In spoken word recognition the future predicts the past

*Laura Gwilliams[1,2], Tal Linzen[4,5], David Poeppel[1,3] & Alec Marantz[1,2]

[1]*New York University*
[2]*NYUAD Institute*
[3]*Max-Planck Institute*
[4]*LSCP & IJN, CNRS, EHESS and ENS, PSL Research University*
[5]*Johns Hopkins University*

* Corresponding Author:

Laura Gwilliams
10 Washington Place
New York,
NY, 10003
United States

Email: laura.gwilliams@nyu.edu
Tel: 1-347-725-5635

Abstract

Speech is an inherently noisy and ambiguous signal. In order to fluently derive meaning, a listener must integrate contextual information to guide interpretations of the sensory input. While many studies have demonstrated the influence of *prior* context, the neural mechanisms supporting the integration of *subsequent* information remain unknown. Using magnetoencephalography, we analysed responses to spoken words with a varyingly ambiguous onset phoneme, the identity of which is later disambiguated at the lexical uniqueness point[1]. Our results uncover a three-level processing network. Subphonemic detail is preserved in primary auditory cortex over long timescales, and re-evoked at subsequent phoneme positions. Commitments to phonological categories occur in parallel, resolving on the shorter time-scale of ~450 ms. Finally, predictions are formed over likely lexical items. These findings provide evidence that future input determines the perception of earlier speech sounds by maintaining sensory features until they can be optimally integrated with top-down information.

Keywords: *Speech; MEG; Lexical access; Auditory processing*

[1] For a demonstration of this retroactive phenomenon, see here

Perceptual interpretation of transient signals presents a fascinating puzzle. Because the system cannot choose to re-receive earlier input, it must either prioritise *accuracy* by allowing sensory evidence to accumulate over time, or prioritise *speed* of processing by committing to an interpretation before it is certain that interpretation is correct. This issue becomes even more problematic when the signal is hierarchically structured, as is true for speech: Comprehension of lower-level units is required to perceive the whole (e.g. phonemes are assembled into words, words into sentences); understanding the whole aids comprehension of its constituents. Therefore, in order to derive the correct interpretation, the speech-comprehension system must balance integration over short and long time-scales, and allow low and high-order representations to interact.

Phonemes (e.g. /b/, /p/, /t/), and their phonetic features (e.g. -voiced, +plosive), are considered the smallest meaningful units of speech. Previous research indicates that phonemes are first processed in terms of low-level spectrotemporal extraction in primary auditory cortex ~50 ms after onset (A1 / Heschl's gyrus), followed by analysis of phonetic feature representations in superior temporal gyrus (STG) ~100 ms (Simos, Diehl et al. 1998, Ackermann, Lutzenberger et al. 1999, Obleser, Lahiri et al. 2003, Papanicolaou, Castillo et al. 2003, Obleser, Lahiri et al. 2004, Mesgarani, Cheung et al. 2014, Di Liberto, O'Sullivan et al. 2015). This is in line with models of auditory processing that assume increasing abstraction as information moves along the auditory pathways (Scott and Johnsrude 2003, Hickok and Poeppel 2004, Liebenthal, Binder et al. 2005, Rauschecker and Scott 2009). In many cases an acoustic signal will be consistent with more than one phoneme, and the system will need to decide which categorisation is the correct one. But it is currently unknown where the recognition and resolution of phonological ambiguity fits relative to these stages of processing. While ambiguity has been shown to modulate fMRI-recorded activation in posterior Heschl's gyrus (Kilian-Hutten, Valente et al. 2011), suggesting there are abstract perceptual sound representations in primary auditory cortex, without timing information it cannot be established whether this region is recruited before, during or after phonological categorisation in higher level areas.

The interpretation of phonemes is not a purely bottom-up process; surrounding context, in the form of words and sentences, also has an influence. So far, integration of these low (phonetic/phonological) and high (lexical/sentential) order representations over long time-scales has primarily been investigated in terms of the influence of preceding context on following phoneme perception. For example, a sound that is acoustically ambiguous between [s] and [f] will be perceived as /f/ in the word "gift" but as /s/ in the word "kiss". The Cohort model accounts for this phenomenon by suggesting that a phonological sequence activates word candidates, which help form predictions about upcoming phonemes (Marslen-Wilson and Tyler 1980, Tyler 1984, Marslen-Wilson 1987). Phonological perception is then biased to be consistent with the sound that was predicted given the context (Warren 1970, Cole 1973, Samuel 1981). Investigations into the neural correlates of this process associate the (dis-)confirmation of predictions with activation in the left superior and transverse temporal gyri (Gagnepain, Henson et al. 2012, Ettinger, Linzen et al. 2014, Gwilliams and Marantz 2015), suggesting that these regions compare the expected with the received information, and any discrepancy between the two is reflected as an error-prediction signal.

One defining characteristic of the Cohort model and other feedforward models of speech processing is that information never flows backwards in time. Consequently, because phonemes unfold more quickly than words, information contained within a word cannot feed back to alter analysis of its previously heard phonemes. However, relevant context often does occur after the sensory input to be interpreted, and still serves to bias perception. For example, a speech sound that is ambiguous between [g] and [k] will be interpreted as /g/ if occurring at the onset of "gift" and /k/ at the onset of "kiss", even though the acoustic signal at the onset of each word is identical (Ganong 1980, McQueen 1991, Gordon, Eberhardt et al. 1993). Functional connectivity analysis of MEG data supports the existence of feedback in such situations, finding evidence that activity linked to lexical processing in supra marginal gyrus affects phonetic processing in STG at a word's point of disambiguation (POD) (Gow, Segawa et al. 2008). The STG has also been implicated in fMRI studies of phoneme ambiguity (Blumstein, Myers et al. 2005, Myers and Blumstein 2008), and with sensitivity to post-assimilation context (Gow and Segawa 2009) (also see (Gow and McMurray 2007) for an overview of related behavioural results). The inferior frontal gyrus is also recruited when ambiguous phonemes distinguish word pairs (e.g. *blade vs. glade*), but not pseudo-words, or pseudo-word/word pairs (Rogers and Davis 2017), further suggesting there is a lexical influence in resolving phonological identity.

If speech processing does not advance in a purely feedforward manner (though see (Norris, McQueen et al. 2000)), what computations license retroactive processing? Some models do permit bi-directional flow of information between phonological and lexical levels of analysis, such as the TRACE model (McClelland and Elman 1986). McMurray and colleagues (McMurray, Tanenhaus et al. 2009) tested the accuracy of the TRACE model in an eye-tracking study, and found that while the model was not able to account for responses in its native form, it was able to accurately predict responses when phoneme-level inhibition was removed. This finding is consistent with psycholinguistic research suggesting that, in the face of ambiguity, sub-phonemic detail is maintained to delay commitments to discrete phonological categories. Exact estimates of how long the system can delay categorisation has varied across studies, although it appears to be on the order of ~1 second (Connine, Blasko et al. 1991, Samuel 1991, McMurray, Tanenhaus et al. 2009, Szostak and Pitt 2013). And eventual commitments are affected by both word frequency and sub-phonemic information (Dahan and Gareth Gaskell 2007). However, here we make the claim that operationalising delayed commitment as sustained sensitivity to sub-phonemic detail may be inaccurate, as these two processes are not necessarily equivalent: Sub-phonemic detail may be maintained in parallel to, or independent from, committing to discrete phonological categories. No study has yet sought to piece these two computations apart.

Three fundamental questions remain regarding the ability of subsequent information to influence phonological perception. First, does the recognition of phonological ambiguity manifest as an early perceptual or higher-order post-perceptual process? Second, how does sub-phonemic maintenance prolong the malleability of phoneme perception? Third, what temporal constraints are placed on the system — what is the limit on how late subsequent context can be received and still be optimally integrated? The current study aims to address these questions.

Whole-head magnetoencephalography (MEG) was recorded across two experiments. In the first, participants performed two-alternative forced choice (2AFC) classification on 11-step syllable continua (e.g., /ba/ <-> /pa/), which provide sensory information about onset phoneme identity but no contextual information. In the second, participants performed a non-metalinguistic auditory-to-visual word-matching task on word to non-word continua ("parakeet" <-> "barakeet"), where the onset of the words contained the syllables used in Experiment 1 (see *Materials and Methods*). These stimuli provided both sensory *and* subsequent contextual information about onset phoneme identity. MEG was used to time-lock analyses to temporally distinct positions during the time course of processing, and to determine the cortical regions demonstrating sensitivity to phonological ambiguity and ambiguity resolution.

2. Results

Analyses on the behavioural responses of Experiment 1 are reported in Supplementary materials (Section 1 & Figure 1). Analysis scripts are available online; data is available on request by contacting the first author.

