Extracting language content from speech sounds: An information theoretic approach

Laura Gwilliams, Matthew Davis

EXTRACTING LANGUAGE CONTENT FROM SPEECH SOUNDS: AN INFORMATION THEORETIC APPROACH

A PREPRINT

Laura Gwilliams
Department of Psychology
New York University
New York, NY 10003
leg5@nyu.edu

Matthew H. Davis
MRC Cognition and Brain Sciences Unit
Cambridge University
Cambridge, UK
matt.davis@mrc-cbu.cam.ac.uk

November 16, 2020

ABSTRACT

Speech comprehension involves recovering the speaker's intended meaning from the speech sounds that they produce. While the sensory-driven components of this process have been widely investigated, the impact of speech content (i.e. linguistic information) on sensory processing is much less understood. Here we summarise the growing body of research demonstrating that neural processing of speech sounds is influenced by morpheme- and word-level statistical properties of the information conveyed. We introduce and review evidence that information theoretic measures such as entropy and surprisal are apparent in neural responses. These findings help uncover fundamental organisational principles of the language system: what units are stored and how they are accessed. Modelling sensitivity to the information content of the speech signal helps explain the interface between (i) auditory processes operating on speech sounds and (ii) the words and meanings that those sounds convey.

***K*eywords** surprisal · entropy · prediction · brain · language processing · information content · lexical access · predictive coding

1 Introduction

Speech is a means of exchanging information. Through verbal communication, humans have the unique ability to convey a potentially limitless number of thoughts and ideas through their utterances [1], and infer the thoughts of others from what they say.

Speaker-listener interactions can be formally described as a *communication system* [2] (Figure 1). The role of the speaker is to conceive of a message and encode it in the auditory signals she produces. These signals are decomposed into elemental time-frequency representations by the cochlea [3, 4], before being passed to auditory cortex. The role of the listener's brain is to decode the intended message from the signals that reach cortex, whereby communication can be considered successful to the extent that the intended conceptual message of the speaker matches the reconstructed conceptual message of the listener.

Although listening to someone talk *feels* like an effortless passive process, speech comprehension involves overcoming some major computational challenges. Not least because the mapping from acoustics to meaning is largely arbitrary [5], different speakers have vastly different ways of pronouncing words depending on biological, regional and incidental factors [6] and external noise, such as the voices of surrounding talkers or non-linguistic noise sources, mask the signal [7], see also van Hedger & Johnsrude (this volume). 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 do not rival the accuracy, speed and robustness to speaker variability demonstrated by human listeners [8, 9].

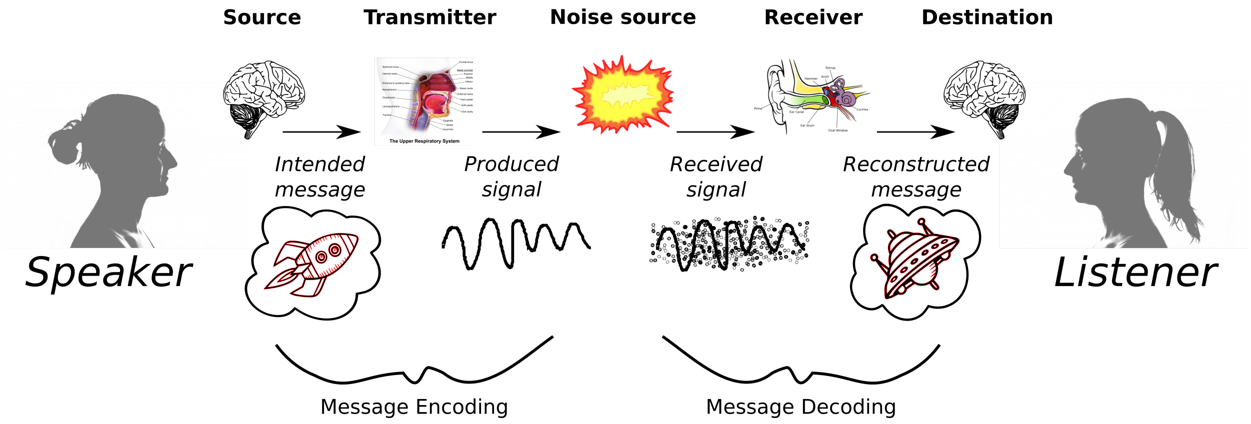

Figure 1: **Schematic diagram of the human communication system.** Verbal exchanges involve a speaker conceiving of a message and encoding that message into complex temporal-spectral patterns through their vocal articulators. As the signal travels through the air, it is contaminated with external sources of sound, such as other people speaking or noises from the environment. This contaminated acoustic signal is received by the auditory system of the listener and passed to auditory cortex for processing. The brain of the listener then needs to decode the original message from the auditory signals that were given as input. Here we see that the intended message that was encoded - a red space rocket - closely resembles but is not identical to the decoded message - a red alien spaceship.

The purpose of speech, in sum, is not to exchange auditory signals but to exchange information content. And within the structure of language, content takes the form of linguistic units such as morphemes, words and phrases (Figure 2). We refer to these information-bearing chunks as 'higher order representations'.

Current models posit that the brain transforms the auditory signal of speech into these higher order representations, which can then interface with stored representations in memory. This is achieved by generating increasingly complex and abstract representations of the acoustic input as neural activity propagates through the auditory pathways [10, 11, 12, 13, 14]. This representational hierarchy is naturally supported by the hierarchical organisation of auditory cortex: Regions of the auditory core (e.g. Heschl's gyrus) are driven by acoustically simple input features (e.g. frequency, amplitude), and surrounding cortical areas (e.g. superior temporal gyrus, left temporal lobe) are sensitive to more complex spectro-temporal features of the input [15, 16, 17, 11, 18, 19]. Regions further along the anterior and posterior inferior temporal lobe in turn contribute to lexical and semantic processing of speech [20, 21, 11, 22]. Generating a hierarchy of progressively more abstract acoustic and linguistic representations serves to convert the sensory input (i.e. the speech that the listener hears) into meaning (i.e. a reconstruction of the message intended by the speaker). A major goal of the brain during speech comprehension is therefore to access the correct chunks from memory, based on the hierarchy of representations that it generates from the auditory signal.

In this chapter we review evidence that neural processing of speech, even at early auditory processing stages, is fundamentally shaped by the goal to correctly and rapidly access higher order information (e.g. words). Before reviewing this evidence we will first consider the processes by which the sounds of speech are unitised into discrete representations (phonemes or similar). We will then introduce quantitative measures of the information content of speech signals; these measures - based on information theory [2] - presuppose that the brain, at some stage during processing, needs to access discrete higher order representations from memory. These units can be straightforwardly assigned probabilities and therefore information value, which can then be used as a proxy measure of processing higher order representations. As we will discuss, the use of information theoretic measures does not require commitment to a single and specific representation. Indeed, one of the strengths of this approach is that it can be equally applied to all levels in the linguistic hierarchy, covering all units between sounds and meanings. The main empirical data reviewed in the sections that follow concentrate on neural responses measured in peri-auditory regions of the superior temporal gyrus and linked to specific information-carrying elements (speech segments, morphemes, words, etc) in single words and in connected speech. We conclude by summarising the computational and neural mechanisms by which the higher level content of speech combines with acoustic signals during comprehension.

2 Discrete and binary representations of speech sounds connect to linguistic units

Articulatory gestures of the speaker, and therefore the acoustic signal she produces, are continuous. Both over time - any given sound can have variable duration - and in terms of content - spectral power can assume any continuous value.

The continuous nature of the speech signal is somewhat at odds with the discrete and binary nature of the higher order representations that need to be ultimately recognised. For example, the speaker is *either* saying 'pair' or 'bare' – she cannot be saying both at the same time.

A key challenge in the perception of speech sounds is therefore to convert the continuously varying acoustic input into discrete units that can be used to interface with higher order representations. While there is some debate as to the specific low-level speech units the brain uses, they seem to resemble phonemes or phonetic features [23]. In order to correctly distinguish between different words, the discrete identity of constituent speech sounds is critical. A spoken consonant such as [p] is defined relative to its manner of articulation (plosive), place of articulation (bilabial) and phonation (voiceless). Each of these distinctive features must be correctly recognised during speech identification - a different speech sound, and hence different words or meanings will be understood if these features are misidentified ('pair' becomes 'fair', 'care' or 'bare' with changes to the manner, place or voicing of the initial consonant). Any measurement of the information conveyed by speech sounds must ultimately operate on, and be calculated relative to, discrete representations that are the product of categorising speech segments. Otherwise, the likelihoods of different words and meanings will be conflated.

