Modeling native phonology and non-native speech perception using EEG signals

4pSC30


Daniel McCloy


Adrian KC Lee

Institute for Learning and Brain Sciences • University of Washington

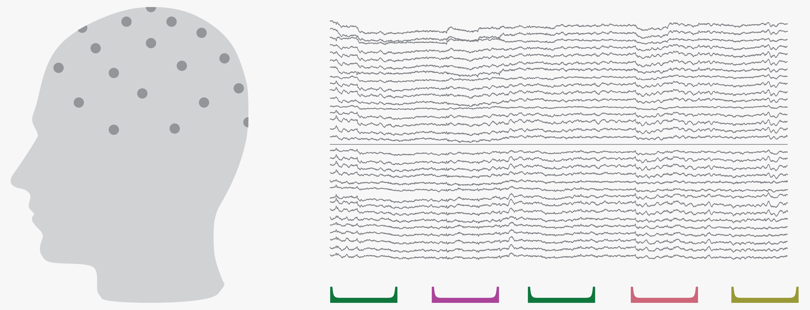
Overview

Listener hears stream of CV syllables

HUN♀ HIN♂ HUN♀ ENG♂ ENG♀

[… dʒa ʋa ɲa ʃa ka …]

EEG responses recorded



Epochs labeled with consonant distinctive feature values

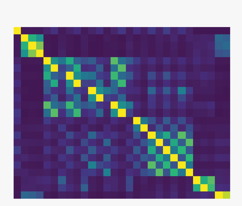
+VOICED +LABIAL −CORONAL −DORSAL +CONTINUANT …

−VOI. −LAB. −COR. +DORS. −CONT. …

Labeled EEG epochs for two ENG talkers are training data for classifiers (1 classifier per distinctive feature)

Trained classifiers are given EEG epochs corresponding to foreign talkers and 2 new ENG talkers; this yields guesses about which features the heard consonant had, and error rates for each classifier.

For each heard consonant type, use error rates to compute joint probability across classifier outputs; do this for different combinations of feature values to get a confusion matrix (see key).



Acknowledgments

NIH T32 DC005361 (Auditory Neuroscience Training Program); NSF Jelenik Speech and Language Technology Workshop 2015 Probabilistic Transcription Team; Mark Hasegawa-Johnson; Preethi Jyothi; Majid Mirbagheri; Nick Foti; Eric Larson.

References

[1] J. Särelä and H. Valpola, "Denoising source separation," *J. Mach. Learn. Res.*, vol. 6, pp. 233–272, 2005.

[2] A. de Cheveigné and J. Z. Simon, "Denoising based on spatial filtering," *J. Neurosci. Methods*, vol. 171, no. 2, pp. 331–339, 2008.

[3] S. Moran, D. McCloy, and R. Wright (eds). *PHOIBLE: Phonetics Information Base and Lexicon Online.* Munich: Max Planck Digital Library, 2013.

Background

- Studying phoneme confusion normally requires stimulus degradation or synthetic intermediate tokens (i.e., "ba - wa" continuum) to pull listener performance away from ceiling

Research Questions

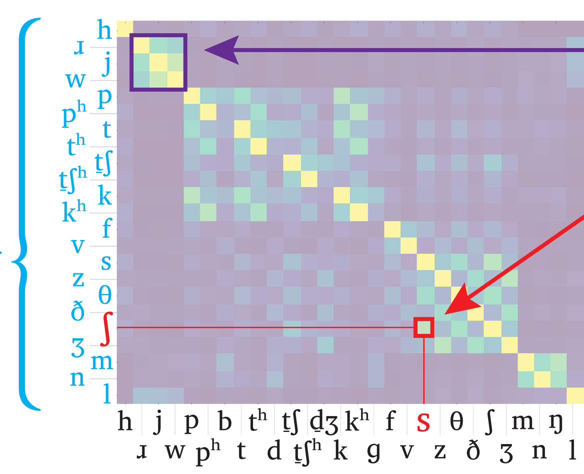
- Can we estimate consonant confusions by observing neural responses to speech sounds?
- Does this method reveal something about the structure of mental representations of phones/ phonemes?

Methods

- **STIMULI**: English & foreign consonant-vowel (CV) syllables; variable consonant, vowel always [a]
- **TRAINING SET (ENGLISH)**: 2 talkers (♂,♀) × 3 recordings × 23 consonants × 20 presentations = 2760 trials
- **TEST SET (ENGLISH)**: 2 new talkers (♂,♀) × 1 recording × 23 consonants × 20 presentations = 920 trials
- **TEST SETS {DUTCH/HUNGARIAN/HINDI/SWAHILI}**: 1 talker {♀/♀/♂/♂} × {18/25/30/30} consonants × 20 presentations = {360/500/600/600} trials
- **RECORDING**: 32-channel BrainVision EEG, left earlobe reference, 1000 Hz sampling rate
- **PREPROCESSING**: bandpass 1-40 Hz, downsample to 100 Hz, align epochs on boundary between C and V, apply denoising source separation[1,2] (DSS), remove time domain autocorrelation with PCA (retains ~20 "time samples"), use only first 4 DSS components
- **SUPERVISED LEARNING**: label each epoch with consonant's distinctive feature values from PHOIBLE[3] database (16 feats. used), train binary classifier (support vector machine with radial basis function) for each distinctive feature (5-fold cross- validation + grid search), set threshold to equalize error rates (false positive rate = false negative rate) to handle class imbalance
- **EVALUATION**: apply classifiers to test data, estimate "probability that listener heard ⟨·⟩" as joint probability of classifier outputs being consistent with the feature values of ⟨·⟩, i.e.:

