When sentences live up to your expectations

Tuennerhoff, Johannes; Noppeney, Uta

DOI:
10.1016/j.neuroimage.2015.09.004

Document Version
Peer reviewed version

Citation for published version (Harvard):
Tuennerhoff, J & Noppeney, U 2016, 'When sentences live up to your expectations', NeuroImage, vol. 124, no. A, pp. 641-653. https://doi.org/10.1016/j.neuroimage.2015.09.004

When sentences live up to your expectations

Johannes Tuennerhoff, Uta Noppeney

PII: S1053-8119(15)00801-0
DOI: doi: 10.1016/j.neuroimage.2015.09.004
Reference: YNIMG 12561

To appear in: *NeuroImage*

Received date: 6 December 2014
Accepted date: 3 September 2015

Please cite this article as: Tuennerhoff, Johannes, Noppeney, Uta, When sentences live up to your expectations, *NeuroImage* (2015), doi: 10.1016/j.neuroimage.2015.09.004

When sentences live up to your expectations

Johannes Tuennerhoff[1*], Uta Noppeney[1,2]

[1]*Max-Planck-Institute for Biological Cybernetics, 72076 Tuebingen, Germany*
[2]*Computational Neuroscience and Cognitive Robotics Centre, Department of Psychology, University of Birmingham, Birmingham B15 2TT, UK*

Abbreviated title: Top-down predictions enable speech comprehension

*Corresponding Author:
Johannes Tuennerhoff
Max Planck Institute for Biological Cybernetics
Spemannstr. 41
72076 Tuebingen
Germany
Phone: +49-(7071)-601601
Fax: +49-(7071)-601616
Email: johannes.tuennerhoff@tuebingen.mpg.de

Abstract

Speech recognition is rapid, automatic and amazingly robust. How the brain is able to decode speech from noisy acoustic inputs is unknown. We show that the brain recognizes speech by integrating bottom-up acoustic signals with top-down predictions.

Subjects listened to intelligible normal and unintelligible fine structure speech that lacked the predictability of the temporal envelope and did not enable access to higher linguistic representations. Their top-down predictions were manipulated using priming. Activation for unintelligible fine structure speech was confined to primary auditory cortices, but propagated into posterior middle temporal areas when fine structure speech was made intelligible by top-down predictions. By contrast, normal speech engaged posterior middle temporal areas irrespective of subjects' predictions. Critically, when speech violated subjects' expectations, activation increases in anterior temporal gyri/sulci signalled a prediction error and the need for new semantic integration.

In line with predictive coding, our findings compellingly demonstrate that top-down predictions determine whether and how the brain translates bottom-up acoustic inputs into intelligible speech.

Keywords: speech recognition; predictive coding; speech intelligibility; priming

1. Introduction

Speech recognition is a seemingly effortless process despite background noise, inter-speaker variability or co-articulation patterns that preclude a simple one-to-one mapping between auditory signal and speech percept. To infer the most likely interpretation of the complex time-varying acoustic signal, the brain is challenged to integrate multiple probabilistic cues with prior expectations. Predictive coding models posit that speech recognition as perceptual inference emerges in the cortical hierarchy by iterative adjustment of top-down predictions against bottom-up sensory evidence (Davis and Johnsrude, 2007; Friston, 2005; Friston, 2010). Specifically, backwards connections provide predictions from higher to subordinate cortical levels. Conversely, forwards connections furnish the prediction error that is computed at each cortical level as the difference between top-down predictions and bottom-up inputs. The cortical architecture may thus recapitulate the hierarchical structure of speech that generates the acoustic inputs. While activations in low level auditory areas reflect prediction errors at the 'acoustic level', activations in higher order auditory areas reflect prediction errors at higher representational (e.g. phonological, semantic) level.

Speech processing is thought to rely on a left-biased frontotemporal system encompassing a dorsal and a ventral stream (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009). The dorsal stream projects to the frontoparietal cortices and is involved in auditory-motor integration to translate the incoming acoustic inputs into articulatory patterns. The ventral stream along the superior temporal sulcus maps the acoustic signals onto semantic representations. Indeed, intelligible normal speech increased activations in the superior/middle temporal gyri and sulci relative to a range of speech-like, yet unintelligible, control stimuli such

as rotated, noise-vocoded or temporally reversed speech (Binder et al., 2000; Crinion et al., 2003; Leff et al., 2008; Peelle et al., 2013; Scott et al., 2000). While the responses in lower level regions adjacent to primary auditory cortex depended on the particular form of speech degradation, they became progressively invariant to the specific stimulus manipulations in higher order areas and reflected primarily the signal's intelligibility (Davis and Johnsrude, 2003). However, since these studies compared intelligible speech with various forms of unintelligible degraded speech, they could not unambiguously dissociate effects of spectrotemporal structure and speech intelligibility.

Only few studies have investigated how speech intelligibility emerges from bottom-up inputs and top-down prior knowledge or expectations. Most of these studies have employed noise or noise-vocoding to render speech partially intelligible thereby allowing prior knowledge to enhance speech comprehension (e.g. (Obleser and Kotz, 2010; Sohoglu et al., 2012)). Yet, these experimental procedures enabled only a small increase in speech intelligibility (e.g. about 20% in (Obleser and Kotz, 2010)). Moreover, as participants were already able to understand 'degraded speech' at least to some extent, participants may have engaged greater attentional resources for comprehension of degraded speech (see (Wild et al., 2012b)). Another study used a written word to render noise vocoded auditory speech intelligible via crossmodal priming (Wild et al., 2012a). To our knowledge, only one very early study (Giraud et al., 2004) manipulated speech intelligibility for sentence stimuli more extensively by presenting participants with broad-band speech envelope noises that were initially unintelligible and rendered intelligible only after extensive practice. However, this experimental design

introduced temporal and training confounds rendering the intelligibility effect more difficult to interpret. Other studies have transformed syllables (Dehaene-Lambertz et al., 2005), single words (Möttönen et al., 2006) or sentences (Lee and Noppeney, 2011b, 2014) into sine wave speech stimuli that were processed as speech or non-speech depending on participants' prior experience. Collectively, these studies emphasized the role of anterior or posterior portions of superior temporal sulci in processing sine wave speech stimuli as speech relative to non-speech. To further investigate the role of prior expectations in speech processing the current study independently manipulated (i) bottom-up inputs by comparing normal and fine structure speech and (ii) subjects' top-down predictions using priming. Fine structure speech preserves the rapidly varying fine structure of the original speech, but lacks the temporal cues of the acoustic envelope. Critically, after additional bandpass filtering fine structure speech signals are generally perceived as noise. Yet, they can be rendered intelligible by prior top-down predictions via immediate priming, i.e. presenting the identical normal sentence directly before the fine structure stimulus (cf. Audio file A.1 in appendix).

To our knowledge this is the first neuroimaging study that generates 'unintelligible speech-like stimuli' by removing the information of the envelope. This approach allows us to compare normal speech with speech-like stimuli that preserve the fine structure information. Moreover, through additional filtering we were able to finetune the fine structure speech stimuli, such that they were only 5% intelligible throughout the entire experiment in the absence of prior knowledge (i.e. unprimed), yet 97% intelligible after priming. This novel auditory pop-out phenomenon enabled us to compare physically identical fine structure stimuli that were or were not intelligible depending on top-down prior predictions.

Based on the notion of predictive coding, we expected prediction errors at multiple levels of the speech processing hierarchy. The regional expression of the prediction error should depend on the prediction that is violated (Lee and Noppeney, 2011a). Further, predictions can be formed at multiple timescales ranging from milliseconds (e.g. online prediction of the next auditory spectrotemporal input) to seconds (e.g. prediction of the next sentence based on prior semantic context). First, we expected that fine structure speech relative to normal speech increases activations in low level auditory areas signaling the brain's failure to anticipate the auditory input in the absence of the temporal envelope. Similar to spatial grouping in the visual modality, we would expect the temporal envelope of speech to enable temporal grouping of auditory signals (i.e. *physical effect*, see (Murray et al., 2002) for a related argument in the visual modality). Thus, the temporal envelope enables moment-to-moment predictions of the incoming auditory signal. Second, prior top-down predictions (i.e. priming) render fine structure speech intelligible and enable speech recognition processes that are not engaged by unprimed fine structure speech. Similar to priming studies in the visual domain (Dolan et al., 1997; George et al., 1999; Henson et al., 2000; Henson, 2003), we would therefore expect enhanced activations for primed intelligible relative to unprimed unintelligible fine structure speech in higher order auditory areas reflecting the formation of novel linguistic representations (i.e. *perceptual effect,* cf. (Davis and Johnsrude, 2003)). From the perspective of predictive coding, these activation increases in higher order areas can be interpreted as prediction error signals that are elicited by the newly formed linguistic representations at a higher hierarchical level. Third, novel (i.e. unprimed) speech that violates subjects' prior semantic (or phonological, syntactic) expectations should elicit a greater response in the anterior superior temporal sulci

signalling a prediction error at a higher representational level. This prediction error at the sentential level emerges at a slower timescale acting from sentence to sentence. It predominantly indicates the need for new semantic integration at the highest cortical level (i.e. *novelty effect*, cf. (Dehaene-Lambertz et al., 2006)). Prediction errors as indexed by physical, perceptual and novelty effects can thus emerge at multiple temporal scales, representational and cortical levels (Werner and Noppeney, 2011).

Finally, we employed effective connectivity analyses (i.e. Dynamic Causal Modelling) to investigate how perceptual and novelty effects emerged from interactions amongst brain regions.

In summary, manipulating top-down predictions and bottom-up physical inputs enabled us to elicit prediction errors at multiple hierarchical levels and temporal scales thereby providing insights into how the brain generates semantic representations at the sentential level from acoustic inputs.

2. Materials & Methods

2.1. Subjects

20 healthy right-handed German native speakers (10 females; 10 males; median age: 24.05) participated in the fMRI study. Seven of those participants took also part in an additional psychophysics study inside the scanner. Eight right-handed German native speakers participated in the additional psychophysics studies performed outside the scanner (Study1: 5 male, 3 female, median age 33 years; Study2: 6 male, 2 female, median age 29.5 years). All

participants gave informed consent to participate in the study. The study was approved by the ethics committee of the university clinic in Tübingen.

