On the Variability of the McGurk Effect: Audiovisual Integration Depends on Prestimulus Brain States

Julian Keil1, Nadia Müller1,2, Niklas Ihssen3 and Nathan Weisz1,2

1Department of Psychology and 2Zukunftskolleg, University of Konstanz, Konstanz 78457, Germany and 3School of Psychology, Cardiff University, Cardiff CF10 3AT, UK

Address correspondence to Julian Keil, Department of Psychology, University of Konstanz, Universitätsstrasse 10, PO Box D25, Konstanz 78457, Germany. E-mail: Julian.keil@uni-konstanz.de.

The McGurk effect demonstrates the influence of visual cues on auditory perception. Mismatching information from both sensory modalities can fuse to a novel percept that matches neither the auditory nor the visual stimulus. This illusion is reported in 60–80% of trials. We were interested in the impact of ongoing brain oscillations—indexed by fluctuating local excitability and interareal synchronization—on upcoming perception of identical stimuli. The perception of the McGurk effect is preceded by high beta activity in parietal, frontal, and temporal areas. Beta activity is pronounced in the left superior temporal gyrus (ISTG), which is considered as a site of multimodal integration. This area is functionally (de)coupled to distributed frontal and temporal regions in illusion trials. The disposition to fuse multisensory information is enhanced as the ISTG is more strongly coupled to frontoparietal regions. Illusory perception is accompanied by a decrease in poststimulus theta-band activity in the cuneus, precuneus, and left superior frontal gyrus. Event-related activity in the left middle temporal gyrus is pronounced during illusion perception. Thus, the McGurk effect depends on fluctuating brain states suggesting that functional connectedness of left STS at a prestimulus stage is crucial for an audiovisual percept.

Keywords: illusion, MEG, multisensory, perception

Introduction

While there is a substantial body of literature about the neural basis of unisensory perception, multimodal information integration has come into focus only recently (Calvert et al. 2004). Integration of information from multiple modalities is crucial and representative for our everyday life. A typical example is speech perception, in which, apart from the actual sound, visual cues from lip movements also have a significant influence on what we actually perceive as being said (Van Wassenhove et al. 2005). A classical demonstration that visual information can significantly impact speech perception is the so-called McGurk effect, first described by McGurk and MacDonald (1976). In this illusion, an auditory syllable is dubbed with a video of lip movements uttering a mismatching syllable (e.g., a video of an actor pronouncing the syllable “ga” is shown together with the audio stream of the syllable “ba”). Participants frequently report having heard a syllable that matches neither the unisensory visual nor the acoustic source (e.g., “da,” see Fig. 1) and do not typically notice the incongruence between the acoustic and visual inputs (Möttönen et al. 2002). Despite being a robust finding on average, the illusory percept does not occur with equal probability in all participants and also fluctuates on a trial-by-trial basis within one participant (~60–80%; illusory “fusion” percepts). One way to conceive of audiovisual integration at a neuronal level is that these perceptions depend on the activity of multisensory cell assemblies, which receive convergent input from multiple sensory modalities. The existence of such multimodal neurons has been shown at several hierarchical levels from midbrain to cortex (Stein and Meredith 1993; Stein et al. 1996; Stein 1998; Bizley et al. 2007;Kayser and Logothetis 2007; Kayser et al. 2007; Ghazanfar et al. 2008; Kayser et al. 2010). Regarding the McGurk illusion, an increasing body of evidence points to the left superior temporal gyrus (ISTG) as a crucial structure (Calvert et al. 2000) of audiovisual information integration (Barracough et al. 2005; Stevenson and James 2009; Dahl et al. 2010), in which the case of nonmatching information leads to an illusory percept. Recent intracranial electroencephalography (iEEG) (Besle et al. 2008), EEG (Van Wassenhove et al. 2005; Cappe et al. 2010), and magnetoencephalography (MEG) (Arnal et al. 2009) studies propose neural routes between auditory and visual areas and STG in speech perception. These studies found that audiovisual interactions—especially of ecologically valid stimuli (i.e., speech)—are expressed in reduced evoked responses mediated by saliency and redundancy of information. These findings suggest rules of audiovisual integration beyond the general principles of response enhancement of multisensory integration established for multisensory cell assemblies (Calvert et al. 2004).

