Published in final edited form as:

Cortex. 2015 July ; 68: 144–154. doi:10.1016/j.cortex.2014.12.014.

The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene

Johanna M. Rimmele[a,b], **Elana Zion Golumbic**[c,d], **Erich Schröger**[e], and **David Poeppel**[a,f]

[a]Department of Psychology and Center for Neural Science New York University New York, NY

[b]Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf Hamburg, Germany

[c]Gonda Center for Brain Research Bar Ilan University, Israel

[d]Department of Psychiatry Columbia University New York, NY

[e]Institute of Psychology, University of Leipzig Leipzig, Germany

[f]Max-Planck Institute for Empirical Aesthetics, Frankfurt, Germany

Abstract

Attending to one speaker in multi-speaker situations is challenging. One neural mechanism proposed to underlie the ability to attend to a particular speaker is phase-locking of low-frequency activity in auditory cortex to speech's temporal envelope ("speech-tracking"), which is more precise for attended speech. However, it is not known what brings about this attentional effect, and specifically if it reflects enhanced processing of the fine structure of attended speech. To investigate this question we compared attentional effects on speech-tracking of natural vs. vocoded speech which preserves the temporal envelope but removes the fine-structure of speech. Pairs of natural and vocoded speech stimuli were presented concurrently and participants attended to one stimulus and performed a detection task while ignoring the other stimulus. We recorded magnetoencephalography (MEG) and compared attentional effects on the speech-tracking response in auditory cortex. Speech-tracking of natural, but not vocoded, speech was enhanced by attention, whereas neural tracking of ignored speech was similar for natural and vocoded speech. These findings suggest that the more precise speech tracking of attended natural speech is related to processing its fine structure, possibly reflecting the application of higher-order linguistic processes. In contrast, when speech is unattended its fine structure is not processed to the same degree and thus elicits less precise speech tracking more similar to vocoded speech.

Corresponding author: Johanna M. Rimmele, PhD Department of Neurophysiology and Pathophysiology University Medical Center Hamburg-Eppendorf Martinistraße 52 D - 20246 Hamburg phone: +49 (40) 7410 - 0 j.rimmele@uke.de. david.poeppel@nyu.edu, elana.zion-golumbic@biu.ac.il, schroger@rz.uni-leipzig.de

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

auditory cortex; fine structure; linguistic; predictions; speech envelope

[1]Listening to speech in multi-speaker situations is challenging, particularly for populations such as older adults or hearing impaired listeners (for review, see Rimmele et al. 2014). These situations require the segregation of speech streams originating from different speakers and the selection of one of these streams for further processing. The neural mechanisms through which attentional selection is achieved and that facilitate the processing of attended speech over competing stimuli are not fully understood. One major question concerns the degree to which the ability to establish a robust representation of speech in auditory cortex (e.g. as required to attend to a particular speaker) is driven by the acoustic properties of the stimulus including both the speech envelope and its fine structure.

Mechanistically speaking, phase-locking of low-frequency neural activity in auditory cortex (“speech-tracking response”) has been proposed to indicate a robust *“object-level”* representation of speech (Luo and Poeppel 2007). The speech-tracking response has been related to both the temporal envelope of the stimulus, which carries information regarding fluctuations in stimulus energy over time, as well as to its fine structure which contains the more detailed spectro-temporal information of speech (Ding and Simon 2014) . Crucially, speech-tracking has been shown to be more robust for attended- compared to unattended speech presented simultaneously (Kerlin et al. 2010; Ding and Simon 2012a, 2012b; Horton et al. 2013, 2014; Zion Golumbic et al. 2013b), suggesting that this mechanism is influenced by attentional selection. The goal of the current study was to clarify the role of two levels of speech acoustics – the temporal envelope and fine structure – in speech-tracking by investigating how they interact with attention in multi-speaker listening situations. Specifically we asked, whether speech's fine structure is utilized by selective attention to enhance speech-tracking of natural speech.

1.1 The speech-tracking response

Phase-locking of neural activity in auditory cortex to the temporal envelope of speech is observed primarily in the theta frequency range (3-7Hz), corresponding to the syllabic time scale in speech. It is well established that the low-frequency fluctuations in the speech envelope, which carry temporal information about syllable onsets/offsets as well as prosodic cues, are crucial for speech intelligibility (Shannon et al. 1995; Giraud and Poeppel 2012; Zion Golumbic et al. 2012; Ghitza et al. 2013; Doelling et al. 2014). Ghitza and Greenberg (2009) showed that the intelligibility of time compressed speech (with a very low intelligibility <50% words correct), increased dramatically when a theta-range “syllabic rate” was artificially induced by adding periods of silence. In light of these results, it has been proposed that the theta-band ‘speech-tracking response’ in auditory cortex plays a role

[1] **Abbreviation:** ECoG, electrocorticography; EEG, electroencephalography; ERF, event-related field; FAR, false alarm rate; fMRI, functional magnetic resonance imaging; HR, hit rate; ICA, independent component analysis; ITPC, intertrial phase coherence; ITPowC, intertrial power coherence; MEG, magnetoencephalography; RMS, Root Mean Square;

in segmenting the speech stream into smaller linguistically-meaningful units (Giraud et al. 2007; Luo and Poeppel 2007; Giraud and Poeppel 2012; Zion Golumbic et al. 2012).

Significant phase-locking to the sound envelope is also observed for unintelligible, time-inverted, or noise-vocoded speech as well as non-speech sounds (Lalor et al. 2009; Howard and Poeppel 2012; Hämäläinen et al. 2012; Peelle et al. 2012; Wang et al. 2012; Millman et al. 2013; Steinschneider et al. 2013; Ding et al. 2014). Nonetheless, speech tracking is more robust for natural compared to noise-vocoded speech, in which the fine structure information is removed but the low-frequency temporal fluctuations contained in the speech envelope are preserved (Luo and Poeppel 2007; Howard and Poeppel 2010; Peelle et al. 2012; Wild, Davis, et al. 2012; Ding et al. 2014). This observation has been interpreted by some as reflecting the application of higher order linguistic processing to natural compared to vocoded speech, since vocoded speech is less intelligible than natural speech (Peelle et al. 2010; Wild, Davis, et al. 2012). Others suggest that these effects may be due to the difference in acoustical information in natural compared to vocoded speech, and that the increased speech-tracking response for natural speech reflects its richer acoustic features (Ding et al. 2014; reviewed in Ding and Simon 2014). It is difficult to distinguish between these alternatives, as differences in speech acoustics and speech intelligibility are inherently confounded when directly comparing natural vs. vocoded speech. Nonetheless, these two interpretations make different predictions regarding the consistency of this phenomenon. Under the acoustic-processing perspective, the advantage for speech-tracking of natural over vocoded speech should remain robust under different cognitive manipulations, since it is primarily due to the acoustic structure of the stimuli. In contrast, under the linguistic-processing perspective, the differences in speech tracking of natural and vocoded speech may be affected by task demands and the degree of linguistic processing applied. Following this rationale, in the current study we tested how the speech-tracking response of natural and vocoded speech was affected by selective attention as a means for investigating the interaction between the acoustic richness of a stimulus and top-down processing demands. As reviewed below, higher order top-down processes, such as linguistic processing or selective attention, influence processing in auditory cortex therefore reconciling their interaction with bottom-up acoustic processing is critical for understanding the neural architecture and hierarchy underlying speech processing.

