TOWARDS A MECHANISTIC ACCOUNT OF SPEECH COMPREHENSION

Précis

LAURA GWILLIAMS
New York University

The challenge of speech comprehension

Although listening to someone talk *feels* like effortless and automatic understanding, speech comprehension involves overcoming major computational challenges. The mapping from acoustics to meaning is mostly arbitrary [1], different speakers have vastly different ways of pronouncing words depending on biological, regional and incidental factors [2] and external noise, such as the voices of surrounding talkers or non-linguistic sources, routinely mask the signal [3]. The extent of this challenge is exemplified by the fact that, despite the vast amounts of money and time invested, current state-of-the-art automatic speech recognition systems such as *Siri*, *Alexa* and *Google Home* do not come close to the accuracy, speed and flexibility demonstrated by human listeners [4].

The ultimate goal of my research program is to develop a theoretically precise, biologically plausible and computationally powerful model of spoken language comprehension. To this end, my work aims to delineate the processing architecture upholding speech comprehension in three primary aspects: (i) what **representations** does the brain generate from the auditory signal; (ii) what **computations** are applied to those representations during the timecourse of processing; (iii) in what **order** do computations unfold. The answers to these questions are key to understanding auditory, speech and language processing, which I strongly believe require complementary insight from linguistics, machine learning and neuroscience in order to be successful.

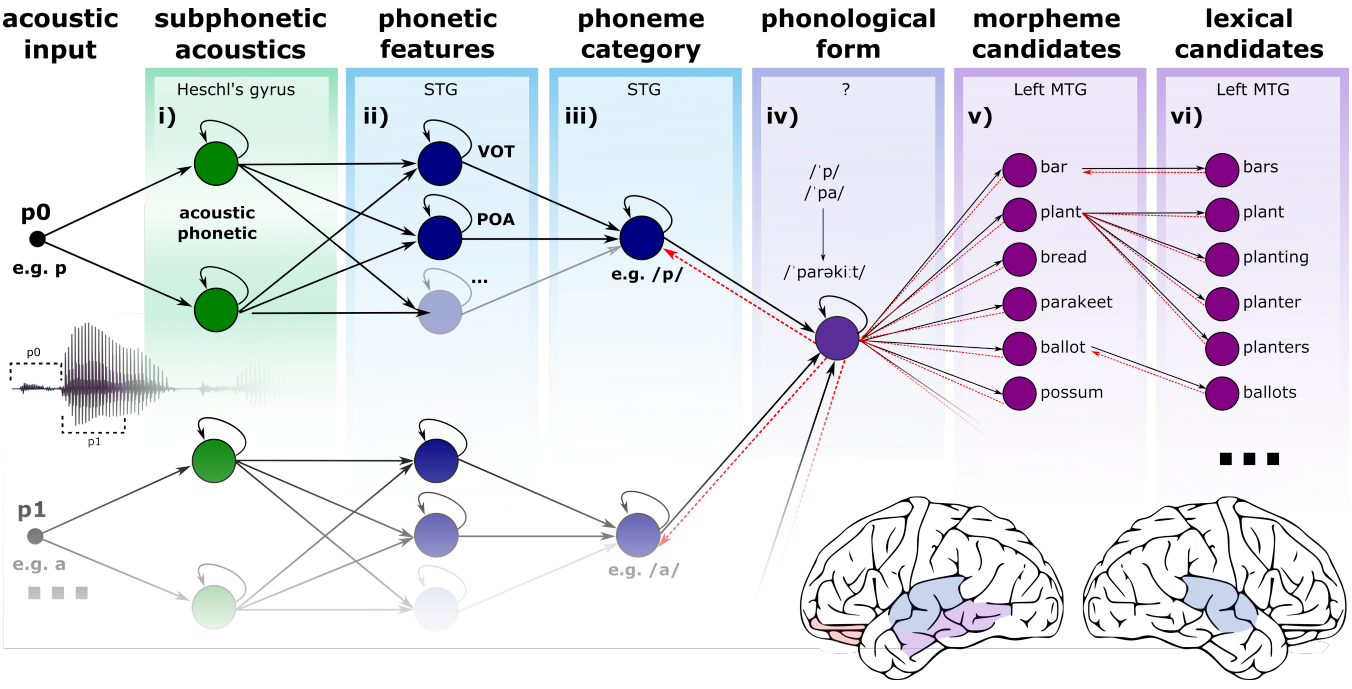

Figure 1: Schematic model architecture. Processing stages are labelled in terms of the representational format they act upon (bold font, above) and their putative location of processing. The example shows the processing of the first phoneme of the word 'parakeet'. *Figure adapted from Gwilliams et al., (2018).*