Exploiting the perceptual variability to the invariant mismatching stimulus, the aim of the current study was to elucidate the factors that determine multisensory integration beyond auditory and visual stimulus properties (e.g., visual speech). The focus was especially on prestimulus oscillatory activity, that is, the brain state in terms of local and interareal synchronization at the time of the mismatching stimulus’ entry into the system. Several recent studies have shown the influence of prestimulus activity on perception in general (and especially near-threshold). Alpha band phase (Mathewson et al. 2009; Busch and Vanrullen 2010) and power (Romei et al. 2010) have been reported to influence the perception of visual stimuli. Growing evidence suggests that alpha rhythms reflect the excitatory–inhibitory balance within sensory and motor regions, with strong alpha activity indicating an inhibitory state (Klimesch 1999; Weisz et al. 2007). However, apart from fluctuations of relatively local activity (i.e., at brain region level), the integration of a region into a distributed network via interregional coupling is also subject to variability. While some evidence exists for poststimulus impacts of interareal coupling on perception (Dehaene et al. 2006; Molloni et al. 2007), prestimulus influences have only recently become a focus of research (Hipp et al. 2011). To date, no study has investigated ongoing cortical prestimulus influences on the McGurk illusion.
Beauchamp et al. (2010) conducted an important study within this context using a “virtual lesion” approach. The authors were able to show that applying transcranial magnetic stimulation (TMS) to lSTG within a window of ~100 ms around the stimulus significantly diminished the proportion of trials in which the illusion was perceived. Importantly, this effect included prestimulus periods, thus implying the importance of the current state of the lSTG (and potentially therefore in connected regions).

A growing amount of empirical evidence suggests that perception involves a widespread neuronal network supplementary to activations of sensory and association regions (Koch 2004; Dehaene et al. 2006). A recent review also proposes that the function of the STG is not only restricted to audiovisual integration but also varies with task-dependent network connections (Hein and Knight 2008). This means that in addition to considering measures of local brain activity, it is also important to investigate functional network states (Buzsáki 2006; Senkowski et al. 2008), frequently manifested in synchronization of phases of oscillatory activity (Tass et al. 1998; Varela et al. 2001; Fries 2005).

Following the literature reviewed above, we hypothesize that prestimulus increases of local power in lSTG at higher frequencies (beta/gamma) or relative desynchronization at lower (theta/alpha) frequencies could be crucial for an illusion following an invariant stimulus. These power increases, as well as increased phase coupling of the lSTG (and thereby a more efficient spreading of its information) with distributed regions relevant for perception could reflect a state of perceptual readiness. Both local power changes and long-range connectivity are expected to correlate with the individual tendency to experience the McGurk illusion—that is, a fusion between auditory and visual information. We used MEG to identify responses in the time-frequency-sensor space, differentiating between subjective perceptions of either one of the presented modalities (i.e., “unimodal,” either the auditory syllable or the visual mouth movement) and the perception of a fusion (i.e., “fusion”) of both sensory modalities within the mismatching trials. We subsequently localized the sources of effects in the pre- and poststimulus interval using adaptive linear spatial filtering (so-called “beamforming,” Van Veen et al. 1997; Gross et al. 2001). Phase synchrony was then computed between the principal region of interest (lSTG) and the whole-brain volume in order to assess functional network states differentiating the perceptual categories. Our data show that special prestimulus conditions are indeed necessary at a local activation level of lSTG (increased beta power) as well as at the level of functional connectivity (increased coupling of lSTG to frontoparietal areas) in order for an illusory percept to subsequently emerge.

Methods

Subjects

Seventeen (6 males/11 females, mean age 24.9 years) paid volunteers participated in this study. All participants gave their written informed consent. All participants were right handed and had normal hearing or corrected-to-normal vision.

Experimental Design and Apparatus

Participants were informed about the experimental procedure and were introduced to the facilities. They were then prepared for the recording session and seated in the magnetically shielded room. The experiment consisted of 390 trials in which we showed videos of an actor pronouncing the syllables “aba,” “ada,” or “aga.” The stimuli were presented via PsyScope X (http://psy.ck.sissa.it/) on a MiniMac (Apple Inc.). Two-thirds of the trials contained a mismatching audiovisual stream (visual “ada”/auditory “aba” or visual “aga”/auditory “aba”) while the videos were the same as in the matching condition. Videos were on average 2.909 s long (standard deviation [SD] = 0.298 s), and sound files were on average 0.479 s long (SD = 0.039 s). Videos were paused for a randomized duration (2000–4000 ms) after the first video frame (showing a neutral face with closed mouth, see Fig. 1A) in order to make the onset of the audio stream and mouth movement unpredictable. Importantly, no visual speech cues preceded differentiating auditory information, as all syllables started on “a-.” Using a forced choice task, participants had to indicate by pressing a button whether they had perceived aba, ada, aga, or something else (other). Thus, the important dependent variable in this investigation was the subjectively perceived content of the audiovisual sensation. The visual stimuli were presented on a screen inside the magnetically shielded MEG acquisition system.

Figure 1. Trial and timing overview with relative responses. (A) Example trial with 2 video frames from a video used in this experiment (top row), the audio trace of the syllable “aba” presented simultaneously to the video (middle row). Mouth movement and audio onset occur at time 0 ms. The interval prior to this is referred to as “prestimulus.” (B) Box plot with responses relative to the total number of trials per category. On average, matching trials were correctly identified in 96.19% of cases. A “fusion” percept was reported in 41.61% of mismatching trials, whereas a “unimodal” percept was reported 48.02% of the time.
room via a video projector (DLA-G11E, JVC, Friedberg, Germany) and a set of mirrors positioned outside the room. The audio streams were presented with an analogue-to-digital converter (Motu 2408) via amplifiers (Servo 200, Samson) and a 6.1-m length, 4-mm diameter tube system (Etymotic Research, ER30). Sounds were corrected for the distortions introduced by the tube system.

