Visual speech speeds up the neural processing of auditory speech

Virginie van Wassenhove[*†‡], Ken W. Grant[§], and David Poeppel[*†‡¶]

[*]Neuroscience and Cognitive Science Program and Departments of [†]Biology and [¶]Linguistics, University of Maryland, College Park, MD 20742; and [§]Auditory-Visual Speech Laboratory, Walter Reed Army Medical Center, Washington, DC 20307

Communicated by Morris Halle, Massachusetts Institute of Technology, Cambridge, MA, December 6, 2004 (received for review September 8, 2004)

Synchronous presentation of stimuli to the auditory and visual systems can modify the formation of a percept in either modality. For example, perception of auditory speech is improved when the speaker's facial articulatory movements are visible. Neural convergence onto multisensory sites exhibiting supra-additivity has been proposed as the principal mechanism for integration. Recent findings, however, have suggested that putative sensory-specific cortices are responsive to inputs presented through a different modality. Consequently, when and where audiovisual representations emerge remain unsettled. In combined psychophysical and electroencephalography experiments we show that visual speech speeds up the cortical processing of auditory signals early (within 100 ms of signal onset). The auditory–visual interaction is reflected as an articulator-specific temporal facilitation (as well as a nonspecific amplitude reduction). The latency facilitation systematically depends on the degree to which the visual signal predicts possible auditory targets. The observed auditory–visual data support the view that there exist abstract internal representations that constrain the analysis of subsequent speech inputs. This is evidence for the existence of an ''analysis-by-synthesis'' mechanism in auditory–visual speech perception.

EEG | multisensory | predictive coding

Studies of auditory–visual (AV) speech highlight several critical issues in multisensory perception, including the key question of how the brain combines signals from segregated processing streams into a single perceptual representation. In the McGurk effect (1), an audio [p_] dubbed onto a facial display articulating [k_] elicits the "fused" percept [t_], whereas an audio [k_] dubbed onto a visual [p_] elicits various "combinations" such as "pk_" or "kp_" but never a fused percept. These results illustrate the effect of input modality on the perceptual AV speech outcome and suggest that multisensory percept formation is systematically based on the informational content of the inputs. In classic speech theories, however, visual speech has seldom been accounted for as a natural source of speech input. Ultimately, when in the processing stream (i.e., at which representational stage) sensory-specific information fuses to yield unified percepts is fundamental for any theoretical, computational, and neuroscientific accounts of speech perception.

Recent investigations of AV speech are based on hemodynamic studies that cannot speak directly to timing issues (2, 3). Electroencephalographic (EEG) and magnetoencephalographic (4–7) studies testing AV speech integration have typically used oddball or mismatch negativity paradigms, thus the earliest AV speech interactions have been reported for the 150- to 250-ms mismatch response. Whether systematic AV speech interactions can be documented earlier is controversial, although nonspeech effects can be observed early (8).

AV Speech as a Multisensory Problem

Several properties of speech are relevant to the present study. (*i*) Because AV speech is ecologically valid for humans (9, 10), one might predict an involvement of specialized neural computations capable of handling the spectrotemporal complexity of AV speech (compared to, say, arbitrary tone–flash pairings), for which no natural functional relevance can be assumed. (*ii*) Natural AV speech is characterized by particular dynamics such as (*a*) the temporal precedence of visual speech (the movement of the facial articulators typically precedes the onset of the acoustic signal by tens to a few hundred milliseconds (Fig. 1) and (*b*) a tolerance to desynchronization of the acoustic and visual signals of ≈250 ms (11), a time constant characteristic of syllables across languages (12) that relates closely to a proposed temporal integration constant underlying perceptual unit formation (13, 14). (*iii*) For speech processing, abstract representations have been postulated. Specifically, linguistic theories dealing with the constituents of the speech signal and the relation to the stored representations build on the central notion of distinctive feature. These abstract building blocks have precise relations to the (intended) motor commands (articulatory gestures) involved in speech production (15, 16) as well as acoustic interpretations (17). (*iv*) Visual speech provides direct but impoverished evidence for particular articulatory targets; in contrast, the auditory utterance alone usually permits complete perceptual categorization (say, on the phone). For instance, although an audio-alone/pa/leads to a clear percept/pa/, its visual-alone counterpart (i.e., seeing a mouth articulating [pa]) is limited to the recognition of a visual place-of-articulation class, or the "viseme" category bilabials, which comprises the possible articulations [p_], [b_], and [m_].

Neurophysiological Basis of Multisensory Integration

Convergent neural pathways onto multisensory neurons (18) have been argued to provide the substrate for multisensory binding (19). A typical signature of multisensory neurons is the enhanced response (supra-additivity) to the presentation of co-occurring events. Consistent with the concept that multisensory neurons mediate the integration of unisensory information into a multisensory representation, functional MRI studies of AV speech show that auditory and polysensory cortices, specifically superior temporal sulcus and superior temporal gyrus, reflect enhanced activation when compared to unimodal speech (20, 21). The involvement of polysensory areas has suggested a possible computational route for AV speech processing: unimodal signals integrated in multisensory cortical sites (say, superior temporal sulcus) feed back onto primary sensory fields (22, 23). The feedback hypothesis predicts the enhanced activation of auditory cortices (24).