Computational processing stages

To address these questions, my PhD research primarily utilised magneto-encephalography (MEG): A non-invasive brain imaging technique with millisecond temporal resolution, capable of tracking the propagation of neural activity across the whole brain. In the sections which follow, I describe specific sub-components of the processing architecture that my dissertation research has characterised, based upon a combination of psychophysics, encoding and decoding analyses of time resolved neural data and computational modelling. The results have led me to develop the schematic processing model shown in Figure 1. Note that although I present stages in sequence, my research suggests that these operations occur largely in parallel, and are initiated in a reverse-hierarchical order during continuous speech comprehension (*Gwilliams et al., submitted, [5]*).

Analytical Framework

Appears in: *King, Gwilliams et al. (2018). The Cognitive Neurosciences.*

One major challenge in developing a neuro-cognitive model of speech comprehension is being able to appropriately decompose dynamic neural signals into an interpretable sequence of operations. This involves overcoming three primary challenges. First, identify what features of the input the brain generates during processing. The question of *what is represented* is common across most, if not all, domains cognitive neuroscience: For example, for vision science, relevant representations include visual angle, spatial frequency and contrast [6, 7]. In speech research, putative representations include things like speech sounds (e.g. /p/, /b/), syllables (e.g. /pa/, /ba/) or word identity (e.g. parakeet, barricade).

Second is to identify *how it is encoded* in neural responses. Different coding schemes have been shown to exist, such neuronal spiking rate [8], spike timing [9] and oscillatory activity [10]. Investigating neural processes in humans typically involves taking non-invasive brain measurements, which aggregate the responses of multiple neurons. Thus, in practice, indirect proxies of neural coding schemes are used, such as strength of activity as measured in femtotesla (in the case of MEG, for example), oscillatory power or phase within a frequency band, or patterns of activity within and across brain regions.

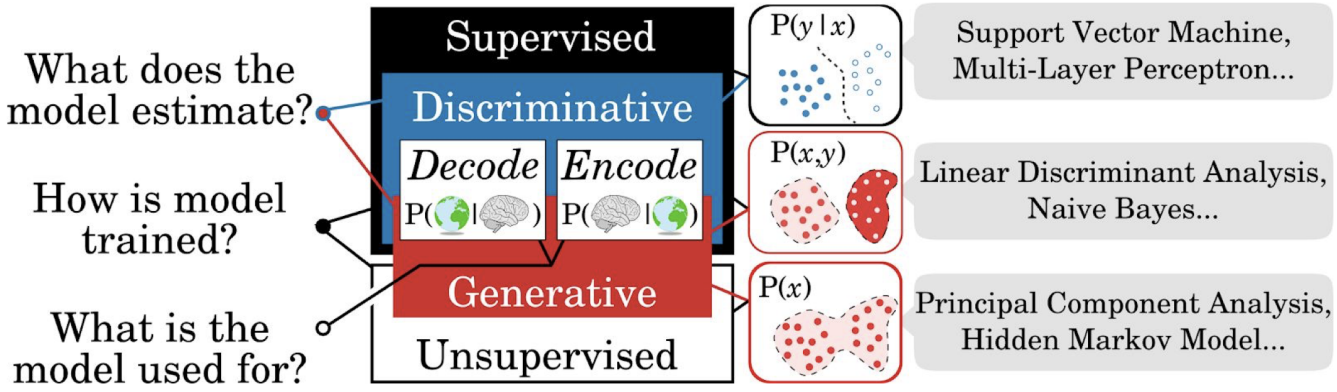

Figure 2: The most common models used in cognitive neuroscience. Left: Models can be distinguished with three main dimensions of the statistical framework. Models that are trained to estimate a conditional probability P(y | x) between two sets of variables (e.g. x=psychological variables, y=neuronal recordings, or vice-versa) are referred to as 'discriminative'. By contrast, models that estimate the joint distribution P(x,y) are considered to be 'generative'. All generative models can thus be used then to derive P(y | x) for prediction. These models are 'supervised' in that they are trained to predict one variable (e.g. y) from another (e.g. x). By contrast, 'unsupervised' models estimate the distribution of a single (possibly multidimensional) variable (x). Finally, a trained model can ultimately be used for different purposes: e.g. decoding or encoding. Right: Examples of classical supervised and unsupervised models. *Figure adapted from King, Gwilliams et al., (2018).*

