The Brain Dynamics of Rapid Perceptual Adaptation to Adverse Listening Conditions

Julia Erb,1 Molly J. Henry,1 Frank Eisner,2 and Jonas Oleser1
1Max Planck Research Group “Auditory Cognition,” Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany and 2Max Planck Institute for Psycholinguistics, 6525 XD Nijmegen, The Netherlands

Listeners show a remarkable ability to quickly adjust to degraded speech input. Here, we aimed to identify the neural mechanisms of such short-term perceptual adaptation. In a sparse-sampling, cardiac-gated functional magnetic resonance imaging (fMRI) acquisition, human listeners heard and repeated back 4-band-vocoded sentences (in which the temporal envelope of the acoustic signal is preserved, while spectral information is highly degraded). Clear-speech trials were included as baseline. An additional fMRI experiment on amplitude modulation rate discrimination quantified the convergence of neural mechanisms that subserve coping with challenging listening conditions for speech and non-speech. First, the degraded speech task revealed an “executive” network (comprising the anterior insula and anterior cingulate cortex), parts of which were also activated in the non-speech discrimination task. Second, trial-by-trial fluctuations in successful comprehension of degraded speech drove hemodynamic signal change in classic “language” areas (bilateral temporal cortices). Third, as listeners perceptually adapted to degraded speech, downregulation in a cortico-striato-thalamo-cortical circuit was observable. The present data highlight differential upregulation and downregulation in auditory–language and executive networks, respectively, with important subcortical contributions when successfully adapting to a challenging listening situation.

Introduction

Humans have the capability to rapidly adapt to degraded or altered speech. This challenge is particularly relevant to cochlear implant (CI) patients adapting to an extremely distorted auditory input delivered by their device (Giraud et al., 2001; Fallon et al., 2008). The current study investigated the short-term neural processes that underlie this adaptation to degraded speech using functional magnetic resonance imaging (fMRI).

We simulated CI-transduced speech in normal-hearing listeners using noise-vocoding (Shannon et al., 1995), which degrades the spectral detail in an auditory signal, and forces the listener to rely more on the (intact) temporal envelope cues for speech comprehension. Listeners with higher sensitivity to envelope fluctuations in an auditory signal, as measured by amplitude modulation (AM) rate discrimination thresholds, adapt more quickly to vocoded speech (Erb et al., 2012). Therefore, we predicted that temporal envelope of non-speech sounds (Giraud et al., 2000) and vocoded speech should be processed by shared neural resources.

Rapid perceptual learning of vocoded speech is a well-established finding of behavioral studies (Rosen et al., 1999; Davis et al., 2005; Peelle and Wingfield, 2005; Bent et al., 2009), but its neural bases are largely unknown. Recently, increased precentral gyrus activity was associated with boosted perceptual learning during the joint presentation of vocoded and clear words (Hervais-Adelman et al., 2012). Further, in a simulated CI rehabilitation program that supplemented presentation of vocoded sentences with simultaneous written feedback, perceptual learning relied on the left inferior frontal gyrus (IFG; Eisner et al., 2010). However, in everyday situations, listeners rarely receive direct feedback. Therefore, the neural dynamics of feedback-free or self-regulated adaptation are investigated here.

Subcortical structures likely play a critical role in adaptation to degraded speech, but the specific contributions of distinct structures are unclear. A recent voxel-based morphometry study demonstrated that gray matter volume in the left pulvinar thalamus predicted how fast listeners adapted to vocoded speech (Erb et al., 2012). However, previous fMRI studies have failed to detect adaptation-related signal changes in subcortical regions, possibly due to these brain areas’ susceptibility to MR artifacts. Here, we implemented a cardiac-gated image acquisition procedure, thereby avoiding heartbeat-related artifacts.

The present fMRI study sheds new light on perceptual adaptation to degraded speech with respect to four points: (1) we investigate the convergence of neural mechanisms underlying effortful speech and non-speech processing; (2) we test feedback-free, short-term adaptation; (3) unlike previous perceptual learning studies (Golestani and Zatorre, 2004; Adank and Devlin, 2010; Eisner et al., 2010), we collect word report scores on every trial and are able to model both trial-by-trial fluctuation in comprehension and adaptation to degraded speech; and (4) we assess subcortical contributions to adaptation. We show that an “exec-
For envelope extraction, we used a second-order, zero-phase bandpass-filtered noise carriers, thereby smearing spectral fine structure. Extracts the amplitude envelope from each band, and reapplies it to vocoding. This procedure divides the raw signal into frequency bands, between 1620 and 2760 ms. Sentences were degraded using 4-band noise which is controlled for the predictability of the final word (high vs low financial compensation of speech in noise (SPIN) sentences (Kalikow et al., 1977; Erb et al., 2012).

Experiment 1.

Stimuli and experimental design

A. Oscillogram (top) and spectrogram (bottom) of the sentence "Gewöhnlich nutzen wir die Räder" (Normally we use the bikes) in degraded (left) and clear speech (right). B. Trial structure in Experiment 1 and 2. After acoustic presentation of a sentence or AM stimulus (lasting for ~2.5 s), participants were prompted to respond when green lights appeared on a screen (lasting for 3 s) by repeating the sentence or pressing a button. Subsequent scan acquisition was initiated by cardiac gating. The onset of auditory stimulation preceded the anticipated scan acquisition by a period of 6.5 s, although the actual scan time was variable due to cardiac gating. C. Experiment 1 comprised 100 degraded speech trials, 24 clear speech trials (inserted after every fifth trial), and 20 randomly interspersed silent trials.

