Phonological (un)certainty weights lexical activation

Anonymous EACL submission

Abstract

Spoken word recognition involves at least two basic computations. First is matching acoustic input to phonological categories (e.g. /b/, /p/, /d/). Second is activating words consistent with those phonological categories. Here we test the hypothesis that the listener's probability distribution over lexical items is weighted by the outcome of both computations: uncertainty about phonological discretisation and the frequency of the selected word(s). To test this, we record neural responses in auditory cortex using magnetoencephalography, and model this activity as a function of the size and relative activation of lexical candidates. Our findings indicate that towards the beginning of a word, the processing system indeed weights lexical candidates by both phonological certainty and lexical frequency; however, later into the word, activation is weighted by frequency alone.

1 Introduction

There is mounting evidence for the predictive nature of language comprehension. Response times and neural activity are reduced in response to more predictable linguistic input. This indicates that the brain forms probabilistic hypotheses about current and future linguistic content, which manifest in expectations of phonemes, morphemes, words and syntactic structures (Connolly and Phillips, 1994; Lau et al., 2006; Lau et al., 2008; Ettinger et al., 2014; Gwilliams and Marantz, 2015).

In speech comprehension, the brain's task is to correctly determine a word's identity as quickly as possible. It is not optimal to always wait until word ending, because the target may be correctly identifiable earlier. For example, after hearing *hippopotamu-* the final /s/ provides very little additional information. Indeed, one could even stop at *hippot-* and still identify the target word correctly most of the time.[1]

How is this done? Research suggests that upon hearing the beginning of a lexical item, the brain activates the cohort of words that are consistent with the acoustic signal. Words in the cohort are activated relative to their match to the phoneme sequence and frequency of occurrence. With each subsequent phoneme, the cohort is reduced as items cease to be consistent with the provided input, until one item prevails (see Figure 1). This process is consistent with the highly influential cohort model of spoken word recognition (Marslen-Wilson and Welsh, 1978; Marslen-Wilson, 1987), and has been associated with activity in left superior temporal gyrus (STG) (Gagnepain et al., 2012; Ettinger et al., 2014; Gwilliams and Marantz, 2015).

In practice though, phoneme identity is often uncertain: the acoustic signal may be consistent with both a [b] and a [p], for example. This phonetic uncertainty, and its effect on lexical activation, is not addressed by the cohort model. However, there is evidence suggesting that phonetic uncertainty affects lexical and sentential processing (Connine et al., 1991; McMurray et al., 2009; Bicknell et al., 2015).

Here we build upon this previous work in order to understand the neural computations underlying lexical activation, in service to spoken word recognition. Concretely, how does fine-grained acoustic information (below the phonological level) serve to activate lexical hypotheses and estimate their

[1] Note that *hippopotomonstrosesquippedaliophobia* ('fear of long words') and *hippopotas* ('a ground-type Pokemon') are also possible lexical items but much less frequent than the target in this case, so less likely to be selected.

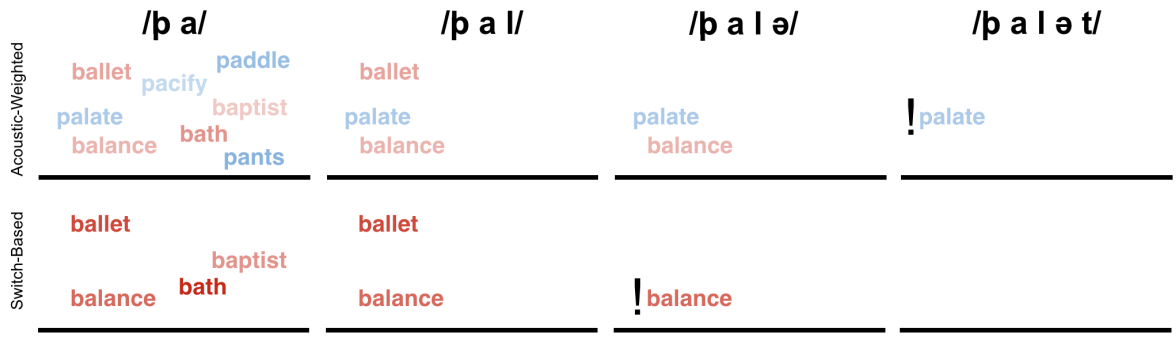

Figure 1: Schematic depiction of cohort activation under each of the two models, for the first five phonemes of the word *palate*. The onset b-p symbol represents that the onset phoneme was 75% consistent with a /b/ and 25% consistent with a /p/. Transparency reflects relative word activation. Note that the change in transparency between the two accounts reflects the actual probabilities predicted by each model — because there are more words activated in the Acoustic-Weighted account, less normalised probability is assigned to each item.

probabilities? And can this integration between phonological and lexical levels of description be read out from activity in auditory cortex?