This explanation has appealing properties, but there are complicating factors. Neurophysiology in nonhuman primates shows that classic multisensory integration sites such as superior temporal sulcus demonstrate little specificity for stimulus attributes (25, 26). Consequently, it is difficult to establish (*i*) what

Abbreviations: AV, auditory–visual; A, audio; V, visual; EEG, electroencephalographic; ERP, event-related potential.

[‡]To whom correspondence may be addressed at: University of Maryland, 1401 Marie Mount Hall, College Park, MD 20742. E-mail: virginie.van.wassenhove@mrsc.ucsf.edu or dpoeppel@deans.umd.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0408949102

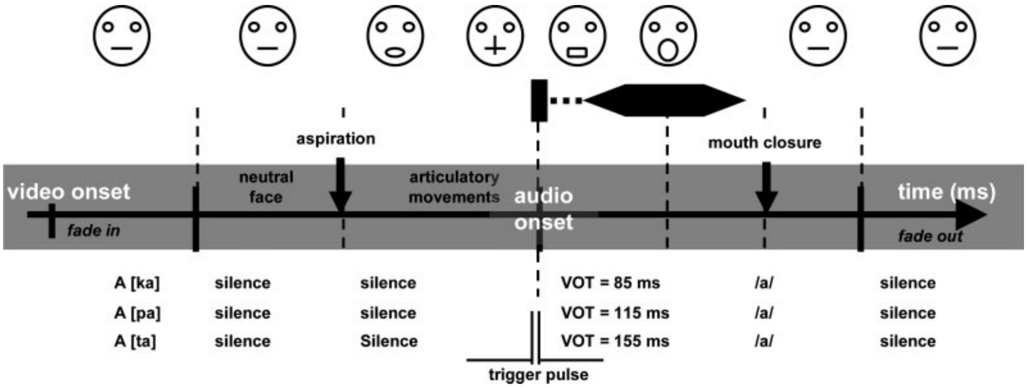
aspiration
mouth closure
video onset
neutral face
articulatory movements
audio onset
time (ms)
fade in
fade out
A [ka] silence silence VOT = 85 ms /a/ silence
A [pa] silence silence VOT = 115 ms /a/ silence
A [ta] silence Silence VOT = 155 ms /a/ silence
trigger pulse

Fig. 1. Timing in natural AV speech. Articulatory movements of the face naturally precede the onset of the audio speech signals by a few tens of milliseconds. The first detectable motion frame demarks the aspiration preceding the production of the consonantal burst in natural speech. Values are for stimuli that were used here. The consonantal burst in the audio portion is the ''audio onset'' and corresponds to the onset or ''index zero'' in all figures and text unless otherwise indicated. VOT, voice onset time.

the nature of information fed back onto auditory cortices may be and (*ii*) whether the generic integration in multisensory sites is sufficient to account for complex rules of integration suggested in AV speech perception. In fact, multisensory integration sites are found throughout the cortex (27), and a fundamental contribution of multisensory neurons may reside in the weighting of one sensory stream against the other, i.e., in reducing stimulus uncertainty (28), rather than in establishing a multisensory perceptual representation. Consistent with this view, complex patterns of activation have been reported that show suppression of sensory-specific cortices (29–31) in conjunction with enhanced activation of multisensory sites. In congruent AV speech, subadditive interactions in polysensory regions have also been observed (32). Additionally, anatomical evidence shows that primary sensory areas are directly interconnected (33–38). Intersensory corticocortical connectivity may mediate cross-modal plasticity when one sensory system is compromised (39), but the role for nonimpaired systems remains unknown.

We investigated the cortical dynamics of perceptual fusion for ecologically natural speech tokens, focusing on the timing of AV integration. We conducted three behavioral and EEG experiments to characterize the influence of visual speech on the most robust auditory event-related potentials (ERPs), N1 and P2, and focused our analysis on systematic variations of the auditory ERP as a function of visual speech information. We used both congruent stimuli (AV syllables [ka], [pa], and [ta]) and incongruent McGurk stimuli (1). In all experiments, participants identified on each trial (three-alternative forced choice) syllables in auditory (A), visual (V), and AV conditions during EEG recording. We show that the visual information systematically influences key timing properties of the auditory responses. We argue for an "analysis-by-synthesis" (15) model for AV speech integration.

Materials and Methods

Participants. Twenty-six native speakers of American English (13 females; mean age, 21.5 years; range, 19–43 years) were recruited. No participant had neurological or audiological problems. They all had normal or corrected-to-normal vision and were right-handed. The study was carried out with the approval of the University of Maryland Institutional Review Board.

Stimuli and Procedure. To preserve the natural timing relations, we used natural speech consisting of a woman's face articulating the syllables [pa], [ta], and [ka]. Movies drawn from a set of stimuli used by Grant and Seitz (40) were rendered into a 640 × 480-pixel movie with a digitization rate of 29.97 frames per s (1 frame = 33.33 ms). Stereo soundtracks were digitized at 44.1 kHz with 16-bit resolution. The incongruent McGurk pair was created by dubbing an audio [pa] (same speaker) onto a video [ka]. The consonantal burst of the digitized audio file [pa] was aligned with the consonantal burst of the underlying audio portion of the video file [ka]. The average duration of the AV stimuli was 2,590 ms, including video fade-in (8 frames), neutral still face (10 frames), place of articulation (variable), and fade out (5 frames). Auditory and visual parameters for each stimulus are illustrated in Fig. 1.