Yet, identifying the cortical signature of discrete processing of speech sounds has been a challenge for research on speech perception (see the chapter by Fox, Oganian and Chang for a review). Neural populations in superior temporal gyrus (STG) around 100 ms after speech onset are sensitive both to the veridical acoustic content of a sound as well as the discrete phonetic categories to which the sound corresponds [24, 25, 26, 27, 28, 29, 12, 30, 31, 32]. Further, representations of speaker-specific details, and other acoustic properties of speech co-exist with categorical representations of linguistic content in auditory areas [33, 19]. Other cortical regions - including motor cortex and frontal regions - are also shown to contribute to coding of the categorical identity of speech sounds [19, 34, 35, 36]. Further evidence suggests that higher-level, non-auditory representations act top-down to constrain and guide lower-level auditory processing of speech signals [37, 38, 39, 40], (see Ullas, Bonte, Formisano, & Vroomen in this volume for a review). Therefore, the categorical responses to speech shown in the STG likely reflect the outcome of a transformation from the continuous auditory signal into discrete phonemic units, as influenced and constrained by higher-level language processing.

In contrast to the established work on speech sound representations, there is less consensus on the representational units that contribute to higher-level processing of speech (e.g. morphemes, words and other meaning-carrying units above the level of the phonetic feature or phoneme). This discrepancy can be partly attributed to the fact that it is simpler to investigate features of the representational hierarchy that are closer to the sensory input than more abstract features that are closer to the meaning content, for at least two reasons.

First, whereas the acoustic sensory signal is easily measured and analysed, higher order representations only exist within the mind of the listener. One significant unresolved issue shared between cognitive neuroscience and engineering, therefore, is *feature discovery*: determining the units under which to represent information that is abstracted from the auditory signal. Modelling these higher order processes can only be as successful as the accuracy of the features being used to define those processes. While engineering approaches to deriving, for example, word meaning representations have been used to predict neural activity during speech comprehension [41, 42, 43] there is currently little evidence to favour specific computational approaches to meaning representation over another. Indeed, it has been argued that current engineering approaches to this problem are missing the key ingredients required for a sufficient representation of meaning required for true comprehension [44].

Second, studying higher order speech structure comes with analytical challenges. Assuming that the correct features have been identified, it is not always straightforward to relate relevant language features to a particular 'moment' in the speech input. For example, if we assume that part of speech (e.g. noun, verb, adjective) is a feature that the brain uses to process words, at what moment in hearing the noun 'hippopotamus' can we say that the brain is processing a noun? Does processing begin at the moment that the part of speech can be identified with 100% certainty, for example, when this word is uniquely specified after 'hippop-', or at word offset? Or are multiple part of speech hypotheses entertained simultaneously until syntactic class can be established beyond some threshold level of certainty? (See [45, 46, 47] for discussions related to this issue). The situation is further complicated by the fact that the timing of relevant neural processes - for different listeners, and for different utterances - will become increasingly variable at higher levels of the processing hierarchy [48]. Thus, even if the correct features have been identified, and even if the optimal latency relative to speech input could indeed be established, the time at which corresponding neural representations are activated will also vary. This significantly reduces the average signal strength associated with higher order processes, making them much more difficult to investigate.

Here we review a set of recent studies which have investigated the effect of higher order linguistic structure on processing of phoneme-by-phoneme information content in speech. The approach is to (i) model responses to phonemic units, which produce clear and well characterised responses in terms of timing and spatial location [24, 26, 29]; (ii) contrast segments that differ in the higher order speech structures they communicate e.g. whether or not specific speech sounds are predictable given the lexical or semantic context they occur within. The rationale is that by testing responses as a function of the *information* that a discrete speech sound provides about higher level representations (e.g. syllables, morphemes, words; Figure 2) it is possible to reverse engineer which higher order representations are relevant to processing, how they are recognised or accessed, and how they interact with other representational units. This approach allows speech research to progress from studying auditory signal processing to information processing in speech. Before we begin to review these studies, we will first provide a brief tutorial on the key quantitative measures of information content that have been employed in the study of speech comprehension.

3 Quantifying information content in speech

Recent studies investigating information processing have capitalised on two properties of language. First, while speakers *can* convey a range of information constrained by the vocabulary and grammar of the language, not all expressions are equally likely: Some phoneme sequences, words and meanings are much more probable than others. For example, it is more likely that I describe myself as 'happy' than 'exultant' or 'jocose'; it is more likely that after hearing /ma/ you will hear the phoneme sequence /t/ (to create the word 'mat') than /lɪs/ (to create the word 'malice'). This difference in likelihood is not directly reflected in the sensory signal - these likelihoods are reversed after hearing /pa/ given that 'palace' is more frequent than 'pat'. This knowledge comes from having an internal model of the language, including the statistical structure (what sound sequences, words or sentences are more or less likely) as well as linguistic regularity (what sound sequences, words, etc are permitted). Our view is that listeners employ both these forms of knowledge during comprehension. We will hence use the term 'statistical regularity' to describe these knowledge sources collectively.

In this chapter we will focus on probabilistic definitions of language knowledge since - based on a second established property of language - a wide range of data shows that listeners are exquisitely sensitive to the statistical structure of speech. Variability in the probability of specific linguistic units leads to differences in behavioural measures of speech comprehension such as response time or accuracy and in the magnitude of neural responses as measured both invasively (ECoG, s-EEG) and non-invasively (EEG, MEG, fMRI) during comprehension. This kind of sensitivity has been demonstrated across many levels of the linguistic hierarchy: at the level of phonemes, morphemes, words, syntactic structures and semantic content. It has also been demonstrated cross-linguistically, in languages with fundamentally different morphological and syntactic structures. These observations therefore suggest that sensitivity to statistical regularity is not only robust, but is common across languages and pervasive across linguistic structures. These observations, however, presuppose that statistical regularities can be directly quantified, which is the focus of this section.

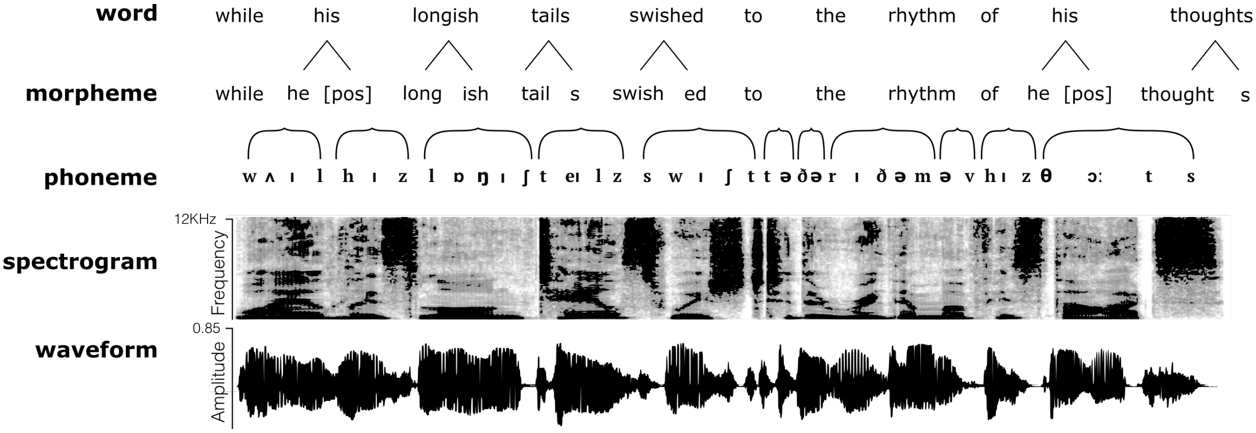
word
while his longish tails swished to the rhythm of his thoughts
morpheme
while he [pos] long ish tail s swish ed to the rhythm of he [pos] thought s
phoneme
w ʌ ɪ l h ɪ z l ɒ ŋ ɪ ʃ t eɪ l z s w ɪ ʃ t t ə ð ə r ɪ ð ə m ə v h ɪ z θ ɔː t s
spectrogram
12KHz
Frequency
waveform
0.85
Amplitude

Figure 2: **Speech hierarchy.** Language is a hierarchically structured stimulus. The acoustic signal can be discretised into a series of phonetic elements, which can be further grouped into morphemes, words, phrases, etc. Here we show an example speech segment, with the raw waveform, the derived spectrogram, and corresponding linguistic annotation. Note that while the acoustic representations are continuous, stepping into linguistic features entails discretisation of the signal. Also note that the linguistic units are hierarchically structured: that words are comprised of morphemes, which are comprised of phonemes. This hierarchical structure is at the core of what allows us to investigate processes at 'higher levels' as a function of neural responses to 'lower levels' - because all of the levels are mutually structurally dependent.