2.2. Stimuli

Stimuli were 9-word-sentences (sentence duration: mean=3.8 sec., STD=0.23 sec., shortest=3.1 sec., longest=4.8 sec.) that were spoken by a male speaker. The sentences were not constrained to follow a particular syntax, include specific lexical items nor were they associated with a particular level of predictability. Sentences were recorded using an external Macintosh microphone (44100 Hz sampling rate). All sentences were low pass filtered at 2200 Hz and high pass filtered at 800 Hz with a 3rd order Butterworth filter in both directions. The fine structure sentences were generated from those bandpass filtered sentences based on the Hilbert transform. The Hilbert fine structure signal is $\cos\varphi\,(t)$, where $\varphi(t) = \tan^{-1}(\frac{s_i(t)}{s_r(t)})$ is the phase of the analytic signal $s(t) = s_r(t) + i\,s_i(t)$, $s_r(t)$ is the band pass filtered signal and $s_i(t)$ is the Hilbert transform of $s_r(t)$ (see (Drullman, 1995; Smith et al., 2002). Since the fine structure sentences were considered uncomfortably loud by participants, we reduced the sound energy (i.e. RMS) of fine structure relative to normal speech signal, so that they were matched in perceptual subjective loudness in an initial psychophysics study of 5 subjects. This adjustment procedure resulted in a RMS of 0.075 for bandpassed sentences and 0.044 for fine structure sentences (mean RMS).

Initial psychophysics studies outside the scanner environment suggested that bandpassfiltered fine structure sentence were intelligible only when preceded by their corresponding normal

sentence throughout the entire experiment. This intelligibility profile was confirmed in an additional behavioural study inside the scanner with the scanner noise being present.

2.3. Experimental Design

2.3.1. Main experiment (inside the scanner)

In the main experiment, subjects listened to normal and fine structure sentences that were arranged in pairs of two sentences. Participants were not informed of the arrangement of sentences into pairs to avoid strategic processing and investigate the role of top-down predictions on signal processing.

The experimental paradigm conformed to a 2 x 2 x 2 factorial design manipulating:

(1) Priming (2 levels): primed vs. unprimed,

(2) Spectrotemporal structure of the first sentence (2 levels): normal sentence vs. fine structure sentence,

(3) Spectrotemporal structure of the second sentence (2 levels): normal sentence vs. fine structure sentence.

This arrangement resulted in 8 conditions, i.e. 4 types of sentence pairs ($N_1 \rightarrow N_2$, $N_1 \rightarrow F_2$, $F_1 \rightarrow N_2$, $F_1 \rightarrow F_2$) that were either primed or unprimed. Critically, only priming with the correct normal sentence has a profound effect on processing of normal and fine structure speech. For normal sentences that are always intelligible, priming increases their processing efficiency (Henson, 2003). By contrast, for fine structure sentences that are unintelligible and sound like noise, priming renders them intelligible like pop out phenomena in the visual modality.

Therefore, our analysis focused primarily on the trials of the central 2 x 2 factorial design component where the 1st stimulus is a normal sentence (i.e. 4 conditions: $N_1 \rightarrow N_2$, $N_1 \rightarrow F_2$ either primed or unprimed). Please note that the 1st sentences and 2nd sentences of primed and unprimed pairs are counterbalanced across subjects. Hence, activation differences between primed and unprimed pairs can only emerge for the 2nd sentence because of priming.

Manipulating the spectrotemporal structure of the 2nd sentence (i.e. fine structure vs. normal sentence) and prior knowledge (i.e. priming: corresponding vs. non corresponding normal sentence) independently enabled us to dissociate the role of bottom-up inputs and top-down prior predictions in speech comprehension.

Subjects silently listened to blocks of six to nine sentences (ISI between sentences: 300ms) interleaved with 6.6s fixation periods. They were engaged in a target detection task where they responded to one particular fine structure and normal sentence that they had learnt prior to the experiment. This target detection task was employed to maintain subject's attention equally to both types of stimuli without confounding the 'speech activations' by task-induced processing (e.g. response selection). 10% of the sentences were target sentences. The target sentences were interspersed between (but not within) the sentence pairs within each block. Therefore, a sentence block could include even and uneven numbers of sentences.

Altogether, there were 320 non-target sentence stimuli (see list of sentences used as stimuli A.1 in appendix). Each stimulus was presented twice in the experiment, once in each form as 1st and 2nd sentence, amounting to 320 sentence pairs. The sentence pairs were equally distributed across $N_1 \rightarrow N_2$, $N_1 \rightarrow F_2$, $F_1 \rightarrow N_2$, $F_1 \rightarrow F_2$ types (i.e. 25% each). 50% of the trials were primed, i.e. they presented two corresponding or even identical 1st and 2nd sentences. The stimuli were

rotated and fully counterbalanced across conditions within and between subjects (i.e. the average sentence duration is identical across conditions). This counterbalancing ensured that differences for primed and unprimed stimuli cannot be caused by differences in stimuli, because the stimuli were identical between priming conditions in each participant.

2.3.2. Additional psychophysics experiment (inside the scanner)

To ensure that the intelligibility profile observed during the initial pilots was maintained inside the scanner with the scanner noise being present. Seven of the subjects that had participated in the main experiment took part in an additional psychophysics experiment at least 2 weeks after the main experiment. The experimental paradigm, scanning sequence (i.e. hence scanner noise) and presentation parameters were identical to the main experiment. Likewise, participants were not informed that the sentences were arranged in pairs. Yet, this time subjects indicated the intelligibility for each sentence via a two choice key press.

2.3.3. Additional psychophysics experiments (outside the scanner)

We characterized the effects and mechanisms of priming further in two additional psychophysics experiments that were performed outside. Each experiment was based on a subset of 40 sentences from the original 320 sentences. These forty sentences were used as the second target sentence where speech intelligibility was explicitly evaluated. For the unprimed sentence pairs the first prime sentence was selected from the remaining set of 280 sentences.

Experiment 1 revisited the question whether participants understood the finestructure sentence when it was primed by its corresponding normal sentence. The experimental

paradigm was basically equivalent to the main experiment. Yet, while in the psychophysics study inside the scanner participants judged speech intelligibility via a two choice key press, in this psychophysics experiment they explicitly typed the sentence that they had understood after the presentation of the second target sentence. Thus, this psychophysics study provided us with an explicit objective measure of speech intelligibility. However, we acknowledge that participants may also have typed words only when they believed they understood them from the finestructure sentence, when it was immediately preceded by the prime sentence. As in the main experiment the forty second target sentences were rotated across conditions across subjects to control for stimulus confounds. The interstimulus interval between 1st and 2nd sentence was 1 second. There was no time limit to respond to the second sentence. After the response the next stimulus pair followed. The percentage of correctly reported words was evaluated automatically via computational methods as well as by human judgments. Both methods provided basically equivalent results.

Specifically, we computed % of words correctly reported for i) $N_1 \rightarrow N_2$ primed, ii) $N_1 \rightarrow N_2$ unprimed, iii) $N_1 \rightarrow F_2$ primed and iv) $N_1 \rightarrow F_2$ unprimed.

The second experiment explored the priming mechanisms by which the corresponding normal sentence facilitates intelligibility of the subsequent finestructure sentence. The corresponding normal sentence provides multiple sorts of information and thus top-down constraints that may facilitate processing of the finestructure sentence. The preceding normal sentence provides the envelope template that has been removed from the finestructure sentence, semantic, phonological and syntactic constraints and is spoken by the same voice. To further determine the information that enables comprehension of finestructure speech we have

therefore manipulated the priming sentence. We included four different types of presentations of the priming sentence:

(i) the corresponding spoken normal sentence; this provides the exact corresponding envelope template, semantic, syntactic, phonological and lexical constraints and identical voice,

(ii) the corresponding written sentence (with same syntax and lexical items); via internal speech this provides a less refined envelope template, semantic, lexical and syntactic constraints, but the voice information is missing

(iii) a written sentence with same lexical items but different syntactic structure; this provides segments of the envelope template, lexical and semantic priming

(iv) a written sentence with synonyms and different syntactic structure; this sentence still provides comparable semantic constraints, yet the temporal envelope template and phonological constraints are removed. Likewise, the exchange of content words precludes lexical priming.

2.4. fMRI

A 3T SIEMENS MAGNETOM TrioTim System (Siemens, Erlangen, Germany) was used to acquire both T1-weighted anatomical image (176 sagittal slices, TR = 1900 ms, TE = 2.26 ms, TI = 900 ms, flip angle = 9°, FOV = 256 x 224 mm, image matrix = 256 x 224, voxel size = 1 x 1 x 1 mm^3) and T2*-weighted axial echoplanar images with blood oxygenation level-dependent (BOLD) contrast (gradient echo, TR = 3080 ms, TE = 40 ms, flip angle = 90°, FOV = 192 x 192 mm, image matrix 64 x 64, 38 slices acquired in ascending direction, voxel size = 3.0 mm x 3.0 mm x 2.6 mm

+ 0.4 mm interslice gap). There were six sessions with a total of 174 volume images per session on average. The first 4 volumes were discarded to allow for T1-equilibration effects.

2.4.1. Conventional SPM analysis

The data were analyzed with statistical parametric mapping (using SPM8 software from the Wellcome Department of Imaging Neuroscience, London; http://www.fil.ion.ucl.ac.uk/spm) (Friston et al., 1994). Scans from each subject were realigned using the first as a reference, unwarped, spatially normalized into MNI standard space, resampled to 2x2x2 mm^3 voxels and spatially smoothed with a Gaussian kernel of 6 mm FWHM. The timeseries of all voxels were highpass filtered to 1/128 Hz. The fMRI experiment was modelled in an event related fashion with regressors entered into the session-specific design matrix after convolving each unit impulse with a canonical hemodynamic response function. In addition to modelling the 1^{st} and the 2^{nd} sentence separately for each type of sentence pair in our 2 x 2 x 2 factorial design (i.e. amounting to 2 x 8 = 16 regressors), the statistical model for each session included the two different types of detection targets (i.e. the fine structure and the normal target sentence) and the realignment parameters as nuisance covariates (to account for residual motion artefacts). Condition-specific effects for each subject were estimated according to the general linear model and passed to a second-level analysis as contrasts.

This involved creating the following contrast images on the first level (summed over sessions) to address our scientific questions of interest directly (see introduction):

(1) Fine structure- and normal speech-preferential activations were identified by comparing fine structure and normal sentences (limited to the first sentence of each sentence pair to avoid

effects of priming; fine structure preferential: all $F_1 > N_1$; normal sentence preferential: all $N_1 > F_1$)

(2) Physical Effects were identified by comparing primed fine structure and primed normal sentences that were nearly matched in terms of their intelligibility, but differed in their spectrotemporal structure (primed F_2 > primed N_2).

(3) Perceptual Effects were identified by comparing primed and unprimed fine structure sentences that were matched in terms of their spectrotemporal structure, but differed in their intelligibility (primed F_2 > unprimed F_2).

