# Heschl's gyrus encoding of abstract speech cues in natural speech perception

Kyle M. Rupp[1], Fernando Llanos[2], Bharath Chandrasekaran[2], Taylor J. Abel[1]

[1]Dept. of Neurosurgery, University of Pittsburgh

[2]Dept. of Communication Science and Disorders, University of Pittsburgh

## INTRODUCTION

- Speech is a continuous, highly variable acoustic signal
- The human brain effortlessly transforms this input into perceptually constant phonemic representations
- Vowels are distinguishable by F1 & F2, but their values vary widely, with overlapping distributions, in connected speech
- Complicated by differences between speakers (vocal tract length) and within speaker (prosody, coarticulation)
- The brain can normalize across these differences to generate single percepts for each vowel, but underlying neural computations are unknown
- We performed direct intracranial recordings in Heschl's gyrus (HG) and planum temporale (PT) while 5 patients listened to natural speech
- High-gamma activity (HGA) was modulated by vowel ID
- Using encoding models, we investigated which acoustic and linguistic vowel representations were encoded by HGA
- Fundamental frequency (f0) and F1 normalized by f0 were encoded most consistently across HG & PT

## METHODS

- HG & PT recorded with sEEG while patients listened to 60 clips of natural speech (each ~1 min clip followed by 2 questions to test comprehension)
- Speech annotated for phoneme identity, on/offsets
- Vowel fundamental frequency (f0) and formants F1-4 extracted (Praat) as the value at the vowel's midpoint