Intertrial intervals were pseudorandomly varied between 500 and 1,500 ms. In experiment 1 ($n = 16$), participants were presented with blocks of 200 AV stimuli (congruent AV [ka], [pa], [ta], and McGurk [ta] presented 50 times per block) and blocks of 240 unimodal stimuli (auditory and visual alone [ka], [pa], and [ta] presented 40 times per block), for a total of 1,000 trials (100 presentations per stimulus). In experiment 2 ($n = 10$), the stimuli used in experiment 1 (A, V, and AV) were presented 100 times per stimulus (total of 1,000 trials) in a fully pseudorandom fashion. Experiment 3 ($n = 10$ subjects who also participated in experiment 1) consisted of 200 incongruent AV stimuli {McGurk fusion and combination pairs (audio [ka] dubbed onto visual [pa]); only fusion responses are reported here}.

Participants were 1 m from the monitor, the movie subtending visual angles of 8.5° (vertical) and 10.5° (horizontal). Videos were displayed centered on a 17-inch Apple G4 monitor on a black background. Sounds were presented through Etymotic ER3A (Etymotic Research, Elk Grove Village, IL) earphones at ≈70 dB sound pressure level. Lights were dimmed. In all conditions, a single-trial three-alternative forced-choice procedure was used. The three choices were [ka], [pa], and [ta]. In the AV conditions (experiments 1 and 2), participants were asked to determine what they heard while looking at the face. In the unimodal conditions (A and V), participants were asked to choose what they heard or saw in the audio or visual conditions, respectively. In experiment 3, participants were asked to report what they saw and neglect what they heard. No feedback was provided.

Recording. EEG recordings were made by using a Neuroscan system (Neurosoft Systems, ACQUIRE 4.2b; Neuroscan Labs, Sterling, VA) with 32 Ag/AgCl sintered electrodes mounted on an elastic cap (Electro-Cap International, Eaton, OH; 10-20 enhanced montage). Data were acquired continuously in ac mode (sampling rate, 1 kHz). Reference electrodes were linked mastoids, grounded to AFz. Four electrodes monitored horizontal and vertical eye movements for off-line artifact rejection. Channel impedances were kept at <5 kΩ.

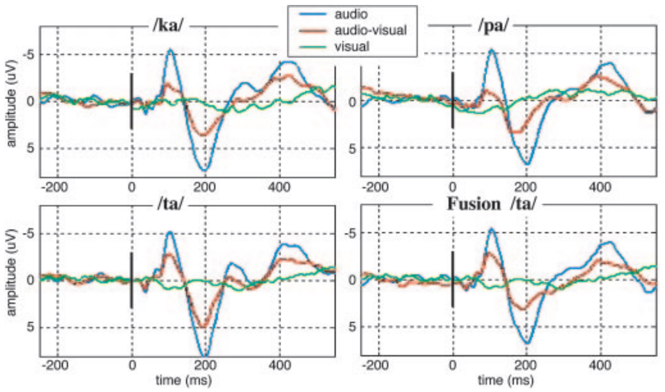

Fig. 2. Average ERPs for four stimulus types. Shown are grand averaged auditory, visual, and AV speech ERPs at a centroparietal recording site illustrating the N1/P2 effects (CPz, data filtered at 1–55 Hz). The black vertical line indicates the onset of the auditory signal. AV speech (red trace) produced faster but smaller auditory ERPs compared to the auditory-alone condition (blue trace). Visual speech (green trace) onset occurred ≈400 ms before auditory onset and did not elicit an auditory ERP but did produce typical visual ERPs at temporo-occipital electrode sites. The three distinct places of articulation as well as the McGurk case are displayed separately.

Analysis. Subsequent to artifact rejection and ocular artifact reduction, epochs were baseline-corrected on a prestimulus interval of 400 ms, chosen before either auditory (A condition) or visual (V alone and AV conditions) onset. Approximately 75–80% of the original recordings were preserved. Individual averages (correct responses only) were made for each signal–response combination. For McGurk conditions, fusion responses [ta] were considered "correct" in all experiments. A zero-phase-shift double-pass Butterworth band-pass filter (1–55 Hz, 48 dB) was applied for ERPs peak analysis. A bootstrapping method (41) was used to resample the data 300 times for each individual, each condition, and each electrode (six electrodes in the present analysis: FC3, FC4, FCz, CPz, P7, and P8). Unprocessed and bootstrapped ERP values were submitted to repeated-measures ANOVAs with the factors modality (two levels: A and AV; in the V condition, no auditory ERP was observed), stimuli (six levels: audio and congruent audiovisual [ka], [pa], and [ta]), ERP component (three levels: P1, N1, and P2), and electrode (six levels). Electrode comparisons were submitted to Greenhouse–Geisser corrections. *t* tests were used to test predicted contrasts. Reported *P* values are for unprocessed ERP values (bootstrapped data lead to similar significant effects).

Results

Fig. 2 shows the grand averaged responses obtained for each place-of-articulation condition tested in A, V, and AV conditions in experiment 1. The presence of visual speech (AV condition) significantly reduced the amplitude of the N1 and P2 auditory ERPs compared to auditory-alone conditions (A), in agreement with the deactivation hypothesis (31, 32) and contrary to the expectation of supra-additivity.