3.1 Introduction to information theory

We begin by detailing key measures of statistical regularity that can be quantified under information theory, focusing on the most prevalent in cognitive neuro-scientific literature. For more complete introductions to information theory and derived metrics see [49, 50].

These computations can equally be applied to units of different types, and to contexts of different sizes. Here we demonstrate how these variables would be computed for the word 'mat' in the sentence 'the cat sat on the mat', and the final phoneme 't' in the word 'mat'.

In the examples below, x refers to a particular outcome; X refers to all possible outcomes of the given unit; C refers to the events preceding x.

Conditional probability The probability of event x is *conditional* upon the preceding events. We compute this probability by dividing the frequency of event x occurring in context C by the frequency of context C alone. Below, we use the wildcard '*' to denote that all continuations contribute to the frequency count.

$$\begin{gathered} P(x) = P(x|C) = \frac{freq(C,x)}{freq(C)} \\ P(t) = P(t|ma) = \frac{freq(mat*)}{freq(ma*)} \\ P(mat) = P(mat|thecatsatonthe) = \frac{freq(thecatsatonthemat*)}{freq(thecatsatonthe*)} \end{gathered} \tag{1}$$

The conditional probability is highest when the two frequencies are similar in magnitude, i.e. when x almost always follows C. The probability is lowest when the frequency of C is much higher than the probability of C followed by x, i.e. when x very rarely follows C. Note that by definition the frequency of C is always equal to or greater than the frequency of C followed by x.

Surprisal The predictability of a particular outcome x is quantified using a measure termed *Surprisal* or *Shannon information content* [2]; strictly speaking this is the opposite of 'predictability', (i.e. surprisal is low for highly predictable sequences). This measure is given as the negative log of the conditional probability of x:

$$\begin{gathered} h(x) = log_2 \frac{1}{P(x)} = -log_2 P(x) \\ h(t) = log_2 \frac{1}{P(t|ma)} = -log_2 P(t|ma) \\ h(mat) = log_2 \frac{1}{P(mat|thecatsatonthe)} = -log_2 P(mat|thecatsatonthe) \end{gathered} \tag{2}$$

Surprisal is thus lowest (approaching zero) when the underlying conditional probability is highest, and highest (approaching infinity) when conditional probabilities are near zero. While such extreme values are rare in natural language, everyday examples are possible. For example, if I hear the syllable /trɒm/ surprisal at hearing the following syllable /bəʊn/ is near zero given that 'trombone' (and derived words like 'trombonist') are among the only English words that contain this syllable (see Figure 3B). If, another different syllable is heard - for instance, when one first learns of a type of mushroom called a 'trompette' the conditional probability for the second syllable given the first (i.e. /pɛt/ following /trɒm/) will be near zero, and surprisal will approach infinity. Surprisal is thus a measure of both the degree to which a particular event is predicted, and the information gained by experiencing that particular event; both are measured in bits. If an outcome is more probable, it is more predictable, and information gain is lower than if the outcome was less likely. If an event outcome has a probability of 1, predictions are entirely confident and no information is gained.

Entropy Uncertainty about the outcome of event x is referred to as *Entropy*. Mathematically, this is equivalent to the expected surprisal of an outcome: i.e. the average surprisal (information gain) of each predicted outcome weighted by the probability of that outcome. In the examples below, *Wo* refers to the entire cohort of possible upcoming words starting with /mae/ and *w* to one instance of the cohort (e.g. 'mat'). *Ph* refers to the entire cohort of possible upcoming phonemes (/t/, /p/, /k/ for 'mat', 'map', or 'mac') and *p* to one specific phoneme instance.

$$
\begin{aligned}
H(X) &= -\sum_{x\in X} P(x)log_2P(x) = \sum_{x\in X} P(x)h(x) \\
H(Ph) &= -\sum_{p\in Ph} P(p)log_2P(p) = \sum_{t\in Ph} P(p)h(p) \\
H(Wo) &= -\sum_{w\in Wo} P(w)log_2P(w) = \sum_{w\in Wo} P(w)h(w)
\end{aligned}
\tag{3}
$$

Entropy is also measured in bits. It will assume the highest value when there are a large number of possible outcomes X each of which have equal probability, and the lowest value when one outcome is much more probable than its competitors. For the example word 'trombone' in the previous section, and shown in Figure 3A, entropy reaches zero at the final /m/.

As show in Figure 3B, both phoneme surprisal and entropy tend to reduce for phonemes later on in spoken words. Although, these measures are quite strongly correlated, they are not unavoidably so - it is thus possible to distinguish between neural responses that correlate with one or other of these measures as we will observe in our review of relevant empirical data.

3.2 Linking information theory to language processing

As we saw in the examples, these information theoretic measures can be readily computed for linguistic units of different types (e.g. phonemes, words) in order to investigate processing at different levels of the linguistic hierarchy. Tailoring the measures for different purposes involves selecting the appropriate levels of representation, in selecting what units be used in place of x, C and X for probabilistic computations. Each selection embeds important theoretical assumptions for the quantification of the processing of that particular representation can therefore be used to make adjudications between different theoretical alternatives. For example, it is possible to measure whether modelling information gain using morphological units better explains neural data as compared to lexical units [51].

Such theoretical choices concerning the following:

x The linguistic event being modelled. This decision determines the units specified as the input. Above we show examples where x assumed a phonological unit p or a lexical unit w, but this could also be defined as any unit of interest - e.g. phonetic feature, syllable, morpheme. In the work outlined below, we will focus on research where the event of interest is a phonological unit.

C The relevant preceding events that influence the probability of linguistic event x. The above examples assume that the relevant context is the set of preceding phonemes in the current word in the case of $P(t|ma)$ and all preceding words of the sentence in the case of $P(mat|thecatsatonthe)$. Note, however, that x and C need not be in the same representational format; for example, it is possible to measure the probability of phoneme p given all preceding words w - indeed, converting between representational formats is necessary if prior words are to constrain processing of word-initial speech sounds. This assumes that the context C is relevant to the processing of event x.

X The alternative outcomes for which event x is informative. Note that x and X are necessarily in the same format because one is a single instance of the full set of the other. For example, phoneme p (e.g /t/) is one instance of the cohort of possible phonemes in cohort P (e.g /t/, /p/, /k/); word w (e.g. 'mat') is one instance of the cohort of possible words in cohort W (e.g. 'mat', 'map', 'mac').

4 Review of studies linking information theoretic measures to neural responses

Here we will review recent studies that have employed information theoretic measures in order to assess the influence of linguistic computations at the level of morphemes, words and phrases on neural activity during speech perception.