Figure 1. Stimuli and experimental design. A. Oscillogram (top) and spectrogram (bottom) of the sentence "Gewöhnlich nutzen wir die Räder" (Normally we use the bikes) in degraded (left) and clear speech (right). B. Trial structure in Experiment 1 and 2. After acoustic presentation of a sentence or AM stimulus (lasting for ~2.5 s), participants were prompted to respond when green lights appeared on a screen (lasting for 3 s) by repeating the sentence or pressing a button. Subsequent scan acquisition was initiated by cardiac gating. The onset of auditory stimulation preceded the anticipated scan acquisition by a period of 6.5 s, although the actual scan time was variable due to cardiac gating. C. Experiment 1 comprised 100 degraded speech trials, 24 clear speech trials (inserted after every fifth trial), and 20 randomly interspersed silent trials.

The sentences were recorded by a female speaker of standard German in a soundproof chamber. The length of the recorded sentences varied between 1620 and 2760 ms. Sentences were degraded using 4-band noise vocoding. This procedure divides the raw signal into frequency bands, extracts the amplitude envelope from each band, and reapplies it to bandpass-filtered noise carriers, thereby smearing spectral fine structure. For envelope extraction, we used a second-order, zero-phase Butterworth lowpass filter with a cutoff frequency of 400 Hz. Noise-vocoding was applied to all sentences in MATLAB 7.11 as described in Rosen et al. (1999) using four frequency bands spanning 70–9000 Hz that were spaced according to Greenwood’s cochlear frequency-position function (Greenwood, 1990; for full settings see Erb et al., 2012). The waveform and spectrogram of the vocoded and clear version of an example sentence are shown in Figure 1A.

On each trial, participants first heard a sentence. They were instructed to repeat as much of the sentence as they had understood when a green light appeared on the screen, but to stop talking when the green light disappeared (after 3 s) to avoid movement during scan acquisition (Fig. 1B). Speech production was recorded for later off-line scoring (Eckert et al., 2009; Harris et al., 2009). Responses were scored as proportions of correctly repeated words per sentence ("report scores"; Peelle et al., 2013). Scoring took into account all words of a sentence including function words; errors in declension or conjugation were accepted as correct.

Clear speech trials were used as a high-level baseline. Clear speech can be assumed to be fully adapted, and therefore to be processed in a stable way over time. This ensured that no neural adaptation would occur in the baseline condition, whereas another type of artificial speech degradation (e.g., rotated speech) might have led to neural adaptation (even in the absence of behavioral adaptation).

In sum, Experiment 1 comprised three conditions: (1) 4-band vocoded sentences ("degraded speech"); 100 trials in total), (2) clear (non-vocoded) sentences ("clear speech"); 24 trials in total), and (3) trials lacking any auditory stimulation ("silent trials"; 20 trials in total). Overall, the experiment comprised 144 trials. Clear speech trials were presented every fifth trial, whereas the silent trials were randomly interspersed (Fig. 1C). Sentences were presented to each participant in one of four different orders; presentation order was counterbalanced across participants.

Participants’ adaptation curves. As in Erb et al. (2012), we modeled each individual’s performance improvement in two different ways: as a power law and as a linear performance increase. To test which function would better describe the data, both curves were fitted to the individual report scores over time using a least-squares estimation procedure in MATLAB 7.11 (for example fits to the scores averaged over participants, see Fig. 2A). We compared goodness of fit by determining the Bayesian information criterion (BIC; Schwarz, 1978) of the linear and the power law fits within each participant.

Experiment 2. Stimuli were 1 s long sinusoidally amplitude-modulated white noises. The standard stimulus was modulated at 4 Hz. Deviant stimuli were modulated at seven different rates that were linearly spaced between 2 and 6 Hz in steps of 0.67 Hz. The middle level was modulated at the same rate as the standard (4 Hz); critically, in this condition participants were unable to distinguish standard and deviant, but still performed the same task. Modulation depth was kept constant at 60%. The onset phase of the sinusoidal modulation was randomly varied for all stimuli, separately for the standard and deviant. Standard and deviant stimuli were presented with an interstimulus interval of 500 ms.

The paradigm was a two-alternative forced choice task: on each trial, participants first heard the standard stimulus (modulated at 4 Hz) followed by one deviant stimulus. After auditory presentation of the sound pair, participants were prompted to respond when a green light appeared on the screen (Fig. 1B). Half of the participants were to indicate which sound had a faster modulation rate, while the other half’s task was to
indicate which sound had a slower modulation rate to counterbalance button presses. Participants responded via a button box in their right hand by pressing the left (for first sound) or right key (for second sound).

In sum, there were seven levels of deviant AM rate (comprising 16 trials each); in addition, we interspersed silent trials lacking any auditory stimulation (16 trials). On the whole, every participant listened to 128 trials in a pseudorandom order where trials of the same condition were never presented subsequently.

**Experimental procedure**

To maximize the comparability between individuals, all participants were tested in the same order, namely degraded speech perception first (Experiment 1), followed by AM rate discrimination (Experiment 2). Before participants went into the scanner, they were familiarized with each of the two tasks; they listened to three 8-band vocoded GSPIN sentences as training for Experiment 1 and three exemplary trials of Experiment 2.

In the scanner, to prevent hearing damage due to scanner noise, participants wore Alpine Musicsafe Pro earplugs, yielding approximately linear 14 dB reduction in sound pressure up to 8 kHz. Auditory stimuli were delivered through MR-Confon headphones using Presentation software. Visual prompts were projected on a screen which participants viewed via a mirror.

