Degraded auditory and visual speech affects theta synchronization and alpha power differently

Anne Hauswald [1, 2, 3], Anne Keitel[4,5], Sebastian Rösch[6], & Nathan Weisz[1, 2, 3]

1. Center of Cognitive Neuroscience, University of Salzburg, Salzburg, Austria
2. Department of Psychology, University of Salzburg, Salzburg, Austria
3. Cimec, University of Trento, Trento, Italy
4. Psychology, School of Social Sciences, University of Dundee, Dundee, UK
5. Centre for Cognitive Neuroimaging, University of Glasgow, Glasgow, UK
6. Department of Otorhinolaryngology, Paracelsus Medical University, Salzburg, Austria

correspondence:
Anne Hauswald@sbg.ac.at

Abstract

Understanding speech can pose a challenge, especially when speech is perceived as degraded, for example when using a hearing aid. Findings on brain dynamics involved in degraded speech comprehension are mixed. We therefore investigated the effects of degraded continuous speech on three measures: intelligibility, theta synchronization, and alpha power. Additionally, we tested another commonly experienced degradation, namely that of blurred vision during visual speech perception (lip reading). Participants listened to unimodal auditory speech and watched unimodal visual speech with three different levels of degradation in a behavioural and an MEG experiment. In the auditory condition, intelligibility declined with declining clarity, implemented by fewer vocoding channels. Theta speech-brain synchronization increased with lower clarity in left auditory regions, while alpha power showed a widespread decrease. We assume that listening effort, which should be strongest for challenging conditions, led to both effects. The idea of a common process driving both measures is also consistent with the finding that increased synchronization (for stronger degradation) was associated with lower alpha power, mainly in right temporal regions. In the visual condition, intelligibility declined with increasing blurriness of the speaker's face. Theta lip-brain synchronization in bilateral visual areas decreased with degradation, while alpha power did not change, indicating that the blurry visual stimulus could not be compensated for by neural mechanisms. Together, these findings illustrate multi-layered neural mechanisms of degraded auditory and visual speech comprehension and suggest fundamental differences between both modalities.

Introduction

Understanding speech can be challenging in normal acoustic environments (e.g. background noise) or due to hearing damage. To compensate for the inferior signal-to-noise ratio of the acoustic information, reaching the central auditory system, people might allocate more attention to the incoming sounds. Accordingly, processing of degraded speech relies on allocated attention while processing of clear speech is not depending on attention (Wild et al. 2012). Likewise, subjective listening effort increases for speech sounds with fewer acoustic details or lower predictiveness (Wöstmann et al. 2015). This top-down process to improve listening performance has been linked to neural oscillations in the alpha (8-12 Hz) frequency range. While attentional demands often lead to alpha power reductions for auditory tasks (Frey et al. 2014; Weisz et al. 2014; Wöstmann et al. 2016), the findings concerning alpha power modulation in response to degraded speech are mixed. Some studies report enhanced alpha power for degraded auditory stimuli (Obleser et al. 2012; Obleser and Weisz 2012), while others report the opposite pattern (McMahon et al. 2016; Miles et al. 2017). These seeming inconsistencies may be due to variations in experimental demands and stimulus material, with simple stimuli (digits or single words) used in the former studies and more complex and longer stimuli (sentences) in the latter.

Besides the neuronal responses caused by task demands, listening to speech also elicits spontaneous temporal synchronization of auditory cortical activity to the speech sound. Different frequency bands of the speech signal are assigned different roles, with theta (4-7 Hz) capturing the syllable structure (Greenberg 1998; Poeppel 2003) and neuronal synchronization with the theta band is a crucial process for the segmentation of the continuous speech sound (Ghitza 2012; Luo and Poeppel 2007; Teng et al. 2018). This synchronization between speech and brain signals is often called speech entrainment or speech tracking, (see however Alexandrou et al. 2018) and can be calculated using different measures such as coherence (Hauswald et al. 2018), mutual information (Gross et al. 2013; A. Keitel et al. 2018), inter-trial correlation (Ding et al. 2014), dissimilarity index (Luo and Poeppel 2007) or temporal response functions (TRF, Crosse et al. 2016; Ding and Simon 2012a). Similar to the alpha rhythm, theta speech tracking is modulated by attention as well as degradation of the speech signal with studies providing mixed findings:

Reduced synchronization in this frequency range is linked to reduced intelligibility either investigated via vocoding (Ding et al. 2014; Luo and Poeppel 2007), time-reversed presentation (Gross et al. 2013; Howard and Poeppel 2010) or externally applied stimulation (Riecke et al. 2018; Zoefel et al. 2018). However, using other measures or experimental procedures show a different pattern: using TFR yields higher M50 and delta synchronization is enhanced for degraded stimuli compared to unaltered stimuli in quiet environment (Ding et al. 2014) and non-native speakers show higher speech entrainment than native-speakers presumably mediated by the higher effort (Song and Iverson 2018). Similarly, multi-speaker and auditory spatial attention studies using sentences or narratives have repeatedly found stronger speech tracking for attended compared to unattended speech (Ding and Simon 2012b; Rimmele et al. 2015) in and nearby auditory cortices (Horton et al. 2013; Zion Golumbic et al. 2013). The enhanced speech representation is positively linked to alpha lateralization, an index of attending to a specific location (Kerlin et al. 2010). Interestingly, shorter (e.g. digits) (Wöstmann et al. 2016) and degraded sounds (Rimmele et al. 2015) do not seem to lead to attention-enhanced tracking.

Besides the many studies dedicated to the effects of degraded acoustic speech signals, the visual modality has received less attention. Although degraded vision is a reality for a large number of people, for example due to a loss of acuity with aging (Rooth 2017), very few studies investigated the effects on visual speech processing. As expected from degraded acoustic speech studies, also blurred visual speech leads to a decline in lip reading performance of unimodal visual speech in younger and older adults (Tye-Murray et al. 2016). Visual mouth movements during speech contain two features, first, the movement onsets, which provide information about the timing of upcoming syllables, and second, the lips, tongue and teeth provide information about place and manner of articulation, thereby constraining lexical selection (Peelle and Sommers 2015). We assume that degradation of the visual stimulus affects these two features differently, with place and manner of articulation being more quickly affected while timing information is preserved even with strongly blurred visual speech.

EEG and MEG studies show that similarly to speech-brain synchronization in auditory regions for acoustic speech, visual cortical regions synchronize to the lip movements in low frequencies (Giordano et al. 2017; Hauswald et al. 2018; Park et al. 2016). To the best of our knowledge, no study investigated the influence of parametrically degraded visual speech signals on synchronization with the brain or the spontaneously eclicited alpha power.