The Amplitude Reduction Is Not Simply Superposition. We compared EEG signals obtained to the presentation of bimodal stimuli (e.g., AV_k) with the estimated sum of the EEG signals obtained to the presentation of the same stimuli in unimodal conditions (e.g., $A_k + V_k$). This method has been used to determine whether AV responses could be solely accounted for by the superposition of auditory and visual evoked potentials (8). A significant deviation from summated unimodal potentials indicates nonadditive interactions. Individuals' average traces were windowed into 50-ms time bins from audio onset to 300 ms postaudio onset. Seven electrodes were chosen for this analysis: left and right parietal (P7 and P8), left and right frontotemporal (FT7 and FT8), and three central locations (FCz, Pz, and Oz). Repeated ANOVAs were performed [factors: electrodes (seven levels), stimuli (k, p, t, and McGurk), response model (bimodal versus summated unimodal), and time window (six levels)]. Greenhouse–Geisser correction was applied. In experiment 1, unimodal and bimodal conditions were run in blocks, and participants knew at the start of a visual trial whether to expect an auditory stimulus. To control for participants' overall expectancy (possible confound with the observed amplitude reduction), the same experimental items were presented pseudorandomly in experiment 2. The same analysis comparing the sum of EEG signals obtained in unimodal conditions with the EEG signals obtained in bimodal condition was performed. It is crucial to note that for both experiments 1 and 2, the observed amplitude reductions at the N1 and P2 cannot be solely accounted for by superposition effects and therefore reflect genuine multisensory interaction.

Characterization of ERPs. In experiment 1, ANOVAs showed a significant effect of modality (A versus AV) on the amplitude of the N1/P2 response component [$F(1.304, 19.553) = 49.53$; $P < 0.0001$]. Additionally, we observed a significant shortening of response peak in AV syllables compared to auditory-alone conditions. No N1/P2 was elicited in visual-alone conditions. Repeated-measures ANOVA testing modality (A and AV) and stimulus identity (k, p, or t) showed a significant interaction [$F(1.834, 27.508) = 14.996$; $P < 0.0001$], with the latency pattern being $p < t < k$. One can observe this effect on the N1, and it is even more pronounced for the P2. These results argue for (*i*) an early and differential AV interaction that is evident as early as the N1 and (*ii*) a manifestation of AV interaction not as response supra-additivity but rather as deactivation and latency shortening. In experiment 2, a similar amplitude reduction was observed affecting the auditory N1/P2 complex in all AV conditions [$F(1.507, 13.567) = 17.476$; $P < 0.0001$]. The temporal-facilitation effect was also observed. Repeated-measures ANOVA showed a significant effect of modality (A and AV) [$F(1, 9) = 21.782$; $P < 0.001$] and a marginally significant interaction of modality and stimulus identity [$F(1.938, 17.443) = 3.246$; $P < 0.06$].

The overall effects of visual speech on auditory ERP amplitude and latency were similar for experiments 1 and 2 (blocked versus randomized designs). It is crucial to note that the temporal-facilitation effect, unlike the amplitude-reduction effect, varied systematically with stimulus identity (p, t, or k), arguing against a generalized attention effect. Whereas visual modulation of auditory ERP amplitude did not significantly vary with stimulus identity [experiment 1: $F(1.884, 28.265) = 1.22$; $P = 0.31$; experiment 2: $F(1.565, 14.088) = 0.033$; $P = 0.94$], the temporal facilitation was a function of stimulus identity [experiment 1: $F(1.908, 28.62) = 13.588$; $P < 0.0001$; experiment 2: $F(1.808, 16.269) = 20.594$; $P < 0.0001$].

As mentioned, articulator movement precedes the auditory signal and may therefore predict aspects of the auditory signal. If a visual input is ambiguous (e.g., visual [k], correctly identified only ≈65% of the time), the predictability of the possible auditory signal should be lower than if the visual stimulus is salient and predictable (e.g., visual [p], ≈100% correct identification), and facilitation effects should vary accordingly: the more salient and predictable the visual input, the more the auditory processing is facilitated (or, the more visual and auditory information are redundant, the more facilitated auditory processing). Consistent with this hypothesis, we observed articulator-specific latency facilitation. Fig. 3 shows the grand averaged visual modulatory effects on N1 and P2 latencies and amplitudes as a function of correct identification in the visual-alone condition. For example, [k] was identified correctly in the visual-alone condition only ≈65% and associated with a 5- to

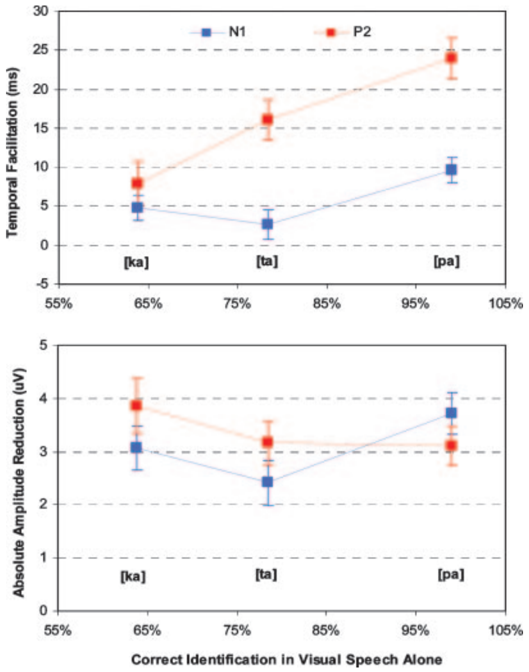

Fig. 3. Latency facilitation and amplitude reduction. Shown is the latency and amplitude difference of N1/P2 in AV syllables as a function of correct identification in the visual-alone condition (experiments 1 and 2, $n = 26$). The latency (*Upper*) and amplitude (*Lower*) differences are the latency (or amplitude) values for the A condition minus the latency (or amplitude) for the AV condition for the N1 (blue) and P2 (red) ERPs. A positive value means that AV is faster than A. The temporal facilitation of the N1 and P2 increased as the saliency (correct identification) of visual inputs improved. The amplitude reduction in AV speech (*Lower*) remained constant across syllables and is independent of visual saliency.