Trial timing was identical in Experiment 1 and 2. Each trial was ~9 s long, but actual trial length varied due to cardiac gating (see below). Trials started with a 1 s silent gap, after which participants heard a sentence (Experiment 1) or two AM-modulated stimuli (Experiment 2) lasting for ~2.5 s. Following stimulus presentation (3.5 s into the trial), a green light (“go”-signal for response) was visually presented and lasted for 3 s. After ~1 s of silence, scan acquisition with a TR of 2 s was triggered using cardiac gating. Thus, the onset of auditory stimulation preceded the anticipated scan acquisition by ~6.5 s (Fig. 1B).

**MRI data acquisition**

MRI data were collected on a 3 T Siemens Verio scanner. Blood oxygenation level-dependent (BOLD) fMRI images were acquired with a 12-channel head coil using an echo-planar imaging (EPI) sequence [TR = 9000 ms, TA = 2000 ms, TE = 30 ms, flip angle = 90°, 3 mm slice thickness, 30 axial slices (ascending), interslice distance = 1 mm, acquisition matrix of 64 × 64, voxel size = 3 × 3 × 3 mm] in two separate runs for Experiment 1 and 2. The acquisition matrix was placed such that the x-axis was in line with the anterior–posterior commissure (AC–PC). We used a sparse-sampling procedure, where TR was longer than TA, allowing for silent periods to play stimuli and record verbal responses (Hall et al., 1999).

Additionally, cardiac gating was applied to avoid movement artifacts caused by the heartbeat in subcortical structures (von Kriegstein et al., 2008), in which we were especially interested. Participants’ heartbeat was monitored using an MR-compatible pulse oximeter (Siemens) attached to their left ring finger. On each trial, after 9 s had elapsed, the scanner waited for the first heartbeat to trigger volume acquisition. Thus, the actual repetition time (TR) was variable but amounted to 9.45 ± 0.27 s (mean ± SEM; across all participants).

- **Figure 2.** **A**, Behavioral adaptation to degraded speech. Linear (red solid line) and power law fit (dashed line) to mean report scores averaged over participants (left). Two examples of participants with a steep slope (top) and a shallow slope (bottom) in their linear fits (“adaptation curves”). Although there is a general increase in performance, it is clearly visible that report scores strongly fluctuate from trial to trial (right). **B**, Effects of auditory stimulation in Experiment 1. Auditory cortices (HG, STG) and the MGB were active when participants listened to sound (degraded and clear speech) compared with silent trials, confirming that the hemodynamic response to auditory stimulation was captured by the scans. L, left; R, right. Error bars indicate SEM.
176 sagittal slices, acquisition matrix of 256 × 240, voxel size = 1 mm\(^3\)). In one participant, we were only able to acquire 136 (as opposed to 144) scans for Experiment 1 due to technical problems with cardiac gating. A second participant had to be excluded from all analyses concerning Experiment 2, because scan acquisition had become desynchronized with stimulus presentation.

Data analysis
Preprocessing. MRI data were analyzed in SPM8 (Wellcome Trust Centre for Neuroimaging, London, UK). Preprocessing was performed separately for Experiment 1 and 2. Structural MRI scans were manually aligned with the coordinate system such that AC–PC was in line with the x-axis and AC in the origin of the coordinate system. Functional images were realigned and unwarped using a field map, coregistered with the structural scan, segmented and normalized to standard space (Montreal Neurological Institute [MNI] space) using the segmentation parameters, and smoothed with an isotropic Gaussian kernel of 8 mm full-width at half-maximum (FWHM).

MR images were statistically analyzed in the context of the general linear model. We set up three different models for Experiment 1 and one model for Experiment 2 to assess the following effects:

**Effects of auditory stimulation.** In a basic model for Experiment 1, we defined three conditions at the single subject level: degraded speech, clear speech, and silent trials. The effect of auditory stimulation was tested by contrasting sound (degraded and clear speech) against silent trials. To avoid overspecification, silent trials were removed from all subsequent analyses.

**Effects of speech degradation and trial-by-trial fluctuation with comprehension.** In a speech degradation model, we included two conditions: degraded and clear speech. Additionally, a parametric modulator of the degraded speech trials was defined, representing the behavioral report scores. A regressor of no interest, containing report latencies, was added to account for differences in speech production (analysis explained in detail below); this regressor was included in all remaining analyses. To assess effects of stimulus clarity, we contrasted degraded against clear speech trials. To reveal effects of trial-by-trial fluctuations in speech comprehension, we assessed correlations with the regressor representing report scores.

**Effects of perceptual adaptation.** To model effects of adaptation, we looked for signal changes over time corresponding to participants’ slow performance increase (“adaptation curves”). However, there are a number of unspecific reasons why the BOLD signal could gradually change over time, for example, scanner drift or fatigue of the participant. Therefore, we compared the changes over time for vocoded speech to the change in activity seen for the clear speech condition, while taking into account the behaviorally observed adaptation to vocoded speech: rather than simply testing a condition × time interaction (which would not consider the adaptation curve of a listener), we tested for the three-way interaction for condition × time × behavior.

To this end, we created two parametric modulators, one for each condition, by multiplying time (i.e., trial number) with the linear adaptation curves. This resulted in a quadratic curve for the degraded speech regressor of which the slope was dependent on the individual adaptation curve (Fig. 5, top). However, since there was no perceptual adaptation in the clear speech condition (all sentences were fully comprehended), the “adaptation” curve was flat such that multiplication effectively left the time regressor unchanged, resulting in a linear curve.

Contrasts tested for the difference in these two regressors, ruling out the possibility that a slow change over time is due to slow unspecific signal drifts (which would be present in both conditions; analysis as in Büchel et al., 1998). Thus, we identified areas where changes over time are more pronounced in the degraded speech condition (where perceptual processing improves) than in the clear speech condition (where perceptual processing remains stable throughout).