Third is how these two levels of description relate to one another. This involves learning a mapping from the content of representations to its coding scheme. Different families of mapping functions can be used for this purpose (see Figure 2), and organised relative to three axes: Is the method supervised, is it generative or discriminative, and does it model neural responses using stimulus properties (encoding) or vice versa (decoding)? Typically, this involves an

explicit decision of whether to prioritise testing *what* is encoded (e.g. syllable versus word identity), while assuming the coding scheme is known (e.g. strength of activity in the temporal lobe). Or, instead, *how* it is encoded (e.g. strength of activity in the temporal lobe versus auditory cortex) for its ability to predict the known representation of interest (e.g. word identity). My PhD research utilises multiple analysis techniques that span this mapping space, as described in the following sections.

How acoustics are mapped to speech sounds

Appears in: *Gwilliams, Linzen, Poeppel & Marantz (2018). Journal of Neuroscience.*

It is posited that the brain encodes representations of linguistically defined properties called *phonetic features* [11] when processing speech [12]. These features are what distinguish a phoneme like /p/ from other phonemes like /b/ or /t/. The neural populations which are selective of particular phonetic features reside in superior temporal gyrus (STG) and respond ∼100 ms after speech-sound onset. While this previous research has been instrumental in helping to understand how speech is processed, it has primarily tested cases where the sounds clearly belong to one phoneme category or another (e.g. /p/ vs. /b/). In real-world speech, by contrast, speech sounds are often ambiguous. That is, it will be unclear, based on the acoustic signal alone at least, what phoneme a speaker intended to pronounce.

The purpose of this study was to understand how the variable and ambiguous acoustic signal of speech gets mapped onto a single and stable set of phonetic features: What series of operations achieve this categorisation?

Participants listened to phoneme continua either at the onset of syllables (Experiment 1) (e.g. /ba/ ↔ /pa/) or words (Experiment 2) (e.g. /barricade/ ↔ /parricade/) while magentoencephalography (MEG) was recorded. Across both experiments, sensitivity to phonological ambiguity occurred at just 50 ms after onset in primary auditory cortex. This illustrates that early stages of processing are sensitive to strikingly complex properties of the signal: The distance between the acoustic signal and the perceptual boundary that distinguishes one speech sound from another. At ∼100ms after syllable onset responses in STG encoded the veridical linear acoustic signal *in parallel to* categorical behavioural reports (Figure 1, boxes ii & iii). This suggests that ambiguity is neutralised, but not necessarily discarded, at this latency. The results therefore support that during speech processing the brain transforms acoustic input into a discrete categorical format that can interface with stored memory representations (e.g. morphemes, words). Simultaneously, properties of the acoustic signal are available to the processing system.

In a related study (*Gwilliams & King, 2020, [13]*) we found similar evidence for a categorisation process in the visual domain. By applying multivariate decoding analyses to MEG data, we found that perceptual decisions are best explained by a joint feedforward and recurrent processing architecture, whereby the system maintains lower-level (linear) representations in parallel to generating and propagating higher-level (non-linear) representations of the input signal. This suggests that non-linear categorisation of inputs may be a 'canonical computation' used across domains, modalities and behavioural task.

How speech sounds activate lexical candidates

Appears in: *Gwilliams, Poeppel, Marantz & Linzen (2018). Cognitive Modeling and Computational Linguistics.*

Having identified phonetic features in the speech input, the system needs to use them to recognise what word is being said. Previous work suggests that this unfolds under a process of lexical competition: As each sound of the input is recognised (e.g. p, a, r, a...) potential lexical candidates are activated (e.g. parakeet, paragraph, paramount...) and inconsistent candidates are inhibited (e.g. pardon, park, parlour...). This process continues until a single candidate 'wins' the competition, its reward being lexical recognition [14] (see Figure 3).