Early studies in this area contrasted fMRI responses to different types of spoken words, e.g. words compared with nonwords, or words with more vs fewer competitors, e.g. [52, 53, 54, 55] and provide evidence for additional activation of superior temporal and frontal regions for more difficult to identify words. There is some evidence of a dissociation of frontal and temporal regions: more unexpected speech (e.g. high surprisal segments in nonwords compared to real

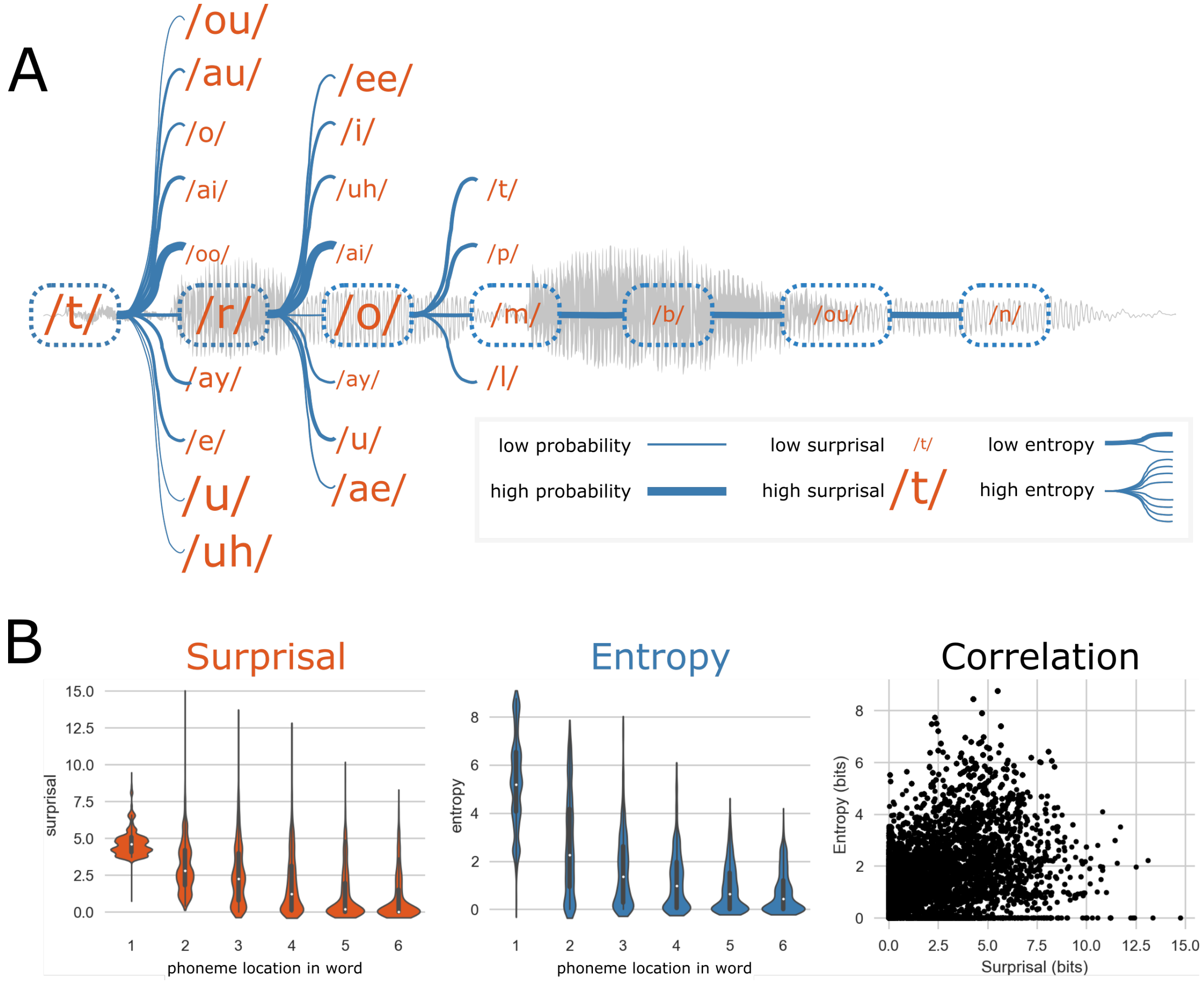

Figure 3: **Information theory metrics for phonemes in spoken words.** A: Waveform of the spoken word 'trombone' superimposed with all the possible phonological alternatives at each phoneme position. Each alternative phoneme is shown with a blue trajectory, for which the thickness of the arrow corresponds to the likelihood of that continuation. Each phoneme continuation is shown in orange, whereby the size of the phoneme also indicates surprisal. As shown in the legend, when there are a number of possible, equally likely, continuations, entropy is higher as compared to when there are fewer, more asymmetrically likely continuations. Note that at the phoneme /m/, the word becomes lexically unique and all of the subsequent phonemes in the sequence are 100% determined, such that entropy and surprisal are zero at /b/, /ou/ and /n/. B: Surprisal and entropy values were computed for around 20,000 phonemes of the audio-book stories used in [29]. Violin plots show quartiles and distributions of surprisal and entropy for different phoneme positions in the word. Correlation strength between the two metrics is r=0.57.

words) activate superior temporal regions [56, 57] whereas words with more competitors (e.g. words like *claim* with an onset-embedded word *clay*) lead to additional activity of inferior frontal regions [54, 55]; yet exceptions to this pattern are also reported [53]. However, these studies did not directly compare neural activity linked to specific statistical properties (such as surprisal or entropy that co-vary). Furthermore, fMRI lacks the time-resolution required to link activation to statistical properties or neural processes that respond to specific speech segments. More consistent evidence that the auditory system is sensitive to the statistical structure of speech input has therefore come from neural measures with higher-temporal resolution (e.g. MEG or EEG). Furthermore, as techniques for estimating language-based statistics from corpora and natural language have improved, so has the application of these computational measures to model neural responses. Here we summarise the main findings demonstrating that phoneme-level metrics of higher order linguistic structure can influence information processing during speech comprehension.

Gagnepain et al. [58] conducted one of the first studies that used neural data to differentiate information theoretic measures of lexical processing of speech. Neural responses in superior temporal gyrus (STG) were recorded using magneto-encephalography (MEG) in response to triples of familiar words (e.g. *formula*), learnt novel words (e.g. *formubo*) and untrained novel words (e.g. *formuty*). Neural responses were time-locked to speech before the divergence-point (i.e. formu-) and after the divergence-point (i.e. -la vs -bo vs. -ty). By comparing responses to item triples containing novel words that were learned and consolidated (and, hence, added to lexical knowledge, cf. [56]), or learned but not consolidated (and hence not lexicalised) the authors could assess the impact of changes to lexical knowledge on neural activity and adjudicate between two information theoretic measures that are ordinarily correlated (see Figure 3B). First, adding a new word to the lexicon (i.e. once the novel word *formubo* has been consolidated) would lead to an increase in lexical entropy, particularly before the divergence point. Yet, phoneme surprisal would decrease during the same pre-DP period due to stronger predictions for shared segments. Conversely, phoneme surprisal will increase in the post-DP window upon hearing phonemes that were not expected (i.e. a phoneme surprisal response will increase for the less expected continuation '-la' once 'formubo' had been consolidated). MEG responses in the pre-DP and post-DP periods changed in line with phoneme surprisal rather than lexical entropy; hence changes in lexical knowledge can modify phoneme-level responses in a way that is consistent with computation of lexically-generated prediction error (this is equivalent to 'phoneme surprisal', though the authors do not use this term). Specifically, responses in STG from 280-350 ms post-DP were increased for word neighbours of the consolidated novel word compared to neighbours of learnt but not consolidated items. Furthermore, consolidated novel words (but not learnt, not consolidated items) showed a reduced response in the same time period. There were no detectable effects of lexical entropy in the pre-DP window as predicted by the lexical inhibition account, indeed response reductions were observed in line with stronger segment predictions. These findings suggest that lexical knowledge (one element of the internal language model) generates predictions for upcoming speech segments that are compared with heard speech leading to STG responses that resemble prediction error. Gagnepain and colleagues further suggest that prediction error signals can be used to update lexical probabilities though they do not provide evidence to show these update mechanisms in operation (see [59] for discussion).

Building from these results, a collection of studies by Marantz and colleagues capitalised on sensitivity to segmental probability to understand the representation and processing of higher-order linguistic units in single spoken words. Specifically, they tested whether internal (morphological) word structure (e.g. a word like 'disappears' is composed of morphemes 'dis', 'appear', 's') influences segment prediction error or surprisal. Using MEG, a study conducted by Ettinger, Linzen and Marantz [60] revealed a main effect of phoneme surprisal in STG responses measured 200 ms after segment onset, which was significantly greater for bi-morphemic words ('bruis-er') as compared to phonologically matched mono-morphemic words ('bourbon'). There were also later effects of phoneme surprisal towards the end of the word (700+ ms after word onset). Furthermore, they found main effects of lexical cohort entropy from 335-377 ms after word onset. The authors conclude that the internal (morphological) structure of words serves to enhance segmental predictions at the phoneme level, and that predictions are delayed under conditions of high lexical entropy. This may suggest that segmental predictions are generated not just at the phoneme-unit level but also at the level of entire morphological units.