Data Acquisition and Analysis
MEG recording was conducted using a 148-channel magnetometer (MAGNES 2500 WH, 4D Neuroimaging, San Diego, CA). A subject-specific headframe coordinate reference was defined by means of 5 anatomical landmarks. These head fiducials, 5 coils, and the subject’s head shape were digitized with a Polhemus 3Space Fastrack at the start of each session. The subject’s head position relative to the pickup coils and the MEG sensors were estimated before and after each session to ensure that no large movements occurred during data acquisition.

Subjects were lying supine in a comfortable position. They were instructed to lie still during the stimulation and to avoid eye movements and blinks as much as possible. Continuous data sets were recorded with a sampling rate of 678.17 Hz (bandwidth 0.1–200 Hz). A video camera installed inside the MEG chamber allowed subjects’ behavior and compliance to be monitored throughout the experiment.

After data acquisition, epochs of 4 s (± 2 s) around speech onset were extracted from the raw data. Epochs were visually inspected for EOG, ECG, or movement artifacts. Trials were categorized according to the combination of type of video, type of audio, and type of response into 2 categories: 1) fusion (mismatch between auditory and visual stimuli and response that matched neither the auditory nor the visual information) and 2) unimodal (mismatch between auditory and visual stimuli and response that matched either the auditory or the visual information).

The numbers of trials for the 2 different categories were equalized for each subject by random omission to ensure comparable signal-to-noise ratios for both perceptual categories. Resulting epochs were filtered with a 1-Hz high-pass filter (zero-phase, Butterworth) before the analysis of oscillatory activity. As the prespeech activation was the main interest of the study, no baseline was defined and outputs of the sensor- and source-space analysis for the conditions were directly compared.

For the analysis of event-related activity, single trials were low-pass filtered with a 30-Hz zero-phase Butterworth filter prior to averaging. For the time-frequency analysis, a multitaper fast fourier transformation time-frequency transformation with frequency-dependent Hanning tapers was computed (time window: Δt = 5/ fj, spectral smoothing: 1/Δf). Average event-related activity was subtracted from the single trials before computing the time-frequency transformation in order to be dominated by the evoked response on ongoing-induced oscillatory activity. This procedure resulted in single-trial estimates of oscillatory power between 2 and 40 Hz in 2-Hz steps.

A linear constrained minimum variance (LCMV) beamformer algorithm (Van Veen et al. 1997) was used to identify the sources of the effects found in the time-series analysis. Source analysis was performed for an activation interval of 550 ms until 650 ms after sound onset based on the effect identified on the sensor level (see Results). The source analysis was separately conducted on the waveforms of the regions of interest and the whole-brain volume was analyzed in terms of group statistics and for illustrative purposes.

The functional connectivity of neuronal activity between cortical regions of interest and the whole-brain volume was analyzed in terms of phase synchrony (Lachaux et al. 1999). Phase synchrony was computed for the time and frequency of interest as identified by the sensor-level analysis and for the regions of interest as identified by the source analysis. If the phase differences between 2 oscillators are constant, these oscillators are likely to interact with each other or share a common driving force. Uniform distributions of phase differences indicate the independence of 2 oscillators. We first Fourier transformed the data at sensor level for the time and frequency range identified in the time-frequency analysis (multitaper analysis, DPSS tapers) and extracted the complex values containing phase information. These complex values were then projected into source space by multiplying them with the accordant beamformer spatial filters. Spatial filters were constructed from the covariance matrix of the averaged single trials at sensor level and the respective leadfield by a LCMV beamformer (Van Veen et al. 1997). Through this, we obtained complex values for each voxel and trial, which were used for later analysis. Frequencies of interest were defined based on the effects found in the time-frequency analysis and confirmed based on a comparison between the fusion and unimodal trials. The complex values were first converted into angles (radians), and the difference was calculated between the reference voxel and all other voxels for each trial. From these values, we calculated the circular mean over all trials and employed a Fisher z transformation in order to ensure normal distribution over subjects. Finally, the fusion-trial values were subtracted from the unimodal-trial values. For a global phase locking estimate, we calculated the absolute value and averaged these over all voxels. This procedure yields a measure reflecting large modulations of phase locking between the trial categories, disregarding precise anatomical information as well as the sign of the changes. By performing a t test, the frequencies that are specifically modulated according to the trial category could be extracted. In a second step, we disclosed the main regions that (de)synchronize their phases with the regions identified with the DICS beamformer at the obtained frequency bands of interest.

