Understanding the neural architecture of speech comprehension

Laura Gwilliams[1]

[1]Stanford University

laura.gwilliams@stanford.edu

Abstract

Humans understand speech with such speed and accuracy, it belies the complexity of transforming sound into meaning. This paper offers an overview of electrophysiology studies that examine neural responses during speech processing, and schematic processing model which synthesises the results into candidate neural computations and representations. The results offer evidence that inputs are rapidly discretised into multiple formats of representation; those representations are maintained throughout the word, and updates are made based on subsequent inputs; phonetic encoding re-configures across space as the word unfolds in order to encode content and order; this dynamic phonetic sequence is used to recognise stored morphological constituents from the input. Overall, the work showcased in this overview demonstrate the utility of combining theoretical linguistics, machine-learning and neuroscience for developing a theoretically grounded, biologically constrained and computationally explicit account of how the human brain achieves the feat of language comprehension.

Keywords: decoding, encoding, language, speech, brain, neural architecture

1 Introduction

Listening to someone talk typically feels like effortless and automatic understanding. With little conscious effort in commanding our mind to comply, an interlocutor can provide information, evoke emotion, or exchange social pleasantries. Despite the ease with which hearing humans verbally communicate, the process of transforming human-articulated sounds into an infinite possibility of novel and complex meanings – i.e., speech comprehension – is a computationally intricate, and currently unsolved, challenge.

What computational solution does the brain implement to overcome this challenge? This is the big question that guides my research. In this article, I will provide an overview of studies which contribute to delineating the processing architecture upholding speech comprehension in terms of what **representations** the brain generates from the auditory signal, and what **computations** are applied to those representations during the timecourse of processing. Describing these components of the human processing architecture is key to understanding auditory, speech, and language processing, which, I believe, require complementary insight from linguistics, machine learning and neuroscience in order to be successful.

What neural measurements are best suited to answering these questions? Currently, the majority of neurolinguistic studies have utilised technologies that record activity from large populations of neurons, either non-invasively using techniques such as functional magnetic resonance imaging (fMRI), magneto-encephalography (MEG), and electro-encephalography (EEG), or invasively using electro-corticography (ECoG). Such methods pool activity across tens to hundreds-of-thousands of neurons, both within and across putative cortical columns. This means that the recorded neural activity is the summation of multiple, functionally distinct, overlapping neural responses. One major challenge is being able to appropriately decompose the dynamic multiplex neural signals obtained from these neural recording technologies into their functional axes, and in turn, associate those axes with an interpretable sequence of operations (King et al., 2020). My work develops and utilises machine learning algorithms to achieve this analytical goal.

The processing of speech signals poses an additional challenge: the speech signal evolves rapidly over time and conveys information across various overlapping timescales. In addition, neural responses evolve rapidly over time (Gwilliams et al., 2022a), both in terms of amplitude at a single location and propagation of activity across the brain. Time-resolved neural recording techniques such as MEG, EEG and ECoG are well suited to capturing this dynamic signal, due to the high sample rates with which they can record brain responses, and their extensive spatial coverage - the entire brain in the case of non-invasive methods, and multiple square-centimetres of cortex in the case of invasive methods. Thus, the rich spatio-temporal dimensionality of these data provide the opportunity to disentangle and track the multiple overlapping neural processes that unfold in parallel during speech comprehension, and in turn link those processes with associated computations and representations. In this brief review, we will

focus on studies that employ such time-resolved methodologies.

Figure 1 offers a schematic model of some of the computations the brain implements (coloured boxes), and representations it generates (bold text), in order to achieve speech comprehension. I focus on neural operations that reside close to the input, serving as the lower level processes which transform the auditory signal into linguistically interpretable sequences. I organise this article into sections, whereby each section corresponds to a sub-component of this schematic processing model. I will provide evidence in favour of each stage and conclude by offering promising next steps to further specify and broaden this model.