To further investigate the influence of morphological units on speech processing, languages with a non-concatenative morphological structure like Arabic and Hebrew are an ideal test case. Whereas in English morphemes are combined one after the other (e.g. dis-appear-s), in Arabic, they are interleaved within one another (e.g. the morphemes [k-t-b] and [a-a-a] are combined to form *kataba*). Thus, the linear order with which the auditory signal unfolds is at odds with the non-linear order that the relevant speech sounds of morphemes are received, allowing the two to be disassociated. Under this rationale, Gwilliams and Marantz [51] assessed neural effects of segmental prediction in order to determine whether spoken Arabic words are processed via their constituent morphemes (k-t-b, a-a-a) or as whole units (e.g. kataba) by opposing two measures of phoneme surprisal. They constructed stimuli that orthogonalised root-based 'morpheme' surprisal (probability of a consonant conditioned on the previous consonants in the root morpheme) and word-based 'linear' surprisal (probability of a consonant conditioned on all previous phonemes in the word). MEG was recorded while Arabic speakers performed a lexical decision task on spoken isolated words. They analysed responses to the final consonant of the words (e.g. kata**b**a) as a function of preceding morphological content and preceding whole-word content. Activity in STG was significantly modulated by morpheme surprisal from 100-250 ms. Word-based linear surprisal modulated later responses, from 250-300 ms in an overlapping set of sources. Thus, the results suggest that words are processed via morphological units before they are processed as wholes. This research showcases the use of phoneme-level responses to understand the representation and processing of higher-order linguistic units in single spoken words, by using language materials for which predictions might come from different linguistic units (e.g. words vs morphemes). By understanding what information is used to constrain predictions of upcoming information, it allows for inferences about what higher level information being accessed, and therefore what information is likely *stored* in lexical memory and deployed in speech perception.

Similar methods have been used in assessing the relationship between lexical and semantic processing of spoken words that refer to specific categories of concrete objects. For example, Kocagoncu and colleagues [61] show lexical uncertainty (quantified as lexical entropy based on participants' responses in a word-gating task) is represented in MEG patterns recorded from superior temporal and inferior frontal regions. These responses precede in time, and partially overlap with frontal and parietal responses that represent the degree of semantic competition (i.e. lexical uncertainty modulated by semantic dissimilarity) among the set of candidate words that are active at specific time-points during word identification. These findings suggest ongoing computations of semantic interpretations of spoken words throughout identification. Access to meaning is not delayed until lexical identification is complete.

Along similar lines, Gwilliams et al. (2017) [40] tested whether activation of lexical candidates is weighted by acoustic evidence in favour of one phoneme or another. They recorded MEG responses of subjects listening to words, where the onset phoneme was manipulated along a 5-step continuum from /b/-/p/, /t/-/d/ and /k/-/g/. The authors quantified two measures of surprisal and entropy: First 'acoustic weighted' metrics consider both the 'b-' and 'p-' cohort of words into the computation of surprisal and entropy, where each cohort is weighted both by word frequency *and* onset acoustic evidence. The second 'switch-based' metrics assume that the brain categorises phonemes before activating lexical candidates, and so in these surprisal and entropy metrics, *either* the 'b-' *or* 'p-' onset words will be included in the information theoretic measures. The authors found that when modelling surprisal and entropy from 200-250 ms after phoneme onset in STG, responses to early phoneme locations were better modelled by the 'acoustic weighted' account, whereas later phoneme locations were better modelled by the 'switch-based' account. The interpretation of these results is that earlier during processing, the brain uses both acoustic detail and lexical frequency equally to activate words, whereas later in processing, the brain favours categorical representations of the input in order to focus more heavily on lexical statistics. These results again showcase the ability to leverage information theoretic measures at the level of phoneme responses to adjudicate specific processing hypothesises at it pertains to higher order structures such as lexical items.

While these studies tested sensitivity to phoneme predictions within isolated words, in natural speech, expectations can also be generated based on previously heard words: Natural speech provides a *continuous* stream of linguistic information, in which preceding words can serve to constrain the probabilities of upcoming inputs. A study conducted by Gaston and Marantz [62] asked the critical question of whether, in minimal phrases, the brain uses preceding words to inform phoneme-level predictions. They tested whether phoneme surprisal and entropy responses could be conditioned *across* word boundaries, based on syntactic constraints provided by the preceding context. In terms of our tutorial above, this would mean contrasting the conditional probabilities that enter into the surprisal and entropy calculations to either include just the prior context within the word, or also prior context across multiple words. MEG was recorded while participants listened to minimal phrases, which were either grammatical (e.g. 'the clash persisted') or non-grammatical (e.g. '*the frown darkly') where the first word (the/to) made deterministic predictions about the part of speech of the subsequent word (noun/verb). The results show that both constrained and unconstrained surprisal metrics significantly accounted for neural responses in STG from around 200-400 ms after each phoneme in the noun/verb target, though no significant effects of entropy were observed. Thus, even when prior context has the *potential* to re-distribute probabilities on the level of 'boundary blind' phoneme-sequences, the brain remains sensitive to the context-free word-internal statistics *in parallel to* the context-sensitive statistics. This important observation suggests that lexical and sub-lexical units are activated based on both sources of information, perhaps aggregating over the predictions at a later stage. Similar findings arise from a study that explored the role of semantic constraints in guiding word identification from Klimovich-Gray et al [63]. For two-word phrases (e.g. 'yellow banana'), MEG response patterns in STG around 150ms after the start of the second word encode the change in entropy (i.e. surprisal) while lexical interpretation is guided both by prior context and by heard speech sounds. Partial correlation analyses confirm that these effects are independent of entropy and overall semantic similarity of word candidates.

A set of recent studies have also demonstrated the ability to use these same calculations of surprisal and entropy, which assume that the word is presented in isolation ('word-internal metric'), to investigate processing of words in natural, continuous speech such as spoken narratives. Brodbeck et al. [64] analysed responses to continuous speech as a function of word-internal information theoretic metrics. They found that phoneme surprisal modulated STG responses peaking around 115 ms after the onset of the relevant speech segments; responses correlated with cohort entropy followed soon after and peaked at around 125 ms. The relative timing of these effects are in line with Gwilliams and Marantz [51], but earlier than those seen in Gagnepain et al. [58] and Gaston and Marantz [62]. It might be that neural responses linked to specific speech segments arise at shorter latencies for words in connected speech than for words heard in isolation.

Another aspect of language context that has been shown to influence phoneme-level predictions are linear transition probabilities across ('blind' to) word boundaries. Donhauser and Baillet [65] modelled MEG responses to continuous speech, as a function of phoneme surprisal and phoneme uncertainty. These measures were computed based on a neural network (cf. [66, 67]) which was trained to predict upcoming segments in speech sequences, including predictions that cross boundaries between higher order units. Put otherwise, unlike the 'word-internal' metrics discussed so far, this metric takes prior lexical context into account. Specifically, their neural network model used a context of the preceding 35 phonemes (around 10 words) sufficient to encode several preceding words and their meaning and syntactic structure. They found that across-word phoneme surprisal modulated responses from around 80-160 ms and 230-420 ms after phoneme onset, and contextual entropy modulated responses 60-120 ms and 230 ms in primary and association auditory cortex. These results further support the notion that early auditory responses reflect sensitivity to higher order structure, and that surprisal and entropy are metrics that tap into distinct neural computations.

These studies (e.g. [64, 65]) tested the contribution of other regressors that are correlated with phoneme surprisal and entropy, and that may otherwise serve as potential confounds. Brodbeck et al [64] tested the contribution of cohort size (how many lexical items are possible given the sequence input) and cohort reduction (how many lexical items are being discarded given the new phoneme that was just heard). Donhauser and Baillet [65] also tested the role of cohort reduction. While Brodbeck et al found no additional contribution of either regressor above the explained variance of the existing analysis factors, Donhauser did find that cohort reduction explained additional variance above the level of phoneme surprisal. This might suggest that previous studies using phoneme surprisal may have actually tapped into two distinct processes. (i) Sensory-surprisal: comparing the predicted input to the received input, which is best modelled using phoneme surprisal proper. (ii) Lexical-update-surprisal: updates in activated lexical items as a *consequence* of the phoneme input, which is best modelled using a cohort reduction measure. When hearing spoken words in isolation, these predictors are perfectly associated (for instance, in Gagnepain, et al, 2012 [58]). Yet, these separable responses for connected speech might suggest multiple processes using both context-specific and more locally computed phoneme probabilities (cf. [62]).

