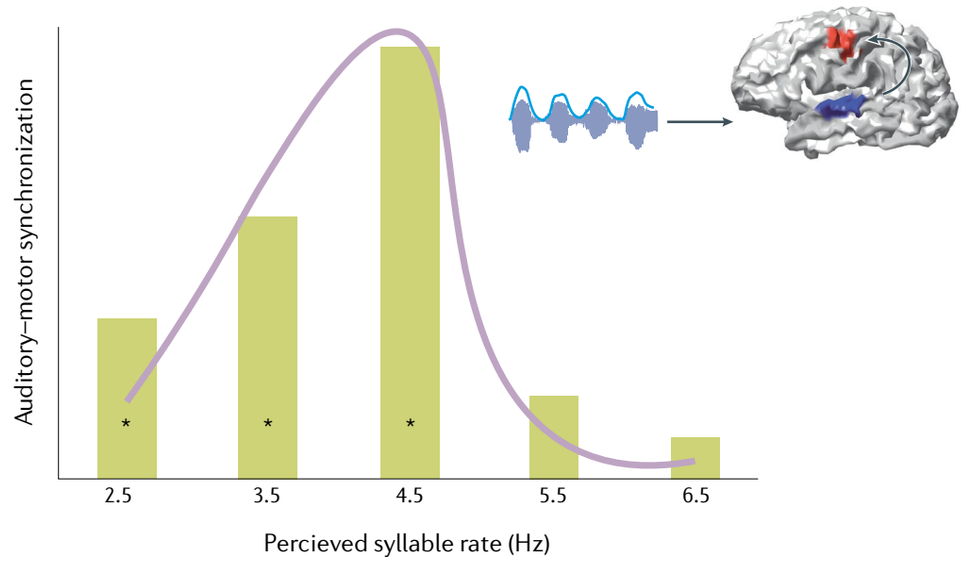
Shared rhythmic structure and synchronization

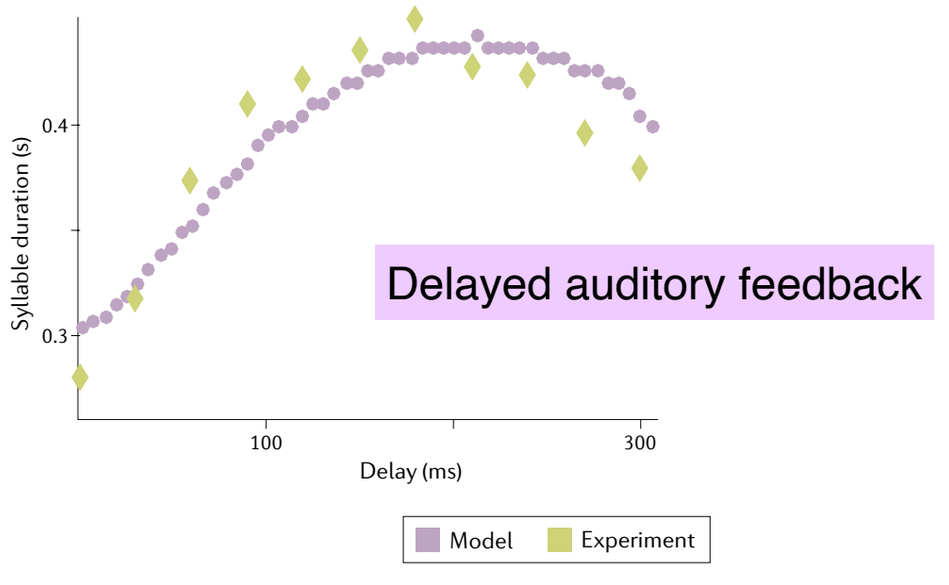
- The MFCC modulation function shares the specific repetitive rhythmic structure of the MBEAM function, to at least some extent.
- This congruity can afford the synchronization (entrainment) of neural oscillations in auditory and motor (and pre-motor) areas.
- This synchronization would be a solution to the binding problem.
- Synchronization has been empirically observed in several studies:
  - Park et al. (2015)
    During listening, oscillations in premotor and motor areas modulated the phase of low-frequency oscillations in the auditory areas more for natural speech than backwards speech.
  - Keitel et (2018)
    Greater entrainment of auditory and motor areas in correctly comprehended sentences than in incorrectly comprehended ones.