**Regressor of no interest for report latency.** Although the present study was designed to image degraded speech perception, parts of the observed activity may be related to speech production or preparation, because participants overtly repeated back what they had understood starting ~3.5 s before scan acquisition (Fig. 1B). In particular, participants’ verbal responses might have been faster for clear relative to degraded speech trials, perhaps leading to partly imaging the BOLD response to speech production, but more so for clear speech trials. Therefore, differences in report latencies might confound the comparison between degraded and clear speech trials. Similarly, adaptation-related signal changes might become confounded, as report latencies likely decrease as participants adapt to degraded speech. To control for this potential confound, we calculated report latency relative to the onset of the visual response cue (Fig. 1B). This measure was included at the first level as one single regressor of no interest in all models concerning Experiment 1 (described above). For trials where participants did not produce an overt response, the subject-specific mean report latency was entered.

*Effects of ∆AM rate.* In Experiment 2 we modeled four conditions; one for each level of AM rate difference between standard and deviant (referred to as “∆AM rate”): ∆AM rate = 0 Hz, ±0.67 Hz, ±1.33 Hz, and ±2 Hz (note: in the ∆AM rate = 0 Hz condition, the deviant was modulated also at 4 Hz). To assess a linear correlation with ∆AM rate, these conditions were weighted with the contrast vector [−3 −1 1 3] for a positive correlation and [3 1 −1 3] for a negative correlation with ∆AM rate. Large effects in these contrasts would mean linear scaling with deviation from the standard AM rate. In an additional conjunction analysis (Friston et al., 1999), we tested for the intersection of the effects of ∆AM rate and speech degradation.

All described analyses were whole brain analyses. Regressors were modeled using a finite impulse response comprising one bin. A high-pass filter with a cutoff of 1024 s was applied to eliminate low-frequency noise. No correction for serial autocorrelation was necessary because of the long TR in the sparse-sampling acquisition.

Second level statistics were calculated using a one-sample t-test. Group inferences are reported at a familywise error (FWE) corrected voxelwise threshold of \( p < 0.05 \), where FWE rate was controlled using random field theory. Only for the adaptation analyses did we use a slightly more lenient threshold of \( p < 0.001 \), where cluster-extent (\( k > 20 \)) was corrected based on a Monte Carlo Simulation (Slotnick et al., 2003). T-statistic maps were transformed to Z-statistic maps using spm.122.m, and overlaid and displayed on the ch2 template in MNI space included with MRICron (Rorden and Brett, 2000).

**Lateralization analysis.** Processes of comprehension are likely left-lateralized (Oblieser et al., 2007; Rosen et al., 2011; McGgettigan et al., 2012a). We therefore tested for lateralization of activity related to trial-by-trial fluctuation in speech comprehension. As in the analyses described above, EPI images first were realigned, unwarped, and coregistered. The images were then segmented using symmetric gray matter, white matter, and CSF templates, which were created by averaging each original template with itself flipped along the y-axis (Salmond et al., 2000). The segmentation parameters were used to normalize the images to MNI-space, resulting in a normalization to a symmetric MNI template (Liegeois et al., 2002) and smoothed at 8 mm FWHM. To conduct a voxel-by-voxel statistical test of laterality, the first level analyses were performed as described above. Resulting maps were flipped along the y-axis and compared with the original maps in a paired t-test (Bociz et al., 2010).

**Regions of interest analyses.** To extract measures of percentage signal change in the regions identified by the whole-brain analyses described above, we defined regions of interest (ROIs) using the SPM toolbox MarsBar (Brett et al., 2002). ROIs were defined as spheres with a radius of 3 mm centered on the identified peak coordinates. Voxels within an ROI were aggregated into a single contrast estimate using the first eigenvariate.

**Results**

**Behavioral results.** In Experiment 1, participants reported on average 51.9 ± 1.4% (mean ± SEM) words correctly per degraded sentence. Performance in clear trials was at 99.7 ± 0.2% correct.

We compared whether a linear or power law function (Fig. 2A) would better describe the report scores’ increase over time by...
calculating the BIC for each fit and each participant. The BIC scores for the linear fits (median 242.44, range 217.41–265.11) were smaller than those for power law fits (median 247.98, range 217.93–269.75), as shown by a Wilcoxon signed-rank test (p < 0.001), indicating that the linear curve better fit the behavioral data. Thus, we chose the linear fit to describe the participants’ improvement (“adaptation curve”).

A one-tailed t-test on the slopes of the adaptation curves showed that they were significantly greater than zero (t(29) = 13.18, p < 0.001), indicating that participants adapted to degraded speech, i.e., that their speech perception performance improved over the course of the experiment (Fig. 2A).

In Experiment 2, all participants performed well on the discriminable AM stimuli (mean ± SEM = 96.9 ± 0.7%). A paired samples t-test showed that there was no improvement from the first half (96.9 ± 0.7%) to the second half (96.7 ± 0.9%) of the experiment (t(29) = 0.34, p = 0.74).

Controlling for speech production-related activations
To dissociate perceptual and response demands, we estimated report latencies. Report latency in degraded speech trials (977 ± 72 ms, mean ± SEM) was significantly longer than in clear speech trials (662 ± 50 ms, p = 0.003). Moreover, report latency decreased over time, based on a t-test on the slopes of linear fits to report latencies as a function of trial (–2.5 ± 0.26 × 10^{-3}, mean ± SEM; t(29) = 9.94, p < 0.001). Therefore, report latency was regressed out in all fMRI analyses. Importantly, analyses without this regressor of no interest yielded very similar fMRI activation patterns (results not reported). We take this as strong evidence that the observed effects are not driven by speech production, but rather perception.

fMRI results
We found extensive activation of the auditory system when contrasting sound against silent trials, in Experiment 1 (Heschl’s gyrus, HG; planum temporale; superior temporal gyrus, STG; medial geniculate body, MGB; Fig. 2B) as well as in Experiment 2 (results not shown), confirming that the BOLD response to auditory stimulation was captured by scan acquisition.