1.2 Effects of linguistic processing on sensory responses

There is much evidence that sensory processing of speech in auditory cortex can be modulated by higher order processing, such as syntactic or semantic analysis (Miller and Isard 1963; Kalikow et al. 1977; Peelle et al. 2012; Peelle 2013), speaker familiarity (Johnsrude et al. 2013) or linguistic expectations set up by visual cues (Jacoby et al. 1988; Zekveld et al. 2008; Sohoglu et al. 2012; for review: Peelle et al. 2010). Sohoglu and colleagues (using EEG and MEG) showed that a visual cue, which provides prior knowledge of the speech content, increases the perceived speech clarity in a similar manner as altering the physical parameters of the stimulus. Furthermore, prior knowledge increased the amplitude of evoked potentials/fields, in inferior frontal cortex, prior to affecting sensory processing in auditory cortex. The authors argue that these findings reflect early top down effects on sensory processing of speech. Imaging research provides further evidence for early effects of higher order processing, related to speech intelligibility, on sensory processing of speech in primary auditory areas (Wild, Davis, et al. 2012).

As mentioned above, several authors have interpreted the increased speech-tracking observed for natural vs. noise-vocoded speech (Peelle et al. 2012; for review: Peelle et al. 2010; see also: Luo and Poeppel 2007) as reflecting the influence of top-down linguistic processes, which are applied to natural but not to vocoded speech (Peelle et al. 2013). It has ben suggested that one way in which linguistic processing may increase the temporal precision of speech-tracking in auditory cortex is by extracting and utilizing predictive information from the speech – e.g. from the lexical context, syllabic/word/clause boundaries, and prosodic cues - which provide temporal cues as to the timing and content of upcoming speech and can be utilized to allocate sensory processing resources to appropriate points in time (Lakatos et al. 2008; Schroeder et al. 2008, 2010; Zion Golumbic et al. 2012).

However, as discussed above, this linguistic-perspective is currently under debate. A recent study showed that speech-tracking of tone-vocoded speech was not altered even when their intelligibility was increased due to training (Millman et al. 2014), suggesting that the mere fact that a stimulus is intelligible does not necessarily enhance its speech-tracking response. Unfortunately, in that study responses to natural speech were not compared to the artificial vocoded stimuli. Thus, at the moment, the debate whether the enhanced speech-tracking of natural speech only reflects its richer acoustic structure (as suggested by Ding et al. 2014) or is additionally affected by linguistic processing remains unresolved.

1.3 Effects of attention on speech-tracking

Attentional focus also exerts a strong top-down effect on sensory processing. Neurophysiological studies in both humans and animals have shown repeatedly that although there are robust responses in sensory cortex to both attended and unattended stimuli across modalities, the magnitude of the sensory response is modulated by attention such that attended stimuli elicit stronger responses (Hillyard et al. 1973; Posner and Driver 1992; Tiitinen et al. 1993; Woldorff et al. 1993; Rimmele et al. 2011). For simple auditory stimuli (tones), attentional effects on sensory evoked responses are generally interpreted as feature-based attention, brought about through local changes in gain and/or width of auditory receptive fields (Ahveninen et al. 2011). Attentional effects have been repeatedly found for neural speech-tracking responses as well, demonstrating more robust tracking of attended vs. unattended speech (Kerlin et al. 2010; Ding and Simon 2012a, 2012b; Horton et al. 2013; Zion Golumbic et al. 2013b). In the case of speech, the mechanistic explanations for attentional modulations are less straightforward, as simple feature-based accounts are probably insufficient. Zion Golumbic and colleagues (2013b) recorded electrocorticography (ECoG) in surgical epilepsy patients while they selectively attended to one speaker and ignored another concurrent speech stream. They found that low-frequency phase as well as high-gamma power in auditory cortex (STG) faithfully tracked the temporal envelope of both the attended and unattended speaker (though speech-tracking was reduced for the unattended speaker); however, in higher-level brain regions (frontal/parietal cortex), low-frequency phase-locking was observed only for attended speech. These findings suggest that brain regions involved in higher-order speech processing and/or attentional focus selectively represent attended speech. These findings are in line with previous studies demonstrating that unattended speech undergoes limited linguistic processing compared to attended speech (Cherry 1953; Dark et al. 1985; Treisman 1986; Bentin et al. 1995; Wood and Cowan 1995; Power et al. 2012) and it is hypothesized that the selectivity of these regions may modulate sensory responses in auditory cortex through top down feedback. According to this view, the observed enhancement of speech-tracking in auditory cortex by attention may reflect a combination of bottom-up sensory processing and additional top-down modulation from higher order brain regions involved in attention (Lakatos et al. 2008; Schroeder and Lakatos 2009).

1.4 Hypothesis

The current study was aimed at understanding the interaction between fine structure acoustics and selective attention and their mutual effect on the speech-tracking response in auditory cortex. To this end, we recorded MEG during a *cocktail party* paradigm where a pair of natural and vocoded speech stimuli were presented concurrently, and participants were instructed to attend to one stimulus and performed an intensity change-detection task while ignoring the other stimulus. We then compared speech-tracking responses to natural and vocoded speech, when they were either attended or ignored.

The attentional manipulation in our experimental setup allowed us to address the issue of the relative contributions of the temporal envelope and fine structure of speech and top-down attention to the precision of the speech tracking response. Specifically, we hypothesized that if the increased speech-tracking previously observed for natural vs. vocoded speech is due primarily to the richness of their acoustic structure, we would expect differences in speech tracking of these speech-stimuli regardless of whether they are attended or unattended. Alternatively, if the increased speech tracking of natural speech is not driven solely by its richer acoustics, but rather is also influenced by top-down processes, we would expect attentional enhancement of speech-tracking to be more prominent for natural compared to vocoded speech.

2 Methods

2.1 Participants

Fifteen healthy participants with normal hearing (self-report) and no history of neurological disorders took part in the experiment. Due to technical problems data from one participant were excluded. Data of the remaining fourteen participants (female: 7; age-range between 19 and 35 years; mean age: 25.7, sd: 4.9) were included in the analysis. All participants were right-handed (Oldfield 1971), native English speakers and provided informed consent prior to the experiment. The experiment was conducted in accordance with the local Institutional Review Board (New York University's Committee on Activities Involving Human Subjects).

2.2 Stimuli and task

Six sentences with a minimum length of 7 seconds (sentence length, mean: 7.52; sd: 0.31; min.: 7.04; max.: 7.84) were selected from a public domain internet audio book website (http://librivox.org). The six sentences differed in their semantic content and were spoken by different, three male and three female, speakers (American English pronunciation; sampling rate of 44.1 kHz). Noise vocoding was used to manipulate stimulus acoustics. Stimuli were generated, similarly as described by Shannon et al. (1995), using the Matlab toolbox “chimeraSoftware” (see Smith et al. 2002 Fig. 1 for an illustration of the procedure). We used noise-vocoded speech stimuli with four channels, which are established as being substantially less intelligible compared to natural speech (Shannon et al. 1995; Smith et al. 2002; Sheldon et al. 2008; Obleser and Kotz 2010), and are similar to the stimuli used in previous studies (Luo and Poeppel 2007; Peelle et al. 2012; Ding et al. 2014). Speech was filtered into four frequency channels using digital 418-point complex finite impulse response (FIR) filters and the efficient FFT-based method of overlap-add (Hamming windowed). Filter cut-offs were determined, according to Liberman's cochlear frequency maps for the cat (Liberman 1982), within a frequency range corresponding to that of the human cochlea (80 Hz - 20 kHz) to be approximately equally spaced at the basilar membrane (channel1: 80-504 Hz; channel2: 504 -1794 Hz; channel3: 1794 - 5716 Hz; channel4: 5716-17640 Hz). The absolute value of the Hilbert Transform was used to extract the envelope in each channel. White noise was similarly filtered and modulated by the speech envelope of each channel. The noise-vocoded speech stimulus was generated by adding the envelope-modulated-noise of all channels. Finally, the signal was normalized (i.e, at each time point it was divided by the maximum of its absolute values). The RMS amplitude of speech stimuli and vocoded speech did not differ significantly (independent sample t-tests comparing the six sentences in both conditions: t(10) = 1.48; p = .148; natural, mean = 0.09, sd = 0.02; vocoded mean = 0.11, sd = 0.02).