In sum, several studies suggest that degradation of sounds will lead to more attention to compensate for the lower signal-to-noise ratio. At the same time, other studies propose that alpha power as well as speech tracking should be modulated by attention. Auditory spatial selective attention produces alpha lateralization (contralateral decrease and ipsilateral increase of alpha due to the crossing of the main fibers) that is stronger for attended sounds. Speech entrainment is similarly modulated as it gets stronger for attended sounds. Taken together, these lines of research suggest degraded sounds elicit more attention which in turns elicits stronger alpha modulation and stronger speech tracking. We investigated this, using degraded but intelligible continuous acoustic speech and magnetoencephalographic recordings. We further investigated effects of degradation of unimodal visual speech using the same measures. We do not statistically compare our measures to these ecologically valid auditory and visual degradations due to the different nature of the manipulations.

Methods

Participants

Twenty-eight people participated in the study (female=17, male=11). Mean age was 23.82 years (*SD* = 3.712), with a range between 19 and 37 years. We recruited only German native speakers and people who were eligible for MEG recordings, i.e. without non-removable ferromagnetic metals in or close to the body. Participants provided informed consent and were compensated monetarily or via course credit. Participation was voluntary and in line with the declaration of Helsinki and the statutes of the University of Salzburg. The study was preregistered at OSF (https://osf.io/dpt34/).

Stimuli

For the MEG recording, uni-modal video and audio files were created from audio-visual recordings of a female speaker reading Goethe's “Das Märchen” (1795) from a teleprompter, which was located behind the camera in a distance that minimized eye movements. The recordings were obtained using a digital camera (FS100, Sony Corporation) with H.264 codec in 1280 * 1080 pixels with a frame rate of 50 Hz. The view on the speaker was frontal. We created 24 unimodal audio files and 24 corresponding unimodal video files. Stimuli lengths varied between approx. 30 s and approx. 3 min, with two stimuli of 15 s, 30 s, 60 s, 90 s, 120 s, 150 s, and 12 of 180 s. Each stimulus ended with a two-syllable noun within the last four words. In order to keep participants' attention on the stimulation, we asked participants after each stimulus to choose from two presented two-syllable nouns the one that had occurred within the end of the last sentence. The syllable rate of the stimuli varied between 4.1 and 4.5 Hz with a mean of 4.3 Hz.

Vocoding of all audio stimuli was done using the vocoder toolbox for Matlab (Gaudrain and Başkent 2015) and we created conditions with 7 and 3 channels (fig. 1A). For the vocoding, the waveform of each audio stimulus was passed through two Butterworth analysis filters (for 7 and 3 channels) with a range of 200-7000 Hz, representing equal distances along the basilar membrane. Amplitude envelope extraction was done with half-wave rectification and low-pass filtering at 250 Hz. The envelopes were then normalized in each channel and multiplied with the carrier. Then, they were filtered in the band, and the RMS of the resulting signal was adjusted to that of the original signal filtered in that same band. To reduce visual detail, videos were blurred using an image-processing filter in Matlab-based Psychtoolbox

(Brainard 1997; Kleiner et al. 2007; Pelli 1997). Therefore, we filtered the videos with a Gaussian smoothing kernel width of 91 and standard deviations of 12 and 18 (fig. 1D).

Unimodal stimuli were presented to the participants in three consecutive audio-only blocks and three consecutive video-only blocks via in-ear-phones and a projector system respectively. The order of video and audio blocks was balanced. Each block contained 4 stimuli which where presented either in an unaltered version or in one of the two degraded versions. The order of the stimuli was random and did not follow the order of the original story. The assignment of stimuli to conditions was controlled in order to obtain similar overall length of stimulus presentation (approx. 400 s) for each modality and degradation levels. We instructed participants to attend to the speech which they would either see or hear. In order to keep participants' attention on the stimulation a behavioral response was required after each stimulus. At the end of each stimulus, a target and a distractor word would appear next to each other. The participants were asked to decide which of the words was presented as the last noun and within the last four words by pressing the button on the side of the response pad that matched the presentation side of the word they chose (fig. 1B). Presentation side of target and distractor words was random. Following the response, they could self-initiate the next trial via a button press. Each block was followed by a short self-determined break. This procedure resulted in only four responses per condition and therefore we added a behavioral experiment following all six blocks, to assess performance. Responses were acquired via a response pad.

For this additional behavioural experiment, we used 48 unimodal video and audio stimuli of a different female speaker reading Antoiné St. Exupery's "The little prince" (1943). Each stimulus contained a single sentence with a two-syllable noun (target word) within the last four words. We created a list of different two-syllable nouns (distractor words) which we also drew from "The little prince" but were not presented during the stimulation. Similar to the main experiment, participants had to choose between two alternatives and the chance level was 50%. The behavioural stimuli were manipulated in the same way as the stimuli for the MEG experiment. Stimulus presentation was controlled using in-house wrapper functions (https://gitlab.com/thht/th_ptb) for the MATLAB-based Psychtoolbox (Brainard 1997; Kleiner et al. 2007; Pelli 1997).

Data acquisition and analyses

Extraction of acoustic speech envelope

We extracted the acoustic speech envelope using the Chimera toolbox by Delguette and colleagues (http://research.meei.harvard.edu/chimera/More.html) where nine frequency bands in the range of 100 – 10000 Hz were constructed as equidistant on the cochlear map

(Chandrasekaran et al. 2009; Gross et al. 2013; Smith et al. 2002). Sound stimuli were band-pass filtered (forward and reverse) in these bands using a 4th-order Butterworth filter. For each band, envelopes were calculated as absolute values of the Hilbert transform and were averaged across bands to obtain the full-band envelope that was used for coherence analysis. We did this for all three conditions (original, 7-chan, 3-chan) resulting in highly similar envelopes for those conditions (fig. 1A).

Extraction of lip area

We extracted the movement of the mouth throughout the nonblurred videos as done in Park et al. (2016). For every video frame, the outer contour of the lips was recognised using the contrast between lip color and the rest of the face. The area of this 2-D shape ("lip area") was then calculated and expressed as a time series for each video. This signal was used to calculate coherence measures with brain activity.

MEG acquisition and preprocessing

Data acquisition and analyses closely resembles with minor exceptions the one described in Hauswald et al. (2018). MEG was recorded at a sampling rate of 1 kHz using a 306-channel (204 first order planar gradiometers) Triux MEG system (Elekta-Neuromag Ltd., Helsinki, Finland) in a magnetically shielded room (AK3B, Vakuumschmelze, Hanau, Germany). The MEG signal was online high-pass and low-pass filtered at 0.1 Hz and 330 Hz, respectively. Prior to the experiment, individual head shapes were digitized for each participant including fiducials (nasion, pre-auricular points) and around 300 points on the scalp using a Polhemus Fastrak digitizer (Polhemus, Vermont, USA). We use a signal space separation algorithm provided by the MEG manufacturer and implemented in the Maxfilter program (version 2.2.15) to remove external noise from the MEG signal (mainly 16.6 Hz, and 50Hz plus harmonics) and realign data to a common standard head position (across different blocks based on the measured head position at the beginning of each block.