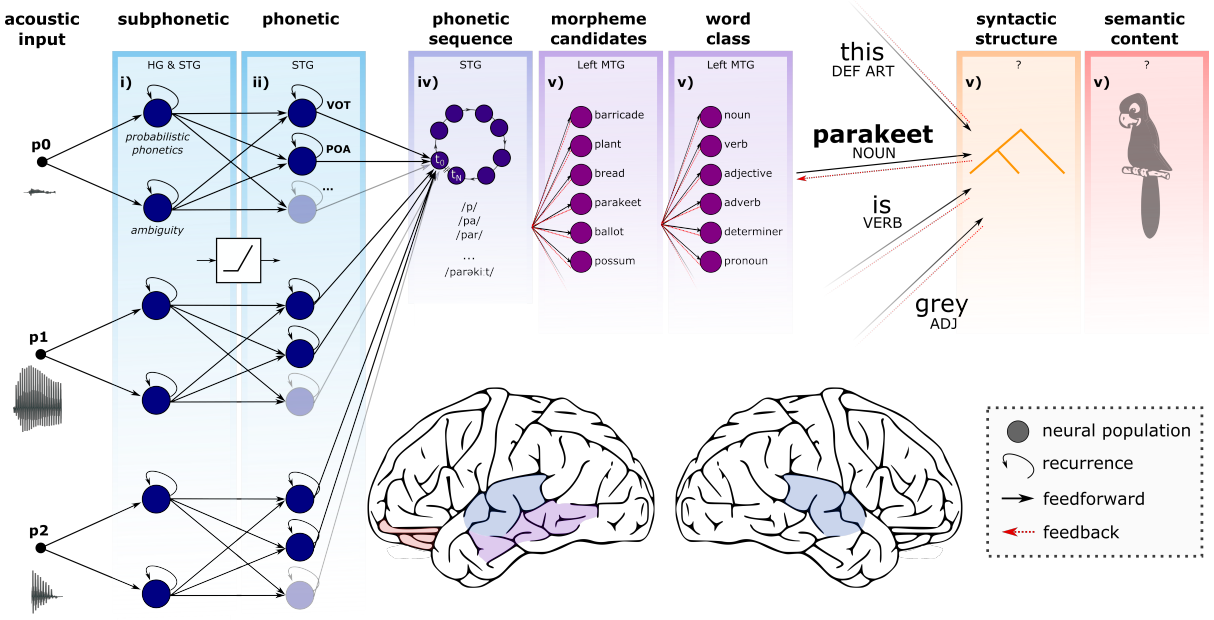

Figure 1: **Schematic processing architecture.** Processing stages are labelled in terms of the representational format they generate (bold font, above) and their putative location of processing (HG = Heschl's Gyrus; STG = Superior temporal gyrus; MTG = Middle temporal gyrus). The example shows the processing of the first three phonemes of the word 'parakeet'.

2 From continuous to discrete

The speech signal, as it reaches auditory cortex from the auditory pathway, is continuous and continuously varying (Ades and Brookhart, 1950). Yet, a listener's perception of a speech utterance is categorical and discrete; organised in terms of phonemes, syllables and words (Marslen-Wilson and Warren, 1994) rather than the spectro-temporal modulations that may be not just orthogonal to, but also a perilous distraction from (Maye et al., 2008, Tsao et al., 2004), the intended meaning of the speaker. One critical breakthrough for understanding speech comprehension processes has been in identifying the discrete format by which speech sounds are represented by neural populations.

What are the candidates for this discretised representation? Looking to linguistic theory and accompanying psycho-linguistic research provides us with some candidate hypotheses, most pertinently phoneme categories and phonetic features (Ali et al., 1999, Cole, 1973, Marslen-Wilson and Warren, 1994).

Phonemes are the minimal speech unit which, if replaced with a different phoneme, transform the word into a different lexical item (e.g., /bat/ versus /pat/ versus /mat/) (Chomsky and Halle, 1968). If phoneme categories were the underlying representation of speech sounds, this would entail a 'one-hot' encoding of each sound, whereby neural populations are tuned to a single phoneme category, firing in response to that phoneme category and otherwise remaining silent.

Phonetic features correspond to properties shared by groups of phoneme categories, which relate to how the speech sound is articulated by the speaker (Liberman and Mattingly, 1985). This spans different dimensions of articulation, such as how the tongue, teeth and lips are positioned in the mouth to generate the sound (place of articulation), how air is allowed to pass through the mouth as the sound is made (manner of articulation) and whether the vocal chords vibrate during the initial articulation of the sound, or not (voicing/phonation). If phonetic features were the underlying neural representation of speech sounds, neural populations would respond to the collection of phonemes which share a phonetic feature – e.g., all of the fricative manner of articulation phonemes such as /s/, /sh/, /f/. In this sense, the phoneme category would be discernible only from the responses across multiple neural populations, and would not be the 'primitive' representation of neural encoding per se.

Here I would like to make clear what I mean by a *speech unit* versus a *speech format*. When I refer to speech units, I refer to the size of the speech sound chunk under scrutiny. For example, phonemes and syllables correspond to different speech units. When I refer to speech format, I refer to the way by which the different realisations of speech units are organised – in the stimulus and, by hypothesis, also in the brain. For example, phoneme categories and phonetic features correspond to different representational formats of the same speech unit.

In a flagship study by Mesgarani and colleagues (Mesgarani et al., 2014), high-density ECoG was used to record auditory responses from participants listening to isolated sentences. The goal was to test whether neural population responses to speech sounds are organised relative to phoneme categories or phonetic features. They analysed responses to each speech sound, for each electrode separately, along the superior temporal gyrus (STG). The results demonstrated that, at a latency of 100-150 ms after speech sound onset, neural populations as recorded from an individual electrode site responded to the groups of speech sounds that share phonetic features, rather than individual phoneme categories.

A similar finding using EEG was published in 2015 from Di Liberto and colleagues (Di Liberto et al., 2015). They compared different representational spaces, including spectrogram, phoneme model and phonetic feature models, and found that the phonetic feature model best explained the non-invasive

scalp recordings. Model weights similarly peaked around 100-150 ms after sound onset, suggesting that the neural sources may also originate from STG.

These two results support that relatively early cortical responses encode the phonetic content of the incoming speech signal, and that binary phonetic features better explain the organisation of these responses than the continuous phoneme categories. The claim, therefore, is that the brain quickly discretises the continuous speech signal into discrete phonetic features for subsequent analysis. This prior body of work forms the empirical bases for the computational stages I explore in the upcoming sections, and is the motivation for choosing phonetic representational bases in Figure 1.

3 Discretisation in the face of ambiguity

In real world speech, the sensory signal is often ambiguous as to the phonetic features it contains. This may be due to the particular way in which the speaker articulated the sound, or because of external sensory signal masking the target speech input. How does the brain decide upon a categorical discretisation when the input itself is consistent with more than one category?

In Gwilliams et al. (2018) (Experiment 1), we investigated how the early 100-150 ms STG response described above encodes speech sounds that vary on a continuum from one phoneme category to another, by varying a single phonetic feature (e.g., /ba/ $\longleftrightarrow$ /pa/). Will responses track the sub-categorical variation present in the speech stimulus, or will it reflect just the most likely category that aligns with the listener's categorical percept?