We employed a *cocktail party* paradigm in which two sentences were always presented simultaneously (Fig. 1). Sentences were paired such that natural speech was always presented simultaneously with vocoded speech, and a female speaker was always presented simultaneously with a male speaker sentence. In each trial, one sentence was designated as *to-be-attended* and started 500ms earlier than the *to-be-ignored* sentence. Participants were instructed to attend to the *early-onset* sentence. Either the *to-be-attended* or the *to-be-ignored* sentence contained a 600 ms-long loudness increase with equal probability. The loudness change could occur at a randomly selected point in time within the last 2-3 seconds of the stimulus. At the end of each trial participants were asked to indicate by button press whether a loudness change had occurred in the attended sentence (yes/no) and received feedback (“hit”, “error”). The next trial followed after an inter-trial interval of 800 ms. To ensure similar high task performance for natural and vocoded sentences the size of loudness increase differed for both types of stimuli and was selected based on a behavioral pilot experiment using an adaptive procedure (natural speech: 10 dB SPL; vocoded: 17 dB SPL).

Behavioral responses were analyzed in both conditions: 1) *attend natural*; 2) *attend vocoded* sentences. Neurophysiological responses were also analyzed for the ignored stimuli: 1) *attend natural*; 2) *attend vocoded*; 3) *ignore natural*; 4) *ignore vocoded*. Each sentence was repeated 18 times per condition and sentences were counter-balanced across conditions.

Participants were instructed to hold their gaze at a fixation-cross. Sentences were presented at normal conversational sound level (~75 dB SPL) with insert earphones (E-ARTONE 3A

50 ohm, Etymotic Research) attached to E-A-RLINK foam plugs that were inserted into the participants' ear canal. The experiment consisted of 12 blocks. The total experiment duration was about 2 hours (45 min recording time). Prior to the experiment participants were familiarized with the task and received a short training. At the beginning of each session we ran an auditory localizer where participants passively listened to a random sequence of pure tones (duration 400 ms; 250 and 1000 Hz).

2.3 MEG recordings

MEG data were recorded on a 157-channel whole-head axial gradiometer MEG system (5 cm baseline axial gradiometer SQUID-based sensors, KIT, Kanazawa Institute of Technology, Japan) in an actively magnetically shielded room. Data were recorded with a sampling rate of 1000 Hz. They were filtered on-line with a notch filter at 60 Hz (to remove line noise) and a 200 Hz analog low pass filter (DC recording). Before and after the experiment, participant's head position was localized via five coils to the MEG sensors. The positions of the coils were determined with respect to three anatomical landmarks (nasion, left and right pre-auricular points) using 3D digitizer software (Source Signal Imaging, Inc.) and digitizing hardware (Polhemus, Inc.). Headshape data were digitized using a 160 three-dimensional digitizer (Polhemus).

2.4 Data Analysis

2.4.1 Behavioral Data Analysis—When participants indicated the detection of a loudness change in the attended sentence (*yes response*) and there indeed was a loudness change in the attended sentence, responses were considered *hits*, otherwise *yes responses* were considered *false alarms* (i.e. when no loudness change occurred in the attended sentence). Hit rate (HR) and false alarm rate (FAR) were calculated separately for each participant and stimulus type (natural/vocoded). Response sensitivity was analyzed by calculating d' according to the signal detection theory (d' = z(HR) - z(FAR)) (Green and Swets 1966). To avoid infinite d' values, 0.5 was added to each individual hit and false alarm score, these values were divided by the number of trials adding one trial (n+1) (Macmillan and Creelman 2005). A paired-sample t-test was used to analyze effects of stimulus type.

2.4.2 MEG Data Analysis—All data were noise-reduced off-line using the Continuously Adjusted Least-Squares Method (CALM; Adachi et al. 2001). The FieldTrip toolbox (http://fieldtrip.fcdonders.nl; Oostenveld et al. 2011) and other custom Matlab toolboxes were used for further data processing.

Preprocessing: The preprocessing procedure for cleaning the MEG data was as follows: the data were filtered (low-pass: 100 Hz) and gross artifacts were rejected based on semi-automatic procedure and visual inspection (localizer data: SQUID jumps, resets and muscle artifacts; experimental data: SQUID jumps and resets; SQUID jumps and resets, z-value cut-off = 100; muscle artifacts, z-value cut-off = 8). The data were then down-sampled to 500 Hz, bad channels were interpolated, and independent component analysis (ICA; runica; Makeig et al. 1996) was used to remove eye-blinks, eye-movements and heartbeat-related artifacts. ICA components were identified as eye-blink and heartbeat related activity by their spectral, topographical and time course characteristics.

Localizer Data Analysis: The M100 component was identified from the auditory localizer data of individual participants by averaging across trials (epoched between -0.2-1 sec) and baseline correcting (pre-stimulus: -0.2-0 sec; Fig. 2a). The top 20 sensors (10 per hemisphere) with maximal M100 amplitude were selected for use in subsequent analyses (Fig. 2b).

Experimental Data Analysis: Intertrial phase/power coherence: After preprocessing, data from the experiment were epoched (post-stimulus: 1-7 sec). The first 1 sec was not included in order to eliminate transient responses due to stimulus onset. Fourier-transform was performed on the data, using a Hann window. Spectral estimates were computed for frequencies *f* from 2 to 100 Hz (1 Hz frequency steps; 1 Hz frequency resolution) across 60 points in time from 1-7 s (1 s window; 0.1 s steps).

Intertrial phase coherence (ITPC) and *Intertrial power coherence* (ITPowC) were calculated separately for individual participants, for each sentence, and for all conditions. These measures quantify the consistency of responses (phase/power time course) across trials. ITPC was calculated as the circular variance of the phase across trials, at frequency and time point (circular statistics Matlab toolbox), and averaged across time points. ITPowC was calculated at each time (*t*) and frequency (*f*) as suggested by Luo and Poeppel 2007, where A_{ntf} is the momentary power at time *t,* frequency *f* and trial *n*:

$$ITPowC_{tf}=\sqrt{\sum_{n=1}^{N}\frac{\left(A_{ntf}^{2}-\bar{A}_{tf}^{2}\right)^{2}}{\overline{A_{tf}}}}$$

Note that while larger ITPC values indicate stronger phase coherence, smaller ITPowC values indicate stronger power coherence across trials. For statistical purposes, we compared the ITPC and ITPowC values calculated across trials in which the same sentence was presented (*within-group*) to values calculated across trials with difference sentences (*between-group*) which serve as an estimation of chance level (Luo and Poeppel 2007). The number of trials used to calculate the between-group signal was matched to the within-group trial number. Overall there were 108 trials per condition (18 repetitions per condition x 6 sentences), resulting in 432 trials total (note that 216 trials were recorded, but trials were differently combined for the attend/ignored condition, resulting in the analysis of 432 trials). After artifact rejection, the following amount of trials was analyzed: attend/ignore natural, mean = 107.14 (sd = 1.7); attend/ignore vocoded: mean = 107.14 (sd = 1.4). Between-group (comparison/control) trials were pseudo-randomly selected from all other trials of the condition. In order to ensure that the between-group signal contained a sufficiently heterogeneous mixture of different sentences, we also ensured that the same sentence was not present in more than 6 trials (either as attended or unattended). In order to focus on responses from auditory cortex, for each hemisphere, data were averaged across the 10 sensors showing maximal M100 amplitude responses in the separate auditory localizer tasks (pure tones). Furthermore, for each participant ITPC and ITPowC values were averaged across all sentences separately for each condition, for both the within- and between-groups. Prior to statistical analyses, mean ITPC and ITPowC values were normalized using a rau transform. Paired-sample t-tests across participants were used to test for significant differences between within- and between-group ITPowC and ITPC, collapsed across conditions, at each frequency level (2-100 Hz). All further analyses were performed on the difference between within-group and between-group ITPC, referred to as the "ITPC dissimilarity values" (Luo & Poeppel, 2007), collapsed across the frequencies where the ITPC was found to be significantly higher than chance, namely 2-8 Hz. Next, a 3-way *Analysis of Variance* (ANOVA) with repeated measures was used to test the modulation of dissimilarity across the different conditions, with the within-subjects factors of *Attention* (attend, ignore), *Stimulus type* (natural, vocoded) and *Hemisphere* (left, right). For all analyses, the Greenhouse-Geisser correction was applied when the assumption of sphericity was violated (Greenhouse and Geisser 1959).