To address these questions, we model neural responses in STG, time-locked to each phoneme in a word, as a function of two theoretically-motivated computational models. One model assumes that activation of a lexical candidate is weighted by the acoustic evidence in favour of that candidate: e.g., "balloon" is activated in proportion to how /b/-like the phoneme is at word onset, even if a different phoneme (e.g., /p/) is more likely. The other assumes that acoustic information serves as a switch either activating or inhibiting that lexical item (henceforth **switch-based**). The latter is what is predicted by the traditional cohort model – the system commits to whichever phoneme is more likely, and this is used to form predictions at the lexical level (see Figure 1). A subset of the data reported here are also published in (Gwilliams et al., 2017).

2 Summary of human data

2.1 Materials

Word pairs were selected such that, apart from the first phoneme, there was an identical phoneme sequence until "point of disambiguation". For example, "palate" and "balance" share their second [æ], third [l] and fourth [ə] phonemes, and diverge on the fifth [t]/[n]. We selected one-hundred and three word pairs with this property. The onset of each word was either a voiced (d, b, g) or voiceless (t, p, k) plosive. A native English speaker was recorded saying each of these 206 words in isolation. The onset of each word was morphed along one phonetic feature, using the TANDEM-STRAIGHT software to create a word (e.g., *direct*) to non-word (e.g., *tirect*) 11-step continuum (see Figure 2). The 11-step acoustic continuum was then re-sampled to form a 5-step perceptually defined continuum, based on the proportion of selections in a behavioural pre-test.

2.2 MEG experiment

Native English participants ($n = 25$) listened to each of the 103×5 words in isolation, and in 20% of trials (randomly distributed) made an auditory-to-visual word matching judgment.

While completing the task, neural responses were recorded using a 208 sensor KIT magneto-encephalography (MEG) system. Data were sampled at 1000 Hz which provided a measure of neu-

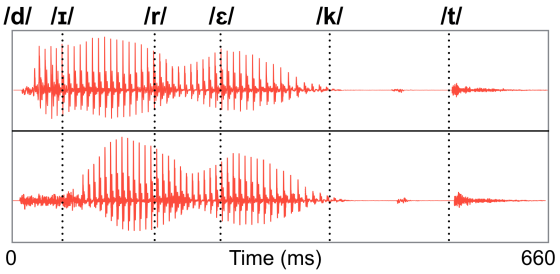

Figure 2: Waveforms of example endpoints of a lexical continuum. The word *direct* is above, and the non-word *tirect* is below. Dashed lines correspond to the timing of each phoneme onset.