Di Liberto et al., [68] analysed electro-encephalography (EEG) signals while subjects listened to continuous speech, and replicated the finding that phoneme surprisal modulates responses at around 110 ms. They linked these responses specifically to the theta band (∼9 Hz), in accordance with [65]. Interestingly, both the theta and delta ranges have been previously linked to acoustic-phonetic processes in previous studies [25], making it unclear whether the spectral characteristics of the signal reflect sensory processing or information processing [69].

Another important insight of [68], although not the focus of the paper, was that phoneme surprisal effects occurred much earlier (∼110 ms) than sensitivity to phonotactic transitions of English (∼300-400 ms). Phonotactics also reflect statistics over phoneme sequences [70], but the source of the statistical regularity is not related to their connection to the mental lexicon. Thus, this result suggests that while both surprisal and phonotactic probability relate to statistical processing of phoneme sequences, those statistics that are informative for lexical access produce an earlier and distinct neural response. Speech perception is optimised for, and prioritises processes that contribute to identification of spoken words.

5 Discussion

The studies reviewed here show consistent evidence that auditory responses to speech sounds are modulated by the higher level information content those sounds communicate, within the first 100-400 ms after speech sound onset. These observations demonstrate how responses to low-level units of speech are shaped by the language system's ultimate goal of linking speech sounds to stored linguistic representations in order to reconstruct the higher level meaning that the speaker intended to communicate.

Two information metrics are most commonly observed to modulate neural responses: phoneme surprisal and lexical entropy. Both metrics modulate neural responses in STG with a similar time-range (see Figure 4A for a schematic summary of this literature). Although these measures are highly correlated and show similar spatio-temporal response profiles, these variables have been shown to make independent statistical contributions to neural data [64, 65] suggesting that these two metrics tap into two distinct neural computations.

How can we place the neural effects of these information theoretical metrics into a computational understanding of speech perception? The goal of speech comprehension can be construed as identifying a sequence of morphemes or words from the auditory signal based on multiple sources of (noisy) information which must be combined with prior knowledge or expectations about the likely words that will occur and their meanings. One popular framework by which to integrate these different sources of information uses Bayesian Inference, and other mathematically similar approaches [72, 73, 74, 75, 76, 77, 78, 79], for a specific treatment of Bayesian inference in speech perception see [80, 81] .

Under a Bayesian formalisation, it is possible to generate an estimate the probability of a specific interpretation (let's say, the identity of the current lexical item) as a function of each incrementally received input (for example, using the identity of each input phoneme). We focus on the relationship between these lexical and phoneme level processes as the input and output (the pink and purple nodes in Figure 4B). However, the Bayesian inference process would

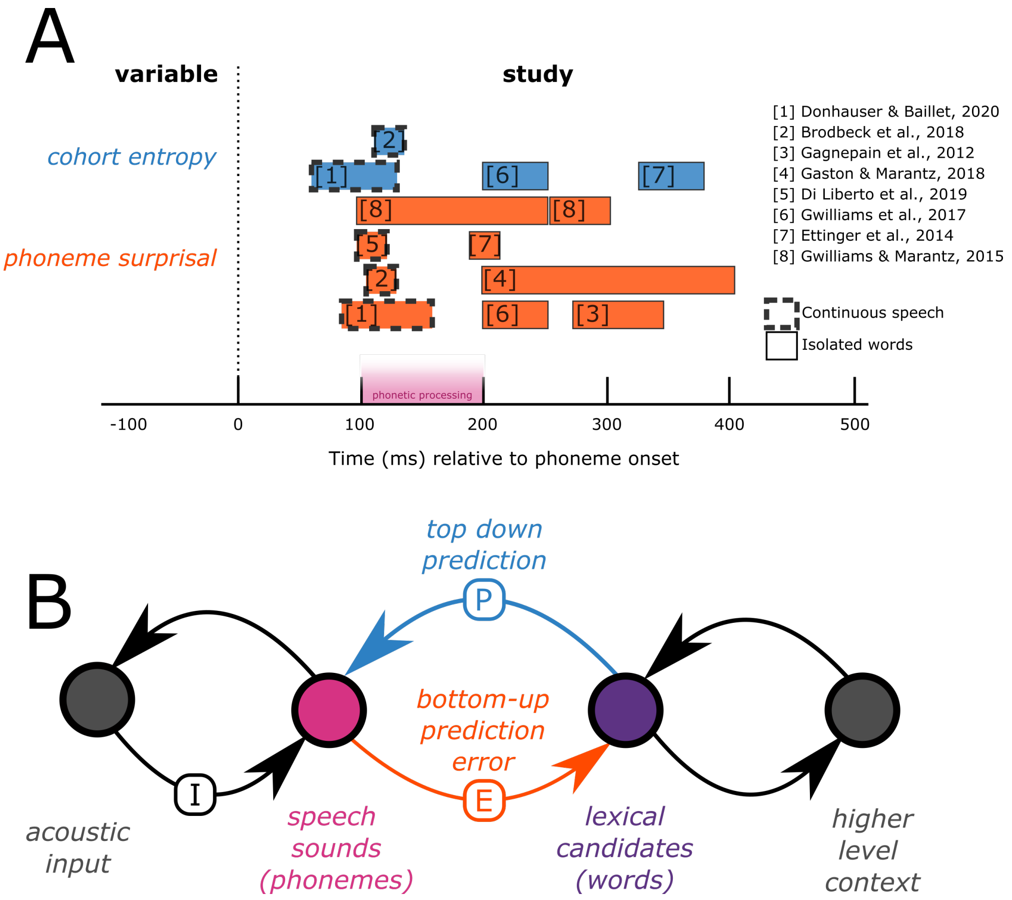

Figure 4: **Information exchange.** (A) Summary timeline of when the studies in our review [65, 64, 58, 62, 68, 40, 60, 71] find significant effects of surprisal and entropy. Boxes with a dashed outline refer to studies using continuous speech; boxes with a solid outline refer to the presentation of an isolated word or a minimal phrase. Pink shading between 100-200 ms corresponds to approximately when phonetic features are processed, for reference. (B) A simple network graph model showing how information is hypothesised to pass between the different of processing during lexical access.

operate similarly for lower-level processes (e.g. recognising phonemes given acoustic signals), or higher-level processes (accessing meaning or syntactic information given the words heard).

Identification of words from phoneme sequences would operate something like this: The listener has an internal language model which is formed based on linguistic experience and includes knowledge of the likelihood of different words, and the identity of the speech sounds that make up those words. This statistical knowledge specifies the prior knowledge that the listener uses to make top-down predictions (P) for the sounds of upcoming words. These predictions can incorporate multiple forms of hierarchically structured knowledge; that is predictions at the phoneme level might be influenced not only by known words and their constituent sounds, but also by higher-level contextual knowledge (semantic, or syntactic representations of the current utterance) that change the likelihood of different words. Thus top-down predictions, or priors provide a probabilistic prediction about the current state of the network (i.e. what word is being said), which is computed based on the frequency with which words have been experienced in the past combined with a representation of the current utterance. The extent to which the system has converged on a single prediction of the word is reflected in the *entropy* metric: Entropy is highest when the prior is uniformly distributed across multiple possibilities, and lowest when all predictions are centred around a single outcome.

The phoneme predictions generated by the prior are then compared to the current input (I), which in our illustration (Figure 4B) would be based on a representation of discrete phonemes. The difference between the phoneme that was predicted, based on the prior of the phoneme sequence of the word, and the phoneme that was heard (P-I) gives the prediction error (E). The magnitude of this prediction error is correlated with phoneme *surprisal*: When hearing an unexpected sound, prediction error and surprisal are higher than when hearing a more strongly predicted sound. When the prediction error is large, this provides additional information to update the probabilities of possible words (orange arrow to the purple nodes in Figure 4B), because the word that was predicted to be the outcome is no longer the best lexical candidate. This iterative updating process happens at successive speech segments throughout the timecourse of word processing, until the optimal candidate can be recognised (see Figure 4B, and [82] for a simple implementation of this model).