In order to identify the neural processes that are sensitive to phonemic ambiguity, a mixed effects regression-based temporal-spatial cluster analysis was applied to source estimates in Heschl's gyrus (HG) and superior temporal gyrus (STG) bilaterally, searching a time-window of 0-200 ms after phoneme onset (see (Gwilliams, Lewis et al. 2016) for more details concerning this analysis technique). The same analysis was conducted on both Experiment 1 & 2 datasets. The 11 acoustically defined steps were re-sampled based on participants' behavioural responses, forming a perceptually defined 5-step continuum (see *Materials & Methods* and Fig. 4). There were three primary variables of interest: Ambiguity (distance from the behaviourally-defined perceptual boundary); Acoustics (position on the 5-step re-sampled continuum); and Feature Type (binary variable coding for place of articulation (POA) or voice onset time (VOT)). Trials were grouped into phoneme categories based on participants' behavioural responses in Experiment 1. Number of trials into the experiment and block number were included in all models as confound variables. We used a cluster forming threshold of $p < .05$, with a minimum of 10 neighbouring spatial samples, and 25 temporal samples. Bonferroni correction for multiple comparisons over time and space was applied following (Maris and Oostenveld 2007).

The results for Experiment 1 are summarised in Table 1 (left) and displayed in Fig. 1A & 1B. In sum, early (~50 ms) Ambiguity effects were left lateralised, and early Acoustic

	Experiment 1: Syllable Onset				**Experiment 2: Word Onset**				**Experiment 2: POD Onset**			
	Ambiguity	Acoustics	VOT	POA	Ambiguity	Acoustics	VOT	POA	Ambiguity	Acoustics	VOT	POA
p-value	< .005*** .029*	0.125	< .001***	< .001***	.034 * .063 ‘	.019*	< .005**	< .005** .028*	.011*	.043*	< .01 **	< .001 ***
Time window (ms)	45-110 105-145	40-75	85-200	90-150	150-182 144-172	106-152	92-138	86-126 88-126	50-84	110-136	98-140	26-96
Hemisphere & Location	LH (HG) RH (STG)	RH (HG)	RH (STG)	LH (STG)	LH (HG) LH (HG)	RH (HG)	RH (STG)	LH (STG) LH (STG)	LH (HG)	LH (HG)	RH (STG)	RH (STG)

Figure 1. Summary of MEG results. VOT = "voice onset time"; POA = "place of articulation"; POD = "point of disambiguation".

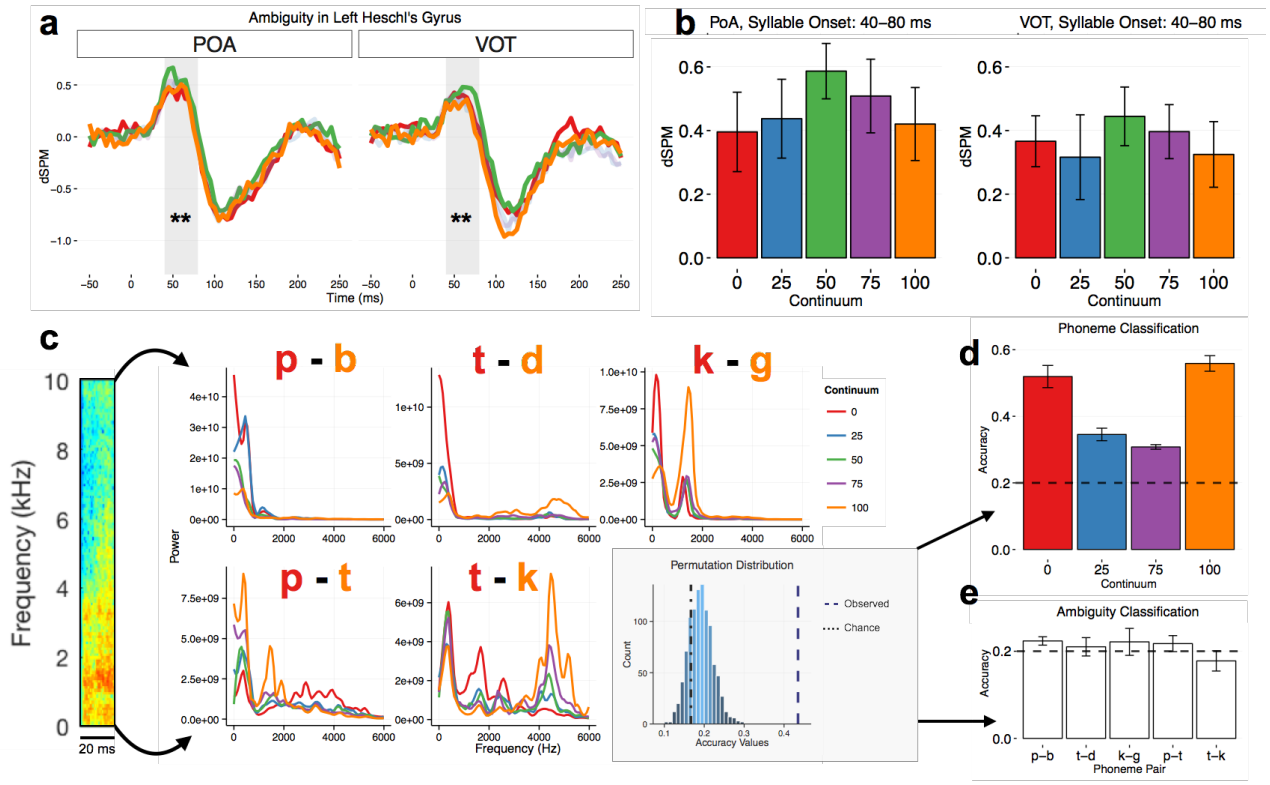

Figure 1. Early responses to ambiguity in left Heschl's Gyrus (LHG) in Experiment 1. (a) Timecourse of responses for each ambiguity level averaged over source-localised responses in LHG. (b) Averaged responses in LHG over the p50m peak, plotted separately for place of articulation (POA) and voice onset time (VOT) continua. (c) FFT decomposition of first 20 ms of the auditory stimuli, plotted for each phoneme continuum. The histogram represents the 1000 permutations used to determine the significance of classification accuracy. (d) Accuracy of the logistic regression classifier in identifying the correct phoneme, based on leave-one-out cross validation — accuracy drops off for more ambiguous tokens. (e) Chance-level accuracy in classifying steps along the continuum.

effects were right lateralised. Later responses (~100 ms) interacted with the phonetic feature being manipulated (i.e., VOT vs. POA). POA modulated left hemisphere responses, and VOT modulated right hemisphere responses.

The results for Experiment 2 are summarised in Table 1 (middle). The lateralisation of effects observed in Experiment 1 was replicated: Sensitivity to Ambiguity and POA in the left hemisphere, and to Acoustics and VOT in the right hemisphere. The Ambiguity cluster was identified at ~150 ms in the lexical context, which is later than the effect found for syllable context. However, when looking at the cluster level *t*-values across time (Fig. 2A), there was a clear peak in sensitivity to Ambiguity at ~50 ms, too. To test if lexical items also elicit early sensitivity to Ambiguity, we ran a post-hoc mixed-effects regression analysis, averaging just in left Heschl's gyrus (the locus of the effect in Experiment 1) at 50 ms post word-onset (the peak of the effect in Experiment 1). Ambiguity, Acoustics, Feature Type and their interaction were coded as a fixed effect and random slopes over items. This revealed a significant interaction between Ambiguity and Feature Type ($X = 5.9$, $p = .015$), and a significant effect of Feature Type ($X = 13.14$, $p < .001$). When breaking the results down at each level of Feature Type, Ambiguity was a significant factor for POA contrasts ($X = 4.84$, $p = .027$) and was approaching significance for VOT contrasts ($X = 3.09$, $p = .078$). This analysis confirms that the early ambiguity effect is replicated in lexical contexts. Interestingly, the direction of the effect was reversed for POA contrasts, whereby more ambiguous tokens elicited less rather than more activity (SM, Fig. 3). This

