Behavioral/Cognitive

Neural Mechanisms Underlying Cross-Modal Phonetic Encoding

Antoine J. Shahin,[1] Kristina C. Backer,[1] Lawrence D. Rosenblum,[2] and Jess R. Kerlin[1]
[1]Center for Mind and Brain, University of California, Davis, California 95618, and [2]Department of Psychology, University of California, Riverside, California 92521

Audiovisual (AV) integration is essential for speech comprehension, especially in adverse listening situations. Divergent, but not mutually exclusive, theories have been proposed to explain the neural mechanisms underlying AV integration. One theory advocates that this process occurs via interactions between the auditory and visual cortices, as opposed to fusion of AV percepts in a multisensory integrator. Building upon this idea, we proposed that AV integration in spoken language reflects visually induced weighting of phonetic representations at the auditory cortex. EEG was recorded while male and female human subjects watched and listened to videos of a speaker uttering consonant vowel (CV) syllables /ba/ and /fa/, presented in Auditory-only, AV congruent or incongruent contexts. Subjects reported whether they heard /ba/ or /fa/. We hypothesized that vision alters phonetic encoding by dynamically weighting which phonetic representation in the auditory cortex is strengthened or weakened. That is, when subjects are presented with visual /fa/ and acoustic /ba/ and hear /fa/ (*illusion-fa*), the visual input strengthens the weighting of the phone /f/ representation. When subjects are presented with visual /ba/ and acoustic /fa/ and hear /ba/ (*illusion-ba*), the visual input weakens the weighting of the phone /f/ representation. Indeed, we found an enlarged N1 auditory evoked potential when subjects perceived *illusion-ba*, and a reduced N1 when they perceived *illusion-fa*, mirroring the N1 behavior for /ba/ and /fa/ in Auditory-only settings. These effects were especially pronounced in individuals with more robust illusory perception. These findings provide evidence that visual speech modifies phonetic encoding at the auditory cortex.

Key words: **audiovisual integration; auditory evoked potentials; cross-modal perception; McGurk illusion; speech perception**

Significance Statement

The current study presents evidence that audiovisual integration in spoken language occurs when one modality (vision) acts on representations of a second modality (audition). Using the McGurk illusion, we show that visual context primes phonetic representations at the auditory cortex, altering the auditory percept, evidenced by changes in the N1 auditory evoked potential. This finding reinforces the theory that audiovisual integration occurs via visual networks influencing phonetic representations in the auditory cortex. We believe that this will lead to the generation of new hypotheses regarding cross-modal mapping, particularly whether it occurs via direct or indirect routes (e.g., via a multisensory mediator).

Introduction

Listeners often rely on visual cues (lip movements) to enhance speech comprehension in difficult listening environments (Sumby and Pollack, 1954). Prior reports on the neural mechanisms mediating audiovisual (AV) integration offer diverging, but not mutually exclusive, theories. One theory posits that AV integration, such as in the McGurk illusion (McGurk and MacDonald, 1976), arises when neural representations from the auditory and visual cortices combine in a multisensory network to produce a fused percept (Calvert et al., 2000; Beauchamp et al., 2004). Another theory postulates that visual representations influence activity in the core and belt regions of the auditory cortex, which are traditionally viewed as unimodal (Sams et al., 1991; Besle et al., 2004; Ghazanfar et al., 2005; Schroeder and Foxe, 2005; van Wassenhove et al., 2005; Saint-Amour et al., 2007; Kayser et al., 2010). Besle et al. (2004) and van Wassenhove et al. (2005) examined the influence of visual speech on the N1 auditory evoked potential (AEP), which reflects sound processing in the auditory cortex (Scherg et al., 1989; Zouridakis et al., 1998). They found that the

Received June 6, 2017; revised Nov. 17, 2017; accepted Dec. 8, 2017.

Author contributions: A.J.S. designed research; A.J.S. performed research; A.J.S., K.C.B., and J.R.K. analyzed data; A.J.S., K.C.B. and L.D.R. wrote the paper.

This work was supported by National Institute of Health/National Institute on Deafness and Other Communication Disorders Grant R01-DC013543 to A.J.S. The original data for this study can be accessed at https://data.mendeley.com/datasets/yydw84284f/1. We thank Hannah Shatzer and Dr. Mark Pitt for providing the audiovisual stimuli.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Antoine J. Shahin, Center for Mind and Brain, University of California Davis, 267 Cousteau Place, Davis, CA 95618. E-mail: ajshahin@ucdavis.edu.