Effects of physical speech degradation: degraded versus clear speech
To reveal regions that are modulated by physical speech degradation or clarity, we compared vocoded with clear speech trials. Areas where degraded relative to clear speech yielded an increased BOLD signal included the left supplementary motor area/anterior cingulate cortex (SMA/ACC), anterior insula, and caudate nucleus bilaterally. In contrast, clear compared with degraded speech yielded higher activations bilaterally in the precentral gyrus spanning the temporal cortices, supramarginal gyrus (SMG), putamen, posterior cingulate cortex, and angular gyrus bilaterally (Fig. 3A, Table 1).

AM rate discrimination and degraded speech processing
The magnitude of the AM rate difference between standard and deviant stimuli (ΔAM rate) correlated positively with activity in right HG, left amygdala, and SMG (Fig. 3B; Table 1), signifying that activity increased with larger deviance from the standard (and thus easier AM rate discrimination). Conversely, ΔAM rate correlated negatively with activity in the ACC/SMA, left insula, and IFG, meaning that the signal increased as modulation rate differences diminished and discrimination became more difficult.

In a series of conjunction analyses, we identified regions commonly involved in processing of degraded speech and of AM stimuli: a conjunction analysis between (1) positive correlation with ΔAM rate and (2) clear > degraded speech was significant in the SMG bilaterally and posterior cingulate. A conjunction of the inverse contrasts (negative correlation with ΔAM rate ∩ degraded > clear speech) yielded a significant cluster in the SMA/ACC, insula bilaterally, and left IFG (Table 1). There were no commonly activated areas for the “cross-over” conjunction “negative correlation with ΔAM rate ∩ clear > degraded speech” or vice versa.

Trial-by-trial fluctuation with speech comprehension
To reveal areas reflecting trial-by-trial fluctuations in speech comprehension, we tested for correlations with the behavioral report scores. The BOLD signal linearly increased with improved comprehension of degraded speech in bilateral temporal cortices comprising HG, superior temporal sulcus (STS), left IFG, pre-central gyrus bilaterally, the putamen, the thalamus including MGB bilaterally, left angular gyrus, frontal medial cortex, posterior cingulate, and cerebellum (Fig. 4). There were no negative correlations between the fMRI signal and report scores. An additional voxel-by-voxel laterality analysis of this speech comprehension effect (see Materials and Methods) revealed that activity in the angular gyrus was significantly left-lateralized (Table 1). This is apparent in Figure 4, where angular gyrus activity is seen in the sagittal slice of the left hemisphere only.

Effects of perceptual adaptation
To identify brain areas that change over time as a function of adaptation, we compared changes over time between degraded and clear speech while taking into account the behaviorally observed adaptation, i.e., we tested the time × condition × behavior interaction (Fig. 5, top). Thus, we separated changes over time due to adaptation (which only occurred in the degraded speech) from unspecific slow signal changes (which would be present in both conditions and hence be canceled out in the contrast). As the behavioral change is marked for vocoded speech but virtually absent for clear speech, the BOLD signal is expected to change at different rates (i.e., more change in the vocoded condition because, on top of scanner drift, adaptation-related signal changes are present). Therefore, we modeled a more pronounced increase in the vocoded (quadratic increase) than in the clear speech condition (linear increase). This model separates uninteresting signal drifts from the effects due to the behaviorally observed adaptation.

At a threshold of p(FWE) < 0.05, we found stronger downregulation over time in degraded relative to clear speech trials as listeners adapted in the anteroventral thalamic nucleus (Morel, 2007). At the slightly more lenient threshold of p < 0.001 (cluster-extent corrected), the caudate, frontal regions and an occipital cluster spanning from the cerebellum to fusiform gyrus up to the precuneus showed the same effect. Conversely, an activity increase over time in degraded more than clear speech with adaptation was found in the left precentral gyrus and posterior cingulate cortex (Fig. 5; Table 1).

Discussion
The present study was designed to reveal neural systems that support rapid perceptual adaptation to degraded speech and contributes three major novel findings. First, degraded more than clear speech activates an “executive” system, which overlaps with the neural substrates of difficult auditory discrimination, namely
in the insula and SMA/ACC. Second, BOLD signal in a “langu-

age” network, comprising auditory, premotor cortices, and left
angular gyrus, depended on speech comprehension rather than


physical clarity of the stimulus. Third, the data provide first evi-

dence that self-regulated perceptual adaptation to degraded

speech co-occurs with a BOLD downregulation of subcortical

structures.

Executive mechanisms in speech and non-speech processing

Degraded more than clear speech evoked enhanced activity in

anterior areas including the insula and the SMA/ACC. This has

been observed consistently for difficult comprehension (Giraud

et al., 2004; Eckert et al., 2009). Clear speech, in contrast, revealed

an expected activity increase in the bilateral temporal cortices
(Scott et al., 2000; Davis and Johnsrude, 2003; Giraud et al., 2004;

Wild et al., 2012a).