Statistical Analysis
In order to define relevant time and frequency windows, a cluster-based (at least 2 sensors per cluster) dependent-samples t test with Monte-Carlo randomization was performed on the sensor data (Maris and Oostenveld 2007). This method allows for the identification of clusters of significant differences in 2D and 3D (time, frequency, and space), effectively controlling for multiple comparisons. Clusters were defined as significant if the probability of observing larger effects from shuffled data was below 5%. The cluster-level test statistic is defined as the sum of the t statistics in 2D or 3D space in the respective cluster. For the identification of the probable neuronal generators of the observed sensor effects, statistical comparisons at the source level were computed using dependent-samples t tests. Results on the source level were thresholded and corrected for multiple comparisons using Alphsim (http://afni.nimh.nih.gov/afni/).

Reaction tendencies were computed as a representation of the individual’s behavior. This relative proportion of fusion reactions (number of fusion trials divided by the number of all mismatching trials; high numbers indicate a large tendency toward a fusion percept) in all mismatching trials was correlated with the individual differences (cortical activity or functional connectivity for the fusion trials vs. unimodal trials) at the source level for the time-frequency analyses. This analysis indicates with which neuronal processes the individual predisposition to perceive the McGurk illusion is associated.

All aspects of offline treatment of the MEG signals were accomplished using fieldtrip (Oostenveld et al. 2011), an open source signal processing toolbox for Matlab (www.mathworks.com). Anatomical structures corresponding to the statistical effects are labeled according to the Talairach atlas.

Results

Behavioral Data
Participants were presented with audio streams and either matching or mismatching videos. After each video, subjects had
to report their perception, which in the case of mismatching audiovisual input could either be a fusion of the auditory and visual input (i.e., fusion) or a perception of only one sensory modality (unimodal). In the analysis of the reaction tendency, which was computed as the relative proportion of fusion responses in all mismatching trials, we found that subjects reported a fusion percept in 41.61%, whereas a unimodal percept was reported in 48.02%. Matching stimuli were correctly identified in 96.19%. This difference between the reaction tendencies toward a fusion versus a unimodal response was not significant ($t = -0.8784$, degree of freedom = 16, $P = 0.39$, for details, see Fig. 1B).

**Event-Related Activity**

Event-related activity was compared between the fusion trials and the unimodal trials and revealed differential activity for both response categories. The amplitude in the fusion trials was significantly more pronounced between 550 and 650 ms after the sound onset ($P < 0.05$, Fig. 2A) in a left frontoparietal sensor cluster (Fig. 2B). Source analysis using the LCMV beamformer (Van Veen et al. 1997) suggests that the left middle temporal gyrus is the source of this difference (MNI coordinates [–56, –47, 2], $P < 0.05$, Fig. 2C). This indicates a differential processing between the 2 perceptual categories arising approximately 100-ms poststimulus offset as the audio streams and mouth movements had an average duration of 0.479 s.

**Oscillatory Activity**

We were specifically interested in the influence of ongoing oscillatory brain activity on varying perception of an invariant physical stimulus. For this purpose, we statistically compared the nonbaseline corrected time–frequency representations of fusion and unimodal trials. This analysis revealed 2 significantly different clusters in time–frequency–sensor space of oscillatory activity between the fusion and unimodal trials: one positive before sound onset (i.e., greater power in the fusion trials) and one negative after sound onset (i.e., less power in the fusion trials).

**Prestimulus Activity**

From $–380$ to $–80$ ms before the sound onset, the trials leading to a fusion percept exhibited greater beta-band (14–30 Hz) power ($P < 0.05$). The nonparametric permutation analysis revealed a sensor cluster comprised of bilateral frontotemporal and parietal sensors in which this difference reached significance (Fig. 3A,B), indicating different perception depending on the prestimulus brain state. Due to the low spatial acuity of topographic sensor maps, a correct interpretation of the results requires the identification of possible cortical generators. Beamformer source analysis (DICS, Gross et al. 2001) suggests that 3 sources are involved in the generation of the effect found at the sensor level: ISTG ([–75, 1, 7], $P < 0.05$), precuneus ([14, –60, 38], $P < 0.05$), and right middle frontal gyrus ([53, 5, 42], $P < 0.05$; Fig. 3C). This underscores the role of the ISTG in the perception of the McGurk effect, as suggested by Beauchamp et al. (2010) and partially overlaps with regions of a beta synchronized functional network proposed by Hipp et al. (2011) consisting of frontal, posterior parietal, and lateral occipital areas. A high positive correlation between the reaction tendency toward the fusion percept and the voxel-wise beta power difference values for the comparison between fusion and unimodal trials was found in the right inferior frontal gyrus (rIFG, $t$ = 0.81, [67, 28, 4], $P < 0.001$, see Fig. 4). This underlines the involvement of frontal processes in this effect. Importantly, while prestimulus beta-band power in ISTG differentiated between the “unimodal” and “fusion” trials, power levels in this region did not linearly correlate with the reaction tendency. This indicates that processes at the level of the ISTG alone may be insufficient in explaining an upcoming illusory percept and suggests that information from this region needs to be efficiently distributed to distant cortical regions. In order to test this hypothesis, the ISTG as the primary source identified by DICS and suggested by the literature was subsequently chosen as the seeding region of interest for the analysis of phase synchrony with all other voxels. In line with Hipp et al. (2011), we found a significant difference between the 2 perceptual categories in the beta band. The analysis of phase synchrony between the ISTG and the whole-brain volume revealed an increase ($P < 0.001$) in phase synchrony for the fusion trials relative to unimodal trials with left middle frontal gyrus and right middle temporal gyrus as well as a decrease ($P < 0.001$) in phase synchrony with medial frontal and bilateral STG as well as left fusiform gyrus (see Fig. 5A). Notably, we found a decrease in phase synchrony with the left BA22, while the beta-band power increase was found in our seeding region in the ISTG. Phase synchrony correlated highly ($P < 0.001$) with the tendency toward the fusion percept in the bilateral superior parietal areas, cingulum, left middle occipital gyrus, and right posterior STG. Negative correlations ($P < 0.001$) were found in the right anterior STG and the right inferior temporal lobe (see Fig. 5B). In sum, these results suggest that the functional connection between the