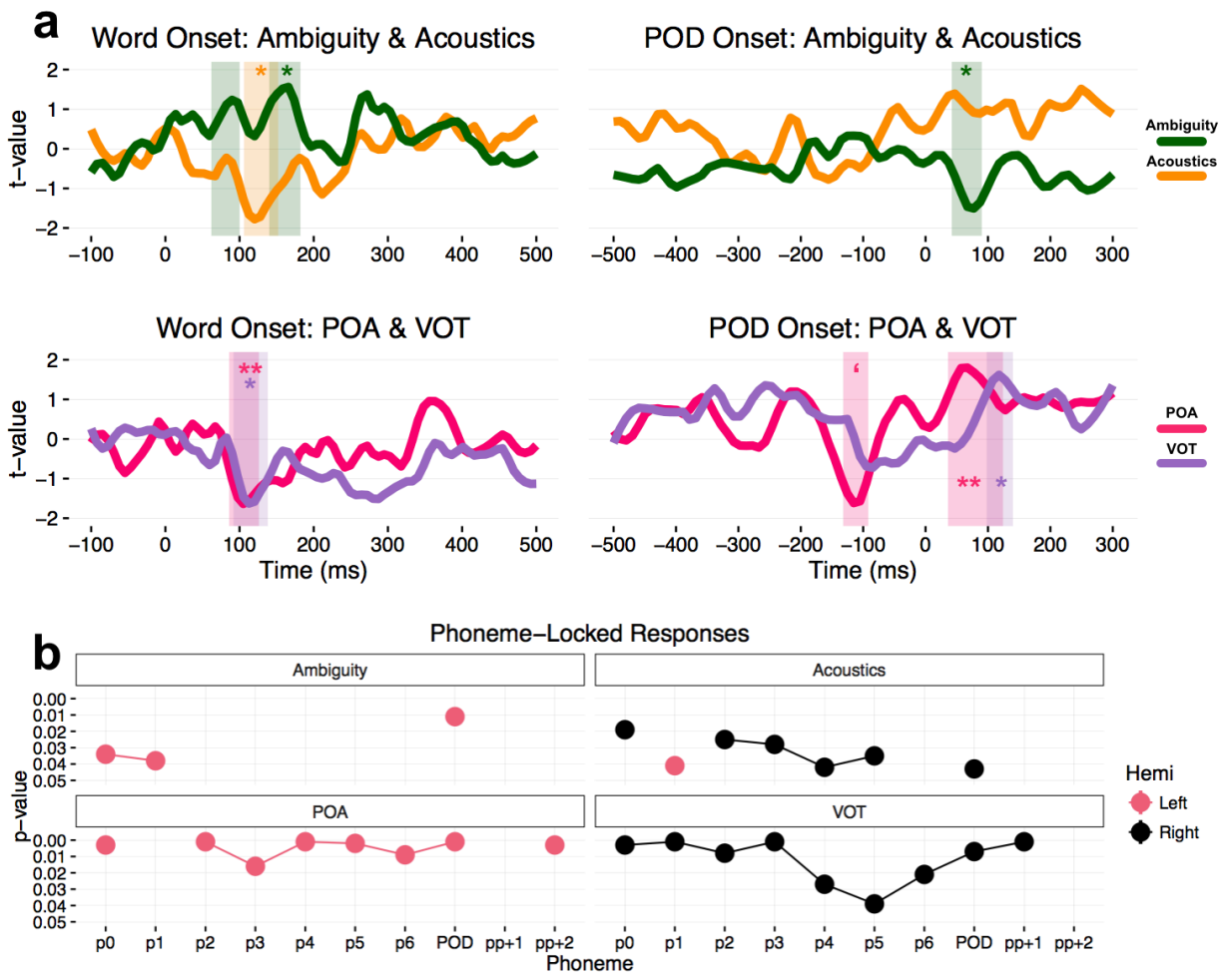

Figure 2. Timecourse of regression analysis for the four primary variables of interest for Experiment 2. (a) Responses time-locked to word onset (left) and point of disambiguation (POD, right). Above: *t*-values of the Ambiguity and Acoustics variables when put into the same regression model. Below: *t*-values of place of articulation (POA) and voice onset time (VOT). Note that *t*-values are extracted from the most significant cluster formed for that variable. (b) Responses time-locked to the onset of each phoneme along the length of the word. p0 = word onset; pp = "post POD". Statistical *p*-values are based on the most significant cluster. If no cluster is formed, the cell is empty for that entry.

interaction may be due to differences in the task or due to processing syllables versus words. More research would need to be conducted to piece these apart.

Integration of top-down lexical information was analysed by time-locking MEG responses to the onset of a word's point of disambiguation (POD). Main results are summarised in Table 1 (right). In sum, sensitivity to Ambiguity, Acoustics, POA and VOT re-emerges at point of disambiguation, with similar lateralisation to that observed at onset (see Fig. 2). The results for Ambiguity and Acoustics tested at each level of Feature Type are in Supplementary Materials 2.1-2.3.

In order to assess whether the results observed at POD were specific to disambiguation or were a general re-activation process, we ran the same analysis time-locked to each phoneme along the length of the word, up to two phonemes post-disambiguation (Fig. 2B). Sensitivity to Ambiguity was observed just at the initial phonemes and at disambiguation; however, sensitivity to Acoustics, VOT and POA was maintained throughout the word (*p*'s < .05, corrected for multiple comparisons across phonemes).

To determine whether the system commits to a phonological category when disambiguation occurs "too late", we tested for an interaction between disambiguation latency and whether the word resolves to the more or less likely word of the pair given acoustics at onset. The rationale is that if the system commits to a /b/, for example, but then the word resolves to a p-onset word, more effort is required to comprehend the lexical item that was thrown away during the commitment process; however, if no commitment has occurred, there should be a minimal difference between word and non-word resolution because both the cohort of p-onset and b-onset words are still active.

First, we applied the spatio-temporal regression to responses 0-300 ms after to point of disambiguation. Because this test assumes higher-level processing, the search area was expanded to include middle temporal gyrus (Fig. 3F). The variables included in the model were lexical resolution (word versus non-word) and its interaction with POD latency, where latency was defined in terms of ms for one test, and in terms of phonemes in the other. No interaction was found between phoneme-defined latency and lexical resolution. However, defining latency in terms of ms did reveal a significant interaction in the left

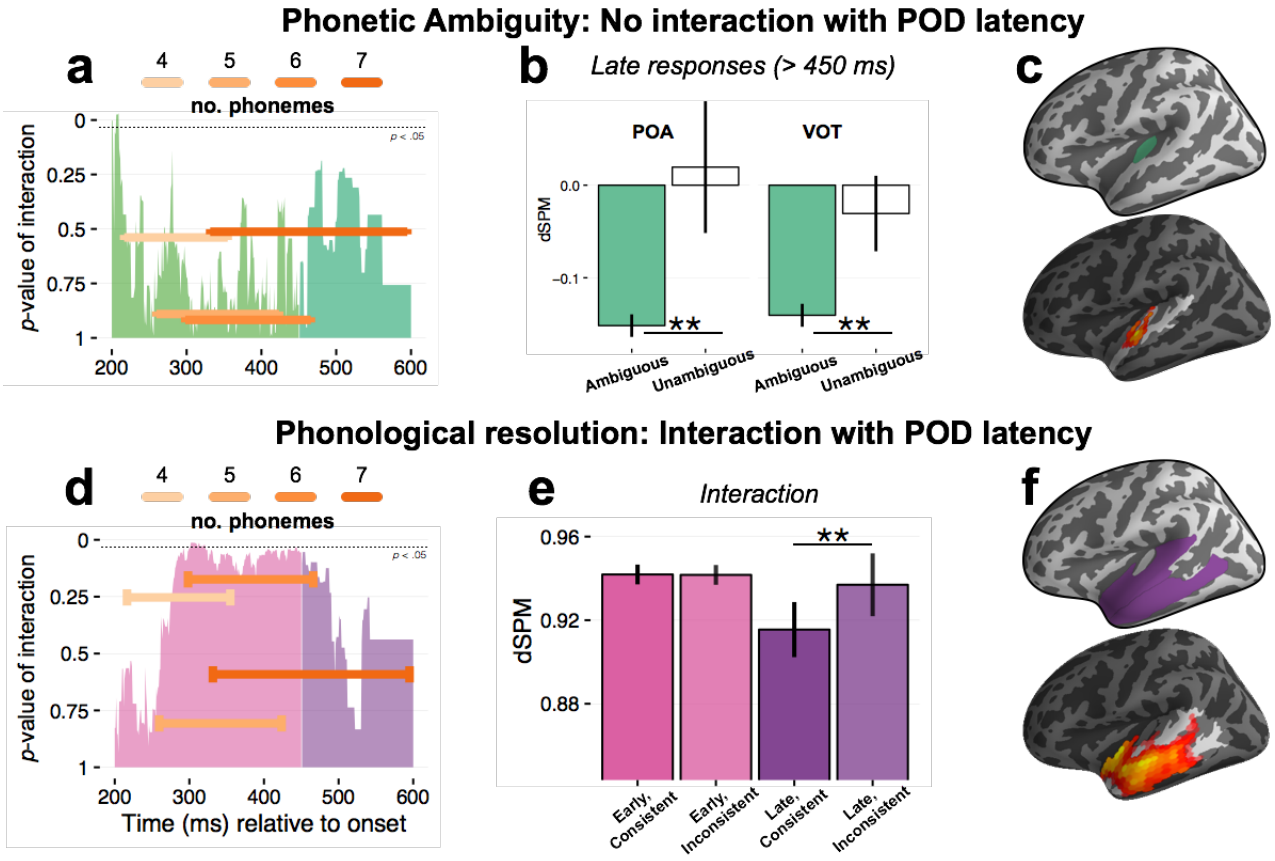

Figure 3. Interaction between point of disambiguation (POD) latency and ambiguity level (above), and with lexical resolution (below). For panels (a) and (d), POD latency defined in milliseconds (200-600 ms) is represented by filled area, and in terms of phonemes (4-7) by horizontal orange bars. The span of the orange bars reflects the range of onset latencies for each phoneme. (a) Non-significant interaction between latency and phoneme ambiguity when splitting latency into "early" vs. "late" incrementally from 200-600 ms post word onset. (b) More activity for more ambiguous sounds at disambiguation, even past 450 ms after word onset. (c) Above: Location of Heschl's Gyrus; Below: Location of significant cluster sensitive to ambiguity. (d) Timecourse of interaction between lexical resolution and latency, again defining latency as "early" and "late" at incremental distances from word onset. (e) Significant interaction when splitting responses at 450 ms. Resolution is "consistent" when the acoustics at onset more closely match the onset of the revealed word of the pair (e.g., 75% /b/ and hear "barricade"), and "inconsistent" when the acoustics better match the onset of the other word of the pair (e.g., 75% /b/ and hear "parakeet"). (f) Above: Location of superior temporal and middle temporal gyrus; Below: Location of cluster sensitive to the interaction between lexical resolution and latency.