The goal of this study was to better understand how the lexical recognition process works. While previous research suggests that multiple candidates of lexical items are considered in parallel, no study had sought to test whether lexical activation also considers multiple phoneme categories. For example, if a speech sound is ambiguous between /p/ and /b/, does that mean that both p-onset words (e.g. parakeet, paragraph) and b-onset words (e.g. barricade, baritone) would participate in the competition? Or, does the brain first commit to a single phoneme category (e.g. /b/), only permitting one cohort of words to compete for recognition - e.g. *just* b-onset words (e.g. barricade, baritone)?

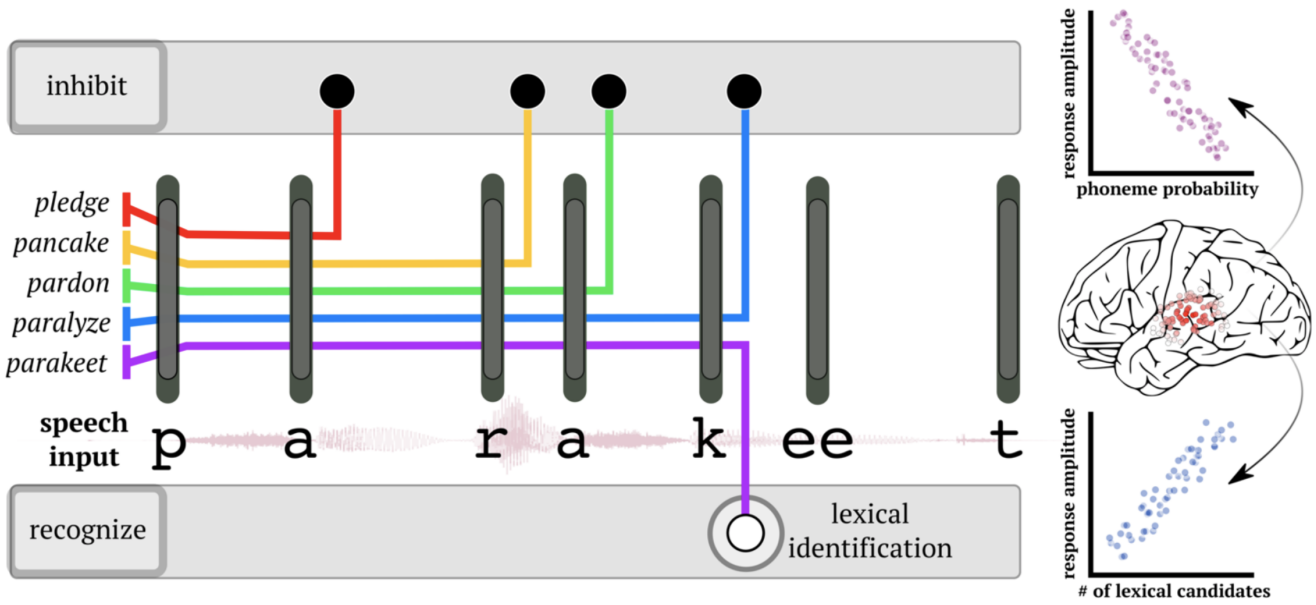

Figure 3: Schematic of spoken word recognition. As the word unfolds, with each phoneme (gray bars), possible lexical candidates for recognition are ruled out. This process continues until only one word remains consistent with the input (e.g. 'parakeet'). Notice that the word can be uniquely identified before word offset – at the 'k' phoneme. Activity in left superior temporal gyrus has been shown to track the probability of each phoneme in a word ("phoneme surprisal"), as well as the relative probability of remaining lexical candidates ('cohort entropy'). *Figure taken from Dikker, Assaneo, Gwilliams et al., (2020).*

To address this question I used supervised encoding models; i.e. I fit linear mixed effects regression to analyse activity in left STG. I compared the results of two models: One that assumes lexical candidates are weighted probabilistically by acoustic-phonetic information (both p-onset and b-onset words could be considered simultaneously in this case, weighted by acoustic evidence); the other assumes that lexical candidates are activated in a categorical manner (either p-onset or b-onset words will be activated: Whichever is more likely given the acoustic signal). The results suggest that at the beginning of a word, lexical hypotheses are activated probabilistically in proportion to the bottom-up acoustic evidence (Figure 1, box ii); towards the end, lexical items are activated relative to the most likely phonological category (Figure 1, box iii). This suggests that the acoustic signal serves to activate lexical items from the very earliest stages of processing, and there exists a 'phoneme commitment' stage, which converges upon categorical phonemes over time.

