Contributions of Bottom-up and Top-down Processes in Speech Cue Encoding: Evidence from EEG and Machine Learning Techniques

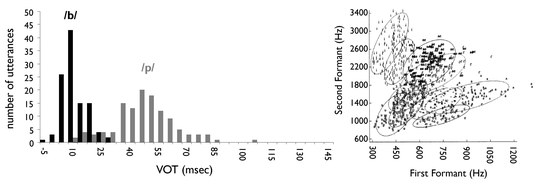
McCall E Sarrett, PhD*
Villanova University

Bob McMurray, PhD
University of Iowa

Joseph C. Toscano, PhD
Villanova University

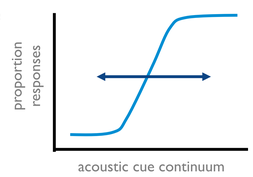
Introduction

The acoustics of spoken language are highly variable:



There is no one-to-one mapping between a given acoustic cue and the phoneme category it corresponds to

Psycholinguistic work has shown
(1) which acoustic dimensions are relevant to categorization
Allen, Miller, & DeSteno (2003); Hillenbrand, Clark, Getty, & Wheeler (1995)
(2) how listeners use top-down expectations to shift categorization responses
Connine (1987), Ganong (1980), Miller, Green, & Schermer (1984)



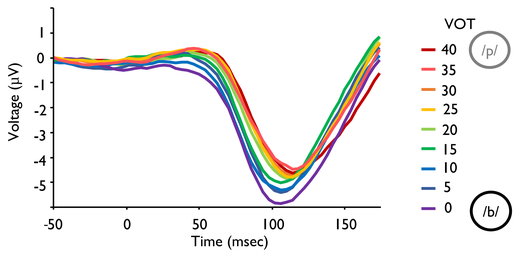
Goals

→ However, the realtime neural mechanisms subserving such processes are not well understood

We present three EEG studies that examine:
(1) which perceptual distinctions are detectable in the neural signal
(2) whether higher-level information influences acoustic encoding directly

Methods

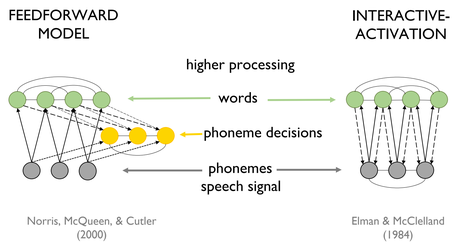
We utilize and extend an EEG paradigm developed in Toscano et al. (2010), which showed that N1 amplitude tracks an acoustic cue of speech sounds (VOT, which distinguishes sounds like /b/ and /p/)



Shorter, more /b/-like VOTs yield a larger amplitude N1 component, whereas longer, more /p/-like VOTs yield a less negative N1

Top-down Influences

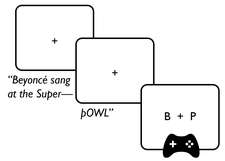
Models of speech perception



Neuroscience methods may help disentangle these accounts…

Two EEG studies used:
- sentence contexts (N=31)
- and visual primes (N=33)

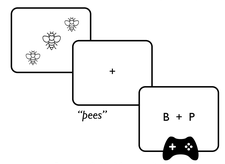
to assess top-down influences on acoustic encoding of a target word



Sarrett, McMurray, & Kapnoula (2020)

Target words varied along a VOT continuum, and sentences/pictures biased listeners to expect one side of the continuum over the other

Participants responded on a gamepad to indicate which phoneme they heard

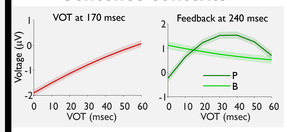


Sarrett & McMurray (in prep)

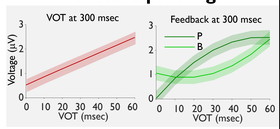
Results

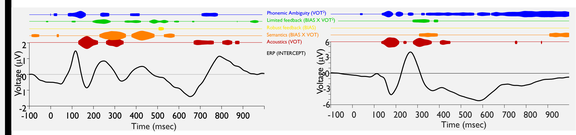
We ran LMEs over time to determine when our factors of interest significantly predicted voltage at frontocentral electrodes, correcting for multiple comparisons Seedorff et al (2020)

Sentence contexts



Picture priming





We find evidence of top-down feedback for both sentence contexts and visual primes, but in limited cases (when acoustic cues are ambiguous); consistent with interactive models of speech perception

Bottom-up Encoding

Pereira et al. (2018) examined N1 amplitude as an index of other phonetic distinctions, using a wide a range of speech sounds

They found that many phonemes were distinguishable by N1 amplitude, however, some phonemes—such as /s/ and /ʃ/—did not show N1 differences using traditional analyses

→ Can machine learning offer a more sensitive measure?
→ Is there enough detail in EEG to decode fine-grained acoustic detail from neural activity at the scalp level?