hemisphere between 196-266 ms after POD (p = .02, Bonferroni corrected for multiple comparisons over time). Second, in order to identify the optimal split between "early" and "late", we averaged activity over the spatio-temporal dimensions of the interaction cluster, and ran a linear mixed-effects regression analysis, testing for an interaction with latency, where latency was shifted incrementally by 1 ms from 200-600 ms after word onset. As can be seen in Fig. 3D, the interaction was maximised when setting the boundary between "early" and "late" between 292-447 ms. When running the same analysis for the Ambiguity variable, no interactions were observed with latency — words that had an ambiguous onset elicited a stronger response at POD regardless of how many ms or phonemes elapsed before disambiguation (Fig. 3A-C).

3. Discussion

The aim of the present study was to determine the neural computations that allow the perception of speech sounds to be affected by subsequent top-down information. Using MEG, we analysed responses to two critical points within a word: 1) varyingly ambiguous phoneme at word onset, and 2) point of disambiguation: onset of the phoneme that reveals the identity of the word, and consequently the identity of the word-initial phoneme. Our results uncover a three-stage processing network: *Maintain subphonemic information* in parallel to the *recognition of phonological categories*, in order to *predict lexical items*. These findings demonstrate how phonetic representations can be moulded by incoming higher-order information while being carried forward through the processing stream.

3.1 Early sensitivity to ambiguity and acoustics

We found evidence for sensitivity to phonological ambiguity very early during processing, at just 50 ms after onset, in left Heschl's Gyrus. This was orthogonal to sensitivity to position on the continuum, i.e., linear acoustic differences, which was right-lateralised at the same latency. While previous studies have found the p50m to be modulated by VOT (Steinschneider, Volkov et al. 1999, Hertrich, Mathiak et al. 2000) and POA (Tavabi, Obleser et al. 2007), and fMRI studies have found sensitivity to ambiguity in primary auditory cortex (Kilian-Hutten, Valente et al. 2011) (see *Introduction*), this is the first evidence of such early responses tracking proximity to perceptual boundaries. This finding supports a hierarchical over reverse-hierarchical processing model cf. (Kilian-Hutten, Valente et al. 2011) because sensitivity comes online before any top-down higher-order influence, and illustrates that early stages of processing are tuned to strikingly complex features of the acoustic signal.

Because of the time it takes the acoustic signal to reach primary auditory cortex, the early ambiguity effect must be reflecting a response to (at most) the first 20 ms of the stimulus. As we were able to decode phoneme category from the spectrotemporal properties of the first 20 ms of the acoustic stimuli (Fig. 1C-E, Supplementary Materials 1.2), it is clear that phoneme category information is present in the signal (also see (Blumstein, Stevens et al. 1977, Stevens and Blumstein 1978) for a similar conclusion in voiced POA contrasts). In light of this, a very plausible explanation for these observations is consistent with an analysis by synthesis model (Halle and Stevens 1962, Poeppel and Monahan 2011). Concretely, responses reflect the number of candidate phonemic representations

generated by the first ~20 ms of acoustic signal. Neurons fire more when the search space over phonemic hypotheses is large, and less when there are fewer possibilities.

Furthering the stimuli decoding analysis, we applied the same logistic classifier to the first 60 ms of acoustic input – the likely amount of information driving the N100m response (see *Introduction*). The classifier was trained either on a single 60 ms spectral segment of the signal, or three sequential 20 ms spectral chunks. The former provides reasonable spectral resolution but poor temporal resolution; the latter provides the opposite. This novel analysis revealed intuitive results: The classifier more accurately distinguished VOT contrasts (a temporal cue) when trained on three 20 ms chunks, and POA contrasts (a spectral cue) when trained on a single 60 ms chunk (see Section 1.2 and Figure 2 in Supplementary Materials). It may be the case that the N100m response is driven by neuronal populations that sample both at fast (~20 ms) and slower (~60 ms) frequencies in order to accurately identity phonemes that vary across each phonetic dimension.

Although not the focus of our study, the data revealed striking asymmetries across hemispheres. In early ~50 ms responses Ambiguity localised to the left hemisphere, and Acoustics to the right. In later ~100 ms responses and along the length of the word, POA was predominantly supported by the left hemisphere, while VOT was supported by the right (see Fig. 2B). Both p50m and N100m asymmetries were replicated across the two studies, suggesting that the right hemisphere does not only contribute in lexical contexts, or only in the presence of metalinguistic judgements cf. (Wolmetz, Poeppel et al. 2011). Rather it appears that, at least in the initial stages of processing, each hemisphere makes a functionally distinct contribution to the discretisation of a speech sound (Gage, Roberts et al. 2002, Poeppel 2003). Concretely, our results suggest that the left hemisphere analyses input relative to abstract phonetic features, whereas the right hemisphere extracts lower level subphonetic information, thus illustrating sensitivity to linearly varying spectral information.

3.2 Re-emergence of subphonemic detail

We observed a re-emergence of sensitivity to the acoustics, POA and VOT of the phoneme heard at onset at each phoneme along the length of the word, at disambiguation point, and at the two phonemes *after* disambiguation. This was specifically time-locked to the onset of each incoming phoneme and was not apparent when analysing based on the time elapsed from word onset (contrast Fig. 2A with Fig. 2B). Sensitivity to onset phoneme ambiguity revealed a different pattern. It faded at intervening positions, resurfaced at the disambiguation point, and decayed after ambiguity resolution. While this may reflect a more subtle response that we have insufficient power to detect at intermediate positions, it is possible that sensitivity to this variable only resurfaces once phoneme identity can be fully resolved. This novel finding is critically important because it supports the hypothesis that the sub-phonemic representation of a speech sound is maintained in superior temporal regions throughout the duration of a word, even while subsequent phonemes are being received; perhaps suggesting that the percept of a speech sound is reassessed at each increment based on the provision of additional input. This finding is also consistent with a recent study using EEG (Khalighinejad, da Silva et al. 2017), which found evidence for continued maintenance of phoneme-category distinctions.

Further, it appears that phonemic reactivation is a general feature of speech comprehension, rather than a specific mechanism recruited in the presence of ambiguity. Specifically, our results indicate that subphonemic information is maintained even when uncertainty about phoneme identity is low, in two capacities. First, re-emergence of phonetic properties was not specific to the ambiguous tokens — it also occurred for the unambiguous phonemes. Second, information about phonetic features continues to be conserved after disambiguating information became available. Overall, these observations are the first to reveal that subphonemic information is maintained, not just in terms of uncertainty about categorisation, but *also* in terms of fine-grained phonetic and acoustic detail of the phoneme under scrutiny. And both sources of information continue to be revisited over long timescales.

3.3 Commitment to phonological categories

Finally, we do see evidence for phonological commitment, resolving on a time-scale of ~300-450 ms (see Fig. 3). The superiority of defining latency in terms of elapsed ms rather than phonemes may indicate that commitment is based on the amount of time or number of completed processing cycles rather than intervening information (Fig. 3C). This process is supported by higher auditory processing regions in anterior STG, a location consistent with a recent meta-analysis of auditory word recognition (DeWitt and Rauschecker 2012). Critically, this seems to be computed in parallel to the maintenance of subphonemic detail in primary auditory regions. Before ~300 ms there is no cost associated with resolution to a lexical item less consistent with word onset: listeners do not get temporarily mislead (garden-pathed) provided resolution comes early enough (Fig. 3D). This suggests that the cohort of words consistent with either phonological interpretation is considered together (e.g., in the presence of b/p ambiguity, both the p-onset and b-onset words are activated). This is fully consistent with previous behavioural studies (Martin and Bunnell 1981, Gow 2001, Gow and McMurray 2007), and a previous eye-tracking study (McMurray, Tanenhaus et al. 2009), which used similar materials and found look-contingent responses to be dependent upon phonetic information at lexical onset until at least ~300 ms (the longest disambiguation delay they tested). However, after ~450 ms a cost begins to emerge when there is a mismatch between the more likely word given word-onset and the resolving lexical information (e.g., “barricade” is more likely if the onset phoneme was more b-like than p-like, so hearing “parakeet” is a mismatch). This plausibly reflects the recruitment of a repair mechanism, a prediction-error response or re-analysis of the input from making an incorrect commitment.