Data were analyzed offline using the Fieldtrip toolbox (Oostenveld et al. 2011). First, a high-pass filter at 1 Hz (6th order Butterworth IIR) was applied to continuous MEG data. Then, trials were defined according to the duration of each stimulus and cut into segments of two seconds to increase signal-to-noise ratio. As we were interested in frequency bands below 20 Hz and in order to save computational power, we resampled the data to 150 Hz. Independent component analysis was applied separately for visual and auditory blocks and we then identified components corresponding to blinks and eye movements and cardiac activity and removed them. On average 3.25 (SD: 1.143) components were removed for auditory blocks and 3.25 (SD: 1.005) for visual blocks. Sensor space data were projected to source space using linearly constrained minimum variance beamformer filters (Van Veen et al. 1997) and further analysis was performed on the obtained time-series of each brain voxel (http://www.fieldtriptoolbox.org/tutorial/shared/virtual_sensors). To transform the data into source space, we used a template structural magnetic resonance image (MRI) from Montreal Neurological Institute (MNI) and warped it to the subject's head shape (Polhemus points) to optimally match the individual fiducials and headshape landmarks. This procedure is part of the standard SPM (http://www.fil.ion.ucl.ac.uk/spm/) procedure of canonical brain localization (Mattout et al. 2007).

A 3D grid covering the entire brain volume (resolution of 1 cm) was created based on the standard MNI template MRI. The MNI space equidistantly placed grid was then morphed to individual headspace. Finally, we used a mask to keep only the voxels corresponding to the grey matter (1457 voxels). Using a grid derived from the MNI template allowed us to average and compute statistics as each grid point in the warped grid belongs to the same brain region across participants, despite different head coordinates. The aligned brain volumes were further used to create single-sphere head models and lead field matrices (Nolte 2003). The average covariance matrix, the head model and the leadfield matrix were used to calculate beamformer filters (regularization factor of 10%). The filters were subsequently multiplied with the sensor space trials resulting in single trial time-series in source space. The number of epochs across conditions was equalized.

We applied a frequency analysis to data of all three conditions of each modality (acoustic: original, 7-chan, 3-chan; visual: original, 12 SD, 18 SD) calculating multitaper frequency transformation (dpss taper: 1-45 Hz in 1 Hz steps, 3 Hz smoothing). These values were used for the analyses of alpha as well as for the coherence calculation between each virtual sensor and the acoustic speech envelope and the lip area. For the three acoustic conditions we used the envelopes of the presented acoustic (original, 7-chan, 3-chan) signal. For the three visual condition we always used the lip area values from the original, unaltered stimuli. As other studies reported the highest correlation between speech signal and brain activity occurs with a lag of brain activity after 80 ms (e.g. Wöstmann et al. 2017), we introduced a lag of 100 ms between the brain activity and acoustic/visual signal (Gross et al. 2013; A. Keitel et al. 2018). Then, the coherence between activity at each virtual sensor and the acoustic speech envelope during acoustic stimulation and lip area during visual stimulation in the frequency spectrum was calculated and averaged across trials. We will refer to the coherence between acoustic speech envelope and brain activity as speech-brain coherence and between lip area and brain activity as lip-brain coherence. As a sanity check, we calculated grand averages of the speech-brain and lip-brain coherence of the original condition (nonvocoded speech, nonblurred videos) and localized the sources of the peak activity (4 Hz).

This shows that temporal, auditory regions synchronize to acoustic speech (fig. 2A) and visual regions to visual speech (fig. 3A).

Statistical analyses

We analysed the behavioural responses from the behavioural experiment within each modality. Due to technical problems, behavioural measures are missing for 3 participants and the responses of the remaining 25 participants were analysed. We used dependent samples t-tests to compare hit rates between conditions and against chance level (50 %).

Many studies of speech-brain entrainment analyze the theta frequency band, and to be able to compare results and set them into existing context, we chose to do the same. As can be seen from the grand average (fig. 2A), the peak was at 4 Hz. We also did all analyses for a frequency band around the peak (3-6 Hz), but the patterns did not change and therefore we will not report this in more detail.

For the MEG alpha power and 4-7 Hz coherence data we applied linear regression (ft_statfun_depsamplesregrt in fieldtrip) a to test linear modulations of neural measures across the three degradation levels. To control for multiple comparisons, a non-parametric Monte-Carlo randomization test was undertaken (Maris and Oostenveld 2007). The t-test was repeated 5000 times on data shuffled across conditions and the largest t-value of a cluster coherent in space was kept in memory. The observed clusters were compared against the distribution obtained from the randomization procedure and were considered significant when their probability was below 5%. Effects were identified in space and the corresponding individual coherence and power values were extracted and statistically compared with two-tailed dependent samples t-tests. Post-hoc *t*-tests between conditions were corrected for multiple comparisons by using the FDR method (Benjamini and Hochberg 1995).

Co-modulation of coherence and alpha:

For each participant we averaged the values of each condition across the voxels that were identified by the significant regression effect for the theta coherence (left temporal for speech-brain coherence; bilateral occipital for lip-brain coherence). We then performed a 1st level statistics by correlating (Pearson) for each subject individually the theta coherence across the three conditions with the alpha power in all voxels. The 2nd level statistics then compared those correlational values against zero using monte-carlo cluster permutation (two-sided, p=0.05, 5000 randomizations) with dependent samples t-tests as test-statistic.

For visualization, source localizations were mapped onto inflated surfaces as implemented in FieldTrip.

Results

Behavioral results

Unimodal auditory stimulation

The mean hit rate for original stimuli was 99% (SD: 0.69%) for the original sound files, for 7-chan vocoded stimuli 92.5% (SD: 2.04%), and for 3-chan vocoded stimuli 69.44% (SD: 3.75%). All conditions showed significant above-chance (50%) hit rates (fig. 1C): for nonvocoded stimuli *t*(24) = 70.787, *p_fdr* = 1.3341e-28, for 7-chan vocoded t(24)=20.821, *p_fdr*=2.1531e-16, and for 3-chan vocoded t(24)=5.333, *p_fdr*=2.1494e-5. Comparing the different vocoding levels with each other showed higher hit rates for nonvocoded stimuli than for 7-chan vocoded (t(24)=3.376, *p_fdr*=0.0025) or 3-chan vocoded stimuli (t(24)=7.632, *p_fdr*=1.437e-7). 7-chan vocoded had higher hit rates than 3-chan vocoded stimuli (t(24)=6.2354, *p_fdr*=2.8733e-6).

Unimodal visual stimulation

The mean hit rate for high clarity stimuli (original) was 70% (SD: 3.15%), for medium clarity (12 SD) 58.5% (SD: 2.86%), and for low clarity (18 SD) 58.0% (SD: 3.74%). Compared to chance level (50%), unaltered videos showed increased hit rates (t(24)=6.358, *p_fdr*=8.5395e-6), as did medium clarity videos (t(24)=2.971, *p_fdr*=0.02), while this was inconclusive for low clarity (t(24)=2.138, *p_fdr*=0.0515). Comparing the different clarity levels with each other showed higher hit rates for high than medium clarity (t(24)=2.588, *p_fdr*=0.0242) or low clarity (t(24)=2.753, *p_fdr*=0.0221). Hit rates did not show a conclusive difference between medium and low clarity (t(24)=0.112, *p_fdr*=0.912, fig. 1E).