10-ms latency facilitation at the N1 and P2; [p], in contrast, was identified correctly >95% and was associated with a latency facilitation of ≈10 ms at the N1 and ≈25 ms at the P2. These results suggest that the degree to which visual speech predicts possible auditory signals affects the amount of temporal facilitation in the N1 and P2 (Fig. 3 *Upper*) but does not affect its amplitude differentially (Fig. 3 *Lower*).

For McGurk fusion, an audio [p] was dubbed onto a visual [k]. If the rules of integration in AV speech are based on the saliency and redundancy of inputs across sensory channels, one predicts that in McGurk fusion, the ambiguity of the visual speech input [k] will not facilitate the latency of the auditory ERP. The amount of latency facilitation observed in McGurk fusion should be less than for a natural AV [p], for which redundant information is being provided. A similar amplitude reduction that is independent from the informational content of visual speech input (as shown in experiments 1 and 2), however, should be observed. Fig. 4 summarizes the latency and amplitude effects observed in experiments 1 and 2 (filled bars) for congruent AV [p] and the McGurk "fusion" token. As predicted, no temporal facilitation was observed for the McGurk condition, whereas the amplitude decrease of the auditory ERP was comparable to that of a congruent AV [p].

One hypothesis to account for the equivalent amplitude reduction across AV conditions (independent of stimulus identity) is that the visual modality divides the attention that participants focus on the auditory modality. This possibility forced a third experiment, in which we tested the effects of attending to the visual modality when auditory and visual inputs were incongruent (to evaluate with which modality the reported percept is associated). If attending to the visual modality underlies the observed amplitude reduction, one predicts that explicitly directing the participants' attention on the visual inputs would further attenuate the auditory ERP (42). Participants were presented with the McGurk stimuli and answered according to what they saw instead of what they heard. Fig. 4*b Right* (fusion in visual attention) shows that there was no amplitude difference between visually attended incongruent stimuli and either congruent (AV [p]) or incongruent AV stimuli tested in experiments 1 and 2, suggesting that in AV speech, the response reduction is automatic and independent of attended modality. Fig. 4*a* shows the temporal facilitation observed for McGurk fusion. It is surprising that under visual attention, the incongruent AV stimulus showed similar temporal facilitation as observed earlier for congruent AV [p], i.e., the auditory ERPs were temporally facilitated despite the ambiguity in the visual domain.

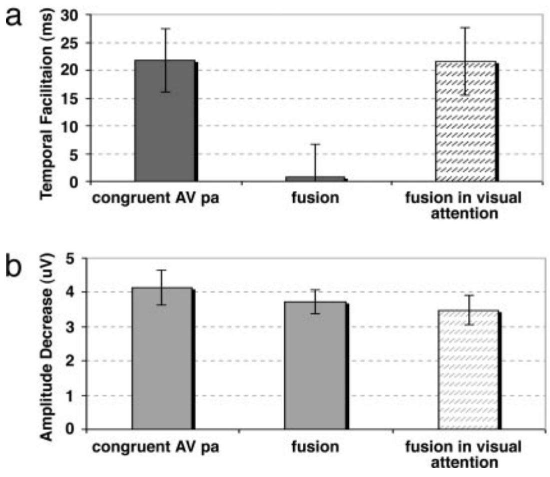

Fig. 4. P2 latency facilitation and intersensory bias. Compared to congruent AV [pa] (*a Left*), no latency facilitation was observed for fusion (*a Center*). When attention is directed to visual inputs in AV conditions, temporal facilitation is recovered in fusion (*a Right*), suggesting that visual attention can enhance the biasing effect of a weak predictor. (*b*) The amplitude decrease was consistent across all stimuli and independent of attended modality, pointing to the automaticity of AV speech integration.

Discussion

Our results show two major electrophysiological features of AV speech integration. First, the degree of perceptual ambiguity in visual speech predicts the processing time in auditory speech, consistent with the use of predictive coding in the neural substrate mediating speech processing. Second, AV speech results in reduced auditory evoked potentials when compared to auditory speech alone. This amplitude reduction is independent of AV speech congruency, participants' expectancy, and attended modality. Our findings suggest that AV speech processing follows specific rules of integration not solely accounted for by general principles of multisensory integration and that (at least) two distinct time scales underlie the integration process.