This result is important for our mechanistic understanding of spoken word recognition for two reasons. First, we show that information below the phonological level (i.e. within-phoneme variation) influences the lexical access procedure. This challenges a strict representational hierarchical view of speech processing [15], by suggesting that low-level variability need not be discarded in order to initiate lexical processing. Second, it suggests that the brain integrates multiple sources of acoustic and linguistic information in order to initiate lexical access: It is not based on a categorical phonological sequence alone, but rather a weighted combination of acoustic, phonological and lexical evidence (Figure 1, box vi).

How words (in turn) aid comprehension of speech sounds

Appears in: *Gwilliams, Linzen, Poeppel & Marantz (2018). Journal of Neuroscience.*
Appears in: *Gwilliams*, Elorrieta*, Marantz & Pylkkanen (2021). Nature Scientific Reports.*

When listening to speech, one seldom hears individual phonemes or syllables in isolation — usually, surrounding context is available to guide correct identification of speech sounds. In this chapter I explore how top-down lexical information (i.e. knowing what word is being said) aids interpretation at the phonetic level (e.g. the word was 'parakeet' so I must have heard a /p/ and not a /b/ at the beginning). We created words where the onset had been

manipulated along a 5-step continuum from one phoneme category to another to create word ↔ non-word continua (e.g. parakeet ↔ barakeet). Participants listened to these words while MEG was recorded.

Overall we found that acoustic properties (which linearly vary along the 5-step continuum) and categorical phonetic features (matching the perceived category of the listener) of the onset phoneme are maintained in parallel, and specifically re-activated within the same brain regions, at subsequent phoneme positions. Sub-phonemic representations therefore seem to be sustained through recurrent connections in superior temporal regions well beyond phoneme offset, while subsequent phonemes of the word are also being processed. We interpret this as a kind of 'safety net' procedure. The brain has the opportunity to reassess the percept of a speech sound based on the provision of each additional input — allowing future context to modulate the perception of previously heard speech sounds, and update the categorisation that may have been derived at an earlier stage if later inputs suggest that the wrong categorisation was made.

A good example of when top-down guidance is critical is when listening to an accented talker. In this case, the listener will find themselves more heavily relying on linguistic and pragmatic context in order to correctly understand what is being said. In *Gwilliams et al., 2021 [16]* we tested the neural processes that aid 'getting used to' accented speech. We created words that contained systematic phoneme substitutions (e.g. travel/trabel). Participants listened to these systematically mis-pronounced words, as well as a typically pronounced baseline, and performed a picture matching task while MEG was recorded. We found that correct word identification in mis-pronounced speech improved as a function of exposure, and this behavioural improvement was associated with neural responses in the inferior frontal gyrus. By contrast, responses in auditory cortex, including STG, were not modulated by exposure - they remained consistently elevated in response to phoneme substitutions.

Together, these sets of results suggest that lexical context is used to guide the processing of individual speech sounds. The brain keeps acoustic ↔ phonemic mapping relatively stable, in favour of repairing and re-categorising based on higher order representations which are invariant to acoustic variability. This adds support to the role of evidence accumulation and perceptual re-evaluation during the timecourse of processing.

How words are recognised in continuous, natural speech

Appears in: *Gwilliams, King, Marantz & Poeppel (under revision). bioRxiv.*

Many of the studies described in this précis demonstrate that the brain remains sensitive to the features of a speech sound way beyond the moment that the sound dissipates from the auditory signal. However, by direct consequence, this also suggests that the brain processes multiple speech sounds at the same time – i.e. not just the sound present in the auditory signal *now* but also the one before that, and the one before that, too. This raises at least two critical questions. First, if the brain is processing multiple phonemes simultaneously, how does it not get confused about the content of those speech sounds? Second, how is it that the brain correctly recalls the order of the sounds it heard in order to select the appropriate lexical item down the line? Phoneme order is crucial for correctly recognising that the word was *teach* and not *cheat*, or *melons* not *lemons*, for example.