Participants listened to syllables that we morphed along an 11-step linear continuum from one phoneme category to another, and performed a two-alternative forced choice decision on each sound they heard, between the two most likely competitors (e.g. /ba/ or /pa/ ?). While participants performed the task, we recorded whole-head MEG, and performed source localisation to allow us to apply our analyses within particular brain regions. Our results showed that neural sources in STG exhibited a summation of both linear sub-categorical responses, in addition to categorical responses that predicted their behavioural report. Both encoding formats of the phonetic input were present at a latency of 100-150 ms after speech sound onset. This result suggests that the brain derives both a linear veridical and a categorical discretisation of the input in parallel, permitting the system access to both formats upon which to derive subsequent perceptual decisions. This aspect of processing is presented in the blue panels of Figure 1, showing that the signal of each phoneme is transformed into a subphonetic and phonetic format in the auditory cortex. This demonstrates that, while the representations extracted from the sensory signals are discrete in time, they are not solely representationally discrete. They are contained both as a probabilistic sub-phonetic (linear) and a rectified-probabilistic (categorical) code.

Is multiplex extraction of properties from ambiguous sensory inputs specific to speech processing,

or is it also found in other domains and modalities? To further our investigation of the processing of ambiguous inputs, in (Gwilliams and King, 2020) we presented participants with visual symbols that were generated along a continuum between letters and digits. The symbols were designed to contain a number of orthogonal properties. Of particular relevance here is the manipulation of ambiguity, perceptual evidence in favour of a letter or digit category, and eventual categorical behavioural report. When fitting decoding models on the neural responses recorded with MEG, we found that all three properties were simultaneously encoded in neural responses, and maintained in parallel until the behavioural response.

This result puts forward the interesting postulation that, across modalities and cognitive domains, the sensory system simultaneously extracts a rich set of properties, in different formats, from the input. Why generate highly redundant representations of the same information? Because certain formats of representation emphasise particular components of the input and de-emphasise others (Marr, 1980), making them better suited to performing different tasks. For example, a linear representation emphasises the relative similarity between different inputs; a representation of ambiguity emphasises distance from canonical examples at categorical endpoints. For the task of categorising speech stimuli, a useful representation is one that emphasises inputs that are most similar to categorical endpoints, while de-emphasising variation within the category. In all, having access to multiple representations allows the brain to flexibly use the formats which are best suited to a particular neural process or behavioural outcome, without having to further transform representations each time it is faced with a different task.

4 Revising discrete percepts

Of course, in the studies above, we are recording neural activity from our participants while they have a very artificial experience: hearing isolated syllables and providing a perceptual report via button press (Hamilton and Huth, 2020). In natural speech, listeners are provided not only with ambiguous inputs, but also with a rich and informative context that can be used to guide processing of the sensory signal towards the correct interpretation (Davis and Johnsrude, 2007, Sohoglu et al., 2012). Furthermore, a listener's perceptual reports are typically an implicit component of understanding the speech input, whereas here we require an explicit play-by-play of what they perceive at each given moment. In these next set of studies we wanted to examine speech processing in a more naturalistic way, in order to understand how surrounding context serves to aid the processing of ambiguous or acoustically compromised speech sounds.

When an ambiguous sound is placed in a word at the end of a biasing sentence fragment such as 'The state governors met with their respective ...', upon hearing the word 'legi#latures', where the # indicates the replacement of a speech sound with white noise, listeners perceive to hear the missing sound, and temporally mis-locate when the noise burst occurred. This was the observation of Warren (1970), which

he referred to as the 'perceptual restoration effect'. This impressive example of top-down influence was later replicated and extended by others (Samuel, 1981), showing it to be a robust perceptual effect.

The neural underpinnings of this process were investigated with an elegant study by Leonard et al. (2016). They presented participants with perceptually ambiguous words such as 'fa#tr' where one of the phonemes were replaced with a noise burst. They asked participants whether they perceived hearing the word 'factor' or 'faster'. By using the high gamma neural responses within STG, as recorded with ECoG, the authors found that the single trial neural responses to an identical sensory input was significantly different depending on whether the individual perceived a 's' or 'k'. The result persisted both when presenting the words in isolation, and when embedding the word at the end of a biasing carrier sentence. Furthermore, the pattern of neural responses was similar to the sensory response evoked by hearing the original unmasked versions of the lexical stimuli. This result suggests that, when a listener perceptually restores an absent speech sound, neural ensembles emulate the pattern of activity that is typically associated with the sound to be 'filled-in'. From the brain's perspective, then, the pattern of activity across STG resembles that of an unabridged sensory sequence. The listener's experience appears to be based upon the sequential neural patterns in STG, gaps in which have been sufficiently repaired during processing to yield a complete percept.

An important extension of this prior work is to consider the fact that relevant information about speech sound identity not only precedes a sound under scrutiny, but can also succeed it. For example, imagine a sentence such as 'When the #ent was repaired in the truck', versus 'When the #ent was repaired in the park' where now the # corresponds to a sound that is ambiguous between /t/ and /d/. The first sentence biases listeners to perceive the ambiguous sound as /d/, in order to create the semantically congruent word 'dent', whereas the latter biases perception towards /t/ to create the word 'tent' (Connine et al., 1991). Connine and colleague conducted a series of behavioural studies using sentences such as the examples provided above, in order to show that perception can be based retroactively – based on information which comes *after* the sensory signal to be interpreted.

Remarkably, it has been found that perceptual reports of an ambiguous sound are affected by information that arrives up to several seconds after the sound of interest (Christiansen and Chater, 2016, Levy et al., 2009, McMurray et al., 2009), which suggests that the percept of a listener remains malleable for an extended period of time. In Gwilliams et al. (2018) (Experiment 2), we investigated the neural computations which make this retroactive update possible.