3 Results

3.1 Behavioral Results

Task performance was moderately high for both conditions: *Attend natural* mean d' = 2.1 (sd = 1.1), mean HR = 0.77 (sd = 0.14); mean FAR = 0.13, (sd = 0.14); *attend vocoded*: mean d' = 1.5 (sd = 1.8), mean HR = 0.71 (sd = 0.28); mean FAR = 0.28 (sd = 0.23). Importantly, performance accuracy (d') did not differ between *attend natural* and *attend vocoded* conditions (no effect of condition, t(13) = 1.78, p = .101).

3.2 MEG Results

Within- and between group ITPowC did not differ significantly at any frequency, consistent with previous research (e.g., Luo and Poeppel 2007; Howard and Poeppel 2010). Within- and between-group ITPC was significantly different in the range from 2-8 Hz (Bonferroni corrected: ps < 0.0005; Fig. 3a; Fig. 3b displays the time course of the speech-tracking response).

A 3-way ANOVA with repeated measures testing within-subjects factors *Attention* (attend, ignore), *Stimulus type* (natural, vocoded) and *Hemisphere* (left, right) (Fig. 4a displays the mean dissimilarity values in all conditions) showed a main effect of Stimulus type, with larger ITPC dissimilarity for *natural* compared to *vocoded* sentences [F(1,13) = 6.62, p < .05 , η^2_p = .34]. Post hoc one-sample Students' t-tests show that the dissimilarity was significantly larger than zero for both *natural* and *vocoded* conditions (averaged across *Hemisphere* and *Attention*; *natural*: p < .001; vocoded: p < .001; Bonferroni corrected alpha = .025). In addition, there was a tendency towards a main effect of *Attention*, with an increase in ITPC dissimilarity in the *attend* compared to *ignore* conditions [F(1,13) = 3.53, p = .083, η^2_p = .21]. Here too, post hoc one-sample Students' t-tests show that the dissimilarity was significantly larger than zero for both attend and ignore conditions (averaged across *Hemisphere* and *Stimulus type*; *attend*: p < .001; *ignore*: p < .001; Bonferroni corrected alpha = .025).

Importantly, there was a significant *Attention* × *Stimulus type* interaction [F(1,13) = 9.95, p < .01, η^2_p = .43] (Fig. 4b). Post hoc paired-sample Students' t-tests show significant differences between *attend* and *ignore* condition only for *natural* (p = .004), but not for *vocoded* sentences (p = .4). The dissimilarity was higher in *attend natural* compared to *attend vocoded* conditions (p = .007), while it was of a similar magnitude in *ignore natural* compared to *ignore vocoded* conditions (p = .508; Bonferroni corrected alpha = .0125). There was no main effect of *Hemisphere* (p = .271) nor significant interactions between *Attention* × *Hemisphere* (p = .178) and *Attention* × *Stimulus type* × *Hemisphere* (p = .51). However, the interaction between *Stimulus type* and *Hemisphere* was significant [F(1,13) = 4.79, p < .05, η^2_p = .27], as differences between response to *natural* and *vocoded* condition were only significant in the left hemisphere (p = .015) but not in the right hemisphere (p = .332; Bonferroni corrected alpha = .025).

4 Discussion

Reliable speech-tracking responses were found in response to all presented stimuli, regardless of their type or attentional status, manifested by significant 2-8 Hz phase-locking in auditory sensors. However, the strength of speech-tracking, reflecting its consistency across trials, differed across experimental conditions. The novel finding of this study is that speech-tracking was modulated by attention only for natural but not for vocoded speech, and that speech tracking of unattended speech was similar to that observed for vocoded speech. This pattern of results has two implications. First, in the absence of attention, the speech-tracking response mostly reflects the cortical representation of the temporal envelope of the stimulus and not analysis of its fine structure, as no differences were found in responses to unattended natural-speech and vocoded speech. Second, attentional effects on speech-tracking responses cannot solely be explained by feature-based attention, as no attentional modulation was observed for vocoded speech; rather they most likely reflect a robust contribution of higher-order linguistic processing, applied more extensively to attended vs. unattended speech. In the following we discuss the implication of these findings on understanding the neural mechanisms of speech processing, particularly we consider a predictive coding interpretation. Furthermore, we also discuss an alternative "acoustic" explanation,

4.1 Bottom-up contributions to speech tracking in auditory cortex

Here we show that the previously reported enhanced speech tracking for natural compared to vocoded speech (Luo and Poeppel 2007; Howard and Poeppel 2010; Peelle et al. 2012; Wild, Davis, et al. 2012; Ding et al. 2014) holds only when these stimuli are attended, whereas we found a similar degree of speech tracking for unattended natural and vocoded speech. The present findings are not consistent with the proposal of Ding et al. (2014) that analysis of the fine structure features also contributes to the speech-tracking response, at least not without attention, as no difference in speech tracking was found for unattended natural vs. vocoded speech. These findings also suggest that, at its core, the speech-tracking response primarily represents the temporal envelope of the acoustics of the entire auditory scene and is not speech-specific (Lalor et al. 2009; Howard and Poeppel 2010; Cogan and Poeppel 2011; Hämäläinen et al. 2012; Wang et al. 2012; Millman et al. 2013;

Steinschneider et al. 2013), however it can be influenced and enhanced by additional top-down processes.

4.2 Selective attention and speech intelligibility

The observed attentional effect on speech-tracking of attended vs. ignored speech is in accordance with several previous neurophysiological studies (Kerlin et al. 2010; Ding and Simon 2012a, 2012b; Horton et al. 2013; Zion Golumbic et al. 2013b). However, in those studies it was difficult to determine whether attentional effects were due to low-level feature-based attention, for example attending to the pitch of the speakers' voice, or whether linguistic processing of the speech content contributed to the attentional effect. In the current experimental design there was great physical difference between the concurrently presented natural and vocoded stimuli, which could have been used to direct attention at the feature-level. Therefore, the lack of an attentional effect on the tracking of vocoded stimuli reported here, suggests that attention was not guided by the global physical attributes of the stimuli. Rather, these findings indicate that a) attentional enhancement of speech tracking depends on the presence of fine-structure in the stimulus, and b) that the fine-structure information in natural speech is utilized to enhance speech tracking only when it is attended. We interpret this selective utilization of fine-structure for attended natural speech as reflecting linguistic processing, as it is well established linguistic processing is substantially reduced for unattended speech (Cherry 1953; Treisman 1986; Bentin et al. 1995; Mulligan 1998; Power et al. 2012; Sklar et al. 2012). Note that linguistic processing of attended speech is thought to occur relative *automatically* independent of the specific task (Warren and Marslen-Wilson 1987; Neely 1991; see also Peelle et al. 2010). Thus, although in our paradigm linguistic processing was not explicitly required, it was most likely applied to attended speech.