When examining to which extent difficult speech and non-

speech perception rely on joint neural substrates, a conjunction

revealed SMA/ACC and the anterior insula bilaterally. This com-

parison pertains directly to contributions of bottom-up (i.e., fi-

delity of AM representations in the ascending auditory pathway)

versus top-down mechanisms (e.g., attentional processes) in

degraded speech perception. For example, frequency discrimina-

tion learning shows partial specificity to the trained frequency
(Amitay et al., 2006), consistent with a bottom-up account. In

contrast, top-down mechanisms of attention are plausibly in-

volved in perceptual learning (Halliday et al., 2011). Training

improves the ability to attend to a task-specific stimulus dimen-

sion, and discrimination learning occurs even in the absence of a
discriminable stimulus difference (Amitay et al., 2006). Similarly,
speech degradation studies have found evidence for bottom-up
accounts (Sebastián-Galleés et al., 2000; Hervais-Adelman et al.,

2008; Idemaru and Holt, 2011) as well as top-down accounts of

perceptual learning, in which lexical information aids perceptual

adaptation (Davis and Johnsrude, 2003; Davis et al., 2005).

The structures commonly recruited for degraded speech pro-

cessing and AM discrimination are clearly not specific to auditory

envelope processing, but have been suggested by a number of

studies to be involved rather in top-down, executive processes
(Adank, 2012). Eckert et al. (2009) demonstrated that the insula

and SMA/ACC are engaged when tasks become increasingly dif-

ficult, independent of modality or task, suggesting that these re-

gions subserve executive processes. Consistently, the anterior

insula and SMA showed an enhanced BOLD signal when listeners

attended to speech (rather than a distracter), and the speech sig-

nal was increasingly degraded (vocoded rather than clear; Wild et

al., 2012b). More specifically, these regions are a resource for

attention and performance monitoring processes (Dosenbach

et al., 2006, 2007; Sadaghiani and D’Esposito, 2012). Thus, we argue

that the anterior insula and SMA/ACC fulfill executive processes,

and that the recruitment of these executive components is pivotal

for a wide range of challenging listening situations.

Figure 3. A, Areas sensitive to speech degradation. Degraded relative to clear speech yielded increased activation in the anterior insula, the caudate, the SMA/ACC, and IFG (in red). Conversely, clear compared with degraded speech revealed increased activity in the temporal cortices, putamen, posterior cingulate, precentral gyrus, SMG and angular gyrus (AG). Estimated percentage signal change for vocoded and clear trials is shown for selected regions. B, Effects of ΔAM rate in Experiment 2 and overlaps with Experiment 1. As AM rate difference (ΔAM rate) between standard and deviant increased, activity reduced in the SMA/ACC and insula (red). Activity increased with higher ΔAM rate in right Heschl’s gyrus, left SMG and amygdala (blue). The contours of the Z-statistic maps showing effects of speech degradation are overlaid to illustrate overlaps in activation. A conjunction analysis between the effects of negative correlation with ΔAM rate and degraded > clear speech was significant in the SMA/ACC, insula bilaterally, and the IFG; conjunction of the inverse contrasts (positive correlation with ΔAM rate ∧ clear > degraded speech) reached significance in the posterior cingulate and SMG bilaterally. L, left; R, right. Error bars indicate SEM.
Table 1. Overview of MRI activation clusters, thresholded at voxel-wise \( p(FWE) < 0.05 \), unless indicated differently