In this work I ask how the brain keeps track of the identity and relative order of phonemes in continuous prose, given that the incoming acoustic-phonetic information is highly overlapping. Participants listened to two hours of short audio-books while MEG was recorded. I annotated the audio for the precise timing, identity and location of the ∼40,000 phonemes that comprise the stories. Using a modified version of the temporal generalisation analysis [17], I trained a decoder on the neural responses to all of the phonemes in the story, at different latencies relative to phoneme onset. Then I evaluated decoding performance separately as a function of the phonemes' position in the word, at different latencies relative to word onset (see Figure 4). This yielded a decoding performance for the first phoneme in the word (p1) separately from the second (p2) and the third (p3), and so on.

We chose the temporal generalisation analysis because it allowed us to assess how neural activity patterns, which encode phonetic information, evolve over time. The analysis revealed that each phoneme was decodable for about 350 ms after onset - about 3 times longer than the duration of the phoneme itself, which is around 80 ms. This confirms that multiple speech sounds are processed at the same time. However, any given spatial pattern of activity was only informative for around 80 ms. This suggests that the speech input is transformed roughly at the rate of phoneme duration (one transformation cycle every ∼80ms), which allows the brain (i) to avoid representational overlap between neighbouring speech sounds; and (ii) to retain the relative order of those sounds in the state of the

