Cortical responses to degraded speech are modulated by linguistic predictions

Jonathan E. Peelle*

*Corresponding author's address: Department of Otolaryngology, Washington University in St. Louis, 660 South Euclid Ave, St. Louis, MO 63110, peellej@ent.wustl.edu

Our perceptual experience is formed by combining incoming sensory information with prior knowledge and expectation. When speech is not fully intelligible, non-acoustic information may be particularly important. Predictions about a degraded acoustic signal can be provided extrinsically (for example, by presenting a written cue) or intrinsically (if the speech is still partially intelligible). Here I review two studies in which the neural response to speech was measured using magnetoencephalography (MEG), with speech clarity parametrically manipulated using noise vocoding. In a study of isolated word processing, accurate predictions provided by written text enhanced subjective clarity and changed the response in early auditory processing regions of temporal cortex. In a separate study looking at connected speech, the phase of ongoing cortical oscillations was matched to that of the acoustic speech envelope in the range of the syllable rate (4-8 Hz). Critically, this phase-locking was enhanced in left temporal cortex when speech is intelligible. Both experiments thus highlight neural responses in brain regions associated with relatively low-level speech perception. Together these findings support the ability of linguistic information to provide predictions that shape auditory processing of spoken language, particularly when acoustic clarity is compromised.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

*Before we began to see or hear or perceive in any way, we must have had a knowledge...*

— Plato, *Phaedo*, c. 360 BCE

Our perception of the world is based on sensory input, but informed to a large degree by our past experience, expectations, and momentary goals. Bringing to bear our knowledge about the past helps us to efficiently process our current environment and make accurate predictions about our future sensory experience (Friston, 2010). The integration of sensory information and prior knowledge is precisely what happens during numerous aspects of spoken language comprehension, and in fact is required if listeners are to have any hope of making sense of the rapid and complex speech signal. The contributions of these two sources of information to perception is particularly evident when sensory information is lacking in detail or degraded in some way, is frequently occurs during natural speech comprehension (Adank, 2012; Mattys *et al*., 2012; Van Engen *et al*., 2012)—in such cases, our perception is necessarily based more strongly on prior knowledge.

The cognitive and neural mechanisms underlying these computations are still unclear, both in their general principles and in their specific manifestation during speech comprehension. A promising framework is one in which information (predictions) generated by non-sensory brain regions can modulate sensory processing. I will discuss two recent studies in which subjects listened to degraded speech while brain activity was recorded with high temporal resolution using magnetoencephalography (MEG). These studies deal with both content (the *what* of speech) and timing (the *when*), ideas conveyed schematically in Figure 1. Although far from exhaustive, the results demonstrate the effect of two types of prediction that come into play during speech comprehension in the context of hierarchical neural and linguistic networks, and how neural activity at any time concurrently reflects influence from multiple sources (Tiesinga *et al*., 2008).

**SPEECH DEGRADATION USING NOISE VOCODING**

In both experiments discussed below, the speech signal was degraded using a noise vocoding technique. During vocoding, the frequency range of the original signal is divided into logarithmically-spaced frequency bands, or channels. The low frequency amplitude envelope (typically below ~30 Hz) is extracted from each channel, used to modulate white noise, and then bandpass filtered to lie in the same frequency range as the original. Using more channels results in a more detailed acoustic reproduction of the speech signal, and thus a more intelligible stimulus. Thus, noise vocoding allows the parametric manipulation of speech intelligibility, while keeping the overall amplitude envelope (and thus the main cues to speech rhythm) consistent.

**LINGUISTIC PREDICTIONS FACILITATE SINGLE WORD PROCESSING**

Sohoglu and colleagues (2012) investigated the degree to which listeners’ subjective experience of a spoken word, and its associated neural signature, differed as a function of prior information. Subjects heard vocoded words that were preceded by a matching, mismatching, or neutral written cue. For example, before the spoken word “clay” subjects might see clay (matching), snail (mismatching), or xxx (neutral). The physical clarity of the spoken words was also manipulated by vocoding them with 2, 4, or 8 channels, resulting in a 3 (cue type: matching, mismatching, neutral) × 3 (speech clarity: 2, 4, 8 bands) factorial design. Following each word, subjects indicated the perceived clarity of the word by indicating a number from 1–8.

Behaviorally, providing a matching cue prior to hearing a degraded word significantly increased its perceived clarity. This effect mimicked the increase in perceived clarity obtained by physically altering the stimulus (see also Jacoby *et al*., 1988).