(4) Novelty Effects were identified by comparing primed and unprimed normal sentences that were matched in terms of their spectrotemporal structure and intelligibility, but differed in their novelty (primed N_2 < unprimed N_2).

Please note that the statistical comparisons for physical, perceptual and novelty effects were limited to the 2nd sentences that were preceded by the corresponding (= primed) or a different (= unprimed) normal sentence (i.e. only primed and unprimed sentence pairs of types $N_1 \rightarrow N_2$ or $N_1 \rightarrow F_2$ were considered).

These contrasts were entered into a second level one-sample t-test and inferences were made at the second level to allow a random effects analysis and inferences at the population level.

Unless otherwise stated, we report activations at $p<0.05$ cluster level corrected for multiple comparisons within the entire brain (with an auxiliary $p<0.001$ voxel threshold). To focus on activations within the sentence processing system, we only report significant activations

(whole-brain corrected) within an implicit mask defined by the normal target sentence relative to fixation at $p < 0.001$ uncorrected).

2.4.2. Effective Connectivity Analysis: DCM

For each subject, 18 bilinear DCMs (Friston et al., 2003) were constructed. Each DCM included 3 regions in a 3-level cortical hierarchy: i) the left Heschl's gyrus that responded more to primed fine structure sentences than primed bandpassed sentences (i.e. physical effect: [HG]; x = -38, y = –32, z = 14), ii) the left posterior superior temporal gyrus that showed increased activation for intelligible (i.e. primed) relative to unintelligible (i.e. unprimed) fine structure speech (i.e. perceptual effect: [post. STG], x = –46, y = –38, z = 10) and iii) the left anterior superior temporal gyrus / sulcus showing increased activations for novel relative to primed normal sentences (i.e. novelty effect: [ant. STS]; x = –60, y = -6, z = -2). The left inferior frontal gyrus / frontal operculum has previously been implicated in speech recognition making it an additional candidate region. However, frontal activations can be enhanced or partly induced by concurrent task demands during speech processing (see (Crinion et al., 2003) for further discussion). This might be the reason why our study that employed a target detection task revealed only weak frontal activations extending from a larger activation cluster in the superior temporal gyrus. Given these weak and unreliable frontal activations, we therefore decided not to include the left inferior frontal gyrus as a key region within the DCM.

The regions were selected using the maxima of the relevant contrasts from our random effects analysis. Region-specific time-series (concatenated over the six sessions and adjusted for confounds) comprised the first eigenvariate of all voxels within a 4 mm radius sphere centered

on the subject-specific peak in the relevant contrast. The subject-specific peak was uniquely identified as the maximum within the relevant contrast (no additional thresholding was applied) in a particular subject in an 8 mm radius sphere centered on the peak coordinates from the group random effects analysis.

In all models, the Heschl's gyrus was bidirectionally connected with the posterior STG and the posterior STG was bidirectionally connected with the anterior STG/STS. Further, fine structure and normal speech stimuli entered as separate extrinsic inputs to Heschl's gyrus. The timings of the onsets were individually adjusted for each region to match the specific time of slice acquisition.

From this basic Dynamic Causal Model we then generated the 18 = 2 x 3 x 3 competing candidate DCMs by factorially manipulating:

i) The presence vs. absence of an additional bidirectional connectivity from Heschl's gyrus directly anterior STG/STS. This allowed us to investigate whether speech comprehension emerges in a serial (i.e. absence of additional connection) or parallel (i.e. presence of additional connection) processing architecture

ii) The connection that is modulated by an intelligible speech percept (forwards, backwards or bidirectional connectivity between Heschl's gyrus and posterior STG)

iii) The connection that is modulated by the novelty of a normal sentence (forwards, backwards or bidirectional connectivity between posterior STG and anterior STG/STS)

2.4.3. Bayesian Model Comparison

To determine the most likely of the 18 bilinear DCMs given the observed data from all subjects, Bayesian model selection was implemented in a random effects group analysis using a hierarchical Bayesian model that estimates the parameters of a Dirichlet distribution over the probabilities of all models considered (DCM 10 implemented in SPM8). At the random-effects level, we report (1) the expectation of this posterior probability and (2) the exceedance probability of one model being more likely than any other model tested (Stephan et al., 2009).

We employed two approaches for Bayesian Model selection: First, we compared the model families separately for each of the three factors that generate our 2 x 3 x 3 model space. Second, we determined the optimal model. For the optimal model, the subject-specific modulatory, extrinsic and intrinsic connection strengths were entered into t-tests at the group level. This allowed us to summarize the consistent findings from the subject-specific DCMs using classical statistics:

Model comparison and statistical analysis of connectivity parameters of the optimal model enabled us to address the following two questions: First, we investigated whether speech is processed in a serial or parallel processing architecture (= presence or absence of additional connectivity between Heschl's gyrus and anterior STG/STS). Second, we investigated whether the perceptual and novelty effects were mediated via increased forwards, backwards or bidirectional connectivity. Based on the notion of predictive coding, we expected that prediction errors (i.e. novelty effect) are furnished by increased forwards connections, while top-down effects induced by the availability of the prior envelope template, lexical and phonological constraints (i.e. perceptual effect) are mediated by increased backwards connections.

3. Results

Subjects were presented with normal and fine structure sentences that –unknown to the subject– were arranged in pairs of two sentences. In the main factorial design (shown in figure 1a and 1b), the 1st sentence was always a normal sentence, while the 2nd sentence could be either a normal or a fine structure sentence. Further, both the fine structure and the normal 2nd sentence could be corresponding (= primed) or unrelated (= unprimed) to the 1st normal sentence.

In short, this factorial design manipulated the bottom-up acoustic signal of the 2nd sentence (normal vs. fine structure) and the top-down constraints using priming (primed vs. unprimed).

In the main fMRI experiment, subjects were engaged in a target detection task (see below), so that the sentence activations were not confounded by task-related processes. Yet, to evaluate the effect of spectrotemporal structure and priming on speech intelligibility, additional psychophysics experiments were carried out outside the scanner and inside the scanner (i.e. with scanner noise present), where subjects had to indicate whether they understood the sentence.

3.1. Speech comprehension (Additional psychophysics experiment inside the scanner)

To ensure that this intelligibility profile was also observed in the context of the scanner noise and preserved throughout the entire experiment, we presented seven subjects with the sentences from the main experiment during scanning and asked them to explicitly judge the intelligibility of the normal and fine structure sentences. As shown in figure 1 c, the fine structure sentences were indeed only intelligible, when primed. In contrast, normal sentences

were equally intelligible irrespective of priming. A repeated measures ANOVA with the factors (1) Spectrotemporal structure of 2nd sentence: Normal vs. fine structure and (2) Priming: Primed vs. unprimed confirmed this impression statistically and revealed a significant interaction between signal type (normal vs. fine structure) and priming ($F(1,6) = 1891.519$, $p<0.001$). Thus, priming altered processing of fine structure and normal speech in different ways. While normal speech was always intelligible irrespective of priming, priming rendered fine structure speech first and foremost intelligible. Given these different priming effects on fine structure and normal speech, our functional imaging analysis directly investigated the neural processes that enable them.

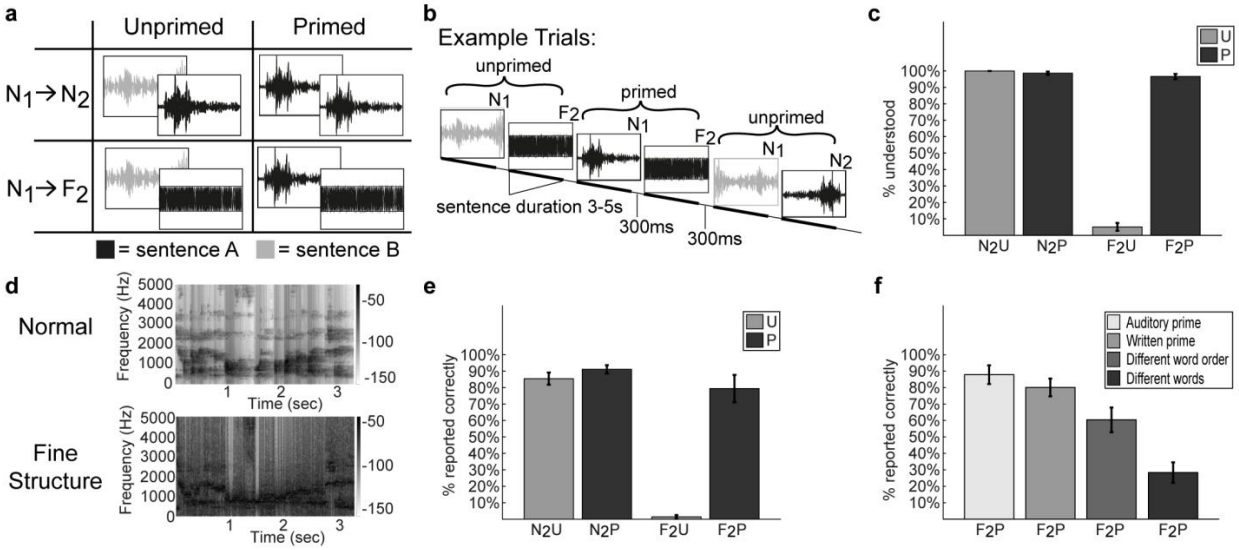

Figure 1

Experimental paradigm, example trial & behavioural results

a. Experimental paradigm. The central 2 x 2 factorial design component manipulated (1) Priming: primed vs. unprimed, (2) Spectrotemporal structure of the 2nd sentence: normal speech vs. fine structure speech sentence. In this central design component, the 1st sentence was always a normal sentence. This generated four trial types: Unprimed $N_1 \rightarrow F_2$, Primed $N_1 \rightarrow F_2$, Unprimed $N_1 \rightarrow N_2$ and Primed $N_1 \rightarrow N_2$. N = normal speech; F = fine structure speech. (n.b. additional four conditions where the 1st sentence is always a fine structure sentence are not shown, as they are not relevant for the main rationale of this study)

b. Example trials and timing of 6 sentences arranged in 3 sentence pairs. Each normal and fine structure

sentence is represented by their characteristic sound waveform. The temporal envelope of the fine structure sentence is flat (i.e. a constant value)
c. Behavioural results (from additional psychophysics experiment inside the scanner). Sentence comprehension: % understood of the 2nd sentence in each of the four conditions (across-subjects' mean ±SEM). The fine structure sentences were only intelligible when primed with the corresponding normal sentence. When unprimed they remained unintelligible throughout the entire course of the experiment.
2nd normal speech sentence unprimed = N_2U; 2nd normal speech sentence primed = N_2P; 2nd fine structure speech sentence unprimed = F_2U; 2nd fine structure speech sentence primed = F_2P.
d. Timefrequency spectrograms of example normal and fine structure sentences.
e. Behavioural results (from psychophysics experiment 1 outside the scanner). Sentence comprehension: % words reported (i.e. typed) correctly for the 2nd sentence in each of the four conditions (across-subjects' mean ±SEM). Replicating the results from the psychophysics study inside the scanner the fine structure sentences were only intelligible when primed with their corresponding normal sentence. When unprimed they remained unintelligible throughout the entire course of the experiment.
2nd normal speech sentence unprimed = N_2U; 2nd normal speech sentence primed = N_2P; 2nd fine structure speech sentence unprimed = F_2U; 2nd fine structure speech sentence primed = F_2P.
f. Behavioural results (from psychophysics experiment 2 outside the scanner). Sentence comprehension: % words reported correctly for the 2nd fine structure sentence (across-subjects' mean ±SEM) that was preceded by four different types of primes: (i) corresponding spoken normal sentence, (ii) corresponding written sentence (i.e. identical syntax and lexical items), (iii) written sentence with identical lexical items (i.e. content words) but different syntactic structure and (iv) written sentence with synonyms and different syntactic structure (i.e. similar semantics).
2nd fine structure speech sentence primed = F_2P.