Finding maintained sensitivity to subphonemic detail in parallel to phonological commitment is very important for the interpretation of psychophysical research, which has implicitly equated insensitivity to within-category variation with phonological commitment (Connine, Blasko et al. 1991, Szostak and Pitt 2013, Bicknell, Tanenhaus et al. 2015). This previous work has largely converged on a processing model whereby phonological commitment can be delayed for over one second after onset. Our results indicate, in contrast, that while subphonemic detail is indeed maintained over large time-scales, this does not implicate that commitment is also put off for this length of time.

3.4 Relationship to models of speech processing

It is unclear which model of speech processing can account for these data. While Shortlist (Norris 1994) and Shortlist B (Norris and McQueen 2008) may be able to model recovery from lexical garden-paths, they do not explicitly model processing of subphonemic detail. While the MERGE model (Norris, McQueen et al. 2000) is capable of modelling such detail, it proposes no feedback from the lexical to phoneme levels of analysis, which is inconsistent with our observation that (sub-)phonemic representations are reactivated when top-down lexical information becomes available. Although it has been demonstrated that TRACE (McClelland and Elman 1986) can be modified to simulate recovery by removing phoneme-level inhibition (McMurray, Tanenhaus et al. 2009), it does not provide the architecture to model initial sensitivity to phoneme ambiguity, or account for how the percept of speech sounds is modulated by past and future linguistic information (see (Grossberg and Kazerounian 2011) for an overview of TRACE limitations). It is also unclear whether this modification would interfere with TRACE's success in accounting for a range of observations in spoken word recognition (see (Gaskell 2007) for a review). One model proposed to deal with TRACE's shortcoming is Adaptive Resonance Theory (ART): each speech sound produces a resonance wave that is influenced by top-down information until it reaches equilibrium and surfaces to consciousness (Grossberg 2003). While this theory is consistent with the idea that there is a critical time-limit to receive top-down information, it suggests that there is a linear decay in subphonemic information as temporal distance from the phoneme increases. Our results do not support that conjecture. Instead, they suggest that subphonemic information is re-evoked later in processing, with a similar magnitude as that experienced at onset. In light of the present results, one shortcoming of these models is their attempt to explain spoken word recognition with a single mechanism, built on the assumption that acoustic-phonetic information is lost once a phonological categorisation is derived. Instead, our results suggest that a three-element processing model is more appropriate, allowing for a dynamic interaction between phonetic, phonological and lexical levels of analysis.

3.5 Conclusion

Later sounds determine the perception of earlier speech sounds through the simultaneous recruitment of acoustic-phonetic and phonological computational pathways. This facilitates contact with lexical items in order to derive the message of the utterance, as well as continued revisitation to the phonetic level of analysis to reduce parsing errors. In this manner, lexical selection can be achieved rapidly, while also reducing the likelihood of mistakes in phonological segmentation. The human brain therefore solves the issue of processing a transient hierarchically structured signal by recruiting complementary computations in parallel, rather than conceding to the trade-off between speed and accuracy.

4. Materials & Methods

4.1 Experiment 1

4.1.1 Participants

Twenty-four right handed native English participants took part in the study (11 female; age: *M*=25.44, *SD*=8.44). They were recruited from the NYUAD community and were compensated for their time. All had normal or corrected vision, normal hearing and no history of neurological disorders.

4.1.2 Materials

Phonological transcriptions and by-phoneme frequency counts were extracted from the English Lexicon Project (ELP) (using scripts available online). We selected all word pairs whose first phoneme was a plosive, and differed either in terms of voicing {t-d, p-b, k-g}, or in terms of one place of articulation feature {t-k, p-t}. Any item that occurred in more than one pair was eliminated, resulting in 53 pairs: 31 VOT and 22 POA. Words ranged from 4-10 phonemes in length (M=6.8; SD=1.33), and from 291-780 ms in length (M=528; SD=97). Latency of disambiguation ranged from 3-8 phonemes (M=5.1; SD=0.97) and 142-708 ms (M=351; SD=92).

A native English speaker was recorded saying the selected words in isolation. Each item pair was exported into TANDEM-STRAIGHT for the morphing procedure (Kawahara, Morise et al. 2008, Kawahara and Morise 2011). In short, the morphing works by taking the follows steps: 1) position anchor points to mark the onset of each paired phoneme; 2) place weights on each anchor point to determine the % contribution of each word at the end-points of the continuum; 3) specify the number of continuum steps to generate.

In long, the process involves decomposing the auditory signal into its spectrotemporal components to make a confusion matrix between each auditory sample of the lexical pair. Then, anchor points are placed along this confusion matrix to mark temporal correspondence between the onset of phonemes in each word. For example, the onset occurs at different latencies for each of the seven phonemes of the words “barricade” and “parakeet”, but a single anchor point is placed to mark the correspondence in timing between the phonemes of the two words. This forms a morphing substrate that can be used to adjust the weighted contribution of the acoustics of each phoneme of each word. Then, the user specifies the anchor position weights at the two ends of the continuum. For example, to generate a “barakeet” <-> “parakeet” continuum, the onset anchor points will be placed at 100% “barricade” at one end of the continuum, and 100% “parakeet” at the other end of the continuum. Until and including disambiguation point (“arak-“), all anchors will be weighted at 50-50 of each word, and after disambiguation (“-eet”), all anchors are placed to be 100% “parakeet”, for both sides of the continuum. In general, for each pair, all anchor points before the POD phoneme were placed at the 50% position, and the first anchor point was positioned either in the congruent position, creating a word (“parakeet”) or with the competitor, creating a non-word (“barakeet”). This ensured that apart from the first phoneme, the acoustic signal remained identical across the two word pairs until the disambiguation point. Eleven continua steps were created for each item. The resulting 1166 auditory files were annotated in Praat for each phoneme onset using the Penn Forced Aligner, making manual adjustments as necessary. The first consonant-vowel sequence was then extracted for each item, and amplitude-normed to 70 dB.

4.1.3 Procedure

Stimuli were separated into eleven blocks using a latin-square design. Each block contained two morphed tokens from the stimuli pair, with the constraint that each item had to be at least three morphed steps away from its paired counterpart. This resulted in 106 trials per block. Items were randomised within each block.

Participants performed a single interval 2AFC classification on each item, while MEG was recorded concurrently. First, two options appeared (e.g., "B or P") on a screen 85 cm away from the participant's face. The auditory stimulus began 500 ms later, and the options remained on the screen until a left or right selection was made via button press using the left hand. After the response, a blank screen was displayed for 500 ms until the onset of the next trial. At each block interval, participants had a self-terminated break. The recording session lasted ~50 minutes.

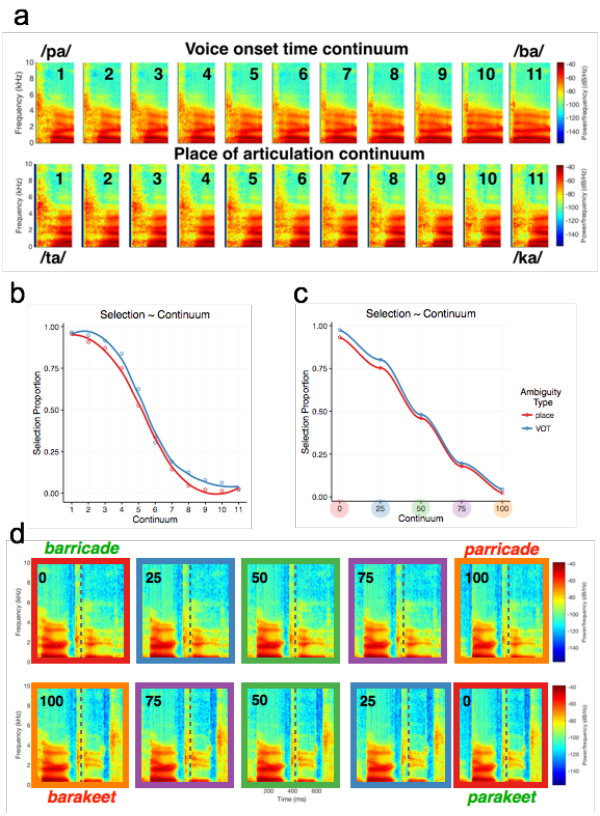

Figure 4. Stimuli production. (a) Example 11-step continuum for voice onset time (above) and place of articulation (below) contrasts used in Experiment 1. (b) Behavioural classification across each continuum step of Experiment 1. (c) Re-sampled continuum at five perception-defined thresholds. (d) Example word (green) - nonword (red) continuum; point of disambiguation is represented with a dashed line.