Neurally, matching written cues resulted in early increases in neural activity, approximately 90–130 ms after spoken word onset. These responses were localized to left inferior frontal cortex. During this time window, activity in left superior temporal gyrus did not discriminate across cue type—this only happened in a later time window (450–700 ms after word onset). During this later time, inferior frontal activity was maintained (matching > other cue types). However, in this case activity for the matching trials was reduced in superior temporal gyrus, relative to the other cue types. These results fit well within a predictive coding framework in which accurately predicted sensory events result in minimal neural activity (hence the reduction in neural activity in superior temporal gyrus for the
matching cue type), but that less-predicted events produce an error signal that may be propagated up the cortical hierarchy. Furthermore, the timing of the activity—inferior frontal cortex preceding superior temporal gyrus—are consistent with a top-down prediction, coming from frontal cortex, that informs sensory areas (here, superior temporal gyrus, near auditory cortex).

**FIGURE 1.** Schematic of predictive neural processes during connected speech comprehension. Top: Spectro-temporal acoustic predictions can be made with knowledge about word identity obtained from, for example, semantic context (or a written cue). This can be compared to incoming acoustic information, and if mismatching, generate a prediction error signal. Bottom: Neural oscillations in the range of syllable rate (~4–8 Hz) entrain to ongoing speech and encode complementary predictions about the temporal occurrence of key acoustic events (e.g., stressed syllables).

**TEMPORAL PREDICTION FOR CONNECTED SPEECH COMBINES ACOUSTIC AND LINGUISTIC CUES**

In an MEG study of spoken sentence processing we again manipulated speech intelligibility using noise vocoding (Peelle *et al.*, in press). Subjects heard sentences that were vocoded using 16 channels (which were highly intelligible), 4 channels (which were moderately intelligible—word report was around 30% accurate), 4 channels but spectrally rotated (unintelligible) and 1 channel (also unintelligible). Thus, although the overall amplitude envelope was preserved across all four conditions, the amount of linguistic information that could be extracted differed substantially.

We used these sentences to investigate the degree to which ongoing brain oscillations phase lock to the acoustic amplitude of spoken sentences. Fluctuations in oscillatory neural activity are associated with different states of neural excitation (Bishop, 1932), and the efficiency of perception can be affected by the phase of underlying neural oscillations (Womelsdorf *et al.*, 2006; Romei *et al.*, 2010). Ensembles of oscillating neurons can thus be viewed as encoding a temporal prediction about when a stimulus is likely to occur (i.e., the period of high excitation/efficiency) (Engel *et al.*, 2001). Given evidence showing that the phase of ongoing oscillatory activity can discriminate across sentences (Luo and Poeppel, 2007; Kerlin *et al.*, 2010), we sought to (a) establish a direct link between the phase of the speech envelope and the phase of neural oscillations, and (b) investigate whether this relationship is modulated by linguistic information.
There was ample evidence for significant phase locking between brain oscillations and the speech envelope in the case of unintelligible (1 channel) speech, localized bilaterally in clusters extending through temporal and motor cortex. When linguistic information was available, this phase locking increased, notably in left temporal cortex (middle temporal gyrus). Although neural oscillations tracking the unintelligible speech would be predicted purely on the basis of the acoustic information present, the increased phase locking for intelligible speech is best explained by a cooperative prediction that integrates both linguistic and acoustic information in anticipating the upcoming acoustic signal. This indicates that rhythmic acoustic information aids normal-hearing listeners during connected speech perception, but also suggests a mechanism through which fluctuating acoustic noise might act to disrupt comprehension (Qin and Oxenham, 2003).

CONCLUSIONS

A guiding principle of sensory processing is the perceptual grouping of acoustic information into auditory objects (Carlyon, 2004; Nelken, 2004). This does not happen in isolation, but is informed in part by the temporal and spectral predictions the auditory system makes based on prior knowledge and expectation. This prior knowledge encompasses information from spectral, temporal, and semantic dimensions along multiple timescales that come together to inform current perception (Holt, 2006; Obleser et al., 2007; Dilley and Pitt, 2010; Idemaru and Holt, 2011; Sohoglu et al., 2012; Peelle et al., in press). The accuracy of prediction can also be thought of as relating to the number of competitors an auditory object (such as a word) has (Zhuang et al., 2011; Gagnepain et al., 2012). Strong predictions can limit the number of possible items, and thus aid identification.

The findings discussed above are informative on their own, but are also important because they begin to strengthen links between cognitive neuroscience and invasive electrophysiological experiments of predictability in primary auditory cortex of nonhuman species, in which probabilistic stimulus adaptation (Ulanovsky et al., 2003), effects of hearing loss on neural timing (Xu et al., 2007), and disruptions to neural timing (Yang et al., 2008) have all been studied at the level of individual neurons.

Complementary sensory and predictive processes thus tie together linguistic and acoustic information during speech comprehension, and allow listeners to increase efficiency by relying on past experience while at the same time being sensitive to potential novelty in sensory input.

ACKNOWLEDGMENTS

This work was supported by NIH grants AG038490 and AG041958. I am grateful to Ed Sohoglu for helpful comments on this manuscript.

REFERENCES