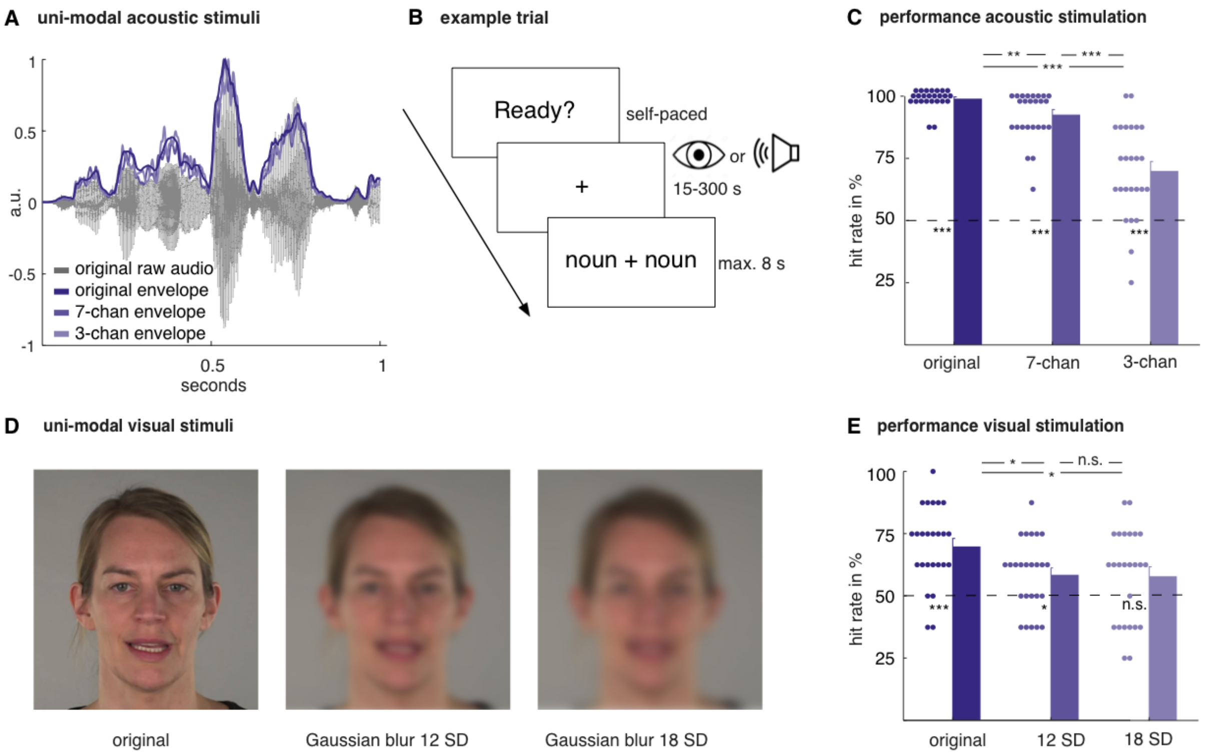

Fig 1. **A**) an exemplary audio file with the corresponding envelope as well as with the envelopes of the vocoded audio stimuli presenting either 7 or 3 channels, used in the MEG experiment. **B**) Example trial of unimodel acoustic or visual stimulation. Participants started the presentation self-paced and listened to the stimulus during the visual presentation of a fixation cross or watched the video. When the stimulus ended, participants were presented with two nouns of which they had to pick the one they perceived in the sentence before. **C**) hit rates in the behavioral experiment during the acoustic presentation. **D**) static example of the different Gaussian blurs applied to the video stimuli used in the MEG experiment. **E**) hit rates in the behavioral experiment during the visual presentation. Dots represent the individual values. Bars represent standard error, *p_fdr*<0.05*, *p_fdr*<0.01**, *p_fdr*<0.001***. Photo printed with permission of the speaker.

MEG data

Unimodal auditory stimulation

Grand average As a first manipulation check, we located the peak synchronization (4 Hz) of the un-altered unimodal signal with brain activity. Highest coherence between speech and brain activity occurred in bilateral temporal regions (fig. 2A), reflecting the tracking of speech in the theta band in auditory and adjacent regions.

Degradation related effects To investigate the effects of reducing the acoustic information, we ran a cluster-corrected regression analysis for the speech-brain coherence (4-7 Hz) of the 3 conditions (original, 7-chan, 3-chan). A positive effect of degradation between 4-7 Hz (p=0.014) was located in left middle temporal and parahippocampal regions. In these areas, the original audio stimuli lead to the weakest speech-brain coherence, while the stimuli with the strongest degradation (3-chan) elicited the strongest speech-brain coherence (Fig. 2Bi). Listening to the original audio files elicited lower coherence than listening to the 7-chan (t(27)=-3.974, *p_fdr*=0.0007) or 3-chan version (t(27)=-4.694, *p_fdr*=0.0002). The two vocoded stimuli classes did not show a conclusive difference (t(27)=-1.437, *p_fdr*=0.1622). Given that alpha is a proxy of listening efforts, we examined the degradation effect on 8-12 Hz power. The regression analysis of alpha power (8-12 Hz) over original, 7-chan and 3-chan revealed a negative effect of degradation (p=0.002, fig. 2Bii), with unaltered stimuli eliciting higher alpha power than 7-chan vocoding (t(27)=3.6552, *p_fdr*=0.0011) and 3-chan vocoding (t(27)=4.864, *p_fdr*=0.0001). And higher alpha for 7-chan than for 3-chan vocoded stimuli (t(27)=4.0017, *p_fdr*=0.0007). The effect was widespread and covered most of the brain (present in 1295 of 1457 voxel) with maxima in left angular/parietal inferior cortex, left temporal middle and inferior cortex, left rolandic operculum, and right cingulum.

Co-modulation of coherence and alpha To address the important question whether and how speech-brain coherence and alpha power are co-modulated by degrading the acoustic stimuli, we calculated correlations: First, we correlated the averaged coherence between 4-7 Hz in the left temporal area that showed a significant degradation effect of coherence with the whole-brain alpha activity first on an individual level across our three conditions and then compared the individual correlation coefficients with zero using cluster-permutation. This indeed showed that increases of coherence over the three conditions in left temporal cortex (original, 7-chan, 3-chan) was co-modulated with decreases of alpha power over the the conditions (original, 7-chan, 3-chan) mainly in right rolandic operculum, temporal and insular cortex (p=0.002, fig. 2C).