Twenty five participants were recruited to listen to phonetically manipulated words while whole-head MEG was recorded. The logic underlying our experimental design is that, if a sound is ambiguous between a /b/ and /p/, but then the word resolves as /?arakeet/, the only interpretation of the onset phoneme that would result in an existent word of English is /p/ (Figure 2).

We selected 53 word pairs where, other than the first phoneme, the phoneme sequence was identical

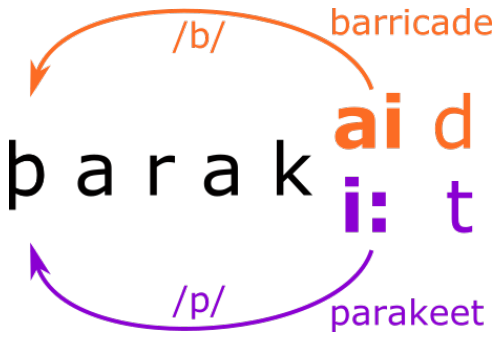

Figure 2: **Stimuli examples.** An example word pair stimulus from (Gwilliams et al., 2018). The two words 'barricade' and 'parakeet' contain an identical phoneme sequence until point of disambiguation ('ai' and 'i:'), and which point, not only is the identity of the word revealed, but also the identity of the ambiguous onset phoneme.

until a disambiguation point, which occurred between 3-7 phonemes after word onset across word pairs. At this critical point, entropy over possible continuations of the word resolved to zero. In other words, upon hearing the 'ee' in '#arakeet', the only word consistent with that sequence of phonemes is 'parakeet'. Simultaneous to revealing word identity at disambiguation point, the identity of the onset phoneme is also disambiguated: given a phoneme sequence '#arakeet', the only phoneme that could be placed at onset in order to create a valid word of English is 'p'. So, at this disambiguation point, not only is the identity of the word revealed, but the identity of the onset phoneme /p/, is also retroactively disambiguated.

We morphed the onset of the word pairs along a 5-step perceptual continuum, whereby in absence of biasing context, participants' average categorisation of the sounds was at 95%, 75%, 50%, 25% and 5%, in a two-class decision paradigm. We then modelled the neural responses in STG to each phoneme in these words as a function of (i) most likely phoneme category given the acoustic input, (ii) the probabilistic phonetic detail, and (iii) the extent of sensory ambiguity. We found that all three types of representation of the onset speech sound were encoded at the onset of each subsequent sound in the word. Concretely, we could read out the phonetic content and the extent of ambiguity experienced at word onset throughout the duration of the word. This information was encoded in neural responses up to at least 750 ms after sound onset (the longest delay that we tested), and remained locally encoded with the same set of neural sources in STG throughout the duration of the word.

This result suggests that the brain maintains the phonetic detail of previously heard speech sounds over time, and updates those phonetic representations based upon the subsequent speech signals it receives. In Gwilliams et al. (2017) (Gwilliams et al., 2017), we explicitly tested this hypothesis by modelling neural responses in the same dataset as a function of (i) probabilistic phonetic detail of the onset phoneme, (ii) probabilistic phonetic categorisation of the onset phoneme given the subsequent sounds of the sequence. We used a Bayesian modelling approach, which allowed us to test the relative contribution of each feature throughout the sequence. Our results demonstrated that neural responses in STG at a 100-150 ms latency are best predicted using a weighted combination of sensory evidence and

lexical evidence in favour of a phonetic categorisation, which suggests that the brain permits retroactive modification of phonetic percepts based upon subsequent inputs.

How does the brain decides upon a categorical discretisation when the input is consistent with multiple categories? We find that the brain both selects the most likely categorisation given the sensory evidence, and maintains a sub-categorical representation of the input for a very long period of time, which can be influenced by the subsequent content that is received. In this way, the percept of the listener is able to benefit from preceding and subsequent information, at both the sensory and lexical levels, to derive comprehension. To visualise these aspects of processing, we added recurrence arrows to the sub-phonetic and phonetic properties of the speech sounds in Figure 1. This indicates that the properties of speech sounds are maintained for a long period of time, within the same neural sources. We also include red feedback arrows from the higher order structures (morphemes, word class), to show that higher order information can serve to influence the maintenance of phonetic information. These feedback connections allows the system to resolve sensory ambiguity (Gwilliams et al., 2018), and even fill in sensory information when the sound is absent from the speech signal entirely (Leonard et al., 2016).

5 Overlapping computations

While the neural longevity of speech sound representations has clear advantages to the processing system in terms of accuracy of eventual percepts, it also comes with the consequence that speech sound information is processed for much longer than the duration of the speech sound unit itself. Concretely, the average duration of a phoneme in continuous speech is around 80 ms; however, our work suggests that the properties of that sound are processed for over half a second. This means that, inevitably, the brain is processing the content of multiple phonemes at the same time.

A similar observation was made in a previous EEG study, which recorded neural responses to continuous speech (Khalighinejad et al., 2017). In their study, phonetic detail could be discerned from the speech signal for around 100 ms, whereas it could be read out from neural responses for around 400 ms. This supports our notion that in continuous speech, the properties of multiple speech sounds are processed simultaneously.

This intriguing observation raises two important questions. First, how does the brain process properties of multiple speech sounds at the same time, without confusing the identity of those speech sounds? For example, if three unique phonemes have been heard, each with their own set of phonetic features, how are the features assigned to the correct phoneme position, without substitutions, which would generate incorrect phoneme sequences. Second, how does the brain recall the order of each speech sound in the sequence? In order to correctly recognise what words a person is saying, the processing system needs to accurately represent both the content and the order of the sounds – the content is what distinguishes

/pat/ from /bat/ and the order is what distinguishes /pat/ from /tap/.