4.2 Experiment 2

4.2.1 Participants

Twenty-five right handed native English participants took part in the study (15 female; age: *M*=24.84, *SD*=7.3). Six had taken part in Experiment 1 two months earlier. All had normal or corrected vision, normal hearing, no history of neurological disorders, and were recruited from the NYUAD community.

4.2.2 Materials

Five out of the initial 11-step continuum were selected for each word pair based on the psychometric responses from Experiment 1. Placing a selection threshold at 0.05, 0.25, 0.5, 0.75 and 0.95, the closest morph position was extracted (see Fig. 4). Items were rejected if the unambiguous endpoints of the continuum were not categorised with at least 80% accuracy for all subjects, or if the position of the ambiguous token was not at least three points away from either end-point of the continuum. This resulted in 49 pairs. All remaining words were amplitude-normed to 70 dB.

4.2.3 Procedure

Each participant's head scan and digitised fiducial points were first collected, as described for Experiment 1.

Participants performed an auditory-to-visual word matching task on 1/5 of the auditory items. The visual word was either the target ("match" response) or the competitor ("mis-match" response). One item of the five morphed tokens would be a "match" trial and another would be a "non-match" trial; randomly assigned using a latin-square procedure. The experiment was split into 5 blocks, and only one of the morphed tokens from each item appeared in each block. The assignment of item-block was also pseudo-randomised in a latin-square fashion. This resulted in 25 unique experimental orders, across which, items were matched for block order, and match-nonmatch assignment.

A fixation cross was displayed for 500 ms, followed by the auditory word. If it was a task trial, the visual word would appear 500 ms after auditory word offset, and remain on screen until participants made a match (left button) or non-match (right button) decision with their left hand. If there was no task, a blank screen was presented and participants could move forward by pressing either button. The recording lasted ~40 minutes.

All participants' head shapes were digitised using a hand-held FastSCAN laser scanner (Polhemus, VT, USA) to allow for co-registration during data preprocessing. Five points on each participant's head were also digitised: just anterior of the left and right auditory canal, and three points on the forehead. Marker coils were later placed at the same five positions to localise each participant's skull relative to the sensors. These marker measurements were recorded just before and after the experiment in order to track the degree of movement during the recording.

MEG data were recorded continuously using a 208 channel axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan), with a sampling rate of 1000 Hz and applying an online low-pass filter of 200 Hz.

MEG data from the two experiments underwent the same pre-processing steps. First, the continuous recording was noise reduced using Continuously Adjusted Least Squares Method (CALM: [53]), with MEG160 software (Yokohawa Electric Corporation and Eagle Technology Corporation, Tokyo, Japan). The noise-reduced data, digital scan and fiducials, and marker measurements were exported into MNE-Python [54]. Bad channels were removed through visual inspection. Independent Component Analysis (ICA) was

computed over the noise-reduced data using FastICA in MNE-Python. Components were removed from the raw recording if they contained ocular or cardiac artefacts, which were identified based on the topography of magnetic activity and time-course response. The data were then epoched from 500 ms pre-stimulus to 1000 ms post-stimulus onset. Any trials whose amplitude exceeded +/- 2000 femto-tesla absolute or peak-to-peak threshold were removed.

In order to perform source localisation, the location of the subject's head was co-registered with respect to the sensory array in the MEG helmet. For subjects with anatomical MRI scans (n=4), this involved rotating and translating the digital scan to minimise the distance between the fiducial points of the MRI and the head scan. For participants without anatomical scans, the FreeSurfer "fsaverage" brain was used, which involved first rotation and translation, and then scaling the average brain to match the size of the head scan.

Next, a source space created, consisting of 2562 potential electrical sources per hemisphere. At each source, activity was computed for the forward solution with the Boundary Element Model (BEM) method, which provides an estimate of each MEG sensor's magnetic field in response to a current dipole at that source. The inverse solution was computed from the forward solution and the grand average activity across all trials. Data were converted into noise-normalised Dynamic Statistical Parameter Map (dSPM) units (see [55]), employing an SNR value of 2. The inverse solution was applied to each trial at every source, for each millisecond defined in the epoch, employing a fixed orientation of the dipole current that estimates the source normal to the cortical surface and retains dipole orientation.

Acknowledgements

This research was supported by ERC-2011-AdG 295810 BOOTPHON, ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC to TL; NIH 2R01DC05660 to DP; NYU Abu Dhabi Institute under grant G1001 to AM. We would like to thank Kyriaki Neophytou for her help with data collection.

Bibliography

Ackermann, H., W. Lutzenberger and I. Hertrich (1999). "Hemispheric lateralization of the neural encoding of temporal speech features: a whole-head magnetencephalography study." Cognitive Brain Research **7**(4): 511-518.

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

Blumstein, S. E., E. B. Myers and J. Rissman (2005). "The perception of voice onset time: an fMRI investigation of phonetic category structure." Journal of Cognitive Neuroscience **17**(9): 1353-1366.

Blumstein, S. E., K. N. Stevens and G. N. Nigro (1977). "Property detectors for bursts and transitions in speech perception." The Journal of the Acoustical Society of America **61**(5): 1301-1313.

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

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

Dahan, D. and M. Gareth Gaskell (2007). "The temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition." J Mem Lang **57**(4): 483-501.

DeWitt, I. and J. P. Rauschecker (2012). "Phoneme and word recognition in the auditory ventral stream." Proceedings of the National Academy of Sciences **109**(8): E505-E514.

Di Liberto, G. M., J. A. O'Sullivan and E. C. Lalor (2015). "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing." Curr Biol **25**(19): 2457-2465.

Ettinger, A., T. Linzen and A. Marantz (2014). "The role of morphology in phoneme prediction: evidence from MEG." Brain Lang **129**: 14-23.

Gage, N. M., T. P. Roberts and G. Hickok (2002). "Hemispheric asymmetries in auditory evoked neuromagnetic fields in response to place of articulation contrasts." Cognitive brain research **14**(2): 303-306.

Gagnepain, P., R. N. Henson and M. H. Davis (2012). "Temporal predictive codes for spoken words in auditory cortex." Curr Biol **22**(7): 615-621.

Ganong, W. F. (1980). "Phonetic categorization in auditory word perception." Journal of Experimental Psychology: Human Perception and Performance **6**(1): 110.

Gaskell, M. G. (2007). "Statistical and connectionist models of speech perception and word recognition." The Oxford handbook of psycholinguistics: 55-69.

Gordon, P. C., J. L. Eberhardt and J. G. Rueckl (1993). "Attentional modulation of the phonetic significance of acoustic cues." Cognitive Psychology **25**(1): 1-42.

Gow, D. W. (2001). "Assimilation and anticipation in continuous spoken word recognition." Journal of Memory and Language **45**(1): 133-159.

Gow, D. W., Jr. and J. A. Segawa (2009). "Articulatory mediation of speech perception: a causal analysis of multi-modal imaging data." Cognition **110**(2): 222-236.

Gow, D. W., Jr., J. A. Segawa, S. P. Ahlfors and F. H. Lin (2008). "Lexical influences on speech perception: a Granger causality analysis of MEG and EEG source estimates." Neuroimage **43**(3): 614-623.

Gow, D. W. and B. McMurray (2007). "Word recognition and phonology: The case of English coronal place assimilation." Papers in laboratory phonology **9**: 173-200.

Grossberg, S. (2003). Adaptive resonance theory, Wiley Online Library.

Grossberg, S. and S. Kazerounian (2011). "Laminar cortical dynamics of conscious speech perception: Neural model of phonemic restoration using subsequent context in noise." The Journal of the Acoustical Society of America **130**(1): 440-460.