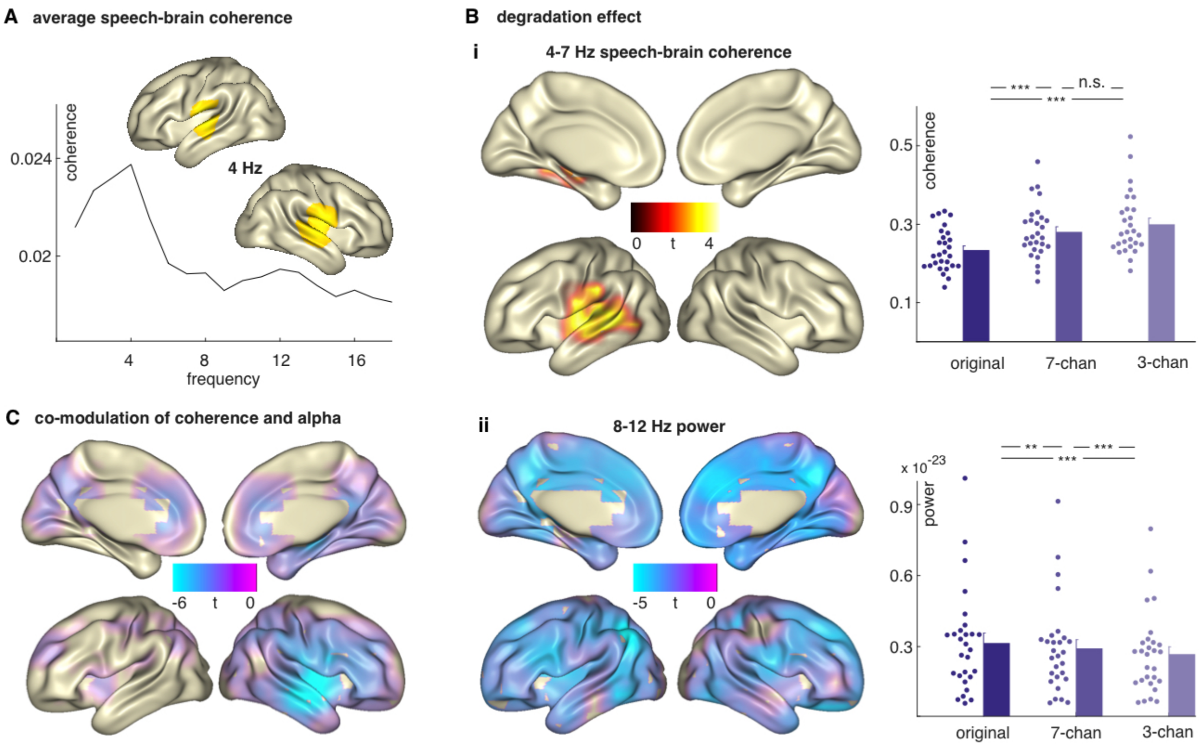

Figure 2: **A**) Frequency spectrum of the speech-brain coherence for original stimuli averaged across all voxels. Source estimations for 4 Hz coherence between original audio files and brain activity masked at 80% of maximum coherence in bilateral temporal superior, Heschl's gyrus and rolandic operculum. **B) i) Left**: source localizations of degradation effects on speech-brain coherence (4-7 Hz) during acoustic stimulation across three conditions (original, 7-chan, 3-chan) in left temporal and parahippocampal regions. **Right**: individual speech-brain coherence values of the three conditions (original, 7-chan, 3-chan) extracted at voxels showing a significant regression effect contrasted with each other. **ii)**) **Left**: source localizations of degradation effects on alpha power (8-12 Hz) during acoustic stimulation across three conditions (original, 7-chan, 3-chan). **Right**: individual alpha power values of the three conditions (original, 7-chan, 3-chan) extracted at voxels showing a significant regression effect contrasted with each other. **C**) Second level statistical contrasting zero with the correlations of individual values of the three condition (original, 7-chan, 3-chan) between whole-brain alpha power (8-12 Hz) and average speech-brain coherence (4-7 Hz) of left temporal and parahippocampal cortex. Dots represent individual values. Bars represent standard error, *p_fdr*<0.05*, *p_fdr*<0.01**, *p_fdr*<0.001***

Unimodal visual stimulation

Grand average As a sanity check, we localized the sources of lip-brain coherence for the peak frequencies (4-5 Hz) of the unaltered visual stimuli. Sources were localized in occipital areas (fig. 3B).

Degradation related effects Running a cluster-corrected regression analysis for the lip-brain coherence (4-7 Hz) of the 3 conditions (original, 12 SD, 18 SD) during visual presentation revealed a negative effect (p=0.002) in bilateral occipital and calcarine regions, cuneus, cerebellum (not displayed in figure) and right middle temporal gyrus (fig. 3D). Watching the original stimuli elicited the strongest coherence between lip area and brain activity compared to the stimuli with medium (12 SD) and low clarity (18 SD). Coherence values extracted at voxels with a significant regression effect showed highest values for watching the original stimuli compared to medium clarity (12 SD) (t(27)=3.956, *p_fdr*=0.0007) and low clarity (18 SD) stimuli (t(27)=5.433, *p_fdr*=2.8608e-5). The medium clear stimuli produced higher coherence than the ones with low clarity (t(27)=2.626, *p_fdr*=0.0141). Regression analyses on alpha power did not reveal conclusive differences (p=0.5694).

Co-modulation of coherence and alpha 1st and 2nd level statistics on the correlation between coherence and alpha did not produce conclusive support of a co-modulation (p=0.67).

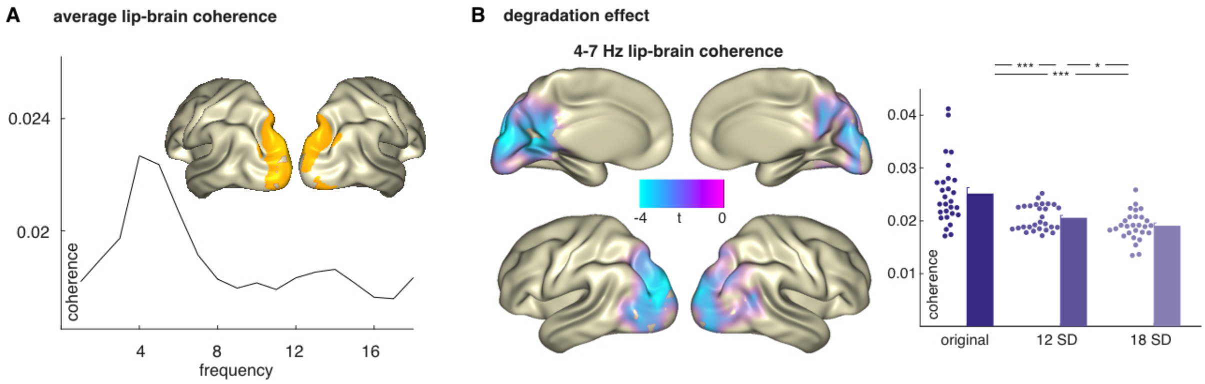

Fig. 3: **A)** Frequency spectrum of the lip-brain coherence for the nonblurred stimuli averaged across all voxels. Inlay: Source estimations for 4 Hz coherence between original video files and brain activity masked at 80% of maximum coherence. **B) Left**: source localizations of degradation effects on lip-brain coherence (4-7 Hz) during visual stimulation across three conditions (original, 12 SD, 18 SD). **Right**: averaged individual lip-brain coherence values of the three conditions (original, 12 SD, 18 SD) extracted at voxels showing an significant degradation effect contrasted with each other. Dots represent individual values. Bars represent standard error, *p_fdr*<0.05*, *p_fdr*<0.01**, *p_fdr*<0.001***

Discussion

As known by previous studies, listening to degraded speech is effortful and increases attentional demands (Wild et al. 2012; Wöstmann et al. 2015). We also know that in investigations of auditory attention, alpha power is modulated (Frey et al. 2014; Weisz et al. 2014; Wöstmann et al. 2016) and speech tracking is enhanced for the attended sounds (Ding and Simon 2012b; Horton et al. 2013; Rimmele et al. 2015; Zion Golumbic et al. 2013). We integrated these findings in the present study, by investigating the effects of degraded speech stimuli on coherence between the speech envelope and brain activity and on alpha power elicited by the different degradation levels (original, 7-chan, 3-chan).