In order to address these questions, we recorded whole-head MEG while participants listened to naturalistic continuous speech (Gwilliams et al., 2022a). We annotated each phoneme (around 40,000 per participant) for its phonetic content and location in the word. We then trained a series of decoders to discriminate between different phonetic features, as a function phoneme position, and as a function of the amount of time elapsed since phoneme onset. The critical question was whether a successful decoder trained to discriminate features of the target phoneme (features_{target}) on activity during the target phoneme (brain_{target}) contained information in the same or different activity pattern from a decoder trained on activity at each subsequent millisecond over time ($\text{brain}_{target+N}$). In other words, for the duration of phonetic feature processing (e.g., 400 ms), does the brain represent that phonetic detail under the same neural ensemble over time, or do the neural ensembles evolve over the course of processing?

Our results show that, with each subsequent phoneme that enters into the ear, the neural pattern which encodes the phonetic information of a target phoneme assumes a different spatial configuration. For example, phoneme_x is initially encoded within pattern_A, which then evolves to pattern_B, then to pattern_C, as each of the subsequent phonemes are received. We found that this movement of information over space occurred locally within STG, rather than a global transposition from auditory to frontal areas. We also found that the movement of activity is systematic, such that the amount of time elapsed since the onset of the speech sound can be reconstructed from the MEG topography at a given time step.

In all, these results suggest that, indeed, the brain processes the properties of speech sounds for an extended period of time – for much longer than the duration of the speech sound itself. And, indeed, this results in the processing of multiple speech sounds at the same time. Here we uncover that parallel processing of speech sounds is made possible by moving the information across space over time, such that any given neural ensemble is not tasked with processing the phonetic content of multiple sounds simultaneously, which avoids errors in read-out of the phoneme identity of speech sounds. In addition, the systematicity of this passage of information permits access to the amount of time elapsed since sound onset, and therefore the relative order with which the sounds occurred in the sequence. Together, this permits the brain with a running N-gram representation of the most recently heard speech sounds, with which the lexical identity of the speech sounds can be uncovered.

These findings are represented in Figure 1 as the 'phonetic sequence' processor, which receives information about each incoming phoneme and passes it through a set path of neural populations over time.

6 Using speech sounds to access stored units

The series of studies that I have over-viewed provide a candidate set of neural computations the brain applies in order to discretise the continuous input into phonetic properties, resolve ambiguity in that discretisation, and faithfully represent the recent history of the phonetic sequence. Of course, there is a higher goal to all of these processes, which is to use this phonetic sequence representation to recognise and understand what the person is saying. In this final empirical section, I turn to research which has investigated how the speech input is used to recognise higher order structures, such as morphemes and words.

A morpheme is considered the smallest unit of standalone meaning, be it functional or semantic in nature. For instance, the item 'disappears' contains three morphemes: the prefix 'dis-', which has a negating function; the root 'appear', which contains the core meaning of the word, and the inflectional suffix '-s' which provides grammatical functional information about the verb (Gwilliams, 2020). The morphological unit is well characterised and understood in linguistic theory (Marantz, 1997, 2013), and has a great deal of precedence in terms of its psychological reality (Schreuder and Baayen, 1995, Taft, 1994).

As is true for much of language research, the majority of studies investigating which of these representations are encoded in neural responses have primarily investigated the case of visual processing during reading (Friederici, 2011). In such reading studies, results support that the brain uses systematic cues in orthographic structure, such as N-gram frequency and transition probabilities, in order to parse the input into morphological constituents (Longtin and Meunier, 2005, Rastle and Davis, 2008, Taft, 2004). Those constituents are then re-combined under morpho-syntactic rules to generate the meaning of the lexical item (Fruchter and Marantz, 2015). In this way, the primary unit of recognition is the morpheme, and the lexical item is the output of online combinatorics.

For speech processing, however, the information is not presented instantaneously to the sensory system. Rather, inputs are received gradually over time, as the signal unfolds (Gwilliams, 2020). Because of this, the brain cannot use the same processing strategy to parse written input as it can auditory inputs because, of course, instantaneous co-occurrence cues do not exist. This also poses analytical challenges, because it is unknown beforehand *when* a neural process of interest will occur *relative to what* unit or property in the speech signal.

One method to investigate how auditory sensory inputs relate to higher order linguistic structure is to model neural responses under an information theoretic framework (Gwilliams and Davis, 2022). Under this approach, the responses to each individual phoneme in speech can be modelled as a function of the information it provides about the resulting higher order unit, such as morpheme identity or word identity. The two metrics that are of particular relevance are surprisal and entropy.

Phoneme surprisal corresponds to how likely a particular speech sound is to occur, conditional upon a

set of preceding speech sounds. Surprisal is measured in bits, and it can be interpreted as how informative a particular sound is – an unpredictable sound carries more information than a highly predictable sound (Lesne, 2014). Lexical entropy corresponds to the extent of statistical certainty about the lexical outcome of a phoneme sequence. If only one word is possible given the sequence of phonemes thus far (e.g., 'parakee-'), certainty about the lexical outcome is high, and so entropy over possible outcomes is low. Whereas if multiple words are possible given a phoneme sequence (e.g., 'para-'), the lexical outcome is less certain, and so entropy is higher. This 'entropy zero' moment has also been referred to as the uniqueness point of the word (Gaskell and Marslen-Wilson, 1997, Luce, 1986).

A number of studies have shown that neural populations in STG at a latency of around 100-150 ms are sensitive to the surprisal and entropy of incoming speech sounds (Brodbeck et al., 2018, Gagnepain et al., 2012, Gwilliams and Davis, 2022, Gwilliams et al., 2017). The interpretation of this finding is that, as the speech sound is received, the brain is computing which sounds are likely to come next, in order to recognise what words are being said. If an unexpected sound occurs, this leads to a neural prediction error response (Sohoglu and Davis, 2016).