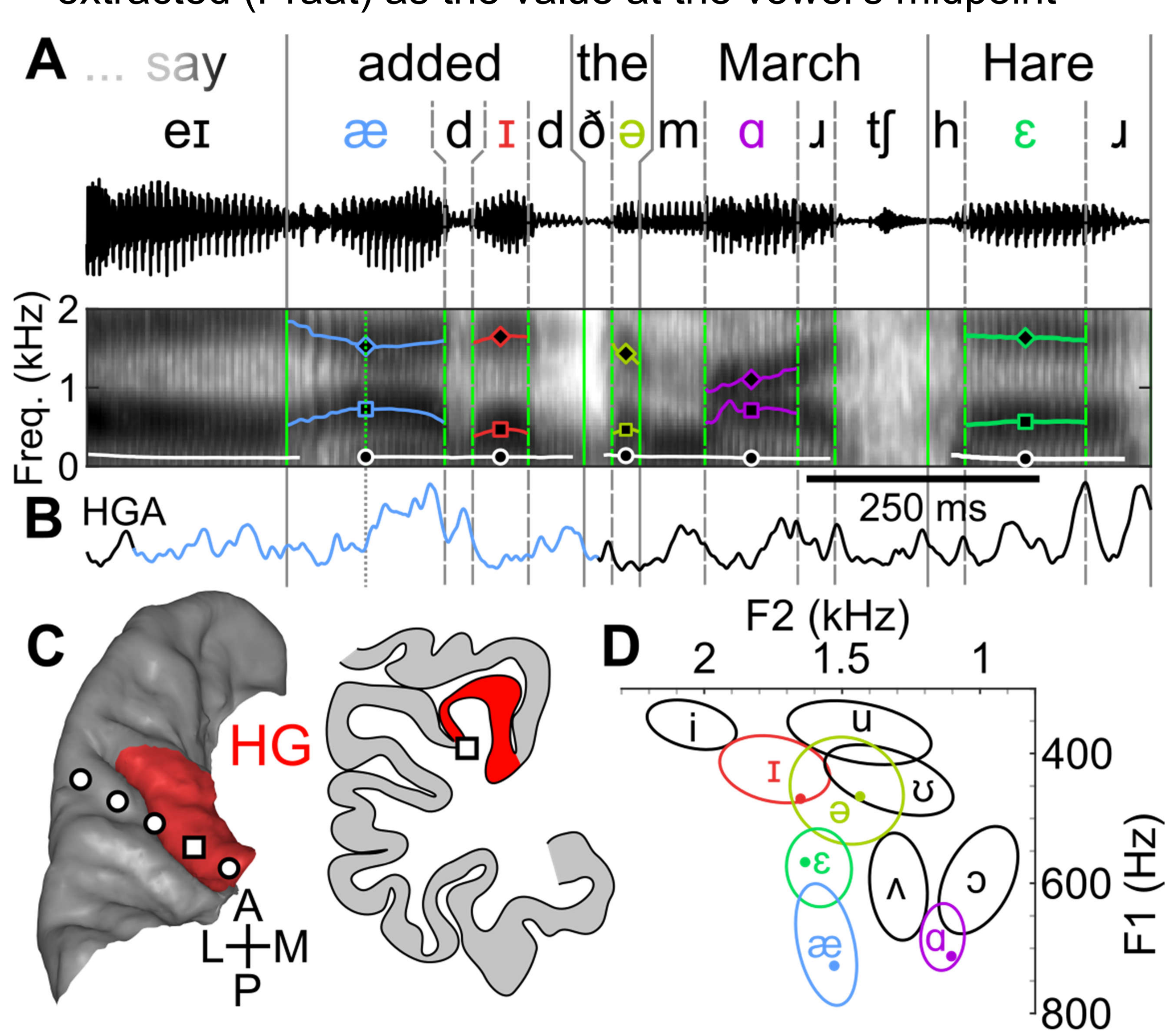


**Figure 1.** (A) Natural speech, annotations, and spectrogram, with fundamental freq. (f0) and formants (F1 & F2) overlaid. Single values (midpoint, see markers) assigned to f0 & F1-4 for every vowel. (B) HGA for a single electrode in left HG, recorded in patient P1. Blue portion: 500 ms window aligned to æ midpoint. (C) Top-down STP view and coronal slice of temporal lobe show electrode location from (B). (D) Formants were extracted across all clips; 2D gaussians fit to each vowel's distribution. Ellipses: 1 standard deviation. Points show the (F1, F2) location of each vowel from (A).

## METHODS

- HGA calculated (Hilbert transform) then extracted by aligning to each vowel's midpoint
- Sliding ANOVA used to identify electrodes modulated by vowel ID
- **Is HG encoding vowel ID, formants, or something else?**
- Encoding models built to predict HGA from acoustic & phonetic features (see Table 1)
- Lasso regularization prevented overfitting and forced sparse feature selection
- Models evaluated by fraction of explained HGA variance ($R^2$)
- Selected features were interpreted as being encoded in HGA

| Type | Features | Type | Features |
|---|---|---|---|
| Formant ratios | F1/F3, F2/F3 [1] | Vowel ID | Binary [i, æ, ə, …] |
| | log(F2/F1), log(F3/F1), log(F4/F1) [2] | Phonetic features | Height, front/back, rounded |
| | log(F1/f0), log(F2/F1), log(F3/F2) [3] | Formants & inverses | F1,…, F4; $F1^{-1}$,…,$F4^{-1}$ |
| | log(F1/f0), log(F2/F1), log(F3/F2) [4] | Fund. freq. & inverses | f0, $f0^{-1}$ |
| | log(F1/F*), log(F2/F*), log(F3/F*) [5] | f0-norm. formants | F1/f0, F2/f0, F3/f0 |
| | F* = geomean(F1, F2, F3) | Acoustic props. | dB, duration |

**Table 1.** List of features included in the encoding model.

## RESULTS

- Some electrodes show graded HGA responses (Fig. 2) that closely match vowel progression along F1-F2 diagonal (Fig. 1D). ANOVA F-stat shows time-dependent separability across all vowel IDs (not just the 5 exemplars in Fig. 2)
- 35/50 electrodes achieved ANOVA significance (α = .01, Bonf. corrected across all patients, channels, & timepoints)
- In 14 electrodes, encoding models could explain >10% of HGA variance (Fig. 3)
- Peak $R^2$ occurred at lags of 30 ms (3 elecs) or 40 ms (11)

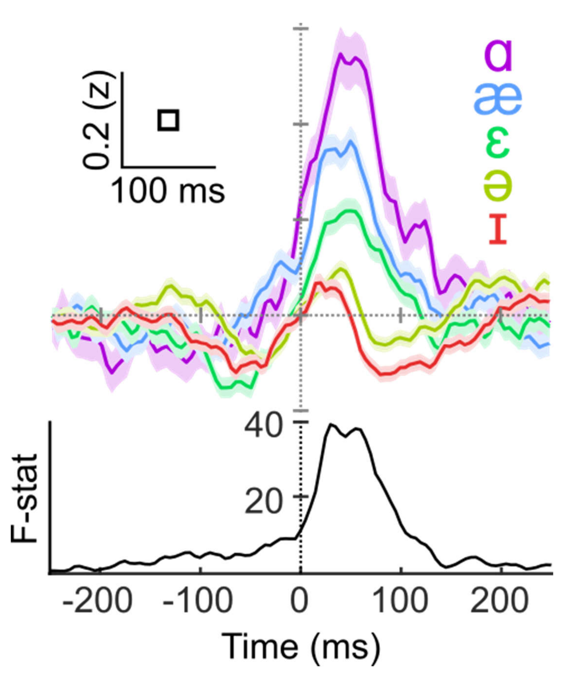


**Figure 2.** Mean HGA (± std err) from Fig. 1 electrode & vowels. Graded HGA closely matches vowel progression along F1, F2 diagonal (Fig. 1D). F-statistics calculated via sliding ANOVA.