3.1.1. Additional psychophysics experiments (outside the scanner)

In the psychophysics study inside the scanner, participants judged the intelligibility of the sentences via a two choice key press. In the first psychophysics study outside the scanner we evaluated whether participants indeed understood primed finestructure speech using an objective procedure where participants typed in the sentence that they understood. This first psychophysics study replicated the results observed inside the scanner. As shown in figure 1 e, the fine structure sentences were indeed only intelligible, when primed (% words correctly reported, across subjects' mean ± SEM for finestructure speech primed: 81.1% ±8.6% and unprimed: 1.4% ± 1.1%). In contrast, normal sentences were similarly intelligible irrespective of priming ((% words correctly reported, across subjects' mean ± SEM for normal speech primed:

94.2% ±2.6%and unprimed: 88.3% ±3.8%). A repeated measures ANOVA with the factors (1) Spectrotemporal structure of 2nd sentence: Normal vs. fine structure and (2) Priming: Primed vs. unprimed confirmed this impression statistically and revealed a significant interaction between signal type (normal vs. fine structure) and priming ($F(1,7) = 94.596$, $p<0.001$).

The second psychophysics experiment further characterized the mechanism of priming by manipulating the presentation mode and structure of the prime sentence. This experiment replicated that the spoken corresponding normal sentence facilitates comprehension of the subsequent finestructure sentence (87.9%, SEM=5.7%). When the prime sentence is presented in a written fashion, the intelligibility decreases to 80.1% (SEM=5.4%) which is significantly less than the intelligibility score obtained for spoken sentence primes ($p=0.024$). Via internal speech the written sentence can provide only a less refined envelope template suggesting that priming emerges partly by the envelope constraining spectrotemporal processing. Indeed, when the semantic content is preserved but expressed using a different syntactic structure (i.e. condition 3), finestructure speech intelligibility decreases to 60.4% (SEM=7.5). Again, the syntactic structure in part determines the temporal envelope. In sentences with different syntactic structures, the temporal envelope is identical between prime and target sentence only for envelope segments pertaining to the lexical items (i.e. preserved content words). Finally, in condition 4 we express similar semantic content using synonyms thereby using prime sentences with a completely different envelope that largely preserve the semantic content of the target sentence. Despite this strong semantic priming the intelligibility now drop to even 28.3% (SEM=6.2%) of the words reported correctly. Collectively, this experiment demonstrates that prime sentences with nearly identical semantic content have only very limited influence on the

intelligibility of finestructure speech sentences. This suggests that the neuroimaging results for finestructure speech are unlikely to result predominantly from semantic priming. Further, the prime manipulations depend on various linguistic and non-linguistic factors such as speaker's voice, syntactic structure and lexical items. Yet, all of them also reduce the availability of the temporal envelope template. Collectively, this study suggests that the availability of the temporal envelope may play a critical role in priming finestructure speech; yet, additional linguistic factors cannot be excluded. In particular the large intelligibility reduction induced by exchanging the lexical items suggests that lexical priming may also be an important contributing factor. Future, more finegrained studies may include all these different priming conditions in the fMRI experiment.

3.1.2. Target Detection (Main experiment inside scanner)

In the main fMRI experiment, subjects were engaged in a target detection task to maintain their attention. Specifically, they responded to presentations of one particular fine structure sentence and one normal sentence that they encoded prior to the fMRI study. A low level target detection task was used to ensure that the activations observed for speech processing were not confounded by task-induced processing (n.b. the target sentences were modelled separately in the general linear model, see methods), whilst controlling subjects' attention at least to some degree.

The % accuracy for the fine structure sentence (0.92 %, STD = 0.11) and the normal sentence (0.96 %, STD = 0.07) were both near ceiling. This performance during the target detection task suggests that subjects attended to the speech and fine structure stimuli comparably.

3.2. Conventional SPM analysis

The conventional SPM analysis was performed in two steps: First, we directly compared fine structure and normal speech. Second, we further characterized activation differences by dissociating physical, perceptual and novelty effects.

Table 1 Activations preferential for fine structure and normal speech

Direct comparison between unintelligible fine structure and normal intelligible speech limited to the first sentence in each sentence pair.

Region		voxel	p_{FWE}-value (cluster)	z-score	MNI Coordinates		
					x	y	z
Fine structure-preferential activations: $F_1 > N_1$							
left Heschl's g. / planum temporale		1382	0.000	7.43	-36	-30	16
left posterior superior temporal g.				6.47	-40	-26	6
right Heschl's g. / planum temporale		1630	0.000	6.83	40	-30	16
right superior temporal g.				5.36	58	-18	12
left middle frontal g.		353	0.001	5.77	-40	46	10
medial prefrontal cortex / pre-supplementary motor area		456	0.000	5.79	8	34	32
right middle frontal g.		570	0.000	4.87	34	50	16
post. cingulate / corpus callosum		526	0.000	5.09	-2	-30	24
right insula		480	0.000	4.65	32	26	8
right inferior parietal g. / supramarginal g.		1231	0.000	5.40	42	-48	40
left inferior parietal g.		941	0.000	5.98	-56	-52	46
right middle frontal g. / superior prefrontal s.		344	0.001	4.42	44	18	34
Normal sentence-preferential activations: $N_1 > F_1$							
left sup. temporal g. / s.		3118	0.000	6.53	-46	16	-20
left inferior frontal g.	orbital part			4.70	-50	26	-8
	triangular part			3.63	-44	28	0
right superior temporal g. / s.		1471	0.000	6.34	48	18	-24
				4.97	54	-24	-4
left putamen		669	0.000	5.45	-24	-2	-6

N_1 = 1st normal sentence; F_1 = 1st fine structure sentence.

3.2.1. Activations preferential for unintelligible fine structure and normal speech

First, we identified activations preferential for either unintelligible fine structure or intelligible normal speech. To ensure that activation differences were not influenced by any priming effect, we directly compared unintelligible fine structure and normal intelligible speech limited to the first sentence in each sentence pair (see methods).

Unintelligible fine structure speech increased activations relative to intelligible normal speech in Heschl's gyri extending into the planum temporale bilaterally. Based on probabilistic cytoarchitectonic maps, parts of the activations were located in subdivisions Te1.0, Te1.1 and Te1.2 of human primary auditory cortex (Morosan et al., 2001). More specifically, 55.9% of TE 1.0 and 99.1 % of TE 1.1 were activated in the left hemisphere and 54.9% of TE 1.0 and 75.5% of TE 1.1 in the right hemisphere (see also Table 1).

Further, fine structure speech increased activations in bilateral (pre)frontal and inferior parietal cortices, posterior cingulate, right insula and supramarginal gyri relative to normal speech. However, these latter activation differences resulted primarily from deactivations for normal sentences (relative to fixation) and will therefore not be discussed further.

Intelligible normal speech relative to unintelligible fine structure speech increased activations primarily in the superior temporal sulci/gyri bilaterally extending into the temporal poles. In addition, we observed activations in putamen and inferior frontal gyri (see also Table 1). Thus, the comparison of normal and fine structure speech revealed activations in the neural systems that have previously been reported for intelligible speech relative to a range of speech-like control stimuli and hence been referred to as 'intelligibility' areas (Binder et al., 2000; Davis and

Johnsrude, 2003; Friederici et al., 2010; Hickok and Poeppel, 2007; Horwitz and Braun, 2004; Narain et al., 2003; Obleser and Kotz, 2010; Obleser et al., 2007a; Scott et al., 2000).

Yet, since fine structure and normal speech differ both in intelligibility and in spectrotemporal structure, these activation differences cannot unambiguously be attributed to speech intelligibility but may also result from differences in bottom-up acoustic inputs. To understand how speech comprehension emerges from interactions between bottom-up acoustic inputs and top-down predictions, we manipulated spectrotemporal structure (i.e. fine structure vs. normal speech) and subjects' prior knowledge (i.e. primed vs. unprimed) independently. Specifically, this enabled us to dissociate the effects of physical structure, intelligible speech percept and novelty.

Table 2 Physical, Perceptual and Novelty effects

Region		voxel	p_{FWE}-value (cluster)	z-score	MNI Coordinates		
					x	y	z
Physical effect: $F_2P > N_2P$							
left Heschl's g. / superior temporal g.		2833	0.000	6.75	-38	-32	14
right Heschl's g. / superior temporal g.		2744	0.000	7.15	60	-18	6
Perceptual effect: $F_2P > F_2U$							
left superior temporal g. / s.	posterior	3576	0.000	5.45	-46	-38	10
left planum temporale				4.81	-54	-26	2
left middle temporal g.	posterior			4.51	-56	-56	2
left inferior parietal cortex				4.58	-52	-36	44
left superior temporal g.	anterior			4.26	-56	8	0
	middle			4.17	-44	-6	-10
left inferior frontal g.				3.44	-56	22	2
right superior temporal g.	posterior	2219	0.000	5.24	66	-22	6
	anterior			3.74	58	4	-12
Novelty effect: $N_2U > N_2P$							
left superior temporal g. / s.	anterior	437	0.002	4.90	-60	-6	-2
right superior temporal g. / s.	middle	244	0.031	4.16	56	-2	-8

N_2U = 2^{nd} normal speech sentence unprimed; N_2P = 2^{nd} normal speech sentence primed; F_2U = 2^{nd} fine structure speech sentence unprimed; F_2P = 2^{nd} fine structure speech sentence primed.