We utilize Pereira's EEG data (N=26) and a multi-class support vector machine (SVM) to answer these questions Chang & Lin (2011)

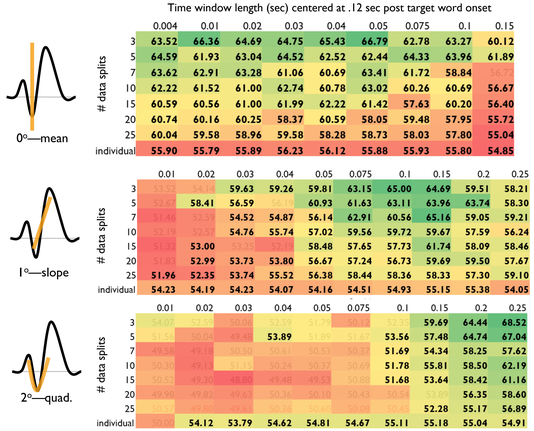
Goals

Sarrett & Toscano (in prep)

(1) To explore what features of the EEG signal maximize machine learning accuracy on a known detectable difference in N1 amplitude—i.e. voicing
(2) To examine the timecourse of decoding for three relevant dimensions of speech—voicing, manner of articulation, and place of articulation
(3) To determine which pairs of phonemes are detectable with machine learning

Results

First, we manipulated feature input along three dimensions (timepoints averaged x number of trials averaged x polynomial fit) on a voicing distinction (i.e. voiced vs. voiceless; chance = 50%)

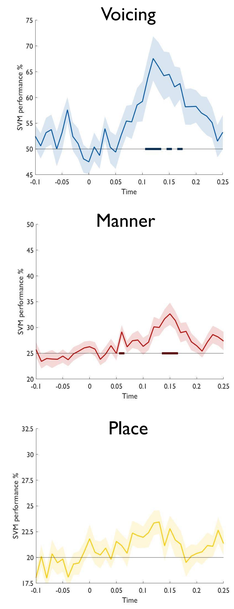


We find that each dimension affects machine learning performance, and that there are different tradeoffs between timepoints averaged and trials averaged, for each polynomial fit

→ Researchers should consider such tradeoffs when determining which aspects are most important for a given research question

Machine Learning

Timecourse of phonetic feature decoding



Second, we apply these techniques (from goal one) in three separate classification jobs (see left)

We were primarily interested in maintaining high temporal resolution, so we used the mean voltage across a .004 sec time window, averaged 3-ways across trials

We found that Voicing & Manner were readily decodeable above chance (chance shown by thin black line; significant windows are indicated with bars), peaking at slightly different times (0.12 and 0.15 sec, respectively)

Place was not decodeable above chance

These results suggest that perceptual representations of speech sounds may be based primarily on acoustic (rather than articulatory) dimensions

Phoneme Pairs

Third, we decoded three sets of phoneme pairs

We wanted to maximize accuracy (since we had fewer trials per category), so we used a quadratic polynomial fit and a .250 sec time window, averaged 3-ways across trials

- /b/ vs. /p/ — 53.87%, t(25) = 4.02, $p<.001$
- /s/ vs. /ʃ/ — 53.57%, t(25) = 5.62, $p<.001$
- /v/ vs. /z/ — 54.25%, t(25) = 4.65, $p<.001$

/b/ vs. /p/ functioned as a "baseline" condition, as we know this distinction shows robust N1 differences

/s/ vs. /ʃ/ and /v/ vs. /z/ did not show N1 differences in Pereira's original analyses

Conclusions

This set of studies demonstrates the complex interplay between perceptual representations and semantic expectations during the cortical processing of spoken language

Top-down influences affect perceptual encoding when acoustic cues are ambiguous—clarifying models of speech perception

The nature of perceptual representations may be elucidated using powerful machine learning techniques, which can detect perceptual distinctions in the neural signal that may be missed by traditional analyses—opening entirely new possibilities for EEG experiments