Gwilliams, L., G. Lewis and A. Marantz (2016). "Functional characterisation of letter-specific responses in time, space and current polarity using magnetoencephalography." NeuroImage **132**: 320-333.
Gwilliams, L. and A. Marantz (2015). "Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words." Brain and language **147**: 1-13.
Halle, M. and K. Stevens (1962). "Speech recognition: A model and a program for research." IRE Transactions on Information Theory **8**(2): 155-159.
Hertrich, I., K. Mathiak, W. Lutzenberger and H. Ackermann (2000). "Differential impact of periodic and aperiodic speech−like acoustic signals on magnetic M50/M100 fields." NeuroReport **11**(18): 4017-4020.
Hickok, G. and D. Poeppel (2004). "Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language." Cognition **92**(1-2): 67-99.
Kawahara, H. and M. Morise (2011). "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework." Sadhana **36**(5): 713-727.
Kawahara, H., M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE.
Khalighinejad, B., G. C. da Silva and N. Mesgarani (2017). "Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech." Journal of Neuroscience **37**(8): 2176-2185.
Kilian-Hutten, N., G. Valente, J. Vroomen and E. Formisano (2011). "Auditory cortex encodes the perceptual interpretation of ambiguous sound." J Neurosci **31**(5): 1715-1720.
Liebenthal, E., J. R. Binder, S. M. Spitzer, E. T. Possing and D. A. Medler (2005). "Neural substrates of phonemic perception." Cereb Cortex **15**(10): 1621-1631.
Maris, E. and R. Oostenveld (2007). "Nonparametric statistical testing of EEG- and MEG-data." J Neurosci Methods **164**(1): 177-190.
Marslen-Wilson, W. and L. K. Tyler (1980). "The temporal structure of spoken language understanding." Cognition **8**(1): 1-71.
Marslen-Wilson, W. D. (1987). "Functional parallelism in spoken word-recognition." Cognition **25**(1): 71-102.
Martin, J. G. and H. T. Bunnell (1981). "Perception of anticipatory coarticulation effects." The Journal of the Acoustical Society of America **69**(2): 559-567.
McClelland, J. L. and J. L. Elman (1986). "The TRACE model of speech perception." Cognitive psychology **18**(1): 1-86.
McMurray, B., M. K. Tanenhaus and R. N. Aslin (2009). "Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition." J Mem Lang **60**(1): 65-91.
McQueen, J. M. (1991). "The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity." Journal of Experimental Psychology: Human Perception and Performance **17**(2): 433.
Mesgarani, N., C. Cheung, K. Johnson and E. F. Chang (2014). "Phonetic feature encoding in human superior temporal gyrus." Science **343**(6174): 1006-1010.
Myers, E. B. and S. E. Blumstein (2008). "The neural bases of the lexical effect: an fMRI investigation." Cereb Cortex **18**(2): 278-288.
Norris, D. (1994). "Shortlist: A connectionist model of continuous speech recognition." Cognition **52**(3): 189-234.

Norris, D. and J. M. McQueen (2008). "Shortlist B: a Bayesian model of continuous speech recognition." Psychol Rev **115**(2): 357-395.
Norris, D., J. M. McQueen and A. Cutler (2000). "Merging information in speech recognition: Feedback is never necessary." Behavioral and Brain Sciences **23**(03): 299-325.
Obleser, J., A. Lahiri and C. Eulitz (2003). "Auditory-evoked magnetic field codes place of articulation in timing and topography around 100 milliseconds post syllable onset." Neuroimage **20**(3): 1839-1847.
Obleser, J., A. Lahiri and C. Eulitz (2004). "Magnetic brain response mirrors extraction of phonological features from spoken vowels." Journal of Cognitive Neuroscience **16**(1): 31-39.
Papanicolaou, A. C., E. Castillo, J. I. Breier, R. N. Davis, P. G. Simos and R. L. Diehl (2003). "Differential brain activation patterns during perception of voice and tone onset time series: a MEG study." Neuroimage **18**(2): 448-459.
Poeppel, D. (2003). "The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'." Speech Communication **41**(1): 245-255.
Poeppel, D. and P. J. Monahan (2011). "Feedforward and feedback in speech perception: Revisiting analysis by synthesis." Language and Cognitive Processes **26**(7): 935-951.
Rauschecker, J. P. and S. K. Scott (2009). "Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing." Nat Neurosci **12**(6): 718-724.
Rogers, J. C. and M. H. Davis (2017). "Inferior Frontal Cortex Contributions to the Recognition of Spoken Words and Their Constituent Speech Sounds." J Cogn Neurosci: 1-18.
Samuel, A. G. (1981). "The role of bottom-up confirmation in the phonemic restoration illusion." Journal of Experimental Psychology: Human Perception and Performance **7**(5): 1124.
Samuel, A. G. (1991). "A further examination of attentional effects in the phonemic restoration illusion." The Quarterly Journal of Experimental Psychology **43**(3): 679-699.
Scott, S. K. and I. S. Johnsrude (2003). "The neuroanatomical and functional organization of speech perception." Trends in neurosciences **26**(2): 100-107.
Simos, P. G., R. L. Diehl, J. I. Breier, M. R. Molis, G. Zouridakis and A. C. Papanicolaou (1998). "MEG correlates of categorical perception of a voice onset time continuum in humans." Cognitive Brain Research **7**(2): 215-219.
Steinschneider, M., I. O. Volkov, M. D. Noh, P. C. Garell and M. A. Howard (1999). "Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex." Journal of neurophysiology **82**(5): 2346-2357.
Stevens, K. N. and S. E. Blumstein (1978). "Invariant cues for place of articulation in stop consonants." The Journal of the Acoustical Society of America **64**(5): 1358-1368.
Szostak, C. M. and M. A. Pitt (2013). "The prolonged influence of subsequent context on spoken word recognition." Atten Percept Psychophys **75**(7): 1533-1546.
Tavabi, K., J. Obleser, C. Dobel and C. Pantev (2007). "Auditory evoked fields differentially encode speech features: an MEG investigation of the P50m and N100m time courses during syllable processing." Eur J Neurosci **25**(10): 3155-3162.
Tyler, L. K. (1984). "The structure of the initial cohort: Evidence from gating." Perception & Psychophysics **36**(5): 417-427.
Warren, R. M. (1970). "Perceptual restoration of missing speech sounds." Science **167**(3917): 392-393.
Wolmetz, M., D. Poeppel and B. Rapp (2011). "What does the right hemisphere know about phoneme categories?" Journal of cognitive neuroscience **23**(3): 552-569.

Supplementary Materials

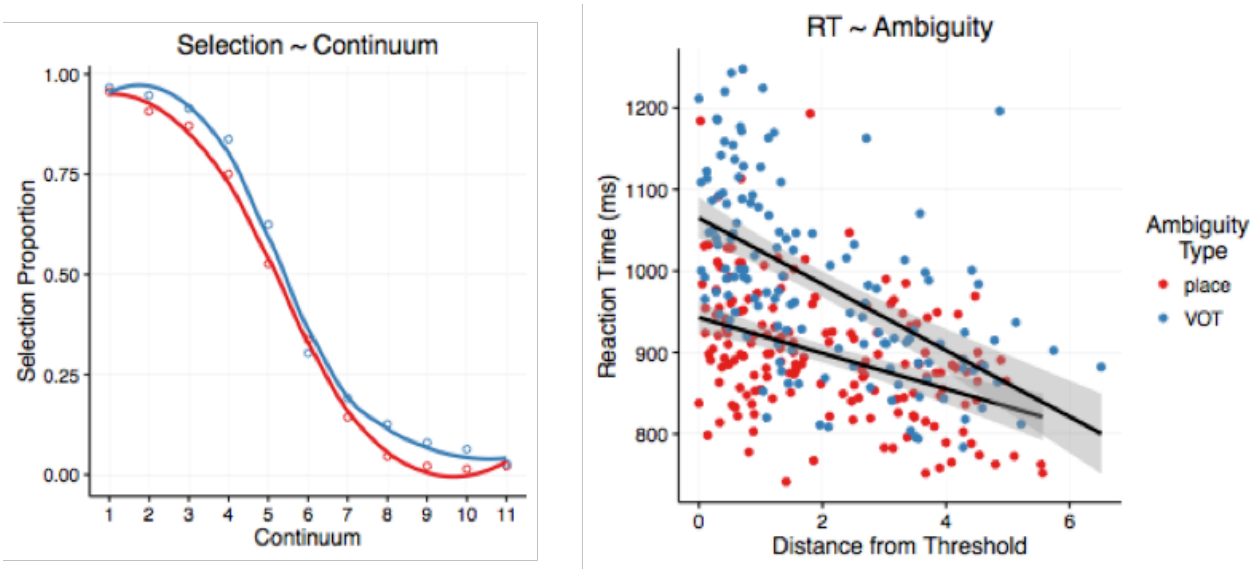

Figure 1. Behavioural phoneme selection and reaction time in experiment 1. Left: Proportion of positive selections as a function of the 11-step continuum. Right: Reaction time as a function of each trial's distance from the 50% selection threshold.