---

**Figure 2.** Summarized results of the time-course analysis. (A) Event-related field trace of the positive sensor cluster for the “fusion” (red) and “unimodal” (blue) trials as well as the $t$-values for the comparison between the 2 conditions (green). (B) Topography ($t$-values, fusion vs. unimodal trials) of the positive sensor cluster (masked for statistical significance) found in the statistical analysis between 550 and 650 ms after sound onset. (C) Source projection ($t$-values, fusion vs. unimodal trials) of the effect found in the event-related field analysis based on the LCMV analysis masked for statistical significance.
lSTG and the frontal and parietal areas as well as the disconnection from inferior temporal areas and the BA22 prior to stimulus onset facilitate subsequent multimodal information integration.

**Poststimulus Activity**

From 200 to 600 ms after sound onset, the fusion trials produced less theta-band (3–7 Hz) power ($P < 0.05$) than the “unimodal” trials. The nonparametric permutation analysis revealed

---

**Figure 3.** Summarized results of the prestimulus time-frequency analysis. (A) Time-frequency representation of the prestimulus interval at sensor level for the comparison between “fusion” and “unimodal” trials. Time 0 ms indicates the onset of mouth movement and audio stream. (B) Topography (14–30 Hz, −380 to −80 ms) of the positive beta-band cluster found in the prestimulus interval at sensor level for the comparison between fusion and unimodal trials. (C) DICS source projection of the beta-band effect found at sensor level. lSTG (left), precuneus (middle), and right middle frontal gyrus (right) were identified as possible generators of the beta-band effect. Displayed are $t$ values of the comparison between trials with a “fusion” percept versus a “unimodal” percept masked for statistical significance.

**Figure 4.** Summary of the correlation between the individual tendency towards a “fusion” percept (0 equals no fusion percept in any trial, 1 equals fusion percept in all trials) and the individual voxelwise power difference values for the comparison between fusion and “unimodal” trials. rIFG (A) displayed a high positive correlation (B) between the individual reaction tendency and the fusion versus unimodal power difference ($r \sim 0.81$). Displayed are correlation indices $r$ masked for statistical significance. High values on the reaction tendency indicate a high tendency toward a fusion percept.
Source analysis was again used to identify possible cortical generators of this effect. Beamformer source analysis (DICS, Gross et al. 2001) suggested cuneus ([9, -84, 5], $P < 0.05$), left superior frontal gyrus ([14, 47, 39], $P < 0.05$), and precuneus ([20, -69, 50], $P < 0.05$) as the sources of the effect found at the sensor level (Fig. 6C). Between the reaction tendency toward the fusion percept and voxelwise power difference values between fusion and unimodal trials, we found a high positive correlation in the right superior frontal gyrus ($r = 0.76$, $P < 0.05$) and a high negative correlation in the rIFG ($r = -0.72$, $P < 0.05$, Fig. 7). Since the theta effects did not point to the involvement of lSTG either at a differential level (i.e., comparing the conditions) or at a correlative level, we refrained from further analysis of functional connectivity for the poststimulus period.

**Discussion**

In the present study, we used MEG to identify cortical responses that differentiate between different perceptions of identical mismatching audiovisual stimuli. We compared “unimodal” perceptions (i.e., no fusion of modalities) and the perception of a fusion of both modalities within identical mismatching trials. The main findings of this investigation are 1) the perception of the McGurk illusion is preceded by relatively increased prestimulus beta activity in distributed cortical regions, in particular the lSTG; 2) compared with
unimodal perceptions, audiovisual integration, as seen in the fusion trials, is characterized by a complex pattern of beta-band coupling and decoupling of the lSTG with frontal and temporal regions; and 3) the individual tendency to “fuse” auditory and visual information is not marked by absolute power increases in the lSTG per se, but by increased right frontal beta activity, increased coupling of the lSTG with frontoparietal areas, and decreased coupling with right temporal areas.