This interpretation in line with two previous brain imaging studies Sabri and colleagues (2008) compared brain activations to speech and acoustically similar non-speech signals (words/pseudo words vs. spectrally rotated speech) when they were attended or ignored. They found that when stimuli were attended natural speech elicited higher activation than rotated signals in postcentral gyri, left supramarginal gyrus and the temporal lobes bilaterally. However when the auditory stimuli were ignored and participants engaged in a demanding visual task, activation in these areas was similar for all auditory stimuli. These findings support the hypothesis that speech undergoes only *low level* sensory processing when it is ignored, whereas higher-order linguistic processing is mostly applied when speech is attended. Another relevant study is work by Wild and colleagues (Wild, Yusuf, et al. 2012) who studied the interaction between attention and speech intelligibility. They reported strong attentional effects on neural responses to highly-intelligible though degraded (six-band noise-vocoded speech; Smith et al. 2002) speech in STS and left inferior frontal gyrus, whereas responses to non-intelligible (noise-vocoded and spectrally rotated) speech were not modified by attention, at least when the competing stimulus was also auditory. The results of Wild and colleagues differ from ours on several accounts, for example they did not find any attentional modulation of responses for clear speech, however this may be explained by the easy nature of the attentional task (attending to a visual signal or an auditory "chirp"), which did not provide enough distraction. In contrast, the highly-intelligible degraded-speech condition used by Wild et al. more closely resembles our cocktail party paradigm, and thus their results are compatible with our findings. The notion that higher order linguistic processing is applied to speech to a larger extent when it is attended vs. ignored is also consistent with findings by Zion Golumbic et al. (2013b) who found that while neural activity in auditory cortex tracked the envelope of both attended and unattended speech, selective tracking only of attended speech was found in higher-order brain regions such as inferior-frontal and parietal areas.

These findings do not preclude the possibility that some degree of linguistic processing is applied to unattended speech, as implied by effects of covert priming (Beaman et al. 2007) and occasional attentional capture (Wood and Cowan 1995). However, as suggested by previous work, semantic processing of unattended speech is sharply reduced and mostly automatic and unavailable for conscious recollection (Cherry 1953; Treisman 1986; Bentin et al. 1995; Mulligan 1998; Power et al. 2012; Sklar et al. 2012).

One seemingly contradictory finding to the present results is reported by Millman et al. (2014) who fail to find any effect of intelligibility on the speech-tracking response to tone-vocoded speech. This null effect, which is also in contrast to several previous studies that found effects of intelligibility on sensory processing of speech (e.g. Sohoglu et al. 2012), does suggest that there may be cases where intelligibility alone is insufficient to drive an increase in speech-tracking. Yet, it is insufficient to negate the claim that when a significant increase in speech tracking is observed – as has been repeatedly found for natural vs. vocoded speech or attended vs. unattended speech – these effects may reflect additional linguistic processing. Indeed, we postulate that the claim that speech-tracking reflects only the acoustic features of a stimulus cannot be reconciled with the present pattern of our results, but rather that the response may be enhance by additional linguistic processing.

4.3 Contributions of top-down predictions to speech-tracking

It is proposed that enhanced speech tracking in auditory cortex reflects more precise temporal coherence between low-frequency neural activity and the speech stimulus, such that neural resources can be optimally allocated at appropriate points in time to process upcoming events in the attended speech stream (Luo and Poeppel 2007; Schroeder et al. 2008, 2010; Zion Golumbic et al. 2012). Arguably, improvement in the temporal precision of speech tracking is afforded by predictive cues about upcoming input provided by sensory information, as demonstrated for audio-visual speech (Arnal et al. 2011; Zion Golumbic et al. 2013a). The precise nature of the processing facilitation that predictive cues offer remains largely unknown. In the case of linguistic processing of speech, several sources of information carry predictive value that can ostensibly be used to enhance temporal expectation for upcoming input in auditory cortex. For example, analysis of lexical structure as well as prosodic cues can improve the precision of temporal prediction as to the timing and duration of upcoming words. Another possibility is that semantic and syntactic analysis of speech generates predictions as to the specific words that will be uttered next (Sohoglu et al. 2012; Dikker and Pylkkanen 2013), allowing for anticipation and more precise tracking of the fine-structure content in narrow-band envelopes (see Peelle 2013).

We have no definitive way of determining whether the increased speech tracking of attended natural speech in the current study is indeed due to some form of temporal predictions.

However, this interpretation converges with theories of *predictive coding* which postulate that predictive knowledge interacts with attention to influence sensory processing (Kok et al. 2012). For example, in a visual fMRI study, Kok et al. showed greater attentional effects on BOLD activity in primary visual cortices when predictive information about an upcoming stimulus was available, compared to when no predictive information was available (neutral cue; see also Tian and Poeppel 2013; Schröger et al. 2014). Thus theories of predictive coding may provide a useful framework for interpreting the current results and for generating predictions for follow-up studies.

4.4 Alternative explanations

Thus far, we have interpreted the interaction between attention and stimulus type (natural vs. vocoded) as reflecting the influence of top-down effects on speech tracking, rather than relying entirely on acoustic input. However, an alternative "bottom-up" interpretation may be considered, which postulates that since natural speech is acoustically richer than vocoded speech, attention can operate more effectively at the sensory level for enhancing the processing of attended natural speech, since it can utilized more acoustic features. This account could, theoretically, lead to a similar pattern of findings as reported here, with larger attentional effects for natural vs. vocoded speech. Under the current experimental design we cannot rule out this possibility, and it is entirely possible that part of the observed effect is also driven by more accurate feature-based attention at the sensory level for natural speech. However, we maintain that this perspective cannot entirely account for our results, and particularly for the lack of a difference in speech tracking for natural and vocoded speech when they are unattended. Future studies are needed, in which responses to unintelligible speech (with complex fine structure) are also measured, in order to unequivocally determine the relative contribution of bottom-up acoustics and top-down linguistic processing to the speech-tracking responses in auditory cortex.

Another set of factors to consider in the interpretation concerns stream segregation. A prerequisite for attentional selection is the ability to segregate speech streams originating from different talkers. For non-speech stimuli it has been shown that informationally rich acoustic structure can facilitate stream segregation (e.g., Bendixen et al. 2010; for review: Winkler 2007; Schröger et al. 2013; Bendixen 2014), therefore, one might wonder whether the attentional effects reported here for natural speech actually reflect a benefit in stream segregation rather than stream selection. However, in the current experiment, natural and vocoded speech were always presented together, thus there was no difference in stream segregation difficulty between conditions. Therefore, the observed effects are probably due to selection, and not segregation.

Finally, our results could be considered in light of the possibility that the higher speech-tracking in the *attend natural* compared to *attend vocoded* condition is due to differential distractibility. According to this view, natural speech always captures more attention due to its higher intelligibility, and therefore the *attend vocoded* condition may suffer from higher distractibility since natural speech is presented in the background (for a similar argument see Sabri et al. 2008). However, since we found no significant difference in the speech tracking response to the *ignore natural* and *ignore vocoded* conditions, it is unlikely that they differed in the degree of distraction they afforded. Moreover, similar task performance in these conditions makes the distraction explanation less likely.

5 Conclusions

We show that while the temporal envelopes of attended and ignored sounds are tracked concurrently in auditory cortex, attended speech is tracked more effectively possibly due to effects of linguistic processing. In contrast, attention does not affect the speech-tracking response for vocoded speech, which lacks the fine-structure of natural speech and is substantially less intelligible. These results are testament to the tight link and mutual influence of high-order linguistic processes and the sensory processing of speech, and to the crucial role of attention in determining the depth of processing applied to incoming stimuli.