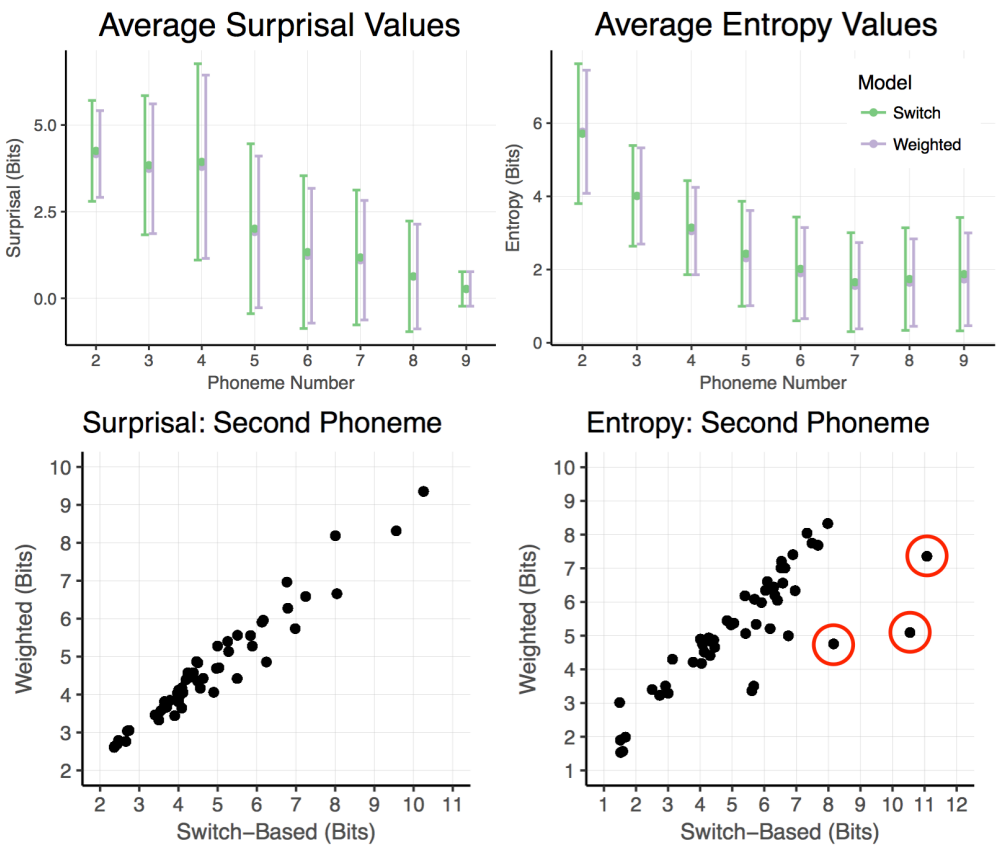

Figure 3: Top: Average surprisal and entropy values at each phoneme along the word. Note that not all words are 9 phonemes long, so phonemes at longer latencies contain fewer entries. Error bars represent one standard deviation from the mean. Bottom: Correlation between the two models' surprisal values and the two models' entropy values, at the second phoneme. Red circles highlight the outliers (from right to left) "topography", "tirade" and "casino".

ral activity at each millisecond. In order to test responses to specific phonemes in a word, the data were cut into a series of 700 ms epochs, where the time at 0 ms corresponds to the onset of a phoneme. The activity recorded from MEG sensors was localised using MNE-Python software (Gramfort et al., 2014), and averaged over the left STG. This provided one datapoint per millisecond (700) per phoneme (4370) per participant (25).

3 Modeling of MEG data

The variables of interest were entropy and surprisal. Entropy quantifies uncertainty about the resulting lexical item. Below, f_w is the frequency of the word (w), and f_C is the total frequency of the cohort (C) (the set of all words consistent with the heard prefix):

$$-\sum_{w \in C} \frac{f_w}{f_C} log_2(\frac{f_w}{f_C})$$

For acoustic-weighted entropy, all words consistent with the input (e.g. all *b*-onset and *p*-onset words) were included in the cohort (C), and weighted by the likelihood of onset-phoneme identity (see Figure 1, top panel). The equation for this weighting term is given below. Switch-based entropy was calculated by simply rounding this leading term to its nearest integer (either 1 or 0; see Figure 1, bottom panel). Thus, the switch-based cohort only included the set of words *most* consistent with the input (e.g. only the *b*-onset words):

$$P(OnsetPhoneme|Acoustics) \times P(CohortFrequency|OnsetPhoneme)$$

Surprisal quantifies how likely the current phoneme (p_0) is to occur, given what has occurred previously (p_{-1}):

$$-log_2(\frac{f(p_0)}{f(p_{-1})})$$

Acoustic-weighted surprisal was calculated by multiplying the conditional probability by onset-phoneme certainty. Again, for switch-based surprisal, the leading term was rounded to its nearest integer.

$$P(OnsetPhoneme|Acoustics) \times P(CurrentPhoneme|PreviousPhonemes)$$

4 Results

The dependent measure was activation of left STG, averaged between 200-250 ms after phoneme onset, following the results of (Ettinger et al., 2014). This activity was modelled time-locked to the second phoneme in the word (mean post-onset latency = 87 ms; SD = 25 ms, 4021 observations) and the sixth phoneme in the word (mean post-onset latency = 411 ms; SD = 78 ms, 3264 observations). We chose the second and sixth phonemes because it allowed a similar number of trials to be included in each model comparison, while also ensuring a substantial difference in latency from word onset. Only responses to partially ambiguous trials were included (0.25 and 0.75), because this is where the predictions of acoustic-weighted and switch-based models are most distinct.