Perceptual Endowment in Multisensory Integration. EEG studies of multisensory integration for artificial AV pairings thus far have supported the response enhancements observed with functional MRI, showing supra-additivity to the presentation of co-occurrent AV stimuli (8). In particular, the amplitude of the auditory N1/P2 complex was increased in AV conditions (tones paired with circles) and preceded by an early enhanced component (40–90 ms poststimulation). In contrast, we did not find supra-additive enhancements for AV speech processing. We believe that the differences in results support recent work on the functional importance of ecologically valid stimulation in mul-

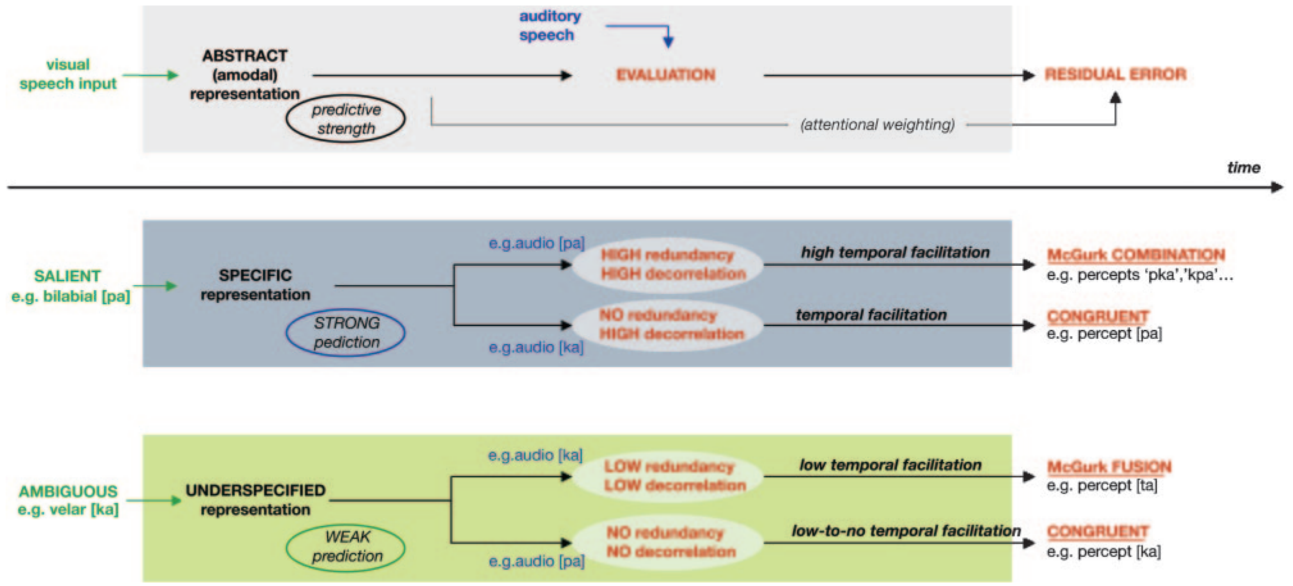

Fig. 5. Analysis by synthesis in AV speech integration. Visual speech inputs typically precede the auditory signals and elicit an abstract speech representation. The predictive value of the abstract token varies as a function of visual saliency and is updated as more visual information is made available. Incoming auditory speech inputs are evaluated against the prediction. Redundancy between predictor and auditory inputs are decorrelated such that greater redundancy leads to greater decorrelation. The stronger the predictor, the faster the auditory speech processing. The N1/P2 complex reflects the residual error of the evaluation process and is observed as amplitude decrease.

tisensory research (9, 10). For example, no clear perceptual categorization of "tones–circles" can be assumed, because no unitary representation tone–circle is available in the absence of specific prior training/learning. Speech perception, however, is a natural property of our species, whether considered in the auditory, visual, or multisensory systems.

Temporal Facilitation. Building on the ecological validity of the signal, we interpret our results as supporting the notion of predictive coding in the context of an analysis-by-synthesis model (15) of AV speech. Fig. 5 illustrates the proposed model, in which the perceptual outcomes depend on (*i*) the saliency of visual inputs and (*ii*) the redundancy of visual and auditory inputs. Predictive coding, first used in motor systems (43), has more recently been tested and extended to sensory systems (44). A central assumption of such models is that an internal representation of the world guides perceptual input and output processes (45). Sensory inputs are not solely processed in a feed-forward fashion but are constrained early on by internal predictions. A major consequence is that early sensory processing can specialize in computing the residual error between the sensory input and the internal prediction, which characterizes the forward nature of the system.

We propose that the natural dynamics of AV speech (e.g., precedence of visual speech inputs) as well as phonological knowledge allow the speech-processing system to build an on-line prediction of auditory signals. The temporal facilitation of auditory ERPs suggests that interactions of AV speech inputs are constrained early on by preceding visual information. In particular, AV speech syllables used in this study naturally provide visible articulatory movements before the acoustic signal. The amount and nature of visual information extracted during this period is proposed to initiate the speech-processing system, in which the formation of an abstract representation is updated continuously through visual inputs, up to the point of explicitly registering auditory input. The set of possible visemic representations initiated in the visual signal provides the context in which auditory inputs are being evaluated. The abstract representations elicited by the visual signals, in turn, provide predictions with precision that correlates with the saliency of visual inputs (i.e., the ease of perceptual categorization in the visual-alone condition) and against which the auditory inputs are being evaluated. What kind of abstract representations of speech can form the basis for such a predictive model and permit a tight mapping between articulatory and acoustic realizations of speech? We suggest that an analysis-by-synthesis model that incorporates the concept of distinctive feature as the elemental representation has the right properties (15, 17).