Acknowledgments

This study was supported by the Erasmus Mundus program of the European Union 2009-5259/003-001-EMA2 to JMR., the German Research Foundation SFB936/B2 to JMR., the I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation (grant No. 51/11)” to EZG, a DFG-Reinhart-Koselleck-grant to ES, and NIH 2R01DC05660 to DP.

References

Adachi Y, Shimogawara M, Higuchi M, Haruta Y, Ochiai M. Reduction of non-periodic environmental magnetic noise in MEG measurement by continuously adjusted least squares method. Appl Supercond IEEE Trans On. 2001; 11:669–672.

Ahveninen J, Hämäläinen M, Jääskeläinen IP, Ahlfors SP, Huang S, Lin F-H, Raij T, Sams M, Vasios CE, Belliveau JW. Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. Proc Natl Acad Sci. 2011; 108:4182–4187. [PubMed: 21368107]

Arnal LH, Wyart V, Giraud A-L. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci. 2011; 14:797–801. [PubMed: 21552273]

Beaman CP, Bridges AM, Scott SK. From dichotic listening to the irrelevant sound effect: a behavioural and neuroimaging analysis of the processing of unattended speech. Cortex J Devoted Study Nerv Syst Behav. 2007; 43:124–134.

Bendixen. Predictability effects in auditory scene analysis: A review. Front Neurosci. 2014 in press.

Bendixen A, Denham SL, Gyimesi K, Winkler I. Regular patterns stabilize auditory streams. J Acoust Soc Am. 2010; 128:3658–3666. [PubMed: 21218898]

Bentin S, Kutas M, Hillyard SA. Semantic processing and memory for attended and unattended words in dichotic listening: behavioral and electrophysiological evidence. J Exp Psychol Hum Percept Perform. 1995; 21:54–67. [PubMed: 7707033]

Cherry EC. Some Experiments on the Recognition of Speech, with One and with Two Ears. J Acoust Soc Am. 1953; 25:975–979.

Cogan GB, Poeppel D. A mutual information analysis of neural coding of speech by low-frequency MEG phase information. J Neurophysiol. 2011; 106:554–563. [PubMed: 21562190]

Dark VJ, Johnston WA, Myles-Worsley M, Farah MJ. Levels of selection and capacity limits. J Exp Psychol Gen. 1985; 114:472–497. [PubMed: 2934499]

Dikker S, Pylkkanen L. Predicting language: MEG evidence for lexical preactivation. Brain Lang. 2013; 127:55–64. [PubMed: 23040469]

Ding N, Chatterjee M, Simon JZ. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. NeuroImage. 2014; 88:41–46.

Ding N, Simon JZ. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci. 2012a; 109:11854–11859. [PubMed: 22753470]

Ding N, Simon JZ. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol. 2012b; 107:78–89. [PubMed: 21975452]

Ding N, Simon JZ. Cortical Entrainment to Continuous Speech: Functional Roles and Interpretations. Front Hum Neurosci. 2014; 8:311. [PubMed: 24904354]

Doelling KB, Arnal LH, Ghitza O, Poeppel D. Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. New Horiz Neural Oscil. 2014; 85:761–768. Part 2.

Ghitza O, Giraud A-L, Poeppel D. Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. Front Hum Neurosci. 2013; 6:340. [PubMed: 23316150]

Ghitza O, Greenberg S. On the Possible Role of Brain Rhythms in Speech Perception: Intelligibility of Time-Compressed Speech with Periodic and Aperiodic Insertions of Silence. Phonetica. 2009; 66(suppl 1-2):113–126. [PubMed: 19390234]

Giraud A-L, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RSJ, Laufs H. Endogenous Cortical Rhythms Determine Cerebral Specialization for Speech Perception and Production. Neuron. 2007; 56:1127–1134. [PubMed: 18093532]

Giraud A-L, Poeppel D. Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci. 2012; 15:511–517. [PubMed: 22426255]

Green, DM.; Swets, JA. Signal detection theory and psychophysics. Wiley; New York, Ny: 1966.

Greenhouse SW, Geisser S. On methods in the analysis of profile data. Psychometrika. 1959; 24:95–112.

Hämäläinen JA, Rupp A, Soltész F, Szücs D, Goswami U. Reduced phase locking to slow amplitude modulation in adults with dyslexia: An MEG study. NeuroImage. 2012; 59:2952–2961. [PubMed: 22001790]

Hillyard SA, Hink RF, Schwent VL, Picton TW. Electrical Signs of Selective Attention in the Human Brain. Science. 1973; 182:177–180. [PubMed: 4730062]

Horton C, D'Zmura M, Srinivasan R. Suppression of competing speech through entrainment of cortical oscillations. J Neurophysiol. 2013; 109:3082–3093. [PubMed: 23515789]

Horton C, Srinivasan R, D'Zmura M. Envelope responses in single-trial EEG indicate attended speaker in a “cocktail party.”. J Neural Eng. 2014; 11:046015. [PubMed: 24963838]

Howard MF, Poeppel D. Discrimination of Speech Stimuli Based on Neuronal Response Phase Patterns Depends on Acoustics But Not Comprehension. J Neurophysiol. 2010; 104:2500–2511. [PubMed: 20484530]

Howard MF, Poeppel D. The neuromagnetic response to spoken sentences: Co-modulation of theta band amplitude and phase. NeuroImage. 2012; 60:2118–2127. [PubMed: 22374481]

Jacoby LL, Allan LG, Collins JC, Larwill LK. Memory influences subjective experience: Noise judgments. J Exp Psychol Learn Mem Cogn. 1988; 14:240–247.

Johnsrude IS, Mackey A, Hakyemez H, Alexander E, Trang HP, Carlyon RP. Swinging at a Cocktail Party: Voice Familiarity Aids Speech Perception in the Presence of a Competing Voice. Psychol Sci. 2013; 24:1995. [PubMed: 23985575]

Kalikow DN, Stevens KN, Elliott LL. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. J Acoust Soc Am. 1977; 61:1337–1351. [PubMed: 881487]

Kerlin JR, Shahin AJ, Miller LM. Attentional Gain Control of Ongoing Cortical Speech Representations in a “Cocktail Party.”. J Neurosci. 2010; 30:620–628. [PubMed: 20071526]

Kok P, Rahnev D, Jehee JFM, Lau HC, Lange FP. Attention Reverses the Effect of Prediction in Silencing Sensory Signals. Cereb Cortex. 2012; 22:2197. [PubMed: 22047964]

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE. Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection. Science. 2008; 320:110–113. [PubMed: 18388295]

Lalor EC, Power AJ, Reilly RB, Foxe JJ. Resolving precise temporal processing properties of the auditory system using continuous stimuli. J Neurophysiol. 2009; 102:349–359. [PubMed: 19439675]

Liberman MC. The cochlear frequency map for the cat: labeling auditory-nerve fibers of known characteristic frequency. J Acoust Soc Am. 1982; 72:1441–1449. [PubMed: 7175031]

Luo H, Poeppel D. Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. Neuron. 2007; 54:1001–1010. [PubMed: 17582338]

Macmillan, NA.; Creelman, CD. Detection theory: A users's guide. 2nd ed.. Erlbaum; Mahwah, NJ: 2005.

Makeig S, Bell AJ, Jung T-P, Sejnowski TJ. Independent component analysis of electroencephalographic data. Adv Neural Inf Process Syst. 1996; 8:145–151.

Miller GA, Isard S. Some perceptual consequences of linguistic rules. J Verbal Learn Verbal Behav. 1963; 2:217–228.

Millman RE, Johnson SR, Prendergast G. The Role of Phase-locking to the Temporal Envelope of Speech in Auditory Perception and Speech Intelligibility. J Cogn Neurosci. 2014 in press.