To be sure that our manipulation actually affects the intelligibility, we let our participants perform a behavioral experiment, after the MEG experiment. This was similar to the MEG experiment (with identical degradation levels) but with shorter stimuli, enabling us to assess more trials. The data showed that participants decline in performance when the stimuli are degraded, which is in line with other studies showing a linear decline in performance

(McGettigan et al. 2012; Strelnikov et al. 2011). The exact number of channels needed for high performance understanding depends on the stimulus material and the specific experimental setup (Dorman et al. 1997; Loizou et al. 1999). For our study, we conclude that even the 3-chan condition was challenging yet not completely unintelligible given that performance is still higher than expected by chance.

Another important manipulation and analysis check was to see if speech tracking actually occurs in and near auditory cortices. An important frequency band of speech tracking is the theta band (4-7 Hz) as it is associated with the syllable rate and segments the speech into chunks (e.g. Ghitza 2012). We localized the peak (4 Hz) of the speech-brain coherence for the unaltered speech stimuli and find that theta coherence elicits bilateral auditory and broader temporal regions with a bias to the right hemisphere. This is consistent with other studies reporting bilateral but asymmetrical speech representation (Giraud and Poeppel 2012; Lam et al. 2018; Poeppel 2003).

To elucidate if the intelligibility, measured by degradation level, affects the speech tracking, measured by speech-brain coherence, we calculated a regression analysis of the 4-7 Hz speech-brain coherence across the three conditions. This revealed a left-lateralized source in auditory-adjacent cortex in which higher entrainment was associated with a higher level of degradation. This finding further highlights the relevance of auditory cortices for speech tracking. Increased activation of left temporal regions for degraded but yet intelligible stimuli compared to unaltered speech was also reported by Davis and Johnsrude (2003) and interpreted as indicating recruitment of compensatory attentional resources. This is also consistent with a study showing non-native speakers produce higher delta/theta speech entrainment than native-speakers and the authors have also proposed this as reflecting the higher effort (Song and Iverson 2018). Further, the M50 of TRF is enhanced for degraded stimuli compared to unaltered ones in quiet environment as is delta entrainment, the latter again suggested to reflect listening efforts (Ding et al. 2014). Although, other studies have reported decreased theta entrainment for degraded speech (Ding et al. 2014; Peelle et al. 2013; Rimmele et al. 2015), synchronization with the speech signal in both frequency bands is enhanced when attended to: multi-speaker and auditory spatial attention studies using sentences or narratives have repeatedly found stronger low-frequency (1-7 Hz) speech tracking for attended compared to unattended speech (Ding and Simon 2012b; Horton et al. 2013; Rimmele et al. 2015; Zion Golumbic et al. 2013).

Another commonly used proxy for cognitive effort is the alpha rhythm. We found widespread decreased alpha power for the stimuli with less acoustic information and this decline was negatively associated with coherence. Although effects are more focal, auditory spatial attention studies also report contralateral alpha decreases (Frey et al. 2014; Weisz et al. 2014;

Wöstmann et al. 2016). The pattern of decreasing alpha power is further consistent with other studies using degradation of complex speech material as e.g. sentences (McMahon et al. 2016; Miles et al. 2017). This decrease even occurs when the sentences are not the target material but distractors (Wöstmann et al. 2017). However, studies using short and simple speech stimuli such as single words (Becker et al. 2013; Obleser and Weisz 2012) or digits (Obleser et al. 2012; Wöstmann et al. 2015) report enhanced alpha for stimuli with more acoustic detail compared to degraded sounds. This difference in findings might be linked to the linguistically more complex nature of the longer speech stimuli as also suggested by (Miles et al. 2017).

Additionally, the long duration of the stimuli (up to 3 min) in our study might also influence their processing via a warm-up effect (Dorman et al. 1997). Experimental investigation of this warm up effect or perceptual learning effect -changes in sensory processing and sensory guided behavior based on prior sensory experience- showed that indeed speech understanding increased over time for degraded stimuli (here, 6-chan vocoding; Davis et al. 2005) and that this increase was bigger for sentences than for single words (Hervais-Adelman et al. 2008) and smallest for very strong (1-chan) or very little (24-chan) vocoding (Sohoglu and Davis 2016). This leaves us to expect that for our vocoded conditions with 7 and 3 channels, perceptual learning took place possibly reflected in the above chance hit rates for the degraded conditions but also in the alpha power effects: Perceptual learning has been linked with alpha activity (Sigala et al. 2014). Reports of stronger alpha decreases have been associated with higher learning success (Freyer et al. 2013) and higher perceptual learning occurs for degraded but still intelligible stimuli (Sohoglu and Davis 2016). A similar relationship with stronger effects for middle range of degradation also occurs for behavioral and neuronal measures: Longest response times were found for difficult to understand acoustic stimuli compared to intelligible and completely unintelligible and proposed to reflect listening effort (Wu et al. 2016). fMRI showed higher activity for degraded (but intelligible) compared to unaltered speech in left temporal cortex (Davis and Johnsrude 2003). The degraded stimuli in our study were difficult to understand as shown in the decline of hit rate over conditions, however, importantly they were still higher than chance and thus not completely unintelligible. Together these findings suggests that our own findings of reduced alpha power for the degraded speech stimuli reflect perceptual learning processes of the speech sounds, a process that is mediated via attention (Sigala et al. 2014) and that needs some time and therefore is more likely to become visible for longer stimulation and is enhanced for the middle range of vocoding (Sohoglu and Davis 2016) as there presumably the strongest benefits of directed attention or listening effort can be observed. This assumption further integrates well with the framework on the characteristics of alpha oscillations in working memory tasks as proposed by van Ede (2018). Here, the idea is put forward that alpha power increases for tasks with sensory disengagement, while it decreases for tasks which recruit the sensory representation. Our task of asking participants to identify which of two presented words did occur within the just heard four last words of a speech stimuli will most likely recruit the sensory representation of exactly those few words, thereby leading to an alpha decrease. Within this framework, also the different findings concerning alpha can be unified by taking the specific task and the resulting demands into account.