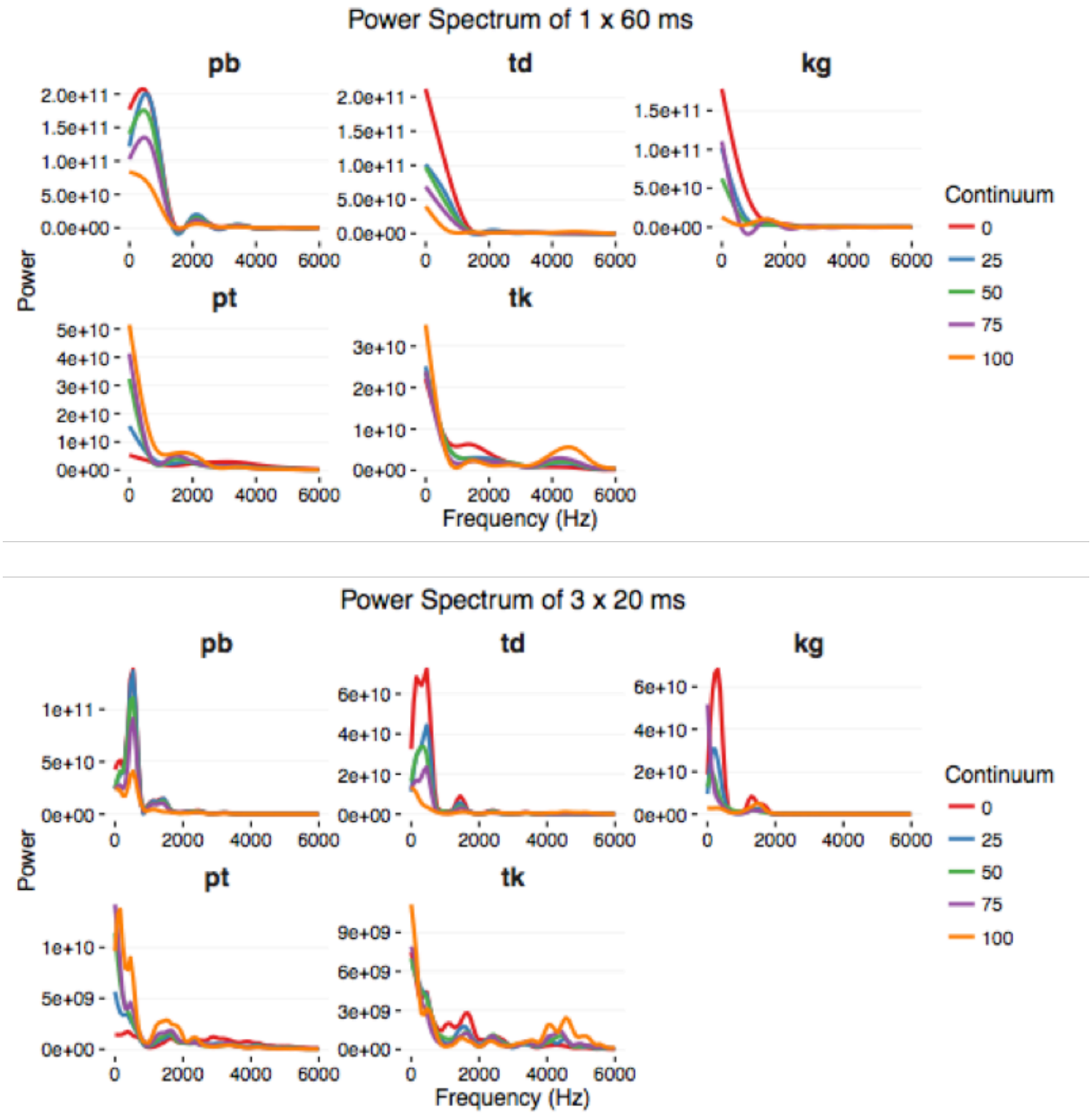

Figure 2. Frequency decomposition of first 60 ms of stimuli items. Above: Signal decomposed over a single 60 ms chunk. Below: Signal decomposed and averaged over three sequential 20 ms chunks.

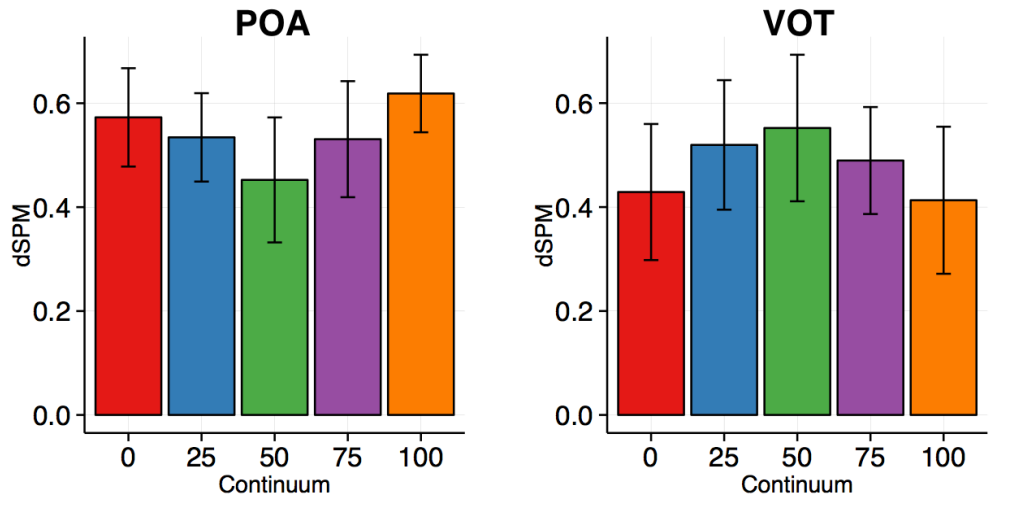

Figure 3. Early ambiguity effects in Heschl's gyrus in Experiment 2. Effects are shown separately for place of articulation continua (left) and voice onset time continua (right).

1. Behavioural results: Experiment 1

To analyse behavioural responses in Experiment 1, we applied a mixed effects regression analysis, coding for Feature Type (place of articulation [POA] vs. voice onset time [VOT]), position on the morphed continuum (henceforth "Acoustics"), Ambiguity (distance from 50% selection point), and their interaction terms (see *Materials & Methods* for detailed information on these variables). Reaction times were significantly slower for VOT continua than POA continua ($p < .001$). Responses significantly slowed with increased Ambiguity ($p < .001$). Ambiguity and Feature Type revealed a significant interaction ($p < .01$). Continuum had a marginal effect on reaction time ($p = .068$). A logistic regression was applied to behavioural selection with the same model structure. Feature Type was a significant predictor ($p < .001$), as well as Continuum ($p < .001$). There were no significant interactions. See Supplementary Figure 1.

2. Ambiguity and Acoustics at each level of Feature Type

2.1 Experiment 1: Syllable Onset

There was a significant interaction between Ambiguity and Feature Type, forming two clusters in the right hemisphere: 110-150 ms ($p = .033$); 155-200 ms ($p = .049$). The interaction between Acoustics and Feature Type also formed two significant clusters, one in the left hemisphere: 100-145 ms ($p = .045$) and one in the right: 105-170 ms ($p = .05$). For POA continua, one marginally significant cluster was found for Ambiguity in the right hemisphere: 80-130 ms ($p = .054$). For VOT, there was a marginally significant effect of Ambiguity 70-110 ms ($p = .057$) and Acoustics 85-115 ms ($p = .06$) both in the left hemisphere.

2.2 Experiment 2: Word Onset

A marginally significant interaction was found between Ambiguity and Feature Type in the left hemisphere: 176-200 ms ($p = .087$), and between Acoustics and Feature Type in the right hemisphere: 84-152 ms ($p < .005$); 84-148 ms ($p = .028$). For POA, no clusters were formed for Ambiguity but Acoustics modulated later responses in the left hemisphere: 178-200 ms ($p = .02$). Responses to VOT continua displayed a significant cluster for Ambiguity in the left hemisphere: 132-180 ms ($p = .021$), and Acoustics revealed two clusters in the right hemisphere: 88-152 ms ($p = .011$); 90-148 ms ($p = .058$).

2.3 Experiment 2: POD

There was a significant interaction between Ambiguity and Feature Type in the left hemisphere: 126-200 ms ($p < .001$), and a significant interaction between Acoustics and Feature Type in the left 60-152 ms ($p < .001$) and right hemisphere 72-168 ms ($p < .01$). Breaking down responses to each level of Feature Type revealed a significant effect of Ambiguity for POA in the left hemisphere: 132-182 ms ($p = .05$). Responses to VOT continua revealed a significant effect of both Ambiguity in the left hemisphere: 44-108 ms ($p < .001$), and Acoustics in the right hemisphere: 116-158 ms ($p = .06$).

3. Classifying phonemes based on first 20 ms of auditory stimuli

Sensitivity to Ambiguity at 50 ms after onset must be reflecting a response to no more than the first 20 ms of the acoustic signal — just the noise burst of the voiceless items, and the initial voicing of the voiced items. To assess what information is available at this latency, we decomposed the first 20 ms of each stimulus into its frequency power spectra using Fast Fourier Transform (FFT). Power at each frequency band from 0-10 KHz, for all stimuli except the fully ambiguous items, was used to train a logistic regression classifier to decode the phonological category (Fig. 1C). Accuracy was significantly above chance level, as determined by 1000 random permutations of phoneme labels ($p < .001$). Accuracy of classification decreased as a function of ambiguity (Fig. 1D), but all continua steps performed greater than chance. Importantly, continua steps themselves could not be decoded from this signal (Fig. 1E), suggesting that this early response indeed scales with distance from the perceptual boundary, and not acoustic properties *per se*.