Millman RE, Prendergast G, Hymers M, Green GGR. Representations of the temporal envelope of sounds in human auditory cortex: Can the results from invasive intracortical “depth” electrode recordings be replicated using non-invasive MEG “virtual electrodes”? NeuroImage. 2013; 64:185–196. [PubMed: 22989625]

Mulligan NW. The role of attention during encoding in implicit and explicit memory. J Exp Psychol Learn Mem Cogn. 1998; 24:27–47. [PubMed: 9438952]

Neely JH. Semantic priming effects in visual word recognition: A selective review of current findings and theories. Basic Process Read Vis Word Recognit. 1991:264–336.

Obleser J, Kotz SA. Expectancy Constraints in Degraded Speech Modulate the Language Comprehension Network. Cereb Cortex. 2010; 20:633–640. [PubMed: 19561061]

Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia. 1971; 9:97–113. [PubMed: 5146491]

Oostenveld R, Fries P, Maris E, Schoffelen J-M. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Intell Neurosci. 2011; 2011:1–9.

Peelle JE. Cortical responses to degraded speech are modulated by linguistic predictions. Proc Meet Acoust. 2013; 19:1–5.

Peelle JE, Gross J, Davis MH. Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. Cereb Cortex. 2012 in press.

Peelle JE, Johnsrude I, Davis MH. Hierarchical processing for speech in human auditory cortex and beyond. Front Hum Neurosci. 2010; 4:51. [PubMed: 20661456]

Posner MI, Driver J. The neurobiology of selective attention. Curr Opin Neurobiol. 1992; 2:165–169. [PubMed: 1638148]

Power AJ, Foxe JJ, Forde E-J, Reilly RB, Lalor EC. At what time is the cocktail party? A late locus of selective attention to natural speech. Eur J Neurosci. 2012; 35:1497–1503. [PubMed: 22462504]

Rimmele J, Jolsvai H, Sussman E. Auditory Target Detection Is Affected by Implicit Temporal and Spatial Expectations. J Cogn Neurosci. 2011; 23:1136–1147. [PubMed: 20146603]

Rimmele J, Sussman E, Poeppel D. The role of temporal structure in the investigation of sensory memory, auditory scene analysis, and speech perception: A healthy-aging perspective. Int J Psychophysiol. 2014 in press.

Sabri M, Binder JR, Desai R, Medler DA, Leitl MD, Liebenthal E. Attentional and linguistic interactions in speech perception. NeuroImage. 2008; 39:1444–1456. [PubMed: 17996463]

Schroeder CE, Lakatos P. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 2009; 32:9–18. [PubMed: 19012975]

Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A. Neuronal oscillations and visual amplification of speech. Trends Cogn Sci. 2008; 12:106–113. [PubMed: 18280772]

Schroeder CE, Wilson DA, Radman T, Scharfman H, Lakatos P. Dynamics of Active Sensing and perceptual selection. Cogn Neurosci. 2010; 20:172–176.

Schröger E, Bendixen A, Denham S, Mill R, Bőhm T, Winkler I. Predictive Regularity Representations in Violation Detection and Auditory Stream Segregation: From Conceptual to Computational Models. Brain Topogr. 2013:1–13.

Schröger E, Marzecová A, SanMiguel I. Attention and Prediction in Human Audition: A lesson from Cognitive Psychophysiology. Eur J Neurosci. 2014 in press.

Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M. Speech Recognition with Primarily Temporal Cues. Science. 1995; 270:303–304. [PubMed: 7569981]

Sheldon S, Pichora-Fuller MK, Schneider BA. Effect of age, presentation method, and learning on identification of noise-vocoded words. J Acoust Soc Am. 2008; 123:476–488. [PubMed: 18177175]

Sklar AY, Levy N, Goldstein A, Mandel R, Maril A, Hassin RR. Reading and doing arithmetic nonconsciously. Proc Natl Acad Sci. 2012; 109:19614–19619. [PubMed: 23150541]

Smith ZM, Delgutte B, Oxenham AJ. Chimaeric sounds reveal dichotomies in auditory perception. Nature. 2002; 416:87–90. [PubMed: 11882898]

Sohoglu E, Peelle JE, Carlyon RP, Davis MH. Predictive Top-Down Integration of Prior Knowledge during Speech Perception. J Neurosci. 2012; 32:8443–8453. [PubMed: 22723684]

Steinschneider M, Nourski KV, Fishman YI. Representation of speech in human auditory cortex: is it special? Hear Res. 2013; 305:57–73. [PubMed: 23792076]

Tian X, Poeppel D. The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. J Cogn Neurosci. 2013; 25:1020–1036. [PubMed: 23469885]

Tiitinen HT, Sinkkonen J, Reinikainen K, Alho K, Lavikainen J, Naatanen R. Selective attention enhances the auditory 40-Hz transient response in humans. Nature. 1993; 364:59–60. [PubMed: 8316297]

Treisman A. Features and objects in visual processing. Sci Am. 1986; 255:114–125.

Wang L, Zhu Z, Bastiaansen M. Integration or predictability? A further specification of the functional role of gamma oscillations in language comprehension. Front Psychol. 2012; 3:187. [PubMed: 22701443]

Warren P, Marslen-Wilson W. Continuous uptake of acoustic cues in spoken word recognition. Percept Psychophys. 1987; 41:262–275. [PubMed: 3575084]

Wild CJ, Davis MH, Johnsrude IS. Human auditory cortex is sensitive to the perceived clarity of speech. NeuroImage. 2012; 60:1490–1502. [PubMed: 22248574]

Wild CJ, Yusuf A, Wilson DE, Peelle JE, Davis MH, Johnsrude IS. Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. J Neurosci. 2012; 32:14010–14021. [PubMed: 23035108]

Winkler I. Interpreting the Mismatch Negativity. J Psychophysiol. 2007; 21:147–163.

Woldorff MG, Gallen CC, Hampson SA, Hillyard SA, Pantev C, Sobel D, Bloom FE. Modulation of early sensory processing in human auditory cortex during auditory selective attention. Proc Natl Acad Sci. 1993; 90:8722–8726. [PubMed: 8378354]

Wood N, Cowan N. The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel? J Exp Psychol Learn Mem Cogn. 1995; 21:255–260. [PubMed: 7876773]

Zekveld AA, Kramer SE, Kessens JM, Vlaming MSMG, Houtgast T. The benefit obtained from visually displayed text from an automatic speech recognizer during listening to speech presented in noise. Ear Hear. 2008; 29:838–852. [PubMed: 18633325]

Zion Golumbic E, Cogan GB, Schroeder CE, Poeppel D. Visual Input Enhances Selective Speech Envelope Tracking in Auditory Cortex at a “Cocktail Party.”. J Neurosci. 2013; 33:1417–1426. [PubMed: 23345218]

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE. Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a “Cocktail Party.”. Neuron. 2013; 77:980–991. [PubMed: 23473326]

Zion Golumbic EM, Poeppel D, Schroeder CE. Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. Brain Lang. 2012; 122:151–161. [PubMed: 22285024]