3.2.2. *Physical, perceptual & novelty effects*

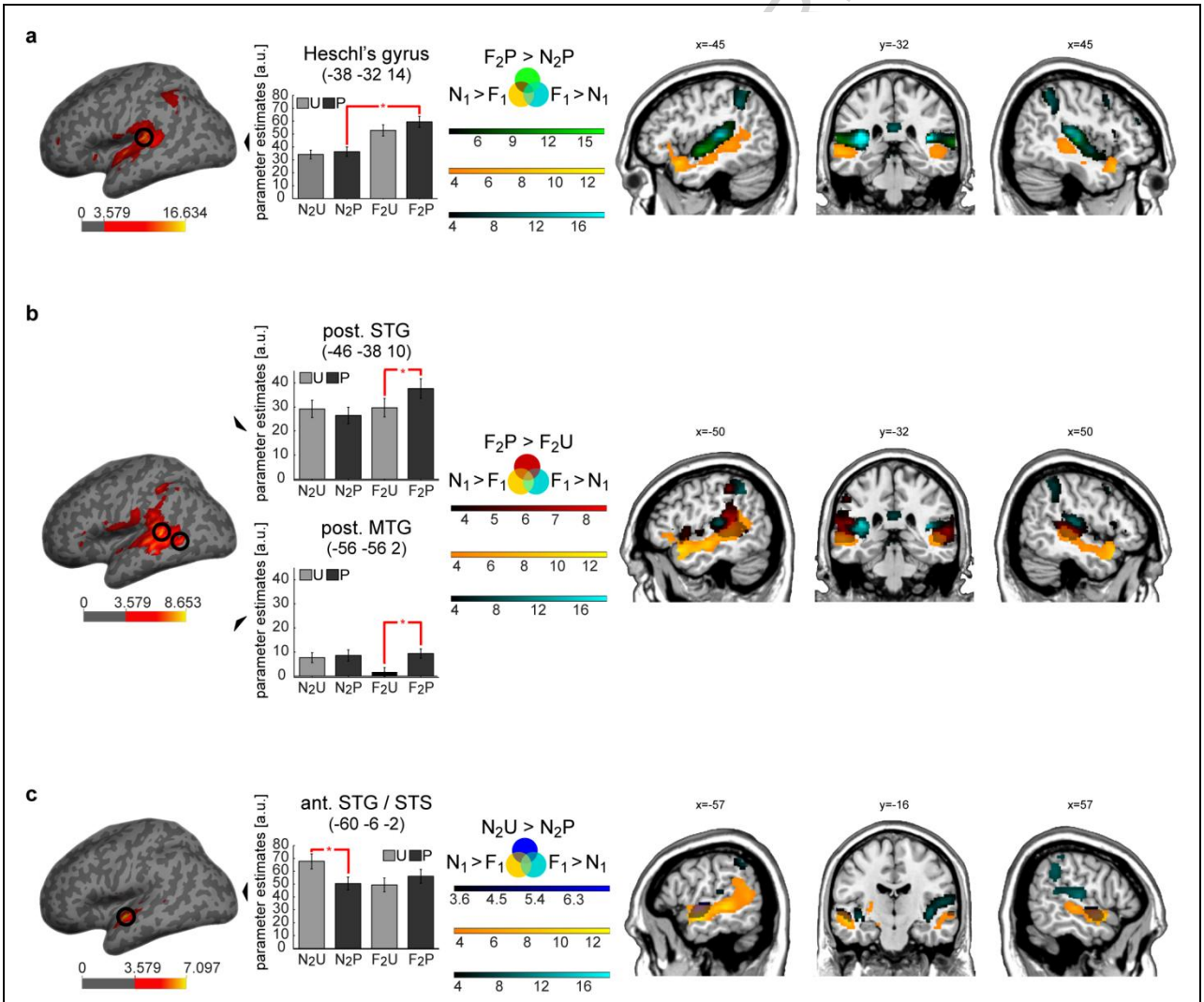

Figure 2

fMRI results – Physical effect (a), perceptual effect (b) and novelty effect (c)

a. Increased activations for primed intelligible fine structure relative to primed normal speech sentences ($F_2P > N_2P$). Primed intelligible fine structure and primed normal sentences are nearly matched in terms of speech intelligibility and novelty, but differ in their bottom-up spectrotemporal inputs.

b. Increased activations for primed intelligible relative to unprimed unintelligible fine structure speech sentences ($F_2P > F_2U$). Primed and unprimed fine structure sentences are matched in terms of bottom-up spectrotemporal inputs, but differ in terms of speech intelligibility and hence top-down predictions.

c. Increased activations for unprimed (novel) normal relative to primed normal speech sentences ($N_2U >$

N_2P). Primed and unprimed normal sentences are matched in terms of spectrotemporal structure and speech intelligibility, but differ in whether a new representation at the sentential level needs to be generated.
Left: Physical effect (a), perceptual effect (b) and novelty effect (c) are rendered on an inflated brain. Height threshold, $p < 0.001$ uncorrected masked with normal target sentence > fixation at $p < 0.001$ uncorrected. Extent threshold > 0 voxels, intensity values represent t-scores.
Middle: Parameter estimates (across subjects' mean ± SEM) for 2nd normal speech sentence unprimed = N_2U; 2nd normal speech sentence primed = N_2P; 2nd fine structure speech sentence unprimed = F_2U; 2nd fine structure speech sentence primed = F_2P. The bar graphs represent the size of the effect in nondimensional units (corresponding to percentage whole-brain mean).
Right: Physical effect (a, green), perceptual effect (b, red) and novelty effect (c, blue) are shown on sagittal and coronal slices of a subject's normalized structural image. To illustrate the functional organization, they are overlaid with F-preferential ($F_1 > N_1$: cyan) and N-preferential ($N_1 > F_1$: yellow) activation. Height threshold $p<0.001$ uncorrected (masked with normal target sentence > fixation at $p < 0.001$ uncorrected), intensity values represent t-scores.

3.2.2.1. Physical effects: We identified *physical* effects by comparing primed fine structure speech and primed normal speech. Primed fine structure speech and primed normal speech are relatively matched in terms of novelty (n.b. both are primed) and speech intelligibility, but differ in their spectrotemporal structure. Importantly, fine structure speech lacks the low frequency temporal cues provided by the acoustic envelope that enables grouping of the continuous auditory signal into higher order units. In the visual domain it is well-established that visual signals that cannot be grouped into higher order spatial representations enhance activations in low level visual areas indicating the brain's failure to predict the incoming visual signals based on higher order representations (Kok and de Lange, 2014; Murray et al., 2002; Murray et al., 2004). Assuming similar functional principles of predictive coding for the auditory and visual systems, we therefore expected a stronger prediction error signal in low level auditory areas for auditory fine structure signals that cannot be structured into larger temporal units relative to normal speech inputs. Indeed, in line with predictive coding, primed

intelligible fine structure speech relative to normal speech increased activations in the primary cortices (see figure 2 a). In the left hemisphere, the activations extended also into the parietal operculum.

For completeness, no additional activations were observed for primed normal speech relative to primed fine structure speech. In other words, normal speech did not elicit any additional activations as compared to fine structure speech, when both signals were nearly matched in terms of speech intelligibility. This contrasts with a previous study demonstrating increased anterior temporal activations for normal speech relative to intelligible broadband speech envelope noises (Giraud et al., 2004). Collectively, these findings suggest that intelligible speech–like signals that preserve the complex fine structure (but not the slow envelope) activate anterior temporal cortices to a similar extent as normal speech. Thus, these anterior temporal areas are rather insensitive to the spectrotemporal characteristics of the acoustic signals and depend more on the particular linguistic processes and representations that are formed.

3.2.2.2. Perceptual effects: We identified *perceptual* effects by comparing primed relative to unprimed fine structure speech sentences. Even though the spectrotemporal structure of primed and unprimed fine structure speech is identical across subjects, prior context and perceptual learning render fine structure speech intelligible when it is primed with its corresponding normal sentence. Hence, primed and unprimed fine structure speech stimuli differ only in the availability of an intelligible speech percept, whilst being matched for their physical properties. In the visual modality, priming enhancement occurs (Dolan et al., 1997;

George et al., 1999; Henson et al., 2000; Henson, 2003), whenever priming enables a new process to be performed on a stimulus. For instance, the fusiform face areas showed enhanced activations when priming enabled face recognition of otherwise unrecognizable degraded visual inputs (e.g. (George et al., 1999)). From the perspective of predictive coding, priming enhancement can be interpreted as the emergence of novel higher order representations that explain away prediction errors in lower level areas and elicit new prediction errors in higher order areas.

Hence, we expected that fine structure speech elicits increased activations in higher order auditory areas, when new intelligible speech representations are formed via top-down prior constraints and/or rapid perceptual learning (= priming) *(e.g. (Doniger et al., 2001; Wiggs and Martin, 1998))*.

Indeed, consistent with our hypothesis, we observed increased activations in the planum temporale and polare, superior/middle temporal gyri extending into inferior parietal and frontal cortices for intelligible relative to unintelligible fine structure speech (see figure 2 b).

Yet, even though superior and middle temporal gyri showed activation increases for intelligible fine structure speech, their activation profiles were very distinct (see parameter estimate plots): The superior temporal areas adjacent to primary auditory cortices showed enhanced activations for primed fine structure speech relative to both unintelligible fine structure and normal speech stimuli. This activation profile suggests that the superior temporal gyri may be involved in a process that is amplified for processing fine structure signals.

In contrast, the posterior middle temporal and inferior frontal gyri were activated for all intelligible stimuli irrespective of their spectrotemporal structure (i.e. equally for intelligible fine

structure and normal speech) suggesting that they are commonly involved in mapping auditory signals onto intelligible representations (Binder et al., 1997; Dronkers et al., 2004; Evans et al., 2013; Geschwind, 1970; Hickok and Poeppel, 2000; Okada et al., 2010; Price, 2012; Rodd et al., 2005; Trebuchon et al., 2013; Turken and Dronkers, 2011).

To further dissociate whether these effects are directly related to the emergence of the speech percept or to a general priming effect per se, we also computed the interaction between sentence type and priming masked with the simple main effect (i.e. the perceptual effect, see table A.2). This interaction contrast basically replicates the results identified by the perceptual effect suggesting that the activations are truly related to the emergence of an intelligible speech percept.