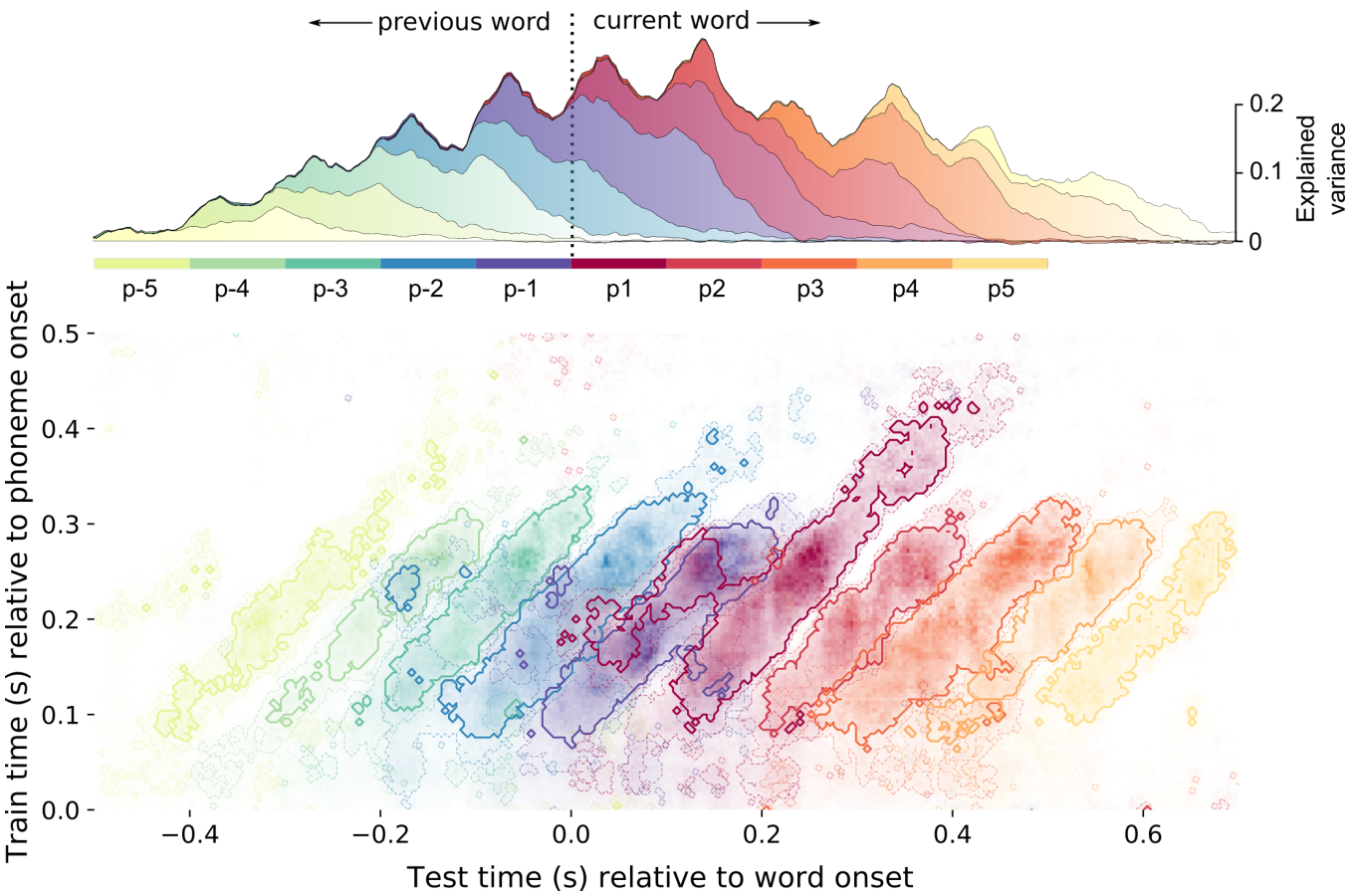

Figure 4: Processing phonetic features in continuous speech. Decoding performance superimposed for 10 phoneme positions. From word onset, forwards (p1, dark red), and from word offset, backwards (p-1, dark blue). Each phoneme position is shifted by the average duration from one phoneme to the next. The y-axis corresponds to the time that the decoder was trained, relative to phoneme onset. The x-axis corresponds to the time that the decoder was tested, relative to word onset. Contours represent a t-value threshold of 4 (darker lines) and 3.5 (lighter lines). Coloured shading represents decoding accuracy. *Figure taken from Gwilliams et al., (2020).*

representation over time. We also found that phonetic features were decodable earlier within predictable sequences, and were decodable for longer when lexical uncertainty was higher. This links to my previous work investigating information theoretic measures, such as surprisal and entropy, for their role in speech comprehension (*Gwilliams & Marantz, 2015; Gwilliams et al., 2017, [18, 19]*; for a review see *Gwilliams & Davis, 2020, [20]*). Overall, the results demonstrate how phoneme sequence order is preserved over long time scales, and how these sequences interface with stored lexical candidates (Figure 1, box iv).

How hierarchical features are generated during naturalistic listening

Appears in: *Gwilliams, Poeppel, Marantz and King (submitted).*

In the final chapter I move beyond phonological and lexical processes to investigate representations across the full linguistic hierarchy: From speech sounds to syntax and meaning. The goal was to test the order with which different features become decodable from neural responses to continuous speech. I annotated spoken narratives for a total of 54 linguistically motivated features, which can be grouped into 6 hierarchical feature families: (1) phonetic features (2) sub-lexical (how many phonemes, syllables and morphemes were contained in a word) (3) part of speech (noun, verb, adjective) (4) syntactic operation (node opening, node closing) (5) syntactic state (tree depth, word position) (6) proxies of semantic processing (principal components of the GloVe word embeddings [21]).

One hypothesis is that because language has compositional structure, language processing is also compositional [22]. If this were the case, we should observe that lower level features of speech (e.g. phonetic features) are decodable

earlier than higher level features (e.g. word class, syntactic properties). An alternative hypothesis is that processing unfolds under a reverse hierarchy [23]. This would predict that higher order properties will be decodable earlier than the lower level ones, because continuous speech affords the construction of sentence structure and preemptive structural predictions. I like to refer to this hypothesis as the 'underdog' because it has been primarily put forward for visual processing, though see [24] for a related proposal in language processing.

Subjects listened to the spoken narratives while MEG was recorded. Using back-to-back regression to control for the feature correlations [25], I derived the spatio-temporal profile of each hierarchical language property. These decoding analyses revealed that each linguistic feature is decodable from the brain signal in a reverse hierarchical sequence, and maintained in parallel for around a second after word offset. This permits the system simultaneous access across multiple levels of representation, allowing the output of one processing stage (e.g. part of speech) to both constrain the solution of subsequent stages, and provide feedback signals to preceding stages. This work suggests that the parallel architecture displayed in Figure 1, in service to phonetic processing in words, also scales to the processing of words within sentences. Furthermore, the reverse hierarchical architecture, which has been primarily put forward for visual processing, also seems to be relevant for processing natural speech. This is suggestive of a canonical architecture which may get re-purposed for different cognitive domains and sensory modalities.

Conclusion

My research aims to integrate Neuroscience, Machine Learning and Cognitive Science in order to uncover the (i) representational format (ii) neural computations and (iii) processing order which upholds successful speech comprehension. I have investigated speech processing at multiple levels of description (phonetic, morphological, lexical, syntactic), applying a range of analytical techniques to MEG data, employing both carefully controlled paradigms and naturalistic approaches.