Concerning the degradation of the visual speech, we could replicate previous findings of declining performance (Tye-Murray et al. 2016) as well as of theta lip-brain synchronization in visual regions (Giordano et al. 2017; Hauswald et al. 2018; Park et al. 2016). In our study, performance declined with degradation but did not differ between the two degraded conditions. We assume that both blurring conditions degraded the relevant features of visual speech in a manner that affected intelligibility comparably.

Interestingly, unlike in the auditory modality, neural lip-brain synchronization decreased with degradation. This linear decrease across conditions suggests that the visual cortex still tracks the lip movements better in the less blurry condition - even though this does not translate to a behavioral advantage. Thus, considering the behavioral pattern and lip-brain synchronization, this indicates that our manipulation affected articulation information (regarding place and manner) already with the mild blurring condition, and that this information is crucial for intelligibility. On the other hand, the timing information provided by the lip movements that fed into the lip-brain synchronization was more robust. Also, unlike in the acoustic modality, no alpha power change was found for the different degradation levels. This suggests that processes that might compensate to some degree the comprehension of degraded acoustic speech are not engaged in the same way for degraded visual speech. For individuals with vision problems, this might have consequences for everyday situations where articulatory visual information could provide comprehension benefits (such as the visual enhancement of speech, Sumby and Pollack 1954).

In sum, prior research shows that i) degraded but intelligible speech enhances attentional demand and subjective effort (Wild et al. 2012; Wöstmann et al. 2015), ii) attention enhances entrainment (Ding and Simon 2012b; Horton et al. 2013; Rimmele et al. 2015; Zion Golumbic et al. 2013) and alpha modulations (Frey et al. 2014; Weisz et al. 2014), iii) listening effort is mostly invested for the conditions where perceptual learning can occur, namely perceptual learning in the middle range of vocoding (Sohoglu and Davis 2016). In this context, our findings of reduced alpha power and enhanced entrainment for degraded but intelligible stimuli, as shown by the above chance hit rates, in left temporal cortex fit nicely with these prior reports and suggest they are indicative of the effort that is needed to attempt to compensate for the degradation. The idea of a common process driving both measures is also consistent with the finding that increasing synchronization (for stronger degradation) is associated with lower alpha power, mainly in right temporal regions. The linear decline in hit rate nevertheless shows that even with the higher efforts these efforts are not sufficient to reach the level of undegraded speech.

For the unimodal visual condition, we find the strongest synchronization for the non-degraded stimulation. This is in line with findings from rhythmic and quasi-rhythmic visual stimulation experiments that show that i) the visual cortex faithfully tracks visual dynamics, and ii) this tracking is reduced with lower visual contrast (Campbell and Kulikowski 1972; C. Keitel et al. 2017). We did not find any consistent alpha modulation in response to the visual degradation. This lets us assume that the stimulus degradation in the unimodal visual condition -the blurring of the lip contour- was not compensated by increasing the attentional demands. It seems that, once the articulatory features of mouth movements are impaired, neural attentional mechanisms are not recruited to change the perception of visual speech. Overall, our results show multi-layered neural mechanisms of degraded auditory and visual speech comprehension, with fundamental differences between both modalities.

Acknowledgements

This study was supported by a FWF Einzelprojekt (P 31230). AK was supported by the Wellcome Trust [204820/Z/16/Z]. We would like to thank Joachim Gross and Hyojin Park for providing their original Matlab script for extracting the lip area. We would also like to thank Siri Ebert and Jonas Heilig for help with data acquisition.

Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2018). Cortical entrainment: what we can learn from studying naturalistic speech perception. *Language, Cognition and Neuroscience*, *0*(0), 1–13. doi:10.1080/23273798.2018.1518534

Becker, R., Pefkou, M., Michel, C. M., & Hervais-Adelman, A. G. (2013). Left temporal alpha-band activity reflects single word intelligibility. *Frontiers in Systems Neuroscience*, *7*. doi:10.3389/fnsys.2013.00121

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436. doi:10.1163/156856897X00357

Campbell, F. W., & Kulikowski, J. J. (1972). The visual evoked potential as a function of contrast of a grating pattern. *The Journal of Physiology*, *222*(2), 345–356. doi:10.1113/jphysiol.1972.sp009801

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7). doi:10.1371/journal.pcbi.1000436

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*. doi:10.3389/fnhum.2016.00604

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *23*(8), 3423–3431.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241. doi:10.1037/0096-3445.134.2.222

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust Cortical Entrainment to the Speech Envelope Relies on the Spectro-temporal Fine Structure. *NeuroImage*, *88*, 41–46. doi:10.1016/j.neuroimage.2013.10.054

Ding, N., & Simon, J. Z. (2012a). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. doi:10.1152/jn.00297.2011

Ding, N., & Simon, J. Z. (2012b). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. doi:10.1073/pnas.1205381109

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *The Journal of the Acoustical Society of America*, *102*(4), 2403–2411. doi:10.1121/1.419603

Ede, F. van. (2018). Mnemonic and attentional roles for states of attenuated alpha oscillations in perceptual working memory: a review. *European Journal of Neuroscience*, *48*(7), 2509–2515. doi:10.1111/ejn.13759

Frey, J. N., Mainy, N., Lachaux, J.-P., Muller, N., Bertrand, O., & Weisz, N. (2014). Selective Modulation of Auditory Cortical Alpha Activity in an Audiovisual Spatial Attention Task. *Journal of Neuroscience*, *34*(19), 6634–6639. doi:10.1523/JNEUROSCI.4813-13.2014

Freyer, F., Becker, R., Dinse, H. R., & Ritter, P. (2013). State-Dependent Perceptual Learning. *Journal of Neuroscience*, *33*(7), 2900–2907. doi:10.1523/JNEUROSCI.4039-12.2013

Gaudrain, E., & Başkent, D. (2015). Factors limiting vocal-tract length discrimination in cochlear implant simulations. *The Journal of the Acoustical Society of America*, *137*(3), 1298–1308. doi:10.1121/1.4908235

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of speech with a manipulated modulation spectrum. *Frontiers in*

Psychology, *3*(238). doi:10.3389/fpsyg.2012.00238

Giordano, B. L., Ince, R. A. A., Gross, J., Schyns, P. G., Panzeri, S., & Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. *eLife*, *6*. doi:10.7554/eLife.24763

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. doi:10.1038/nn.3063

Greenberg, S. (1998). A syllable-centric framework for the evolution of spoken language. *Behavioural and Brain Sciences*, 518.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology*, *11*(12). doi:10.1371/journal.pbio.1001752

Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., & Weisz, N. (2018). A Visual Cortical Network for Deriving Phonological Information from Intelligible Lip Movements. *Current biology: CB*, *28*(9), 1453-1459.e3. doi:10.1016/j.cub.2018.03.044

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460–474. doi:10.1037/0096-1523.34.2.460

Horton, C., D’Zmura, M., & Srinivasan, R. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology*, *109*(12), 3082–3093. doi:10.1152/jn.01026.2012

Howard, M. F., & Poeppel, D. (2010). Discrimination of Speech Stimuli Based on Neuronal Response Phase Patterns Depends on Acoustics But Not Comprehension. *Journal of Neurophysiology*, *104*(5), 2500–2511. doi:10.1152/jn.00251.2010

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, *16*(3). doi:10.1371/journal.pbio.2004473

Keitel, C., Thut, G., & Gross, J. (2017). Visual cortex responses reflect temporal structure of continuous quasi-rhythmic sensory stimulation. *NeuroImage*, *146*, 58–70. doi:10.1016/j.neuroimage.2016.11.043

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional Gain Control of Ongoing Cortical Speech Representations in a “Cocktail Party.” *Journal of Neuroscience*, *30*(2), 620–628. doi:10.1523/JNEUROSCI.3631-09.2010

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What’s new in psychtoolbox-3. *Perception*, *36*(14), 1–16.