For completeness, no areas showed increased activations for unprimed relative to primed fine structure stimuli. From the perspective of predictive coding this may at first be surprising. As priming enables the formation of higher order representations, they should 'explain away' the bottom-up auditory fine structure signals and thereby reduce prediction errors at lower hierarchical levels. Hence, we would expect reduced activations indexing prediction error signals in low level, i.e. primary auditory areas. By contrast, activations in low level auditory areas were not significantly modulated by priming. This surprising finding can be explained by the fact that priming as 'rapid perceptual learning' may not only suppress prediction errors, but also increase the precision of those prediction errors (Feldman and Friston, 2010), which in turn may be associated with activation increase. In short, priming may induce two counteracting effects, suppression of prediction errors and amplification of the precision of prediction errors.

As these two effects may cancel each other out, we may not have observed any net change in activation.

3.2.2.3. Novelty effect: We identified *novelty* effects by comparing unprimed relative to primed normal speech sentences. Novel or unprimed sentences violate subjects' predictions formed based on the preceding sentence. Therefore we expected them to evoke prediction error signals along the anterior superior temporal gyri / sulci indicating the need to generate a novel 'intelligible' representation at the sentential level. Indeed, unprimed relative to primed normal speech increased activations in the anterior portions of superior temporal gyri / sulci extending into the temporal poles (see figure 2 c).

To further dissociate whether these effects are directly related to the novelty effect for intelligible speech or to a general priming effect per se, we also computed the interaction between sentence type and priming masked with the simple main effect (i.e. the novelty effect, see table A.2). This interaction contrast basically replicates the results identified by the novelty effect suggesting that the activations are truly related to the processing of novel intelligible speech.

For completeness, activations encompassing the left supramarginal gyrus, posterior superior temporal gyrus, inferior parietal gyrus and the left insula extending into the frontal operculum were increased for primed relative to unprimed sentences (see also table A.1 and figure A.1 in appendix).

3.3. Effective connectivity analysis: dynamic causal modelling

Finally, we employed dynamic causal modelling and Bayesian model comparison to investigate how speech comprehension emerges from interactions amongst brain areas. Since the physical, perceptual and novelty effects were primarily located in the temporal cortices, we limited the DCMs to regions in the ventral stream that is thought to process auditory signals for meaning. In particular, we did not include the left inferior frontal gyrus as an additional node, because we observed significant activations at $p<0.05$ uncorrected in the 8 mm sphere centered on the group peak in only 13 out of 20 subjects.

Figure 4 shows the 18 candidate DCMs that factorially manipulated (i) the intrinsic connectivity structure (serial vs. parallel processing), (ii) the connection modulated by the perceptual effect (forwards, backwards, bidirectional) and (iii) the connection modulated by the novelty effect (forwards, backwards, bidirectional). This factorial model space allowed us to evaluate the role of each of these three effects by comparing the relevant DCM families using Bayesian model comparison at the random effects level.

First, we investigated whether speech processing emerges in a serial or parallel architecture by comparing the two model families that differed in their intrinsic structure. Bayesian model comparison demonstrated an increased model evidence for the DCM family that included the bidirectional connection between anterior superior temporal sulcus and Heschl's gyrus (expected posterior probability: 0.95; exceedance probability: 1), even though these DCMs were more complex and had a higher number of parameters (see figure 4 a). These results suggest that speech is processed in parallel rather than serial processing streams along the ventral temporal cortices.

Second, we investigated whether the perceptual and novelty effects were mediated via modulation of the forwards, backwards or bidirectional connectivity. Our results demonstrated an increased model evidence for the DCM families where the perceptual effect modulated the backwards connections from posterior superior temporal gyrus to Heschl's gyrus (expected posterior probability: 0.58; exceedance probability: 0.86) and the novelty effect modulated the forwards and backwards connections from posterior to anterior portions of the superior temporal gyrus (expected posterior probability: 0.59; exceedance probability: 0.93).

Finally, we compared all 18 models (i.e. without grouping the models into model families) to determine the optimal DCM. The results of this analysis converged with those obtained from the family comparisons. The winning model was again characterized by a 'parallel processing' architecture. Further, the perceptual effect modulated the backwards connections and the novelty effect affected both the forwards and backwards connections (expected posterior probability: 0.24; exceedance probability: 0.78).

The winning model also allowed us to estimate the strength of each connection across subjects using classical statistics. As shown in figure 4 b, the perceptual effect significantly increased the backwards connection from posterior STG to Heschl's gyrus, whereas the novelty effect enhanced the forwards connection from posterior STG to anterior STG/STS. Even though the optimal model included a modulatory effect of the novelty effect on the backwards connections from anterior STG/STS to posterior STG, this effect was not significant across subjects.

Thus, in line with the notion of predictive coding, the top-down predictions of a prior envelope template, lexical and phonological priming are conveyed via increased backwards connectivity.

By contrast, the prediction errors for a novel (i.e. unprimed) sentence that violates prior semantic expectations are furnished predominantly by increased forward connectivity.

4. Discussion

The neural mechanisms that enable robust speech recognition are poorly understood. Our results support models of predictive coding where speech is decoded by integrating bottom-up inputs (or prediction error signals) and top-down prior predictions. In the following, we will discuss the activation results in relation to the three hypotheses posed in the introduction based on predictive coding (Davis and Johnsrude, 2007; Friston, 2005; Friston, 2010).

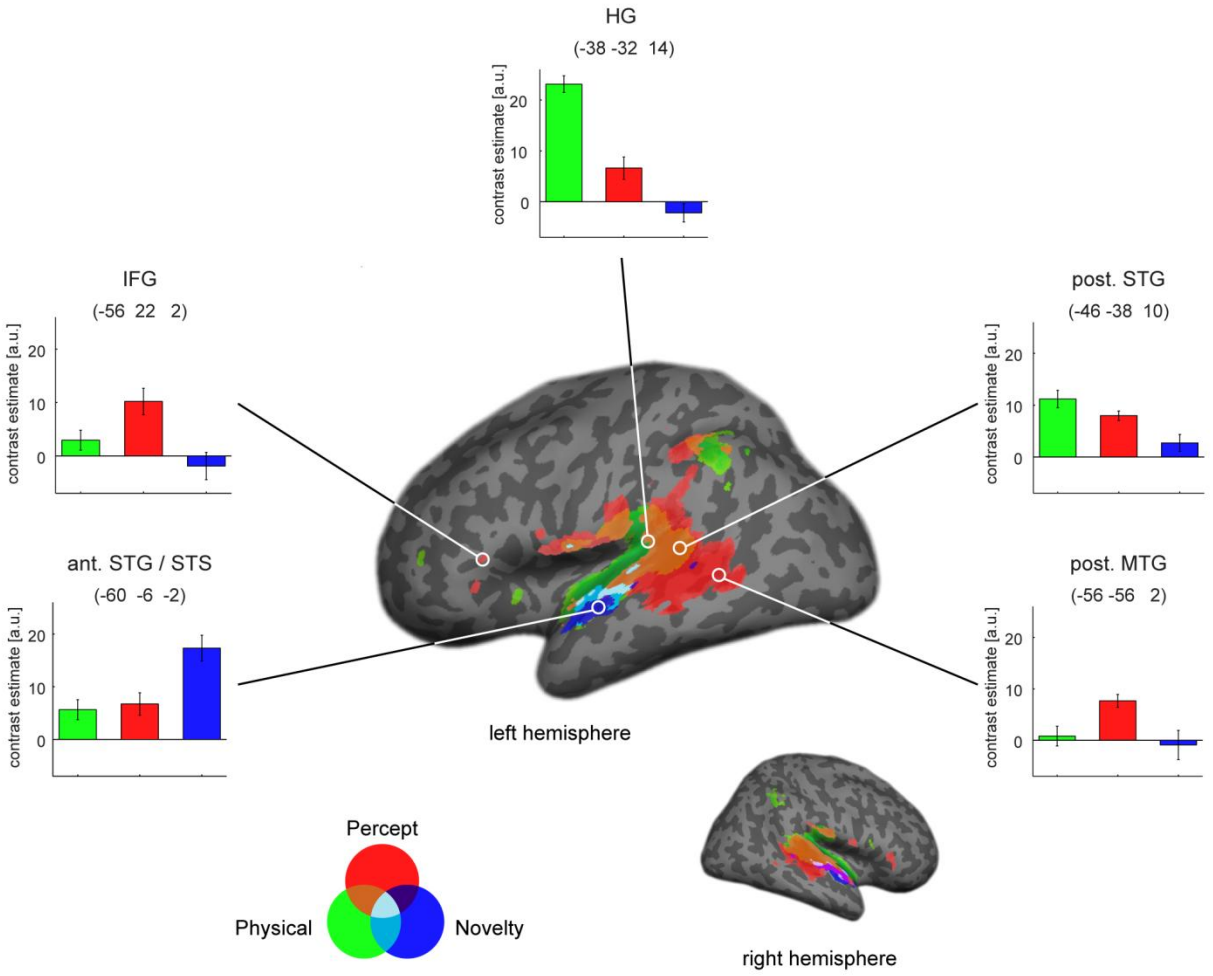

Figure 3
fMRI results – Cortical hierarchy of speech recognition:
The Physical ($F_2P > N_2P$: green), Perceptual ($F_2P > F_2U$: red) and Novelty ($N_2U > N_2P$: blue) effects are schematically projected on an inflated brain. Height threshold, $p < 0.001$ uncorrected, masked with normal target sentence > fixation at $p < 0.001$ uncorrected. Contrast estimates (across subjects' mean ± SEM) for Physical, Perceptual and Novelty effects are shown for Heschl's gyrus, posterior superior

temporal gyrus, posterior middle temporal gyrus, anterior superior temporal gyrus/sulcus and inferior frontal gyrus at the given coordinate locations. The physical effect decreases progressively from Heschl's gyrus to anterior superior temporal gyrus/sulcus. The novelty effect is present only in the anterior temporal gyrus/sulcus.
The bar graphs represent the size of the effect in nondimensional units (corresponding to percentage whole-brain mean).

4.1. Failure to predict the temporal evolution of acoustic signals: Physical effects

In line with our first hypothesis, we observed activation increases in primary auditory cortices for fine structure relative to normal speech. These activation increases are thought to signal a physical prediction error indicating the brain's failure to predict the temporal evolution of fine structure signals that lack the temporal cues of the acoustic envelope (Drullman, 1995). By contrast, despite having more sound energy than fine structure speech, normal speech that allows temporal grouping of the acoustic inputs suppressed neural activity in low level auditory cortices, but increased activations along the superior temporal gyri / sulci. This seesaw relationship between lower and higher order sensory areas has previously been described for the visual system (Murray et al., 2002) where spatial grouping induced neural suppression in primary visual areas and amplification in the lateral occipital complex. Consistent with the principles of predictive coding, spatial (vision) and temporal (audition) grouping may thus enable the formation of higher order representations that explain away the bottom-up sensory inputs and thereby suppress prediction error signals in lower level sensory areas (Kersten and Yuille, 2003).