So far, my work has identified candidate processing motifs that are apparent across multiple sub-components of the processing architecture, and are even observable in other cognitive domains and sensory modalities. Such computations include non-linear transformation of veridical inputs, parallel and 'greedy' processing of lower- and higher-level representations, information maintenance through recurrent connections and reverse hierarchical processing. These are good candidates for canonical computations which get re-purposed for different cognitive tasks and behavioural goals.

Moving forward, I plan to continue developing the schematic processing model shown in Figure 1, to be more specific in terms of the neural coding scheme upholding each computation (e.g. which brain areas are involved, is information encoded within distributed population-level code or a more focal code) and in terms of expanding the processes to better account for sentence level phenomena. This will necessitate the continued integration of engineering and natural language processing (NLP) tools to generate a sufficient suite of linguistic features to test, and to provide the analytical infrastructure to test them. By continuing my research under a fully unified approach, it will be possible to develop a functional architecture of speech processing that can continue to be refined, and ultimately make measurable progress in understanding the human faculty for speech comprehension.

References

1. De Saussure, F. *Course in general linguistics* (Columbia University Press, 2011).
2. Stevens, K. N. & Blumstein, S. E. The search for invariant acoustic correlates of phonetic features. *Perspectives on the study of speech*, 1–38 (1981).
3. Mattys, S. L., Davis, M. H., Bradlow, A. R. & Scott, S. K. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes* **27,** 953–978 (2012).
4. Graves, A., Mohamed, A.-r. & Hinton, G. *Speech recognition with deep recurrent neural networks* in *2013 IEEE international conference on acoustics, speech and signal processing* (2013), 6645–6649.
5. Gwilliams, L., Marantz, A., Poeppel, D. & King, J.-R. Transforming acoustic input into a (reverse) hierarchy of linguistic representations. *bioRxiv* (2020).
6. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology* **160,** 106–154 (1962).
7. Wandell, B. & Thomas, S. Foundations of vision. *Psyccritiques* **42** (1997).

8. Shadlen, M. N. & Newsome, W. T. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of neuroscience* **18,** 3870–3896 (1998).
9. Gerstner, W. & Kistler, W. M. *Spiking neuron models: Single neurons, populations, plasticity* (Cambridge university press, 2002).
10. Singer, W. & Gray, C. M. Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience* **18,** 555–586 (1995).
11. Chomsky, N. & Halle, M. The sound pattern of English. (1968).
12. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343,** 1006–1010 (2014).
13. Gwilliams, L. & King, J.-R. Recurrent processes support a cascade of hierarchical decisions. *Elife* **9,** e56603 (2020).
14. Marslen-Wilson, W. D. Functional parallelism in spoken word-recognition. *Cognition* **25,** 71–102 (1987).
15. Davis, M. H. & Johnsrude, I. S. Hierarchical processing in spoken language comprehension. *Journal of Neuroscience* **23,** 3423–3431 (2003).
16. Gwilliams, L., Blanco-Elorrieta, E., Marantz, A. & Pylkkanen, L. Neural adaptation to accented speech: prefrontal cortex aids attunement in auditory cortices. *Nature Scientific Reports,* 852616 (In Press).
17. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences* **18,** 203–210 (2014).
18. Gwilliams, L. & Marantz, A. 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 (2015).
19. Gwilliams, L., Poeppel, D., Marantz, A. & Linzen, T. Phonological (un) certainty weights lexical activation. *arXiv preprint arXiv:1711.06729* (2017).
20. Gwilliams, L. & Davis Matthew, H. Extracting language content from speech sounds: An information theoretic approach. *HAL Preprint Archive* (2020).
21. Pennington, J., Socher, R. & Manning, C. D. *Glove: Global vectors for word representation* in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), 1532–1543.
22. Gaskell, M. G. & Marslen-Wilson, W. D. Integrating form and meaning: A distributed model of speech perception. *Language and cognitive Processes* **12,** 613–656 (1997).
23. Hochstein, S. & Ahissar, M. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* **36,** 791–804 (2002).
24. Ferreira, F., Bailey, K. G. & Ferraro, V. Good-enough representations in language comprehension. *Current directions in psychological science* **11,** 11–15 (2002).
25. King, J.-R., Charton, F., Lopez-Paz, D. & Oquab, M. Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *NeuroImage* **220,** 117028 (2020).