Subjects reported an illusory perception, thus a subjective perception that represents a fusion of auditory and visual stimuli in 41.61% of mismatching trials. This ratio is considerably lower than the 60–80% illusory perceptions reported previously (McGurk and MacDonald 1976). We attribute this difference to the reduced quality of audiovisual stimulation inside the magnetically shielded acquisition room. However, it was decisive for our data analysis to have a sufficient number of illusion trials; thus we do not think this inconsistency is critical for our claims.

Increasing evidence demonstrates that conscious perception requires brain states marked by specific patterns of oscillatory brain activity expressed in local modulations of synchronous activity and synchronized activity between brain regions. Most of the support for this notion come from studies of visual modality that examine whether or not a stimulus was perceived (e.g., near-threshold stimuli; Hanslmayr et al. 2007; Kranzioch et al. 2007; Romei et al. 2010). The overwhelming majority of these studies show a relationship between visual cortex alpha activity and behavior. They indicate that for “simple” perceived versus not perceived distinctions, low-level visual cortical regions must be in a relatively desynchronized alpha state, reflecting an increased excitability of visual regions. In contrast to these experiments with unisensory visual near-threshold stimuli, we found no effects in the alpha band. However, our study embraces the notion of the relevance of prestimulus states, but surpasses previous studies in 2 important regards. First, we investigated the relevance of prestimulus brain states with respect to a conceptually more complex type of perception (i.e., audiovisual integration) and contrasted 2 categories of perception. We specifically aimed at identifying neurophysiologic processes that upon invariant mismatching stimulation differentiate occasions when participants perceive an illusion from those when they do not. Thus, the distinction relevant to our study concerned the content of the percept, rather than whether or not a stimulus was perceived. A popular notion within perception research is that increasingly complex types of perception (e.g., of objects or faces) require the activation of distinct cortical association regions (Kanwisher 2000), also known as “essential nodes” (Zeki and Bartels 1998; Zeki 2003; Koch 2004). The lSTG is one brain region that has frequently been considered as such an essential node with respect to the McGurk illusion and audiovisual integration in general (Beauchamp et al. 2004; Barraclough et al. 2005;

![Figure 6. Summarized results of the poststimulus time-frequency analysis. (A) Time–frequency representation of the poststimulus interval at sensor level for the comparison between “fusion” and “unimodal” trials. Time 0 ms indicates the onset of mouth movement and audio stream. (B) Topography (3–7 Hz, 200–600 ms) of the negative theta-band cluster found in the poststimulus interval at sensor level for the comparison between fusion and unimodal trials. (C) DICS source projection of the theta-band effect found at sensor level. Left superior frontal gyrus (left), precuneus (middle), and cuneus (right) were identified as possible generators of the theta-band effect. Displayed are t values of the comparison between trials with a fusion percept versus a unimodal percept masked for statistical significance.](https://academic.oup.com/cercor/article-abstract/22/1/221/369613)
Stevenson and James 2009; Dahl et al. 2010). Secondly, our study also surpasses previous efforts since it focuses on the influence of prestimulus functional network states on complex speech perception at the source level rather than at the surface sensor level. Trial-by-trial fluctuations have been reported at the sensor level (Kranczioch et al. 2007; Hanslmayr et al. 2007), but on top of missing statements about which brain regions (de)couple, the sensor-level approach suffers from the confounding factor of volume conduction. This confounding factor is strongly attenuated at the source level, particularly with the use of adaptive linear spatial filtering ("beamforming," Schoffelen and Gross 2009). The influence of oscillatory synchronization has recently been demonstrated (Hipp et al. 2011) but only with respect to a simple ambiguous audiovisual stimulus.

**The Perception of the McGurk Illusion Is Preceded by Relatively Increased Prestimulus Beta Activity in Distributed Cortical Regions**