4.2. Transforming acoustic signals into intelligible speech: Perceptual effects

While 'unprimed' fine structure sentences were rather unintelligible throughout the entire course of the experiment, they became readily intelligible when preceded (i.e. primed) by their corresponding normal sentence. Priming may have provided subjects with an envelope template, lexical and phonological constraints as top-down predictions that allowed them to segment the acoustic fine structure signal into larger intelligible temporal units alike spatial 'pop out' phenomena in visual object recognition (Dolan et al., 1997). Under predictive coding, we expected that this form of rapid perceptual learning induces the formation of higher order representations in adjacent auditory areas that in turn exert top-down constraints onto lower auditory areas (Kiebel et al., 2009). Indeed, while processing of unprimed fine structure speech was confined to low level auditory cortices, intelligible (i.e. primed) fine structure speech induced activations in the posterior middle temporal gyrus. This rapid neural plasticity was associated with increased backwards connectivity from the middle temporal gyrus to primary auditory areas (c.f. Dynamic Causal Modelling, figure 4) suggesting that top-down constraints from novel representations in posterior MTG transformed fine structure inputs into intelligible speech. Yet, our additional psychophysics study suggests that potentially priming of finestructure speech does not only rely on the temporal envelope template, but be supported potentially by multiple mechanisms of which lexical and semantic priming may be particularly important.

Indeed, our functional imaging results also suggest that speech intelligibility evolves in two stages (see parameter estimate plots in figure 2 b). In a first stage, areas adjacent to primary auditory cortex (i.e. post. STG in figure 2 b) match the bottom-up fine structure signals onto a prior envelope or phonological templates (Hickok et al., 2011). As this temporal segmentation

process is enhanced for the noise-like fine structure relative to normal speech inputs, the posterior superior temporal gyrus shows increased responses preferentially for intelligible fine structure speech relative to all other stimulus classes. In a second stage, the posterior STS/MTG and inferior frontal gyrus map these segmented auditory signals onto a meaningful representation (e.g. lexical access). As this second stage is generic to comprehension of fine structure and normal speech, the posterior MTG shows comparable responses to normal speech and intelligible fine structure signals (cf. figure 2 b: parameter estimate plots for post. MTG). To further substantiate this interpretation future fMRI studies are need that prime finestructure speech by sentences that preserve semantic content, but employ a different syntactic structure and different lexical items (e.g. synonyms).

The role of posterior STS in speech recognition also converges with previous work comparing sine wave speech signals (i.e. syllables, single words) in speech and non-speech mode (Dehaene-Lambertz et al., 2005; Möttönen et al., 2006). By contrast, when sine wave speech signals were spoken sentences rather than individual words or syllables (i.e. (Lee and Noppeney, 2011b)), a more anterior mid-STS region showed enhanced activations for processing audiovisual sine wave signals in speech relative to non-speech mode. Collectively, these results suggest that posterior STS/MTG regions may be more involved in lexical access rather than integrating speech signals into sentential representations.

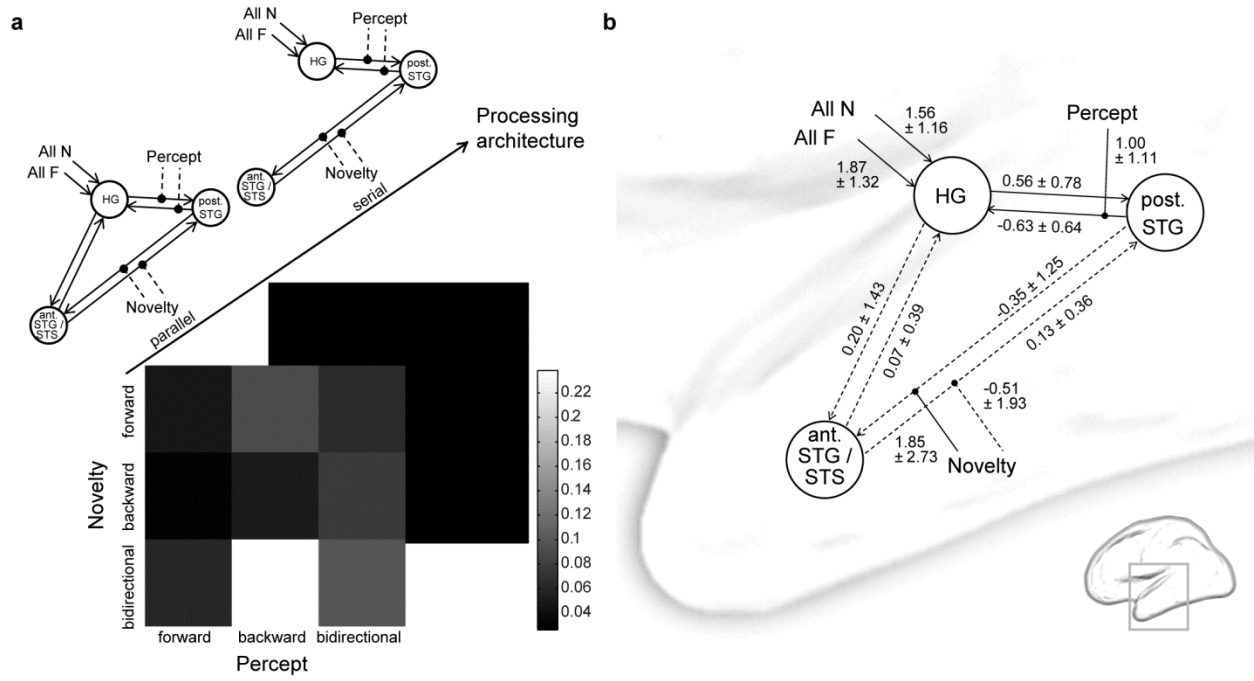

Figure 4
Dynamic Causal Modeling
a. 18 candidate dynamic causal models were generated by factorially manipulating (i) the intrinsic connectivity structure (serial, parallel), (ii) modulatory effect of intelligible speech percept (forwards, backwards, bidirectional) and (iii) modulatory effect of novelty (forwards, backwards, bidirectional). The matrix image shows the expected posterior probability from the random-effects Bayesian model comparison in a factorial fashion.
b. In the optimal model (i.e. model with highest expected posterior probability) (i) the processing architecture is parallel, (ii) the speech percept modulates the backwards connections and (iii) the sentence novelty affects both forwards and backwards connections.
Values are the across-subjects mean ±SD representing the strength of intrinsic connections, modulatory or extrinsic effects ($p < 0.05$ indicated by solid lines). The bilinear (i.e. modulatory) parameters quantify how experimental manipulations change the values of intrinsic connections.
In the optimal model, an intelligible speech representation as prior top-down prediction significantly increases the backwards connection from posterior STG to Heschl's gyrus. In contrast, novel normal sentences significantly increase the forwards connection from posterior STG to anterior STS signaling a prediction error and the need for new semantic integration.

4.3. Creating a novel representation: Novelty effects

Finally, priming generates high level semantic expectations that constrain and thereby facilitate the interpretation also of normal speech input. From the perspective of predictive coding, we

therefore expected novel (i.e. unprimed) speech signals to elicit increased responses signaling a prediction error at the highest semantic or sentential level in anterior temporal areas - previously implicated in speech intelligibility and sentence processing (Dehaene-Lambertz et al., 2006; DeWitt and Rauschecker, 2012; Evans et al., 2013; Friederici et al., 2010; Humphries et al., 2006; Obleser et al., 2007b; Scott et al., 2000; Scott et al., 2006). Indeed, unprimed speech sentences that violate subjects' semantic expectations increased responses in the anterior portions of the superior temporal gyri/sulci. As normal sentences provide phonological and semantic constraints even when unprimed, we did not observe prediction errors at lower levels of the cortical hierarchy (e.g. see (Wacongne et al., 2011) for a very thoughtful study that dissociated prediction errors at multiple hierarchical levels using a mismatch negativity paradigm). Dynamic Causal Modelling suggested that these prediction error signals were mediated via increased forwards connections from posterior superior temporal gyrus to anterior superior temporal gyrus/sulcus. Thus, consistent with the notion of predictive coding, prediction errors - here violations of semantic-sentential expectations - were furnished by the forwards connections from lower to higher cortical levels. Future studies that orthogonally manipulate semantics, syntax and phonology at the sentential level are needed to further disentangle and characterize the linguistic nature of these prediction errors at the highest representational level.

5. Conclusions

5.1. Implications for a neuroanatomical model of speech recognition

Generally, our results highlight the importance of top-down predictions in speech recognition. They support models of predictive coding where speech recognition emerges in the cortical hierarchy via iterative adjustment of top-down predictions against bottom-up acoustic signals.

In brief, activations for unintelligible fine structure speech were confined to primary auditory cortices; yet when top-down predictions made fine structure speech intelligible, they propagated into posterior middle temporal areas to enable lexical access and speech recognition (Binder et al., 1997; Hickok and Poeppel, 2007; Kotz et al., 2002; Wise et al., 2001). By contrast, normal speech engaged posterior middle temporal areas irrespective of subjects' predictions. Critically, when normal speech violated subjects' top-down semantic predictions, activation increases in anterior temporal gyri/sulci signalled a prediction error and the need for new semantic integration.

This double dissociation suggests that sentences that live up to our expectations can already be recognized by posterior temporal cortices, while sentences that are novel and violate our expectations require additional semantic integration processes in anterior temporal cortices. It refines current models of speech processing that attribute speech intelligibility predominantly to anterior portions of the superior temporal sulcus (Giraud et al., 2004; Liebenthal et al., 2005; Scott et al., 2000). Instead, our results suggest that posterior and anterior portions of the superior/middle temporal sulcus/gyrus make distinct contributions to speech intelligibility: Posterior temporal cortices can already sustain speech intelligibility, when a primed semantic representation needs to be re-instantiated. By contrast, more anterior portions of the STG/STS are invoked when a novel intelligible representation needs to be formed at the sentential level.