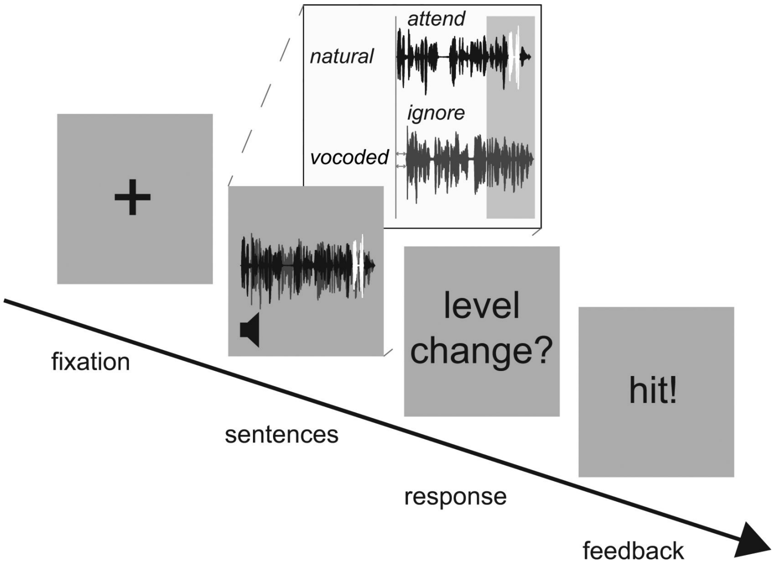

Figure 1. Schematic of the paradigm

Prior to the onset of sentence presentation, participants hold their gaze at a fixation-cross. In each trial a pair of sentences – one natural and one vocoded - were presented simultaneously and monaurally through earphones. One stimulus started 500ms before the other, and participants were instructed to attend to the early onset stimulus (black waveform) and ignore the other stimuli (gray waveform; the onset delay is indicated by gray arrows). A brief loudness increase could occur in either the attended or ignored sentence (displayed in white), at a randomly chosen time point in the last 2-3 seconds of the sentence (within the time window highlighted in gray). After sentence offset participants indicated whether there was a loudness increase in the attended sentence (*response*) and received *feedback*. Each trial was followed by an 800 ms intertrial interval.

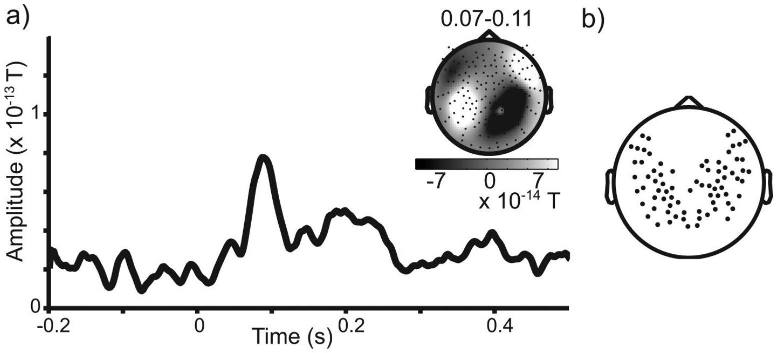

Figure 2. Localization of auditory sensors

a) The event-related field (ERF) to simple tones in the auditory localizer task is demonstrated for one participant. The peak time of the M100 component is identified as maximum amplitude around 0.1 s post stimulus onset (calculated on the root-mean-square of the data across sensors). The topography of the M100 at its peak window (0.07-0.11 s) is displayed above the ERP. Sensors with maximal M100 amplitude were selected individually for each participant (10 for the left and 10 for the right hemisphere) and used in subsequent analysis; b) The figure shows all M100 sensors across participants (black dots indicate a sensor that was beneath the maximum sensors of at least one participant).

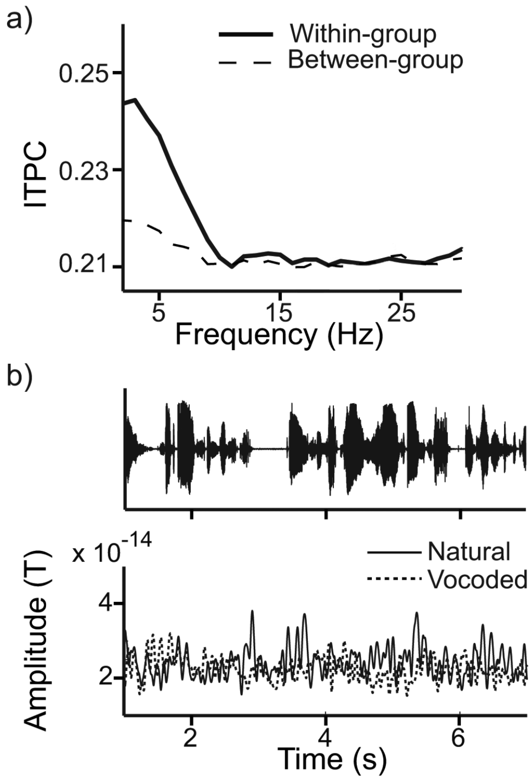

Figure 3. ITPC dissimilarity – Spectral and temporal course

a) The within-group ITPC (thick solid line) and between-group ITPC (thin dashed line) for each frequency, averaged across participants and conditions. ITPC was significantly larger than chance for frequencies between 2-8Hz; **b)** The time course (x-axis: 1-7 s) of the speech–tracking response is displayed for one sentence. The preprocessed MEG data were filtered in the relevant frequency range (high-pass cut-off: 2 Hz; low-pass cut-off: 8 Hz; two-pass butterworth filter; filter order: 6). For each participant and acoustic condition (*natural*, *vocoded*) the data was averaged across all trials were the sentence was attended (~18 trials per condition) and across maximal M100 sensors of the left hemisphere. The grand average across participants is displayed for the *attend natural* (solid line) and *attend vocoded* (dashed line) condition. The waveform of the sentence (*natural*) is displayed above the MEG signal (x-axis: 1-7 s).

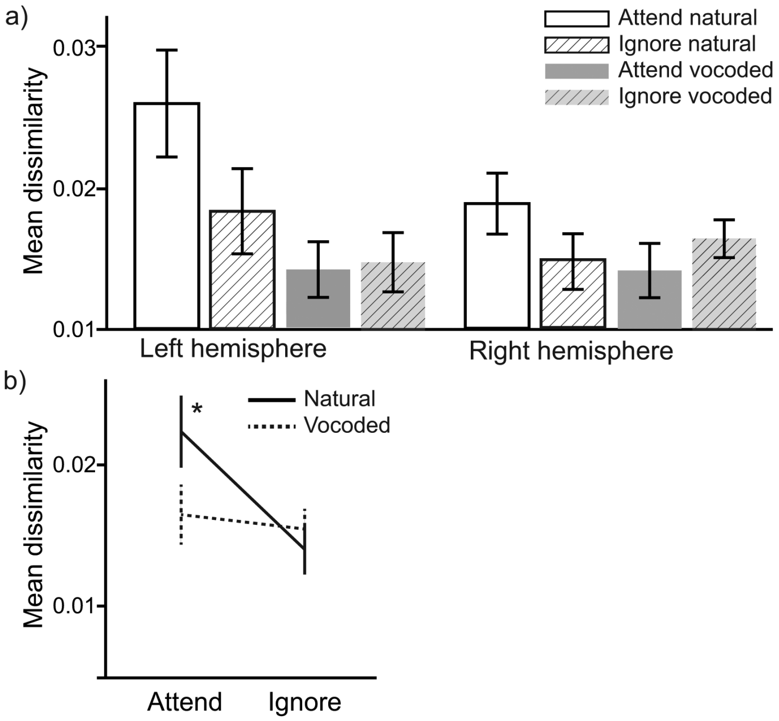

Figure 4. Effects of attention and acoustic on ITPC dissimilarity

a) The mean dissimilarity values for natural (white bar) and vocoded (gray) speech, in both the attend (plain) and ignore (shaded) conditions, averaged across auditory sensors separately for the left and right hemisphere. Error bars indicate the standard error of the mean (+− 1); **c)** The mean dissimilarity values for original (solid line), vocoded (dashed line), in the attend (left) and ignore (right) conditions, averaged across auditory sensors and collapsed over hemispheres. Error bars indicate the standard error of the mean (+− 1). The dissimilarly was significantly higher in the *attend original* compared to the *ignore natural* condition, whereas there was no attentional effect for vocoded speech.