We evaluated the fit of the predictions of each model to the neural measurement using a linear mixed effects regression model. The full model contained switch-based and acoustic-weighted surprisal, switch-based and acoustic-weighted entropy, phoneme latency, trial number, block number, stimulus amplitude of the first 30 ms, phoneme pair and ambiguity as fixed effects. By-subject slopes were included for all entropy and surprisal predictors. This full model was compared to a model where either acoustic-entropy and surprisal, or switch-based entropy and surprisal, were removed as fixed effects (but remained as by-subject slopes). This gave a statistical assessment of the amount of variance the acoustic-weighted and switch-based models were accounting for.

At the second phoneme, the acoustic-weighted variables explained a significant amount of variance ($\chi^2 = 5.02$, $p = .025$), whereas the switch-based variables did not ($\chi^2 = 2.62$, $p = .1$). At the sixth phoneme, the opposite pattern was true: the switch-based variables explained a significant amount of variance ($\chi^2 = 5.26$, $p = .022$) and the acoustic-weighted variables had only marginal explanatory power ($\chi^2 = 3.46$, $p = .06$).

5 Discussion

We have found evidence that the brain indeed uses fine-grained acoustic information to weight lexical predictions in spoken word recognition. At the beginning of a word, lexical hypotheses are activated in proportion to the bottom-up acoustic evidence; towards the end, acoustic evidence acts as a switch-like function, to either fully activate or deactivate the word, bounded by its frequency of occurrence. This finding has two primary implications.

First, it suggests that the system does not wait until phonological categories have been disambiguated before activating lexical items. Rather, uncertainty about phonological classification is used to modulate higher level processes. This supports interactive models of speech processing, because it suggests that the output of one stage does not need to be determined before initiating the following. In particular, this finding is inconsistent with the Cohort model of speech perception (Marslen-Wilson and Welsh, 1978), which assumes that the system first commits to the most likely phoneme before making lexical predictions.

Second, it suggests that the same processing strategy is not heuristically applied in all situations. Rather, phonological information appears to be used more when processing the beginning of a word than the end. There are two explanations for this. This could reflect that the system commits to a particular phonological category, and so the phonological weights are themselves converging to a decision point. Or perhaps phonological detail of earlier sounds becomes less informative as the word progresses and so its content is given less predictive power by the processing system. A simple way to tease these alternatives apart in future work is to manipulate the ambiguity of phonemes within a word, not just in initial position.

References

K. Bicknell, M.K. Tanenhaus, and T.F. Jaeger. 2015. Listeners can maintain and rationally update uncertainty about prior words. *Manuscript submitted for publication.[KB].*

C. Connine, D. Blasko, and M. Hall. 1991. Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. *Journal of Memory and Language*, 30(2):234–250.

J. Connolly and N. Phillips. 1994. Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, 6(3):256–266.

A. Ettinger, T. Linzen, and A. Marantz. 2014. The role of morphology in phoneme prediction: Evidence from meg. *Brain and language*, 129:14–23.

P. Gagnepain, R. Henson, and M. Davis. 2012. Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, 22(7):615–621.

A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and Matti S Hämäläinen. 2014. MNE software for processing MEG and EEG data. *Neuroimage*, 86:446–460.

L. Gwilliams and A. Marantz. 2015. Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken arabic words. *Brain and language*, 147:1–13.

L. Gwilliams, T. Linzen, D. Poeppel, and A. Marantz. 2017. In spoken word recognition the future predicts the past. *bioRxiv*, page 150151.

E. Lau, C. Stroud, S. Plesch, and C. Phillips. 2006. The role of structural prediction in rapid syntactic analysis. *Brain and language*, 98(1):74–88.

E. Lau, C. Phillips, and D. Poeppel. 2008. A cortical network for semantics:(de) constructing the N400. *Nature Reviews Neuroscience*, 9(12):920–933.

W. Marslen-Wilson and A. Welsh. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1):29–63.

W. Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2):71–102.

B. McMurray, M. Tanenhaus, and R. Aslin. 2009. Within-category vot affects recovery from lexical garden-paths: Evidence against phoneme-level inhibition. *Journal of memory and language*, 60(1):65–91.