5.2. Predictive coding and alternative interpretations

In this study, we interpreted the physical, perceptual and novelty effects from the perspective of predictive coding as the brain's efforts to predict the incoming acoustic signals based on prior experience. From the perspective of predictive coding unprimed stimuli that violate participants' predictions elicit prediction error signals (Davis and Johnsrude, 2007; Friston, 2005; Friston, 2010). Yet, we acknowledge that these neural responses could also be explained in more traditional terms as differences in spectrotemporal structure (i.e. physical effect), speech intelligibility (i.e. perceptual effect) or repetition suppression (i.e. novelty effect). In particular, activation increases for primed intelligible relative to unprimed unintelligible fine structure speech may also reflect the access to higher order linguistic representations. Likewise, as has often been discussed in the literature repetition suppression can be explained by models of neuronal fatigue, response sharpening or facilitation (Grill-Spector et al., 2006). From those perspectives the increased activations for unprimed stimuli do not reflect prediction error signals but less refined representations. Nevertheless, following the Occam's razor principle (Penny et al., 2004), we offer predictive coding as a coherent framework that helps to explain three distinct neural responses via one underlying explanatory principle which may be corroborated in future studies.

Conflict of Interest: There is no conflict of interest.

Funding: This work was supported by Max Planck Society.

Acknowledgements: We thank Mario Kleiner for help with recording the stimuli.

Corresponding Author:

Johannes Tuennerhoff

Max Planck Institute for Biological Cybernetics

Spemannstr. 41

72076 Tuebingen

Germany

Email: johannes.tuennerhoff@tuebingen.mpg.de

References:

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10, 512-528.

Binder, J.R., Frost, J.A., Hammeke, T.A., Cox, R.W., Rao, S.M., Prieto, T., 1997. Human brain language areas identified by functional magnetic resonance imaging. J Neurosci 17, 353-362.

Crinion, J.T., Lambon-Ralph, M.A., Warburton, E.A., Howard, D., Wise, R.J., 2003. Temporal lobe regions engaged during normal speech comprehension. Brain 126, 1193-1201.

Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J Neurosci 23, 3423-3431.

Davis, M.H., Johnsrude, I.S., 2007. Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear Res 229, 132-147.

Dehaene-Lambertz, G., Dehaene, S., Anton, J.-L., Campagne, A., Ciuciu, P., Dehaene, G.P., Denghien, I., Jobert, A., Lebihan, D., Sigman, M., Pallier, C., Poline, J.-B., 2006. Functional segregation of cortical language areas by sentence repetition. Hum Brain Mapp 27, 360-371.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., Dehaene, S., 2005. Neural correlates of switching from auditory to speech perception. NeuroImage 24, 21 - 33.

DeWitt, I., Rauschecker, J.P., 2012. Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci U S A 109, E505-514.

Dolan, R.J., Fink, G.R., Rolls, E., Booth, M., Holmes, A., Frackowiak, R.S., Friston, K.J., 1997. How the brain learns to see objects and faces in an impoverished context. Nature 389, 596-599.

Doniger, G.M., Foxe, J.J., Schroeder, C.E., Murray, M.M., Higgins, B.A., Javitt, D.C., 2001. Visual perceptual learning in human object recognition areas: a repetition priming study using high-density electrical mapping. NeuroImage 13, 305-313.

Dronkers, N.F., Wilkins, D.P., Van Valin, R.D., Jr., Redfern, B.B., Jaeger, J.J., 2004. Lesion analysis of the brain areas involved in language comprehension. Cognition 92, 145-177.

Drullman, R., 1995. Temporal envelope and fine structure cues for speech intelligibility. The Journal of the Acoustical Society of America 97, 585-592.

Evans, S., Kyong, J.S., Rosen, S., Golestani, N., Warren, J.E., McGettigan, C., Mourao-Miranda, J., Wise, R.J., Scott, S.K., 2013. The Pathways for Intelligible Speech: Multivariate and Univariate Perspectives. Cereb Cortex.

Feldman, H., Friston, K.J., 2010. Attention, uncertainty, and free-energy. Front Hum Neurosci 4, 215.

Friederici, A.D., Kotz, S.A., Scott, S.K., Obleser, J., 2010. Disentangling syntax and intelligibility in auditory language comprehension. Hum Brain Mapp 31, 448-457.

Friston, K., 2005. A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360, 815-836.

Friston, K., 2010. The free-energy principle: a unified brain theory? Nat Rev Neurosci 11, 127-138.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273-1302.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: A general linear approach. Human Brain Mapping 2, 189-210.

George, N., Dolan, R.J., Fink, G.R., Baylis, G.C., Russell, C., Driver, J., 1999. Contrast polarity and face recognition in the human fusiform gyrus. Nat Neurosci 2, 574-580.
Geschwind, N., 1970. The organization of language and the brain. Science 170, 940-944.
Giraud, A.L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M.O., Preibisch, C., Kleinschmidt, A., 2004. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cereb Cortex 14, 247-255.
Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn Sci 10, 14-23.
Henson, R., Shallice, T., Dolan, R., 2000. Neuroimaging evidence for dissociable forms of repetition priming. Science 287, 1269-1272.
Henson, R.N., 2003. Neuroimaging studies of priming. Prog Neurobiol 70, 53-81.
Hickok, G., Houde, J., Rong, F., 2011. Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69, 407-422.
Hickok, G., Poeppel, D., 2000. Towards a functional neuroanatomy of speech perception. Trends Cogn Sci 4, 131-138.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat Rev Neurosci 8, 393-402.
Horwitz, B., Braun, A.R., 2004. Brain network interactions in auditory, visual and linguistic processing. Brain Lang 89, 377-384.
Humphries, C., Binder, J.R., Medler, D.A., Liebenthal, E., 2006. Syntactic and semantic modulation of neural activity during auditory sentence comprehension. J Cogn Neurosci 18, 665-679.
Kersten, D., Yuille, A., 2003. Bayesian models of object perception. Curr Opin Neurobiol 13, 150-158.
Kiebel, S.J., von Kriegstein, K., Daunizeau, J., Friston, K.J., 2009. Recognizing sequences of sequences. PLoS Comput Biol 5, e1000464.
Kok, P., de Lange, F.P., 2014. Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. Curr Biol 24, 1531-1535.
Kotz, S.A., Cappa, S.F., von Cramon, D.Y., Friederici, A.D., 2002. Modulation of the lexical-semantic network by auditory semantic priming: an event-related functional MRI study. NeuroImage 17, 1761-1772.
Lee, H., Noppeney, U., 2011a. Long-term music training tunes how the brain temporally binds signals from multiple senses. Proc Natl Acad Sci U S A 108, E1441-1450.
Lee, H., Noppeney, U., 2011b. Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. J Neurosci 31, 11338-11350.
Lee, H., Noppeney, U., 2014. Temporal prediction errors in visual and auditory cortices. Curr Biol 24, R309-310.
Leff, A.P., Schofield, T.M., Stephan, K.E., Crinion, J.T., Friston, K.J., Price, C.J., 2008. The Cortical Dynamics of Intelligible Speech. J. Neurosci. 28, 13209-13215.
Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., Medler, D.A., 2005. Neural substrates of phonemic perception. Cereb Cortex 15, 1621-1631.
Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., Zilles, K., 2001. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. NeuroImage 13, 684-701.

Möttönen, R., Calvert, G.A., Jääskeläinen, I.P., Matthews, P.M., Thesen, T., Tuomainen, J., Sams, M., 2006. Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. NeuroImage 30, 563-569.
Murray, S.O., Kersten, D., Olshausen, B.A., Schrater, P., Woods, D.L., 2002. Shape perception reduces activity in human primary visual cortex. Proc Natl Acad Sci U S A 99, 15164-15169.
Murray, S.O., Schrater, P., Kersten, D., 2004. Perceptual grouping and the interactions between visual cortical areas. Neural Netw 17, 695-705.
Narain, C., Scott, S.K., Wise, R.J.S., Rosen, S., Leff, A., Iversen, S.D., Matthews, P.M., 2003. Defining a left-lateralized response specific to intelligible speech using fMRI. Cereb Cortex 13, 1362-1368.
Obleser, J., Kotz, S.A., 2010. Expectancy constraints in degraded speech modulate the language comprehension network. Cereb Cortex 20, 633-640.
Obleser, J., Wise, R.J.S., Dresner, M.A., Scott, S.K., 2007a. Functional integration across brain regions improves speech perception under adverse listening conditions. J Neurosci 27, 2283-2289.
Obleser, J., Zimmermann, J., Van Meter, J., Rauschecker, J.P., 2007b. Multiple stages of auditory speech perception reflected in event-related FMRI. Cereb Cortex 17, 2251-2257.
Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.H., Saberi, K., Serences, J.T., Hickok, G., 2010. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex 20, 2486-2495.
Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 23, 1378-1387.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. NeuroImage 22, 1157-1172.
Price, C.J., 2012. A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. NeuroImage 62, 816-847.
Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12, 718-724.
Rodd, J.M., Davis, M.H., Johnsrude, I.S., 2005. The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cereb Cortex 15, 1261-1269.
Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123 Pt 12, 2400-2406.
Scott, S.K., Rosen, S., Lang, H., Wise, R.J.S., 2006. Neural correlates of intelligibility in speech investigated with noise vocoded speech--a positron emission tomography study. J Acoust Soc Am 120, 1075-1083.
Smith, Z.M., Delgutte, B., Oxenham, A.J., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature 416, 87-90.
Sohoglu, E., Peelle, J.E., Carlyon, R.P., Davis, M.H., 2012. Predictive top-down integration of prior knowledge during speech perception. J Neurosci 32, 8443-8453.
Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. NeuroImage 46, 1004-1017.
Trebuchon, A., Demonet, J.F., Chauvel, P., Liegeois-Chauvel, C., 2013. Ventral and dorsal pathways of speech perception: An intracerebral ERP study. Brain Lang.
Turken, A.U., Dronkers, N.F., 2011. The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. Front Syst Neurosci 5, 1.

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., Dehaene, S., 2011. Evidence for a hierarchy of predictions and prediction errors in human cortex. Proceedings of the National Academy of Sciences 108, 20754-20759.
Werner, S., Noppeney, U., 2011. The contributions of transient and sustained response codes to audiovisual integration. Cereb Cortex 21, 920-931.
Wiggs, C.L., Martin, A., 1998. Properties and mechanisms of perceptual priming. Curr Opin Neurobiol 8, 227-233.
Wild, C.J., Davis, M.H., Johnsrude, I.S., 2012a. Human auditory cortex is sensitive to the perceived clarity of speech. NeuroImage 60, 1490-1502.
Wild, C.J., Yusuf, A., Wilson, D.E., Peelle, J.E., Davis, M.H., Johnsrude, I.S., 2012b. Effortful listening: the processing of degraded speech depends critically on attention. J Neurosci 32, 14010-14021.
Wise, R.J.S., Scott, S.K., Blank, S.C., Mummery, C.J., Murphy, K., Warburton, E.A., 2001. Separate neural subsystems within `Wernicke's area'. Brain 124, 83-95.