Lam, N. H. L., Hultén, A., Hagoort, P., & Schoffelen, J.-M. (2018). Robust neuronal oscillatory entrainment to speech displays individual variation in lateralisation. *Language, Cognition and Neuroscience*, *33*(8), 943–954. doi:10.1080/23273798.2018.1437456

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *The Journal of the Acoustical Society of America*, *106*(4), 2097–2103. doi:10.1121/1.427954

Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*, *54*(6), 1001–1010. doi:10.1016/j.neuron.2007.06.004

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi:10.1016/j.jneumeth.2007.03.024

Mattout, J., Henson, R. N., & Friston, K. J. (2007). Canonical source reconstruction for MEG. *Computational Intelligence and Neuroscience*, *2007*. doi:10.1155/2007/67613

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: behavioural and neural interactions. *Neuropsychologia*, *50*(5), 762–776. doi:10.1016/j.neuropsychologia.2012.01.010

McMahon, C. M., Boisvert, I., de Lissa, P., Granger, L., Ibrahim, R., Lo, C. Y., et al. (2016).

Monitoring Alpha Oscillations and Pupil Dilation across a Performance-Intensity Function. *Frontiers in Psychology*, *7*. doi:10.3389/fpsyg.2016.00745

Miles, K., McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., Graham, P., & Lyxell, B. (2017). Objective Assessment of Listening Effort: Coregistration of Pupillometry and EEG. *Trends in Hearing*, *21*, 2331216517706396. doi:10.1177/2331216517706396

Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. *Physics in medicine and biology*, *48*(22), 3637–3652. doi:10.1088/0031-9155/48/22/002

Obleser, J., & Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cerebral Cortex*, *22*(11), 2466–2477. doi:10.1093/cercor/bhr325

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse Listening Conditions and Memory Load Drive a Common Alpha Oscillatory Network. *Journal of Neuroscience*, *32*(36), 12376–12383. doi:10.1523/JNEUROSCI.4908-11.2012

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*. doi:10.1155/2011/156869

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife*, *5*(e14521). doi:10.7554/eLife.14521

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cerebral Cortex (New York, NY)*, *23*(6), 1378–1387. doi:10.1093/cercor/bhs118

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181. doi:10.1016/j.cortex.2015.03.006

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time.' *Speech Communication*, *41*(1), 245–255. doi:10.1016/S0167-6393(02)00107-3

Riecke, L., Formisano, E., Sorger, B., Başkent, D., & Gaudrain, E. (2018). Neural Entrainment to Speech Modulates Speech Intelligibility. *Current Biology*, *28*(2), 161-169.e5. doi:10.1016/j.cub.2017.11.033

Rimmele, J. M., Zion Golumbic, E., Schröger, E., & Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex*, *68*, 144–154. doi:10.1016/j.cortex.2014.12.014

Rooth, M. A. (2017). The Prevalence and Impact of Vision and Hearing Loss in the Elderly. *North Carolina Medical Journal*, *78*(2), 118–120. doi:10.18043/ncm.78.2.118

Sigala, R., Haufe, S., Roy, D., Dinse, H. R., & Ritter, P. (2014). The role of alpha-rhythm states in perceptual learning: insights from experiments and computational models. *Frontiers in Computational Neuroscience*, *8*. doi:10.3389/fncom.2014.00036

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90. doi:10.1038/416087a

Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, *113*(12), E1747–E1756. doi:10.1073/pnas.1523266113

Song, J., & Iverson, P. (2018). Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents. *Cognition*, *179*, 163–170. doi:10.1016/j.cognition.2018.06.001

Strelnikov, K., Massida, Z., Rouger, J., Belin, P., & Barone, P. (2011). Effects of vocoding and intelligibility on the cerebral response to speech. *BMC Neuroscience*, *12*(1). doi:10.1186/1471-2202-12-122

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215. doi:10.1121/1.1907309

Teng, X., Tian, X., Doelling, K., & Poeppel, D. (2018). Theta band oscillations reflect more than entrainment: behavioral and neural evidence demonstrates an active chunking process. *European Journal of Neuroscience*, *48*(8), 2770–2782. doi:10.1111/ejn.13742

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. (2016). Lipreading and Audiovisual Speech Recognition across the Adult Lifespan: Implications for Audiovisual Integration. *Psychology and aging*, *31*(4), 380–389. doi:10.1037/pag0000094

Van Veen, B. D., van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, *44*(9), 867–880. doi:10.1109/10.623056

Weisz, N., Müller, N., Jatzev, S., & Bertrand, O. (2014). Oscillatory Alpha Modulations in Right Auditory Regions Reflect the Validity of Acoustic Cues in an Auditory Spatial Attention Task. *Cerebral Cortex*, *24*(10), 2579–2590. doi:10.1093/cercor/bht113

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. *Journal of Neuroscience*, *32*(40), 14010–14021. doi:10.1523/JNEUROSCI.1528-12.2012

Wöstmann, M., Fiedler, L., & Obleser, J. (2017). Tracking the signal, cracking the code: speech and speech comprehension in non-invasive human electrophysiology. *Language, Cognition and Neuroscience*, *32*(7), 855–869. doi:10.1080/23273798.2016.1262051

Wöstmann, M., Herrmann, B., Maess, B., & Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences*, *113*(14), 3873–3878. doi:10.1073/pnas.1523357113

Wöstmann, M., Herrmann, B., Wilsch, A., & Obleser, J. (2015). Neural Alpha Dynamics in Younger and Older Listeners Reflect Acoustic Challenges and Predictive Benefits. *Journal of Neuroscience*, *35*(4), 1458–1467. doi:10.1523/JNEUROSCI.3250-14.2015

Wu, Y.-H., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (2016). Psychometric functions of dual-task paradigms for measuring listening effort. *Ear and hearing*, *37*(6), 660–670. doi:10.1097/AUD.0000000000000335

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a “Cocktail Party.” *Neuron*, *77*(5), 980–991. doi:10.1016/j.neuron.2012.12.037

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. *Current Biology*, *28*(3), 401-408.e5. doi:10.1016/j.cub.2017.11.071