In this type of model, the temporal facilitation observed in the auditory N1/P2 complex under AV speech conditions reflects the residual errors of the auditory inputs matched against the internal predictor. Based on the neural generators of the N1/P2 (46) and prior evidence of multisensory localization (22, 32), a possible locus executing the relevant computations is the superior temporal gyrus.

The neutralization of the temporal facilitation observed in the McGurk condition points to a possible role of attention in the model. In particular, perhaps the weight of the visually initiated predictor can be regulated by attention. It has been suggested that in conflicting multisensory presentation (such as McGurk), directing attention to a particular modality tends to increase the bias of the attended modality over the unattended modality (47). In our study (experiment 3), this attentional biasing effect is observed as temporal facilitation regardless of the degree of saliency, i.e., the visual-based prediction is proposed here to dominate the auditory input in the evaluation process.

Supra-Additivity. The amplitude reduction of the auditory N1/P2 responses complicates the discussion on supra-additive effects reported for multisensory events in the brain-imaging literature. Recent functional MRI work (32) also shows decreased activation of sensory cortices when stimulation targeted a different modality (decreased auditory cortex activation to presentation of visual stimuli). This finding was proposed to result from a deactivation mechanism in which stimulation of one modality inhibits the nonstimulated modality (30). Consistent with this proposal, deactivation mechanisms may provide a way to minimize the processing of redundant information cross-modally. In our model, the internal prediction deriving from visual input narrows the informational content to place of articulation (viseme). In the incoming acoustic signal, information pertaining to place of articulation is confined, roughly, to the second and third formants. On the assumption that the system acts on incoming inputs to reduce signal uncertainty to extract novel information, the deactivation of auditory cortices by preceding visual inputs could reflect the auditory neural population extracting information only in the relevant frequency range.

Temporal Integration and Early Interaction. The effect of visual speech input on early auditory evoked responses raises the issue of the temporal locus of AV speech integration. Previous electrophysiological studies using the mismatch negativity paradigm in the context of AV speech reported that the mismatch negativity paradigm, typically peaking between 150 and 250 ms, could be elicited when a visual signal incongruent with the auditory speech syllables was presented, suggesting that visual speech accesses auditory sensory memory (4–6). The type and timing of first speech-specific cross-modal interaction, however, has remained speculative. We observe that the processing of auditory speech depends on visual inputs as early as 100 ms (N1 effects both in amplitude and in time), suggesting that the first systematic AV speech interaction occurs before N1 elicitation, at least when predictive context is provided.

Additionally, amplitude and temporal effects evolve on two different time scales: whereas the latency facilitation occurs in the 25-ms range and depends on visual saliency (thereby informational content), the amplitude reduction is independent of visual speech information and spreads over ≈200 ms. These time constants have been hypothesized to underlie feature extraction and perceptual unit formation (14). Our results suggest that at least two computational stages of multisensory interactions are in effect in AV speech integration: first, as reflected in the auditory ERP latency facilitation, a featural stage in which visual information enables the prediction of the auditory input; and second, as reflected in the amplitude decrease, a perceptual unit stage in which the system is in a bimodal processing mode, independent of the featural content and attended modality. The range of temporal phenomena observed electrophysiologically (≈20 ms of temporal facilitation and ≈200 ms of amplitude reduction) may relate to speech features associated with (sub)-segmental-based analysis and syllabicity, respectively.

We show that visual speech differentially modulates early stages of auditory processing (≈50–100 ms). This observation is in line with early integration. The early interaction is manifested as a latency shortening of the N1/P2 responses, conditioned by the salience of visual inputs, which suggests that visual inputs carry a specific predictive value for the auditory utterance. The findings are in line with recent evidence showing that learning cross-modal associations leads to facilitated responses as a function of predictability of the association (47, 48) and with theoretical analyses that highlight the extent to which prior knowledge constrains the internal construction of multisensory perceptual representations (49). In our view, the data are most naturally interpreted in the context of speech perception theories that incorporate an analysis-by-synthesis component.

We thank the reviewers and Frank Guenther for thoughtful and constructive comments on the manuscript and Anita Bowles, Jonathan B. Fritz, and Désirée Gonzalo for comments on earlier versions. This work was supported by National Institutes of Health Grant R01 DC05660 (to D.P.). D.P. acknowledges the support of the Wissenschaftskolleg zu Berlin and the DaimlerChrysler Fellowship at the American Academy in Berlin.