| Patient | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Comprehension (%) | 85 | 40 | 58 | 88 | 83 |
| Total number of electrodes | 6 | 11 | 15 | 7 | 11 |
| Significant ANOVA | 3 | 8 | 9 | 5 | 10 |
| Encoding model ($R^2$ > 0.1) | 1 | 0 | 3 | 4 | 6 |

**Table 2.** Summary of results. Last 3 rows are electrode counts.

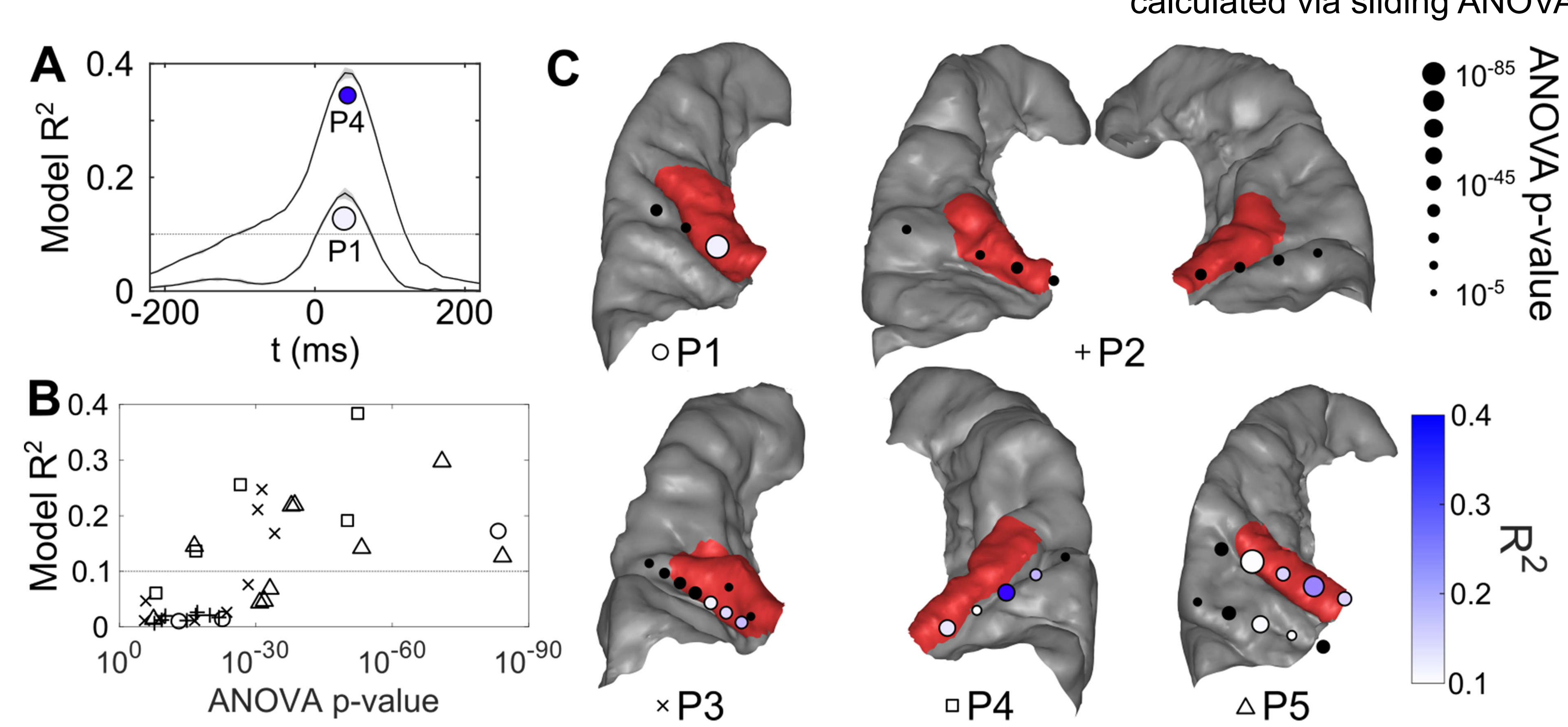


**Figure 3.** (A) Encoding model $R^2$ for 2 electrodes (mean ± std err across CV folds) . (B) 35/50 elecs had significant HGA modulation by vowel ID (ANOVA); a subset of these were well-explained by encoding models ($R^2$>0.1, dashed lines). Marker type corresponds to patient ID (see labels in C). (C) Anatomical locations of significant elecs. Black elecs were significant via ANOVA but did not achieve the $R^2$ cutoff.