Several studies have already been performed on the comparison between matching and mismatching audiovisual stimuli with regard to poststimulus activity (Senkowski et al. 2008). However, this does not take into account the current brain state at the time when sensory stimulation impinges on ongoing and constantly fluctuating brain activity. Whereas the STG is more strongly activated by matching than by mismatching stimuli—thus pointing to a role in integrating highly correlated information (Beauchamp et al. 2004)—we found more pronounced activity in the IMTG posterior to the STG in the event-related field analysis in the fusion trials, suggesting a stronger activation for a perceived match. Involvement of the IMTG as well as the ISTG has been reported in the processing of audiovisual stimuli in hemodynamic (Calvert et al. 2000; Beauchamp et al. 2004) and electrophysiological (Van Wassenhove et al. 2005; Besle et al. 2008; Arnal et al. 2009; Cappe et al. 2010) studies. It has been shown that presenting auditory stimuli along with mismatching visual information elicits auditory mismatch responses in temporal cortical areas (Möttönen et al. 2002; Saint-Amour et al. 2007). More importantly, we identified the relative prestimulus increase of oscillatory activation in lSTG prior to a fusion percept compared with a unimodal percept within the identical mismatching stimulus category. This suggests that it is not only the congruency that activates the ISTG and possibly IMTG but also that effective integration of multiple sensory information streams depends on prior activation. Local beta-band power might reflect the predisposition of the left STG for integrating multimodal information. Furthermore, whereas ISTG activation putatively reflects a predisposition toward multisensory integration, the perception of this integration, which we see in the present illusion, might also depend on interareal coupling of this region at prestimulus stages via phase-coupled oscillations in the beta range. Thus, STG could indeed be an essential node for audiovisual integration, whose
output however needs to spread to "workspace" regions (Dehaene et al. 2006; see below). A recent review (Senkowski et al. 2008) presents the hypothesis that poststimulus multimodal processing in natural environments depends on a complex network involving frontal, parietal, and temporal regions as well as primary sensory areas rather than a direct synchronization between early sensory areas. However, regardless of the precise functional cause, our study clearly argues for the importance of prestimulus "states" in the case of the lSTG. Without any direct experimental control of the "baseline" period, it is not possible to state to what extent fluctuating levels of selective attention could promote illusory "fusion". Importantly, by showing that the illusion percepts depend on the prestimulus integration of the multisensory region (lSTG) into a distributed corticol network, our study implies that these mechanisms form a predisposition prior to stimulation rather than being elicited by the mismatching stimulus itself. It is however worth noting that, at poststimulus intervals, differences in event-related activity between the conditions were identified in the left MTG ~100 ms following stimulus offset, whereas the relatively decreased level of induced theta activity in the illusion condition was mainly localized to the superior frontal gyrus, cuneus, and precuneus. The present poststimulus results of theta-band modulation as well as the difference in the event-related activity could represent a more general mismatch detection process (Keil et al. 2010). The timing of this effect, the location of the generators of the theta-band modulation, as well as the areas correlating with behavior point to a processing of mismatching information. This has already been reported numerous times in the analysis of the McGurk effect (Möttönen et al. 2002; Saint-Amour et al. 2007). Notably, we found a prestimulus beta-band effect as well as a poststimulus theta-band effect in the precuneus. This indicates a possible link between the prestimulus brain "state" and subsequent mismatch detection. This mismatch detection process likely works at higher cognitive levels rather than at the sensory level, as conflicting information must be coordinated with behavior. Calvert et al. (2000) argues that although integration of modality-specific information occurs in the STG, it may be followed by more elaborate processing in upstream heteromodal areas (MTG). Importantly, we compared brain responses within a stimulus category and thereby excluded strong mismatches of the physical stimulus features, which can be found when comparing between stimulus categories. Our results point to a differential processing of perceived stimulus quality in absence of changes in stimulus material.

Audiovisual Integration Is Characterized by a Complex Pattern of Beta-Band Coupling and Decoupling of ISTG with Frontal and Temporal Regions

In addition to hinting at the importance of functional coupling of the lSTG with parietal and frontal areas, our data indicate a prestimulus functional decoupling of the lSTG from regions processing voice (BA22) and facial information (left fusiform region). This counterintuitive result could suggest a relevant representational state that favors a McGurk illusion: a stronger integration of both sensory streams, resulting in an illusory percept in the case of mismatching information, could be a consequence of "filling in" mechanisms required in the case of degraded unisensory information to the lSTG (Pessoa and De Weerd 2003). Reduced functional coupling between left multimodal integration regions and lower level sensory areas at prestimulus intervals may constitute a predisposition for a subsequently degraded representation of the separate unisensory information. Since the sensory cues from the individual modalities will not suffice in developing a coherent representation, the enhanced prestimulus beta activity in the ISTG could then be interpreted as an adaptive mechanism for more efficient information integration from both modalities in circumstances in which multisensory information is expected. Conflicting or ambiguous information will then be integrated to a novel percept. Interestingly, we observed a decoupling from left BA22 despite the local power increase in the lSTG and an approximately 2 cm (MNI space) distance between the sources. This indicates not only independence between phase synchrony and power as well as the potency of beamformer source analysis to segregate activity from sources in spatial proximity (see Gross et al. 2004) but also underscores the role of the lSTG as a multisensory region that integrates information from lower level sensory areas.