To investigate what higher order units the brain recognises from the speech input, we used Arabic as a test case (Gwilliams and Marantz, 2015). Arabic has an interesting morphological structure, because the morphemes combine in an interleaved fashion, rather than the concatenative structure found in languages such as English. For example, the Arabic word *kataba* comprises the root morpheme [k-t-b], which contains semantic information, and the pattern [-a-a-a], which contains morpho-syntactic information (see Figure 3). We tested whether, for native speakers of Arabic, their neural predictions about upcoming sounds is based on the preceding speech sounds of the morphological constituent (e.g., [k-t-b]), or the preceding speech sounds of all sounds in the word (e.g., [kataba]). The logic is that, if the morphological constituent is the target representation of the speech processing system, predictions of upcoming sounds should be based on the phonemes present in the morpheme, *despite* the fact that those phonemes do not occur in a linear uninterrupted sequence.

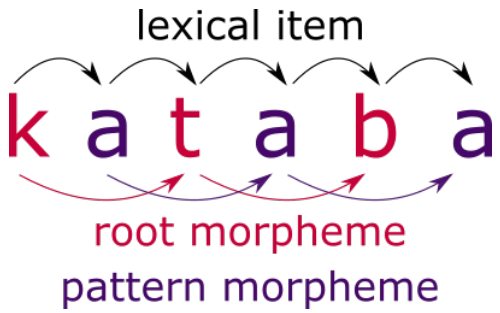

Figure 3: **Morphological structure and phoneme predictions of an example stimulus Arabic word.** Arabic words contain both a root morpheme (pink) and a pattern morpheme (purple). Here we show the order of phoneme sequence predictions under a morphological target (pink arrows), and a lexical target (black arrows).

We used MEG to record neural activity while the participants made lexical decisions on CVCVCV (C=consonant, V=vowel) Arabic nouns. We orthogonalised the predictability of the final consonant of the items based on the preceding consonants only (the root morpheme) and based on all preceding sounds (the whole word), and modelled responses as a function of these two types of predictability. Overall, our results show that the brain is sensitive to the likelihood of a speech sound under a morphological regime earlier and stronger than it is sensitive to the likelihood under a lexical regime. This suggests that the brain first uses the speech signal to recognise morphological constituents, and later recognises the full lexical item which those constituents create.

In order to represent these findings in Figure 1, we add the 'morphological candidates' column in purple. The idea is that, as the phonetic sequence develops over time, it is used to activate different possible morphological units. These units are the format by which words are stored in memory, in terms of their semantic and morpho-syntactic properties (Gwilliams, 2020). The transition between the phonetic sequence to morpheme candidates implicates a second transition: from representations generated in real time to those which have been stored through language learning. It is at this transition stage that we would expect effects of language proficiency, for example, to play a particularly crucial role.

7 Future Directions

Overall, the results of the studies outlined here provide insight into some of the computations the brain implements, and representations it generates, in order to transform the auditory signal of speech into an interpretable sequence of lexical items. A picture emerges whereby the inputs are rapidly discretised into multiple formats of representation; those representations are maintained throughout the word, and updates are made based on subsequent inputs; phonetic encoding re-configures across space as the word unfolds in order to encode content and order; this dynamic phonetic sequence is used to recognise stored morphological constituents from the input.

While this is an informative set of operations, there remains much to be delineated. Of particular priority is furthering our understanding of how higher order structure, such as grammatical relations and semantic content, is derived from the speech input, and in what format it is represented. In Figure 1, I have included a processing stage whereby word class (e.g., noun, verb, adjective) is derived from the morphological candidates in the feedforward pass of information, and from the syntactic structure in a feedback pass. I see word class as a potentially crucial level of representation, because it has the ability to act as a intermediary state to transform information between a sensory phonetic sequence and into an abstract syntactic structure. There is much precedence for the psychological reality of word class, both from imaging studies of healthy adults (Berlingeri et al., 2008, Moseley and Pulvermüller, 2014, Sahin et al., 2006) and from aphasia studies of individuals with acquired language disorders (Caramazza et al.,

1981, Hillis et al., 2004). It is also closely linked with the theorised representation of morphemes and associated morpho-syntactic structure (Gwilliams, 2020, Marantz, 1997).

In a recent study, we investigated the processing of word class in participants listening to stories while MEG was recorded (Gwilliams et al., 2022b). We used decoding analyses to investigate whether word class was generated from the syntactic structure (feedback) or from the phonetic sequence (feed-forward), when words occurred in contexts that biased interpretations in either direction. We found that, first of all, word class was decodable from neural responses for multiple seconds – spilling over into the processing of subsequent words. Second, in cases where the sentence structure and phonetic sequence provide contradictory predictions, the feedback sentence-level interpretation bore out. This finding suggests that, similar to how lexical information guides sensory processing through prolonged representational maintenance (Gwilliams et al., 2017, 2018), sentence structure also guides lexical processing. An important next step will be to assess whether the feedback-through-maintenance process is a common computational strategy that is recycled across multiple domains of language processing, perhaps also extending into other cognitive domains.

Accounting for higher level structures also implicates accounting for information that operates over speech units of a much larger size, and therefore over a much longer timescale. Whereas the representations of lower level units like phonemes can be derived with relatively few time samples from the input, properties of the sentence structure such as long-distance dependencies or WH-movement requires accumulating information over long timescales, and integrating that information non-adjacently (Newport and Aslin, 2004, Wilson et al., 2020). The current version of the schematic model I outline here only deals with a small aspect of this problem, in dealing with individual phonemes, full phoneme sequences and morphological units. Moving forward, any comprehensive model of speech comprehension should be able to readily deal with the dynamic structure of speech, both in terms of the temporal structure of the input and of the target structures that need to be generated from that input.