## RESULTS

- For each model, $\tilde{\beta} = |\beta|/\text{sum}(|\beta|)$
- f0 was most strongly encoded in HGA
  - 13/14 models: largest $\tilde{\beta}_i$ was $\tilde{\beta}_{1/f0}$
  - $\tilde{\beta}_{f0}+\tilde{\beta}_{1/f0}$=0.63 (mean across 14 models)
- F1/f0 was 2nd most strongly encoded
  - 11 models: 2nd or 3rd largest feature
  - 12 models: $\tilde{\beta}_{F1/f0} > \tilde{\beta}_{F1} + \tilde{\beta}_{1/F1}$
- Other encoded features:
  - Duration
  - Loudness (dB)

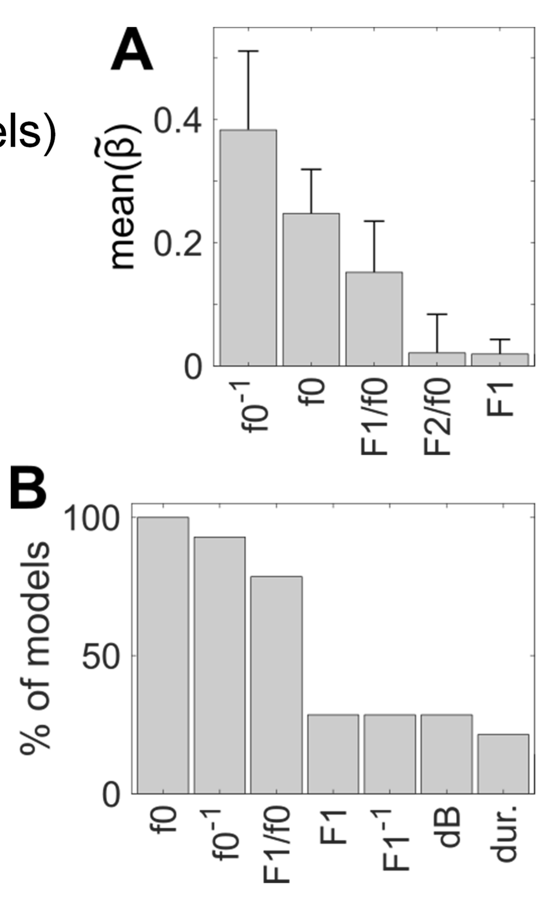


**Figure 4.** (A) Coeff magnitudes for each model were scaled to sum to 1. Mean scaled coeff mag (±std dev) for top 5 features is shown. (B) For each model, scaled coeff mags were sorted, and the first N features that sum to 0.9 were kept. Bars represent the percent of total models (out of 14) that kept that feature.

## DISCUSSION

- HGA on Heschl's gyrus is differentially activated across vowels during naturalistic listening conditions
- At some sites, HGA encoded acoustic features
  - Raw:
    - Duration & loudness (less perceptually relevant)
    - Fundamental frequency, 1st formant (more relevant)
  - Normalized: formants normalized to f0
- f0-normalized formants may be perceptually relevant for normalization across speakers or contexts (e.g. coarticulation)
- Limitations
  - Only 1 speaker
  - Results are dependent on user-defined input features
    - E.g. both f0 & $f0^{-1}$ chosen in same models: in HGA~F(f0), F may be unknown
    - Only explored intrinsic cues; future work will also explore extrinsic contextual cues

## REFERENCES

1) Monahan, P. J., & Idsardi, W. J. (2010). Auditory Sensitivity to Formant Ratios: Toward an Account of Vowel Normalization. *Language and cognitive processes*, 25(6), 808–839.
2) Peterson, G.E. (1961) Parameters of vowel quality. *Journal of Speech and Hearing Research* 4, 10-29.
3) Syrdal, A.K. & Gopal, H.S. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. Acoust. Soc. Am.* 79, 1086-1100.
4) Miller, J.D. (1989) Auditory-perceptual interpretation of the vowel. *J.Acoust. Soc. Am.* 85, 2114-2134.
5) Sussman, H.M. (1986) A neuronal model of vowel normalization and representation. *Brain Lang.* 28, 12-23
6) Johnson, K. (2008). 15 Speaker Normalization in Speech Perception. *The handbook of speech perception*, 363.