Taken together, our results of pre- and poststimulus effects underscore the importance of state dependency on multimodal integration as well as perception. This finding confirms and extends a recent TMS report (Beauchamp et al. 2010) showing that application of TMS to the lSTG is most efficient in reducing illusory McGurk percepts if applied during a time window between ~100 ms and 100 ms relative to stimulus onset. Crucial for our interpretations, we did not apply so-called "baseline corrections" to our data but directly compared the output of the time-frequency calculation. By applying "baseline corrections," the strong prestimulus effect would have falsely led to effects in the poststimulus interval. Whereas it is likely that the prestimulus frontal and parietal areas found here represent nonsensory higher order brain regions, we cannot rule out the possibility that these regions are also directly activated by multimodal input since most studies in this area recorded single or multiunit activity only from a spatially restricted region of interest (usually the STG, based on anatomical connections to the auditory and visual cortex; Ghazanfar et al. 2008) or blood oxygen level-dependent activity with a low temporal resolution (Calvert et al. 2000). Hipp et al. (2011) have recently suggested the involvement of a frontoparietooccipital beta-band network in audiovisual perception, but more studies using time-sensitive methods with broad spatial coverage (e.g., EEG, MEG, iEEG, or TMS) are required in order to shed light on this issue.

Fusion between Auditory and Visual Information Is Marked by Power Increases in Right Frontal Cortex and Modulated Coupling of ISTG with Frontoparietal and Temporal Networks

In accordance with the data presented above and the existing literature on hemodynamic studies (Beauchamp et al. 2004; Ghazanfar et al. 2008; Stevenson and James 2009), intracranial recordings (Barracough et al. 2005; Besle et al. 2008; Dahl et al. 2010), EEG studies (Van Wassenhove et al. 2005; Cappe et al. 2010), and MEG studies (Arnal et al. 2009), we argue that while the STG is a locus of multimodal information integration, the subsequent perception of this integration depends on a larger distributed network that communicates via phase synchronized activity in the beta band. Recently, Hipp et al. (2011) have reported on perception-related beta-synchrony network in
audiovisual perception, although no power effects were found. The correlation between local beta activity and the tendency to subsequently perceive the McGurk effect as well as the enhanced communication with a frontoparietal system may therefore signify a top-down initiated, enhanced predisposition to integrate upcoming information, and a more efficient transfer of this multisensory information to higher order brain regions.

The impact of functional connectivity states at prestimulus intervals has rarely (Beauchamp et al. 2010) been reported with specific regards to the content of perception, yet poststimulus functional network involvement has been indicated several times (Dehaene et al. 2006; Melloni et al. 2007). Recent notions suggest that we only become aware of representations in sensory and association areas if these engage a distributed frontoparietal (“workspace”) system (Dehaene et al. 2006). We argue that network processes reflected by modulation of both local power and long-range synchrony could already systematically determine multimodal integration at a prestimulus level. Local computations in the STG might be insufficient in eliciting an integrated percept and require an efficient transfer of processed information to a frontoparietal network, reflected in the focal frontal correlation of beta power and the individual tendency to perceive an audiovisual fusion. Whereas coupling between the lSTG and frontal and parietal areas might represent integration into a higher order network, decoupling of sensory areas from the lSTG can be equally as important for illusory perceptions as discussed above. This is underlined by correlations with the individual tendency to perceive a fusion. Parietal and frontal areas as well as areas on the border among temporal, parietal, and occipital cortices show a high correlation between beta-band coupling and the individual perception tendency. Right temporal areas show a decreased beta-band coupling with lSTG, that is negatively correlated to the individual tendency to perceive a fusion. Hein and Knight (2008) recently proposed that the function of superior temporal areas varies depending on the task. The present results could indicate that relative decoupling in the fusion trials as compared with the unimodal trials rather represents an increased coupling in the latter trials favoring the unisensory perception. In this way, a unimodal versus a fused percept depends on the current state of coupling within a larger perception-related network and the lSTG.

Conclusions

Previous research has demonstrated the role of the lSTG in multisensory information processing. In this study, we show for the first time that the (illusory) percept of a fusion between auditory and visual information, as seen in the McGurk effect, critically depends both on prestimulus local beta-band activity as well as on the current functional state of a distributed information-processing network. In particular with regards to the lSTG, which has been the focus of a lot of research as a region of audiovisual integration, our results imply that ongoing prestimulus fluctuations of oscillatory activity as well as fluctuating integration of this region into a distributed network form predispositions whether different sensory streams will be integrated or not. For the lSTG, this predisposition effect appears to be more important than the actual processes elicited by the delivery of the stimulus. A hypothesis derived from our results is that the McGurk illusion is promoted when the functional state prior to stimulus onset favors a degraded representation of unisensory information in the lSTG (decoupling effects). This stimulates a more efficient integration of the degraded individual sensory streams in order to produce a coherent percept (local power effect). Furthermore, in order to perceive the illusion, “incorrectly” integrated information from the lSTG has to be transferred to frontoparietal system (coupling effects).

Funding

Deutsche Forschungsgemeinschaft reference number 4156/2-1, Tinnitus Research Initiative, and Zukunftskolleg of the University of Konstanz.

Notes

We thank Professor Thomas Elbert and Dr Sabine Heim for support and input. Conflict of Interest: None declared.

References