<table>
<thead>
<tr>
<th>Location</th>
<th>MNI coordinates</th>
<th>Z-Score</th>
<th>Extent (mm(^3))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degraded clear speech</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L SMA/ACC</td>
<td>-6</td>
<td>7.19</td>
<td>3060</td>
</tr>
<tr>
<td>L anterior insula</td>
<td>-30</td>
<td>7.05</td>
<td>3825</td>
</tr>
<tr>
<td>R anterior insula</td>
<td>33</td>
<td>5.84</td>
<td>576</td>
</tr>
<tr>
<td>L caudate nucleus</td>
<td>-9</td>
<td>5.18</td>
<td>162</td>
</tr>
<tr>
<td>R caudate nucleus</td>
<td>12</td>
<td>5.45</td>
<td>171</td>
</tr>
<tr>
<td>Clear degraded speech</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R superior frontal gyrus</td>
<td>6</td>
<td>6.55</td>
<td>7866</td>
</tr>
<tr>
<td>R precentral gyrus</td>
<td>54</td>
<td>6.5</td>
<td>28647</td>
</tr>
<tr>
<td>R putamen</td>
<td>27</td>
<td>7.19</td>
<td></td>
</tr>
<tr>
<td>R MTG</td>
<td>63</td>
<td>6.79</td>
<td></td>
</tr>
<tr>
<td>R angular gyrus</td>
<td>51</td>
<td>7.94</td>
<td></td>
</tr>
<tr>
<td>L precentral gyrus</td>
<td>-54</td>
<td>7.59</td>
<td>24183</td>
</tr>
<tr>
<td>L putamen/globus pallidus</td>
<td>-27</td>
<td>6.87</td>
<td></td>
</tr>
<tr>
<td>L HG</td>
<td>-54</td>
<td>7.18</td>
<td></td>
</tr>
<tr>
<td>L angular gyrus</td>
<td>-51</td>
<td>7.24</td>
<td></td>
</tr>
<tr>
<td>L MTG</td>
<td>-60</td>
<td>6.69</td>
<td>1044</td>
</tr>
<tr>
<td>L posterior cingulate</td>
<td>-6</td>
<td>7.47</td>
<td>20700</td>
</tr>
<tr>
<td>Positive correlation with ΔAM rate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R HG</td>
<td>51</td>
<td>5.95</td>
<td>270</td>
</tr>
<tr>
<td>L amygdala</td>
<td>-24</td>
<td>5.37</td>
<td>135</td>
</tr>
<tr>
<td>L SMG</td>
<td>-57</td>
<td>5.17</td>
<td>279</td>
</tr>
<tr>
<td>Negative correlation with ΔAM rate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L SMA/ACC</td>
<td>-6</td>
<td>6.19</td>
<td>1143</td>
</tr>
<tr>
<td>L anterior insula</td>
<td>-30</td>
<td>4.85</td>
<td>54</td>
</tr>
<tr>
<td>L IFG, pars opercularis</td>
<td>-48</td>
<td>5.16</td>
<td>270</td>
</tr>
<tr>
<td>Positive correlation with ΔAM rate ( \cap ) clear degraded speech</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L SMG</td>
<td>-57</td>
<td>5.22</td>
<td>477</td>
</tr>
<tr>
<td>R SMG</td>
<td>60</td>
<td>5.08</td>
<td>261</td>
</tr>
<tr>
<td>L posterior cingulate cortex</td>
<td>0</td>
<td>5.14</td>
<td>486</td>
</tr>
<tr>
<td>Negative correlation with ΔAM rate ( \cap ) degraded clear speech</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L SMA/ACC</td>
<td>-6</td>
<td>7.15</td>
<td>1602</td>
</tr>
<tr>
<td>L anterior insula</td>
<td>-30</td>
<td>5.65</td>
<td>324</td>
</tr>
<tr>
<td>R anterior insula</td>
<td>33</td>
<td>5.25</td>
<td>360</td>
</tr>
<tr>
<td>L IFG, pars opercularis</td>
<td>-48</td>
<td>5.17</td>
<td>315</td>
</tr>
<tr>
<td>Positive correlation with trial-by-trial fluctuations in comprehension</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L frontal medial cortex</td>
<td>-3</td>
<td>6.62</td>
<td>765</td>
</tr>
<tr>
<td>L IFG</td>
<td>-51</td>
<td>5.75</td>
<td>243</td>
</tr>
<tr>
<td>L STS</td>
<td>-57</td>
<td>7.47</td>
<td>18279</td>
</tr>
<tr>
<td>L precentral gyrus</td>
<td>-54</td>
<td>6.58</td>
<td></td>
</tr>
<tr>
<td>L HG</td>
<td>-57</td>
<td>6.67</td>
<td></td>
</tr>
<tr>
<td>L putamen</td>
<td>-30</td>
<td>6.21</td>
<td></td>
</tr>
<tr>
<td>L angular gyrus</td>
<td>-51</td>
<td>7.50</td>
<td></td>
</tr>
<tr>
<td>R precentral sulcus</td>
<td>57</td>
<td>6.74</td>
<td>11475</td>
</tr>
<tr>
<td>R HG</td>
<td>57</td>
<td>6.7</td>
<td></td>
</tr>
<tr>
<td>R putamen</td>
<td>30</td>
<td>6.34</td>
<td></td>
</tr>
<tr>
<td>L thalamus</td>
<td>-12</td>
<td>5.56</td>
<td>885</td>
</tr>
<tr>
<td>R thalamus</td>
<td>12</td>
<td>5.57</td>
<td>306</td>
</tr>
<tr>
<td>L posterior cingulate</td>
<td>-3</td>
<td>6.22</td>
<td>2376</td>
</tr>
<tr>
<td>R cerebellum</td>
<td>15</td>
<td>6.08</td>
<td>3141</td>
</tr>
<tr>
<td>L cerebellum</td>
<td>-12</td>
<td>5.73</td>
<td></td>
</tr>
<tr>
<td>Lateralization of comprehension network</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L angular gyrus</td>
<td>-45</td>
<td>5.83</td>
<td>378</td>
</tr>
<tr>
<td>Decrease with time × behavior for degraded clear speech, ( p &lt; 0.001, k &gt; 20 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L superior frontal gyrus</td>
<td>-27</td>
<td>3.79</td>
<td>666</td>
</tr>
<tr>
<td>R MFG</td>
<td>33</td>
<td>4.89</td>
<td>729</td>
</tr>
<tr>
<td>R anteroventral thalamus</td>
<td>3</td>
<td>5.07</td>
<td>7929</td>
</tr>
<tr>
<td>R caudate</td>
<td>18</td>
<td>4.48</td>
<td></td>
</tr>
<tr>
<td>L fusiform gyrus</td>
<td>-24</td>
<td>4.84</td>
<td>19116</td>
</tr>
<tr>
<td>Increase with time × behavior for degraded clear speech, ( p &lt; 0.001, k &gt; 20 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L precentral gyrus</td>
<td>-54</td>
<td>3.8</td>
<td>189</td>
</tr>
<tr>
<td>L posterior cingulate</td>
<td>0</td>
<td>4.05</td>
<td>333</td>
</tr>
</tbody>
</table>

L, left; R, right; ACC, anterior cingulate cortex; HG, Heschl’s gyrus; IFG, inferior frontal gyrus; MFG, middle frontal gyrus; MTG, middle temporal gyrus; SMA, supplementary motor area; SMG, supramarginal gyrus; STG, superior temporal gyrus; STS, superior temporal sulcus.
Revisiting the speech comprehension network

A number of earlier imaging studies manipulated physical stimulus features to vary speech intelligibility and found sensitivity to these manipulations along the superior temporal gyrus and sulcus, often extending into prefrontal and inferior parietal regions (Scott et al., 2000; Davis and Johnsrude, 2003; Zekveld et al., 2006; Obleser et al., 2007; Obleser and Kotz, 2010). In contrast, the present study held physical stimulus properties constant (i.e., 4-band vocoding). Therefore, we were able to identify regions where activation varied with actual speech comprehension (i.e., behavioral report scores), independent of acoustic differences.

For a given trial, activity in a perisylvian network was tightly coupled to comprehension. Importantly, these areas (Fig. 4) overlapped largely with those activated by clear compared with degraded speech (Fig. 3A, blue Z-statistic maps). Together, this is strong evidence that the observed network supports sensorimotor operations involved in successful comprehension (and, hence, successful repetition) rather than simply indexing sensitivity to physical stimulus characteristics.