$P(\langle ð \rangle) = P(+\text{VOI.}) \times P(-\text{SON.}) \times P(+\text{COR.}) \times \ldots \times P(+\text{CONT.})$

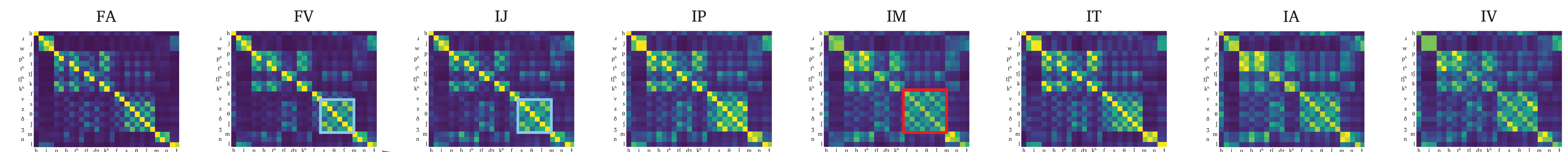
Key



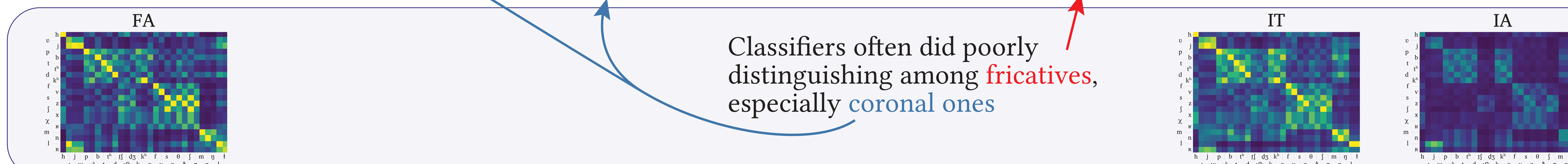
Results

- Classifier performance varies greatly across the 12 listeners even for ENG test data (8 best listeners shown). Worst cases still show some diagonal structure (i.e., at least *some* of the classifiers somewhat learned to detect natural classes from EEG).
- Overall the technique looks promising for studying human phone classification / confusion.
- Results for foreign talkers are less consistent; some look as good as ENG results, others look terrible; probably need better SNR (more foreign talkers / tokens).

English test stimlui



Dutch

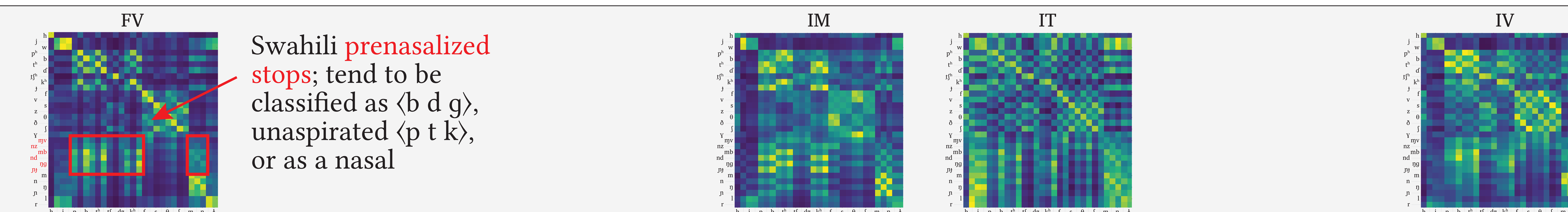


Classifiers often did poorly distinguishing among fricatives, especially coronal ones

Hungarian

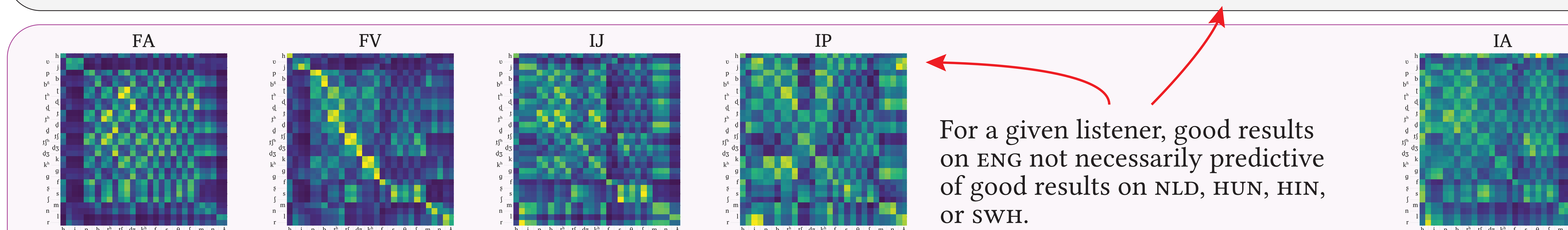
This "checkerboard" pattern shows discrimination of voicing but failure to learn manner/place of articulation differences

Swahili



Swahili prenasalized stops; tend to be classified as ⟨b d g⟩, unaspirated ⟨p t k⟩, or as a nasal

Hindi



For a given listener, good results on ENG not necessarily predictive of good results on NLD, HUN, HIN, or SWH.

Future directions

- Vary both consonants and vowels
- More languages / speech sound types (airstream and phonation contrasts, tone)
- Increase SNR: more talkers/tokens, different classifier strategies
- Unsupervised learning: derive optimal, perceptually based distinctive features
- Simultaneous MEG + EEG experiments to connect confusion patterns to cortico-spatial patterns
- Other applications of this method: diagnostic use for hearing / language impairments?