1. McGurk, H. & McDonald, J. (1976) *Nature* **263,** 747–748.
2. Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. (2001) *Nature* **12,** 150–157.
3. Attwell, D. & Iadecola, C. (2002) *Trends Neurosci.* **25,** 621–625.
4. Sams, M., Aulanko, R., Hämäläinen, M., Hari R., Lounasmaa, O. V., Lu, S. T. & Simola, J. (1991) *Neurosci. Lett.* **127,** 141–145.
5. Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F. & Deltenre, P. (2002) *Electroencephalogr. Clin. Neurophysiol.* **113,** 405–206.
6. Möttönen, R., Krause, C. M., Tiippana, K. & Sams, M. (2002) *Brain Res. Cogn. Brain Res.* **13,** 417–425.
7. Fingelkurts, A. A., Fingelkurts, A. A., Krause, C. M., Mottonen, R. & Sams, M. (2003) *Brain Lang.* **85,** 297–312.
8. Giard, M.-H. & Peronnet, F. (1999) *J. Cogn. Neurosci.* **11,** 473–490.
9. De Gelder, B. & Bertelson, P. (2003) *Trends Cogn. Sci.* **7,** 460–467.
10. De Gelder, B., Vroomen, J. & Pourtois, G. (2003) in *Handbook of Multisensory Processes*, eds. Calvert, G., Spence, C. & Stein, B. E. (MIT Press, Cambridge, MA).
11. Munhall, K., Gribble, P., Sacco, L. & Ward, M. (1996) *Percept. Psychophys.* **58,** 351–362.
12. Greenberg, S. (1999) *Speech Commun.* **29,** 159–176.
13. Näätänen, R. (1992) *Attention and Brain Function* (Lawrence Erlbaum Associates, Hillsdale, NJ).
14. Poeppel, D. (2003) *Speech Commun.* **41,** 245–255.
15. Halle, M. (2002) *From Memory to Speech and Back* (Mouton de Gruyter, Berlin).
16. Liberman, A. M. & Mattingly, I. G. (1985) *Cognition* **21,** 1–36.
17. Stevens, K. N. (2002) *J. Acoust. Soc. Am.* **111,** 1872–1891.
18. Stein, B. E. & Meredith, A. M. (1993) *The Merging of the Senses* (MIT Press, Cambridge, MA).
19. Meredith, A. M. (2002) *Brain Res. Cogn. Brain Res.* **14,** 31–40.
20. Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D. & Anthony, S. D. (1999) *NeuroReport* **10,** 2619–2623.
21. Calvert, G. A., Campbell, R. & Brammer, M. (2000) *Curr. Biol.* **10,** 649–657.
22. Calvert, G. A. (2000) *Cereb. Cortex* **11,** 1110–1123.
23. Driver, J. & Spence, C. (2000) *Cur. Biol.* **10,** R731–R735.
24. Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D. & David, A. S. (1997) *Science* **276,** 593–596.
25. Bruce, C., Desimone, R. & Gross, C. G. (1981) *J. Neurophysiol.* **46,** 369–383.
26. Watanabe, J. & Iwai, E. (1991) *Brain Res. Bull* **26,** 583–592.
27. Wallace, M. T., Ramachandran, R. & Stein, B. E. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 2167–2172.
28. Patton, P., Belkacem-Boussaid, K. & Anastasio, T. J. (2002) *Brain Res. Cogn. Brain Res.* **14,** 10–19.
29. Raij, T., Uutela, K. & Hari, R. (2000) *Neuron* **28,** 617–625.
30. Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y.-F., Field, A. S. & Stein, B. E. (2002) *J. Cognit. Neurosci.* **14,** 420–429.
31. Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kanasaku, K. & Hallett, M. (2003) *Nat. Neurosci.* **6,** 190–195.
32. Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J. & McCarthy, G. (2003) *Cereb. Cortex* **13,** 1034–1143.
33. Falchier, A., Clavagnier, S., Barone, P. & Kennedy, H. (2002) *J. Neurosci.* **22,** 5749–5759.
34. Rockland, K. S. & Hisayuki, O. (2003) *Int. J. Psychophysiol.* **50,** 19–26.
35. Foxe, J. J., Wylie, G. R., Martinez, A., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., Ritter, W. & Murray, M. M. (2002) *J. Neurophysiol.* **88,** 540–453.
36. Fu, K. M., Johnston, T. A., Shah, A. S., Arnold, L., Smiley, J., Hackett, T. A., Garraghty, P. E. & Schroeder, C. E. (2003) *J. Neurosci.* **23,** 7510–7515.
37. Werner-Reiss, U., Kelly, K. A., Trause, A., Underhill, A. M. & Groh, J. M. (2003) *Curr. Biol.* **13,** 554–562.
38. Schroeder, C. E., Lindsler, R. W., Specht, C., Marcovici, A., Smiley, J. F. & Javitt, D. C. (2001) *J. Neurophysiol.* **85,** 1322–1327.
39. Bavelier D. & Neville, H. J. (2002) *Nat. Rev. Neurosci.* **3,** 443–452.
40. Grant, K. W. & Seitz, P. F. (1998) *J. Acoust. Soc. Am.* **104,** 2438–2450.
41. Efron, B. (1979) *Ann. Statist.* **7,** 1–2.
42. Woods, D. L., Alho, K. & Algazi, A. (1992) *Electroencephalogr. Clin. Neurophysiol.* **82,** 341–355.
43. Wolpert, D. M., Gharamani, A. & Jordan M. I. (1995) *Science* **269,** 1880–1882.
44. Rao, R. P. N. & Ballard, D. H. (1999) *Nat. Neurosci.* **2,** 79–87.
45. Barlow, H. (1994) in *Large Scale Neuronal Theories of the Brain*, eds. Koch, C. & Davis, J. L. (MIT Press, Cambridge, MA).
46. Näätänen, R. & Picton, T. (1987) *Psychophysiology* **24,** 375–425.
47. Welch, R. B. & Warren, D. H. (1980) *Psychol. Bull* **88,** 638–667.
48. Gonzalo, D. & Büchel C. (2004) in *Attention and Performance*, eds. Kanwisher, N. & Duncan, J. (Oxford Univ. Press, New York), Vol. XX, pp. 225–240.
49. Ernst, M. O. & Bülthoff, H. H. (2004) *Trends Cogn. Sci.* **8,** 162–169.