Assaneo & Poeppel model

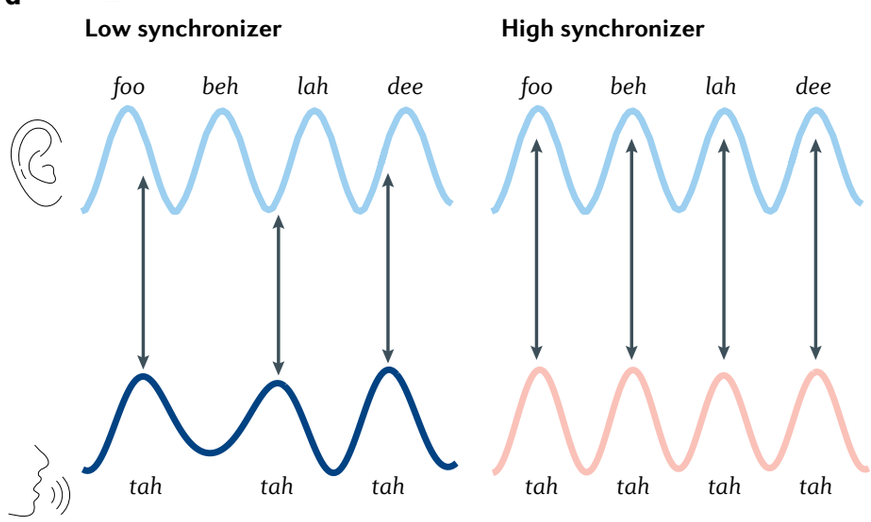


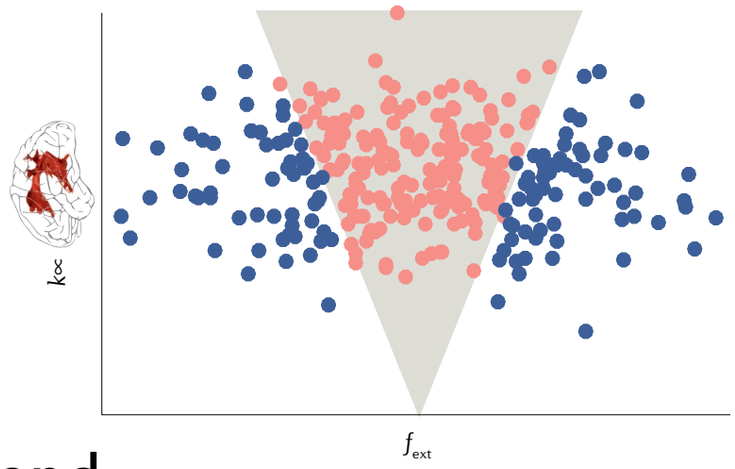
- Neural rhythm in auditory area entrains the neural rhythm in the motor area.
- Critical component of model is that there is a neural oscillation in the motor area and the articulatory modulation functions provide that.
- Model predicts the results of several experiments.
  - Strength of synchronization varies as a function of auditory syllable rate.
  - Lengthening of syllables in delayed auditory feedback.
- Needs to go in both directions.



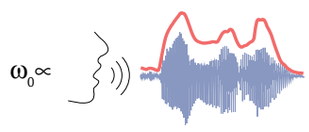


Spontaneous Synchronization





- Participants hear sequence of syllables and produce only /ta/.



- Bimodal behavior of participants:
  High synchronizers entrain productions to perceived syllable sequence

- High synchronizers also show more brain-to-envelope synchronization during passive listening to syllable sequences.

- High synchronizers have more white matter in the dorsal pathway.

- Model makes prediction about synchronization as a function of coupling strength (~white matter) and frequency of auditory sequence,