Under this framework, our interpretation of the surprisal response is a reflection of the extent to which the relative activation of lexical candidates needs updating on receiving each new piece of phonological input. If heard phonemes are strongly expected, there is little information gained, and therefore lexical activations, and subsequent predictions will go unaltered. If the phoneme was unexpected, this requires a big shift in which lexical candidates are most likely, which is reflected in the surprisal signal. This surprisal response may therefore reflect the extent to which the internal state of the system needs to be updated (orange arrow, Figure 4B), leading to changes to the predictions generated for subsequent inputs in the phonological sequence (blue arrow, Figure 4B [58, 65, 29]).

Given this iterative updating of predictions based on prediction error, entropy will reflect the current state of uncertainty about which lexical candidate will ultimately ’win’ the recognition process. The current entropy (as in domain-general accounts [83]) serves to boost or dampen the information that is likely to be gained from subsequent sensory signals (blue arrow, Figure 4B). Time points at which entropy is low (i.e. one candidate word is much more likely than the rest) permits easy identification and subsequent sensory input is not so critical; the likely outcome of word recognition is already known. However, in cases of high entropy (i.e. if multiple candidate words are activated to a similar degree) then resolving which lexical candidate is correct will require processing of incoming sensory input to ensure that the correct lexical item is selected. This interpretation is in line with Bayesian accounts of predictive coding [84, 59]: when predictions for upcoming input are uncertain (high entropy), sensory processing plays a more important role in disambiguating the input and prediction error will tend to be higher to compensate (see Figure 4B).

As can be seen in the summary timeline in Figure 4A and Table 1, the estimates for *when* these different metrics matter for neural processing are highly varied. Responses have been reported as early at 60 ms and as late at 400 ms, for both surprisal and entropy, with variability between studies in terms of both the onset latency and duration of neural effects. While some of this variation might be due to differences in statistical power or thresholds in specific studies, other variation may also be due to properties of the speech stimuli used. For example, one consistent effect that we do observe is that latencies are shorter when words are presented in the context of continuous speech rather than in isolation. Studies using naturalistic stimuli find effects of entropy and surprisal at around 120 ms after phoneme onset on average (simultaneous with processing the phonetic features of the speech sound itself), whereas studies using isolated words find sensitivity to the same features around 250 ms after phoneme onset on average. Nonetheless, even allowing for this variation we also observe that in some studies surprisal onsets earlier than entropy [64, 60], and in other studies, the reverse is observed [65]. Understanding whether any reliable temporal difference between surprisal and entropy exists, or whether they are better described as simultaneous processes, promises to provide significant insight into the computational operations being applied to the speech signal. In addition, how those responses can be changed with the provision of higher-level context which allows for predictions of upcoming lexical input, or with changes to the sensory quality of the speech signal which might permit more rapid or slower speech processing (grey nodes, Figure 4B) .

Even though both surprisal and entropy reflect higher order processes, it is noteworthy that their neural correlates are not located in the cortical areas that are typically associated with lexical access, such as middle temporal gyrus or inferior frontal gyrus [85, 21, 20, 11]. Instead, all of the studies we have described broadly localise these responses to auditory brain regions – including the transverse temporal gyrus and superior temporal gyrus – overlapping with the processing of acoustic and phonetic features are known to be processed [24, 86, 14]. Overall this indicates that local, perhaps recurrent, processing of speech sounds in auditory cortex is influenced by higher order structure, such as sequence statistics, and higher order computations, such as lexical access. While we and others have assumed functional and anatomical hierarchies in speech processing this does not imply that higher-level features of speech do not influence lower-level auditory responses. Further investigation, perhaps by taking advantage of the joint spatio-temporal resolution of intracranial recordings, will be required to fully specify the spatial location of sensitivity to phonetic features, surprisal and entropy, and the extent to which the are supported by the same versus neighbouring neural populations.

6 Conclusion

Overall, the evidence presented here suggests that the brain applies Bayesian-inference(-like) computations in order to decode meaning from the speech signal. The acoustic input (the likelihood) is weighted by probabilities over *what the speaker could say* (the prior) in order to derive *what the speaker is saying* (the posterior).

As we saw in this review, neural responses illustrate how multiple sources of information are potentially computed in parallel, including context-free and context-sensitive measures of prior probability distributions. These context-sensitive measures allow lexical items to be activated based on current syntactic and semantic context [87], whereas context-free measures rely on within-word phoneme-sequence statistics alone. Aggregating over both measures allows the brain to jointly estimate the best interpretation of upcoming input for cases in which predictions across multiple information sources converge, and to be more sceptical when multiple information sources do not converge. These situations can cue revisions to perceptual interpretations, demand on semantic reinterpretation or the detection of lexical novelty or speech errors [59].

Any valid account of how the brain achieves speech comprehension needs to explain not just how the acoustic signal is processed, but how this signal is used to identify the words being said. Information theoretic measures, such as surprisal and entropy, provide excellent tools for examining such 'higher order' processes. For example, what priors are used to form predictions of upcoming information, and what linguistic units these predictions may comprise, shed light on what representations are accessed and composed online during comprehension. As techniques for estimating such probability measures become increasingly precise, as does our ability to model how the brain uses them for language understanding.

References

[1] Noam Chomsky et al. *New horizons in the study of language and mind*. Cambridge University Press, 2000.

[2] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[3] Shihab A Shamma. Speech processing in the auditory system i: The representation of speech sounds in the responses of the auditory nerve. *The Journal of the Acoustical Society of America*, 78(5):1612–1621, 1985.

[4] Brian CJ Moore. Basic auditory processes involved in the analysis of speech sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):947–963, 2008.

[5] Ferdinand De Saussure. *Course in general linguistics*. Columbia University Press, 2011.

[6] Kenneth N Stevens and Sheila E Blumstein. The search for invariant acoustic correlates of phonetic features. *Perspectives on the study of speech*, pages 1–38, 1981.

[7] Sven L Mattys, Matthew H Davis, Ann R Bradlow, and Sophie K Scott. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7-8):953–978, 2012.

[8] Douglas O'Shaughnessy. Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979, 2008.

[9] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[10] Milene Bonte, Tiina Parviainen, Kaisa Hytönen, and Riitta Salmelin. Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cerebral Cortex*, 16(1):115–123, 2006.

[11] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

[12] Edward F Chang, Jochem W Rieger, Keith Johnson, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428, 2010.

[13] Tatyana O Sharpee, Craig A Atencio, and Christoph E Schreiner. Hierarchical representations in the auditory cortex. *Current opinion in neurobiology*, 21(5):761–767, 2011.

[14] 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.

[15] Sophie K Scott, C Catrin Blank, Stuart Rosen, and Richard JS Wise. Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123(12):2400–2406, 2000.

[16] CM Wessinger, J VanMeter, Biao Tian, J Van Lare, J Pekar, and Josef P Rauschecker. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of cognitive neuroscience*, 13(1):1–7, 2001.

[17] Matthew H Davis and Ingrid S Johnsrude. Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, 23(8):3423–3431, 2003.

[18] Kayoko Okada, Feng Rong, Jon Venezia, William Matchin, I-Hui Hsieh, Kourosh Saberi, John T Serences, and Gregory Hickok. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex*, 20(10):2486–2495, 2010.

[19] Samuel Evans and Matthew H Davis. Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. *Cerebral cortex*, 25(12):4772–4788, 2015.

[20] Ellen Lau, Colin Phillips, and David Poeppel. A cortical network for semantics:(de) constructing the N400. *Nature Reviews Neuroscience*, 9(12):920–933, 2008.

[21] Matthew H Davis. The neurobiology of lexical access. In *Neurobiology of language*, pages 541–555. Elsevier, 2016.

[22] Josef P Rauschecker and Sophie K Scott. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, 12(6):718–724, 2009.

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

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

[25] 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.

[26] 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.

[27] Laura Gwilliams, Tal Linzen, David Poeppel, and Alec Marantz. In spoken word recognition the future predicts the past. *bioRxiv*, page 150151, 2017.