Such linguistic processes of comprehension have been proposed to be largely processed in the left hemisphere (McGettigan et al., 2012a; Peelle, 2012). Concordantly, we found evidence for a left-lateralization of activity in the angular gyrus, a structure that is associated with semantic processing (Ferstl and von Cramon, 2002; for review see Price, 2012) and has been suggested to facilitate speech comprehension when signal quality declines (Obleser et al., 2007; Obleser and Kotz, 2010).

Cortical contributions to perceptual adaptation

One important objective here was to identify and track on-line the neural systems supporting adaptation to degraded speech. Depending on adaptation, activity increased over time in cortical areas of the premotor cortex and posterior cingulate and decreased in frontal and occipital areas.

The premotor cortex has been suggested to mediate successful perceptual learning of degraded speech by mapping the unfamil-
iar auditory signal onto existing articulatory representations of speech sounds. Consistent with this idea, adaptation to time-compressed speech is reflected in a downregulation of activity in the ventral premotor cortex (Adank and Devlin, 2010). Hervais-Adelman et al. (2012) observed precentral sulcus activation, when a vocoded word was paired with its clear representation, a type of feedback that is known to enhance perceptual learning (Davis et al., 2005). Similarly, we note in the present study, although with feedback-free learning, that activity in the premotor cortex increases as a listener adapts to vocoded speech. Thus, perceptual learning might be rooted in sensorimotor integration mediated by speech–productive areas.

Enhanced activity in the posterior cingulate cortex has been noticed by Obler et al. (2007) when semantic context helped comprehension of degraded speech at an intermediate signal quality (8-band vocoded speech). This is commensurate with the current data where posterior cingulate activity not only correlated with successful speech comprehension at a given trial (Table 1), but increased as a listener adapted to degraded speech. Hence, adaptation-related increase of posterior cingulate activity might directly relate to its facilitating role in degraded speech comprehension.

Although the present experiment did not involve visual manipulations, we found occipital cortex activity (fusiform gyrus, precuneus) to correlate negatively with adaptation. While the fusiform gyrus has been implicated in audiovisual speech perception (Stevenson et al., 2010; Nath and Beauchamp, 2011; McGtigan et al., 2012b), von Kriegstein et al. (2003) observed modulation of fusiform gyrus activity during auditory speech perception, even in the absence of visual stimuli. Similarly, Giraud et al. (2001) have established that visual cortex contributes to auditory speech processing in CI patients but also in normal-hearing listeners, especially when listening to meaningful sentences (Giraud and Tuy, 2002). They hypothesized that auditory-to-visual cross-modal interaction contributes to semantic processing. In line with these studies, we speculate that initial recruitment of the visual cortex might help a listener extract meaning when first confronted with a novel form of speech degradation.

A previous study, which used a feedback-based vocoded-speech learning paradigm simulating CI rehabilitation programs (Eisner et al., 2010), found the IFG to be involved in successful adaptation. The authors attributed to the IFG a role in the “specific use of simultaneous written feedback to enhance comprehension,” in line with the view that the IFG serves as integration site for different sources of information necessary for speech comprehension (Hagoort, 2005; Rauschecker and Scott, 2009). In the present study, we did not observe IFG involvement in adaptation. Likely due to the absence of feedback, listeners might have relied on substantially different neural mechanisms to adapt to degraded speech.

Subcortical contributions to perceptual adaptation

A major novel contribution of the present study was the detection of subcortical involvement in perceptual adaptation. In previous imaging studies, such subcortical contributions might have been obscured due to heartbeat-related artifacts, which we avoided by use of a cardiac-gated scanning protocol.

We identified the anteroventral nucleus of the thalamus to be functionally involved in adaptation (Fig. 5); note that this structure is proximal to but not identical to the pulvinar, where structural differences have been reported as predictive of degraded speech adaptation before (Erb et al., 2012). Adaptation-related decrease in activity also encompassed the caudate. Although the basal ganglia have primarily been implicated in motor function, there is accumulating evidence that they play an important role in language processing (Lieberman et al., 1992; Kotz et al., 2002; Fiebach et al., 2003; Kotz, 2006). Presumably, they exert their function in language processing through their high connectivity to the cortex (Croson, 1999): The anteroventral thalamic nucleus and the caudate form part of the cortico-striato-thalamo-cortical loop, which is proposed to collect cortical information and funnel and converge it at cortical output areas, thereby reconfiguring cortical activation patterns (Kemp and Powell, 1970; O’Connell et al., 2011). In the context of adaptation to degraded speech, naïve listeners might at first be forced to rely more on information-selection processes supported by the cortico-striato-thalamo-cortical loop. Engagement of this pathway is likely to sharpen the cortical representation of a stimulus and ultimately lead to a convergence of the degraded speech signal onto a clear-speech “representation,” allowing for enhanced comprehension.

Conclusions

The present work elucidates the central neural mechanisms of rapid adaptation to acoustic speech degradation with respect to three points. First, when listening tasks become increasingly difficult, in the speech as well as the non-speech domain, listeners rely on a common executive network for “effortful listening” (Eckert et al., 2009), involving the SMA/ACC and anterior insula. Second, a persylivinary network subserves speech comprehension and fluctuates with actual comprehension rather than physical stimulus features. Finally, the present data advance the understanding of how a listener adapts to a degraded speech input, demonstrating that rapid adaptation is partly explained by hemodynamic downregulation in subcortical structures.

References


Eckert MA, Menon V, Walczak A, Ashburn J, Denckl S, Horwitz A, Dubno


Stevenson RA, Altioglu NA, Kim S, Pison DB, James TW (2010) Neural processing of