Along similar lines, the semantic meaning of utterances has been primarily modelled in terms of the meaning of individual words (Huth et al., 2016), rather than the contextual meaning of the entire utterance, or even the contextual meaning of the non-linguistic environment that a verbal exchange is taking place within. There are some efforts to leverage the impressive advances seen in artificial intelligence language models. For example, by building candidate representations of the meaning of utterances with larger context windows (Caucheteux et al., 2022), or by using the representations from a language model which has been jointly trained on linguistic and image data (Zhang et al., 2020, 2021). A crucial future direction will be to understand how complex meanings are created and represented in neural activity, not just from a single word, but of a full utterance as would be encountered in natural speech. This is, of course, one of those holy grail questions. The ability to combine words flexibly is one of the crucial abilities of human language, permitting the user to convey any infinite possibility of novel

meanings (Chomsky, 2009, Pylkkänen, 2019).

As a final point, I would like to mention some of the exciting recent advances in neural recording technologies. Historically, language neuroscience research entailed recording neurophysiological activity from large populations of neurons – on the scale of hundreds of thousands for MEG and EEG, and tens of thousands for ECoG and sEEG. Because language does not seem to be present in any animal other than the human, the application of animal models for better understanding language processing has been limited (Fitch and Tallal, 2003). This presents the challenge that the population-level recordings are ambiguous as to the true neural configuration that gives rise to the summation of activity that can be recorded from the skull or cortical surface. This has had sizeable consequences on the size and speed of research advances that have been possible, as compared to neuroscientific inquiry into cognitive capacities that are shared with other animals, such as lower level visual processing or decision making. However, neural recording techniques that were originally developed for use in non human mammals are now being adapted for safe and successful use in humans (Bullard et al., 2020, Chung et al., 2022, Paulk et al., 2022). I believe that having access to single neuron ensemble activity will represent a major paradigm shift in the type of questions that can be asked, and the precision with which we can answer them. And in turn, serve to constrain our interpretation mass population-level activity based on our understanding of the underlying single-cell generators.

It is an exciting time for speech neuroscience. Advances in neural recording techniques, analytical approaches and sophisticated language models will be invaluable tools in further delineating the neural architecture of speech processing. Drawing upon different areas of expertise fosters rich and innovative reasoning, expands and advances analytical approaches, and promotes parsimonious explanations that satisfy multiple angles of constraint. I believe that breakthroughs in language research will be borne from the intentional coalition of neuroscience, data science and linguistics, and I am excited to see what insights the next decade brings.

References

Harlow W Ades and John M Brookhart. The central auditory pathway. *Journal of Neurophysiology*, 13 (3):189–205, 1950.

AM Abdelatty Ali, Jan Van der Spiegel, Paul Mueller, Gavin Haentjens, and Jeffrey Berman. An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech. In *1999 IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 3, pages 118–121. IEEE, 1999.

Manuela Berlingeri, Davide Crepaldi, Rossella Roberti, Giuseppe Scialfa, Claudio Luzzatti, and Eraldo Paulesu. Nouns and verbs in the brain: Grammatical class and task specific effects as revealed by fmri. *Cognitive neuropsychology*, 25(4):528–558, 2008.

Christian Brodbeck, L Elliot Hong, and Jonathan Z Simon. Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, 28(24):3976–3983, 2018.

Autumn J Bullard, Brianna C Hutchison, Jiseon Lee, Cynthia A Chestek, and Parag G Patil. Estimating risk for future intracranial, fully implanted, modular neuroprosthetic systems: a systematic review of hardware complications in clinical deep brain stimulation and experimental human intracortical arrays. *Neuromodulation: Technology at the Neural Interface*, 23(4):411–426, 2020.

Alfonso Caramazza, Rita Sloan Berndt, Jerry J Koller, et al. Syntactic processing deficits in aphasia. *Cortex*, 17(3):333–347, 1981.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, 12(1):1–10, 2022.

Noam Chomsky. Syntactic structures. In *Syntactic Structures*. De Gruyter Mouton, 2009.

Noam Chomsky and Morris Halle. The sound pattern of english. 1968.

Morten H Christiansen and Nick Chater. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, 39, 2016.

Jason E Chung, Kristin K Sellers, Matthew K Leonard, Laura Gwilliams, Duo Xu, Maximilian E Dougherty, Viktor Kharazia, Sean L Metzger, Marleen Welkenhuysen, Barundeb Dutta, et al. High-density single-unit human cortical recordings using the neuropixels probe. *Neuron*, 2022.

Ronald A Cole. Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics*, 13(1):153–156, 1973.

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

Matthew H Davis and Ingrid S Johnsrude. Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing research*, 229(1-2):132–147, 2007.

Giovanni M Di Liberto, James A O'Sullivan, and Edmund C Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465, 2015.

R Holly Fitch and Paula Tallal. Neural mechanisms of language-based learning impairments: insights from human populations and animal models. *Behavioral and Cognitive Neuroscience Reviews*, 2(3): 155–178, 2003.

Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011.

Joseph Fruchter and Alec Marantz. Decomposition, lookup, and recombination: Meg evidence for the full decomposition model of complex visual word recognition. *Brain and language*, 143:81–96, 2015.

Pierre Gagnepain, Richard N Henson, and Matthew H Davis. Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, 22(7):615–621, 2012.

M Gareth Gaskell and William D Marslen-Wilson. Integrating form and meaning: A distributed model of speech perception. *Language and cognitive Processes*, 12(5-6):613–656, 1997.