[28] Han Gyol Yi, Matthew K Leonard, and Edward F Chang. The encoding of speech sounds in the superior temporal gyrus. *Neuron*, 102(6):1096–1110, 2019.

[29] Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. *bioRxiv*, 2020.

[30] Niclas Kilian-Hütten, Giancarlo Valente, Jean Vroomen, and Elia Formisano. Auditory cortex encodes the perceptual interpretation of ambiguous sound. *Journal of Neuroscience*, 31(5):1715–1720, 2011.

[31] Charlotte Jacquemot, Christophe Pallier, Denis LeBihan, Stanislas Dehaene, and Emmanuel Dupoux. Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. *Journal of Neuroscience*, 23(29):9541–9546, 2003.

[32] Neal P Fox, Matthew Leonard, Matthias J Sjerps, and Edward F Chang. Transformation of a temporal speech cue to a spatial neural code in human auditory cortex. *Elife*, 9:e53051, 2020.

[33] Elia Formisano, Federico De Martino, Milene Bonte, and Rainer Goebel. " who" is saying" what"? brain-based decoding of human voice and speech. *Science*, 322(5903):970–973, 2008.

[34] Jessica S Arsenault and Bradley R Buchsbaum. Distributed neural representations of phonological features during speech perception. *Journal of Neuroscience*, 35(2):634–642, 2015.

[35] Joao M Correia, Bernadette MB Jansma, and Milene Bonte. Decoding articulatory features from fmri responses in dorsal speech regions. *Journal of Neuroscience*, 35(45):15015–15025, 2015.

[36] Connie Cheung, Liberty S Hamilton, Keith Johnson, and Edward F Chang. The auditory representation of speech sounds in human motor cortex. *Elife*, 5:e12577, 2016.

[37] Niclas Kilian-Hütten, Jean Vroomen, and Elia Formisano. Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. *Neuroimage*, 57(4):1601–1607, 2011.

[38] Hyojin Park, Robin AA Ince, Philippe G Schyns, Gregor Thut, and Joachim Gross. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Current Biology*, 25(12):1649–1653, 2015.

[39] 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.

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

[41] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

[42] Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13, 2018.

[43] Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018.

[44] Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proc. of ACL*, 2020.

[45] Lee H Wurm. Auditory processing of prefixed english words is both continuous and decompositional. *Journal of memory and language*, 37(3):438–461, 1997.

[46] Laura Winther Balling and R Harald Baayen. Morphological effects in auditory word recognition: Evidence from danish. *Language and Cognitive Processes*, 23(7-8):1159–1190, 2008.

[47] Laura Winther Balling and R Harald Baayen. Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, 125(1):80–106, 2012.

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

[49] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[50] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[51] 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.

[52] Jeffrey R Binder, Julie A Frost, Thomas A Hammeke, Patrick SF Bellgowan, Jane A Springer, Jackie N Kaufman, and Edward T Possing. Human temporal lobe activation by speech and nonspeech sounds. *Cerebral cortex*, 10(5):512–528, 2000.

[53] Kayoko Okada and Gregory Hickok. Identification of lexical–phonological networks in the superior temporal sulcus using functional magnetic resonance imaging. *Neuroreport*, 17(12):1293–1296, 2006.

[54] Mirjana Bozic, Lorraine K Tyler, David T Ives, Billi Randall, and William D Marslen-Wilson. Bihemispheric foundations for human speech comprehension. *Proceedings of the National Academy of Sciences*, 107(40):17439–17444, 2010.

[55] Jie Zhuang, Billi Randall, Emmanuel A Stamatakis, William D Marslen-Wilson, and Lorraine K Tyler. The interaction of lexical semantics and cohort competition in spoken word recognition: an fmri study. *Journal of Cognitive Neuroscience*, 23(12):3778–3790, 2011.

[56] Matthew H Davis and M Gareth Gaskell. A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536):3773–3800, 2009.

[57] Jie Zhuang, Lorraine K Tyler, Billi Randall, Emmanuel A Stamatakis, and William D Marslen-Wilson. Optimally efficient neural systems for processing spoken language. *Cerebral Cortex*, 24(4):908–918, 2014.

[58] 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.

[59] Matthew H Davis and Ediz Sohoglu. Three functions of prediction error for bayesian inference in speech perception, 2019.

[60] Allyson Ettinger, Tal Linzen, and Alec Marantz. The role of morphology in phoneme prediction: Evidence from meg. *Brain and language*, 129:14–23, 2014.

[61] Ece Kocagoncu, Alex Clarke, Barry J Devereux, and Lorraine K Tyler. Decoding the cortical dynamics of sound-meaning mapping. *Journal of Neuroscience*, 37(5):1312–1319, 2017.

[62] Phoebe Gaston and Alec Marantz. The time course of contextual cohort effects in auditory processing of category-ambiguous words: Meg evidence for a single "clash" as noun or verb. *Language, Cognition and Neuroscience*, 33(4):402–423, 2018.

[63] Anastasia Klimovich-Gray, Lorraine K Tyler, Billi Randall, Ece Kocagoncu, Barry Devereux, and William D Marslen-Wilson. Balancing prediction and sensory input in speech comprehension: The spatiotemporal dynamics of word recognition in context. *Journal of Neuroscience*, 39(3):519–527, 2019.

[64] 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.

[65] Peter W Donhauser and Sylvain Baillet. Two distinct neural timescales for predictive speech processing. *Neuron*, 105(2):385–393, 2020.

[66] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[67] Paul Cairns, Richard Shillcock, Nick Chater, and Joe Levy. Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, 33(2):111–153, 1997.

[68] Giovanni M Di Liberto, Daniel Wong, Gerda Ana Melnik, and Alain de Cheveigné. Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *Neuroimage*, 196:237–247, 2019.

[69] Laura Gwilliams. Hierarchical oscillators in speech comprehension: a commentary on meyer, sun, and martin (2019). *Language, Cognition and Neuroscience*, pages 1–5, 2020.

[70] Peter W Jusczyk, Paul A Luce, and Jan Charles-Luce. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5):630, 1994.

[71] Laura E Gwilliams, Philip J Monahan, and Arthur G Samuel. Sensitivity to morphological composition in spoken word recognition: Evidence from grammatical and lexical identification tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(6):1663, 2015.

[72] David Mumford. On the computational architecture of the neocortex. *Biological cybernetics*, 66(3):241–251, 1992.

[73] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[74] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.

[75] Michael W Spratling. Reconciling predictive coding and biased competition models of cortical function. *Frontiers in computational neuroscience*, 2:4, 2008.

[76] Lucia Melloni, Caspar M Schwiedrzik, Notger Müller, Eugenio Rodriguez, and Wolf Singer. Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *Journal of Neuroscience*, 31(4):1386–1396, 2011.

[77] Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

[78] Anil K Seth. Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, 17(11):565–573, 2013.

[79] Lisa Feldman Barrett and W Kyle Simmons. Interoceptive predictions in the brain. *Nature reviews neuroscience*, 16(7):419–429, 2015.

[80] Dennis Norris and James M McQueen. Shortlist b: a bayesian model of continuous speech recognition. *Psychological review*, 115(2):357, 2008.

[81] Dave F Kleinschmidt and T Florian Jaeger. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2):148, 2015.

[82] Helen Blank and Matthew H Davis. Prediction errors but not sharpened signals simulate multivoxel fmri patterns during speech perception. *PLoS biology*, 14(11):e1002577, 2016.

[83] Harriet Feldman and Karl Friston. Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 4:215, 2010.

[84] Rick A Adams, Klaas Enno Stephan, Harriet R Brown, Christopher D Frith, and Karl J Friston. The computational anatomy of psychosis. *Frontiers in psychiatry*, 4:47, 2013.

[85] Rik Vandenberghe, Cathy Price, Richard Wise, Oliver Josephs, and Richard SJ Frackowiak. Functional anatomy of a common semantic system for words and pictures. *Nature*, 383(6597):254–256, 1996.

[86] Matthew K Leonard, Kristofer E Bouchard, Claire Tang, and Edward F Chang. Dynamic encoding of speech sequence probability in human temporal cortex. *Journal of Neuroscience*, 35(18):7203–7214, 2015.

[87] William D. Marslen-Wilson and Alan Welsh. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10(1):29–63, 1978.