Laura Gwilliams. How the brain composes morphemes into meaning. *Philosophical Transactions of the Royal Society B*, 375(1791):20190311, 2020.

Laura Gwilliams and Matthew H Davis. Extracting language content from speech sounds: the information theoretic approach. In *Speech Perception*, pages 113–139. Springer, 2022.

Laura Gwilliams and Jean-Remi King. Recurrent processes support a cascade of hierarchical decisions. *Elife*, 9:e56603, 2020.

Laura Gwilliams and Alec Marantz. 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.

Laura Gwilliams, David Poeppel, Alec Marantz, and Tal Linzen. Phonological (un) certainty weights lexical activation. *arXiv preprint arXiv:1711.06729*, 2017.

Laura Gwilliams, Tal Linzen, David Poeppel, and Alec Marantz. In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, 38(35):7585–7599, 2018.

Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature communications*, 13(1):1–14, 2022a.

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Remi King. Top-down information flow drives lexical access when listening to continuous speech. *bioRxiv*, 2022b.

Liberty S Hamilton and Alexander G Huth. The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, cognition and neuroscience*, 35(5):573–582, 2020.

Argye E Hillis, Sangjin Oh, and Lynda Ken. Deterioration of naming nouns versus verbs in primary progressive aphasia. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 55(2):268–275, 2004.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

Bahar Khalighinejad, Guilherme Cruzatto da Silva, and Nima Mesgarani. Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience*, 37(8):2176–2185, 2017.

Jean-Rémi King, Laura Gwilliams, Chris Holdgraf, Jona Sassenhagen, Alexandre Barachant, Denis Engemann, Eric Larson, and Alexandre Gramfort. Encoding and decoding framework to uncover the algorithms of cognition. *The Cognitive Neurosciences,*, 6:691–702, 2020.

Matthew K Leonard, Maxime O Baud, Matthias J Sjerps, and Edward F Chang. Perceptual restoration of masked speech in human cortex. *Nature communications*, 7(1):1–9, 2016.

Annick Lesne. Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Mathematical Structures in Computer Science*, 24 (3), 2014.

Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090, 2009.

Alvin M Liberman and Ignatius G Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985.

Catherine-Marie Longtin and Fanny Meunier. Morphological decomposition in early visual word processing. *Journal of Memory and Language*, 53(1):26–41, 2005.

Paul A Luce. A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, 39(3):155–158, 1986.

Alec Marantz. No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. *University of Pennsylvania working papers in linguistics*, 4(2):14, 1997.

Alec Marantz. No escape from morphemes in morphological processing. *Language and cognitive processes*, 28(7):905–916, 2013.

David Marr. Visual information processing: The structure and creation of visual representations. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):199–218, 1980.

William Marslen-Wilson and Paul Warren. Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological review*, 101(4):653, 1994.

Jessica Maye, Daniel J Weiss, and Richard N Aslin. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental science*, 11(1):122–134, 2008.

Bob McMurray, Michael K Tanenhaus, and Richard N Aslin. Within-category vot affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of memory and language*, 60(1):65–91, 2009.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

Rachel L Moseley and Friedemann Pulvermüller. Nouns, verbs, objects, actions, and abstractions: Local fmri activity indexes semantics, not lexical categories. *Brain and language*, 132:28–42, 2014.

Elissa L Newport and Richard N Aslin. Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive psychology*, 48(2):127–162, 2004.

Angelique C Paulk, Yoav Kfir, Arjun R Khanna, Martina L Mustroph, Eric M Trautmann, Dan J Soper, Sergey D Stavisky, Marleen Welkenhuysen, Barundeb Dutta, Krishna V Shenoy, et al. Large-scale neural recordings with single neuron resolution using neuropixels probes in human cortex. *Nature Neuroscience*, 25(2):252–263, 2022.

Liina Pylkkänen. The neural basis of combinatory syntax and semantics. *Science*, 366(6461):62–66, 2019.

Kathleen Rastle and Matthew H Davis. Morphological decomposition based on the analysis of orthography. *Language and cognitive processes*, 23(7-8):942–971, 2008.

Ned T Sahin, Steven Pinker, and Eric Halgren. Abstract grammatical processing of nouns and verbs in broca's area: evidence from fmri. *Cortex*, 42(4):540–562, 2006.

Arthur G Samuel. Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General*, 110(4):474, 1981.

Robert Schreuder and R Harald Baayen. Modeling morphological processing. *Morphological aspects of language processing*, 2:257–294, 1995.

Ediz Sohoglu and Matthew H Davis. Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, 113(12):E1747–E1756, 2016.

Ediz Sohoglu, Jonathan E Peelle, Robert P Carlyon, and Matthew H Davis. Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, 32(25):8443–8453, 2012.

Marcus Taft. Interactive-activation as a framework for understanding morphological processing. *Language and cognitive processes*, 9(3):271–294, 1994.

Marcus Taft. Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology Section A*, 57(4):745–765, 2004.

Feng-Ming Tsao, Huei-Mei Liu, and Patricia K Kuhl. Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child development*, 75(4):1067–1084, 2004.

Richard M Warren. Perceptual restoration of missing speech sounds. *Science*, 167(3917):392–393, 1970.

Benjamin Wilson, Michelle Spierings, Andrea Ravignani, Jutta L Mueller, Toben H Mintz, Frank Wijnen, Anne Van der Kant, Kenny Smith, and Arnaud Rey. Non-adjacent dependency learning in humans and other animals. *Topics in cognitive science*, 12(3):843–858, 2020.

Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, 11(1):1–13, 2020.

Yizhen Zhang, Minkyu Choi, Kuan Han, and Zhongming Liu. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. *Advances in Neural Information Processing Systems*, 34:18513–18526, 2021.