DOI:10.1523/JNEUROSCI.1566-17.2017

 0270-6474/18/381835-15$15.00/0

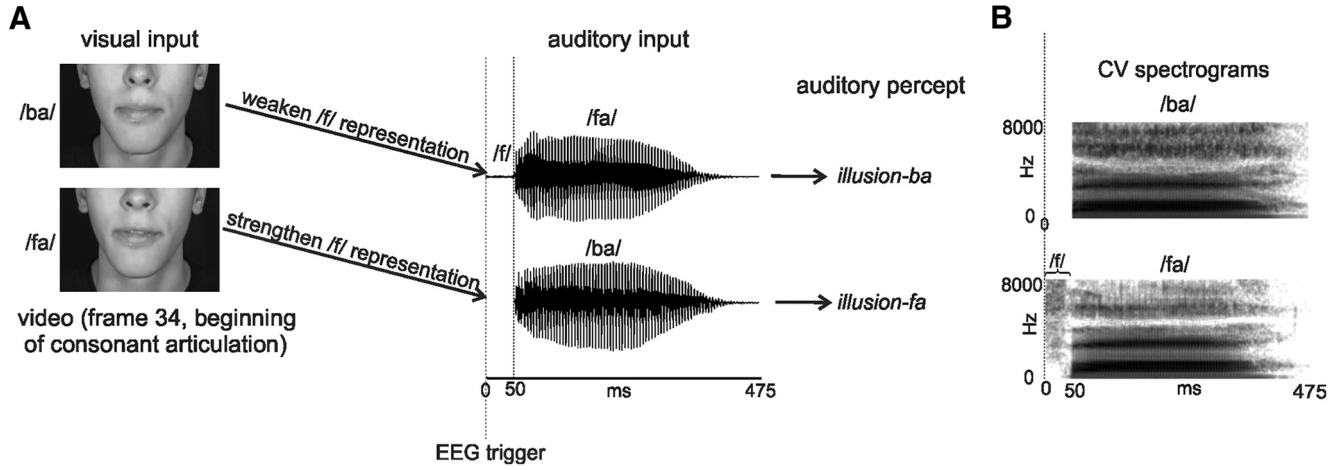

Figure 1. Hypothesis and stimuli. ***A***, Visual depiction of our hypothesis. When perceivers are presented with visual /ba/ and acoustic /fa/ and hear /ba/ (*Illusion-ba*), visual networks weaken the weighting of the phone /f/ auditory representation. When perceivers are presented with visual /fa/ and acoustic /ba/ and hear /fa/ (*Illusion-fa*), visual networks strengthen the weighting of the phone /f/ representation. ***B***, Spectrograms of the acoustic CVs.

N1 amplitude was suppressed during AV versus Auditory-only speech perception. They posited that the preceding information conveyed by vision makes the corresponding auditory information redundant, leading to suppression of the N1 AEP (Besle et al., 2004; van Wassenhove et al., 2005).

In this study, we build on these theories and propose that AV integration of spoken language involves the visual modality acting upon the auditory modality, via strengthening or weakening the weighting of phonetic representations. Although previous studies may have indirectly suggested that the suppressive N1 effect reflects visually induced phonetic encoding (e.g., Besle et al., 2004; van Wassenhove et al., 2005; Pilling, 2009), they were not designed to test for visual modulation of phonetic encoding, nor did they provide explicit evidence to that effect. More relevant to the current research question, however, is an electrocorticography study by Smith et al. (2013). They showed that spectral activity in the parabelt region of the auditory cortex in response to McGurk stimuli was more similar to the spectral activity of the visually conveyed phonemes than the spoken phonemes, indicating that visual context impacts phonetic encoding.

To assess the visual modality's influence on the relative weighting of phonetic representations, we used a play on the McGurk illusion, underscoring the visual system's intricate ability to excite (strengthen) and inhibit (weaken) phonetic representations to alter auditory perception. Thus, we constrained the visual influence to a specific phonetic cue; enhancement or weakening of this cue in the auditory cortex, indexed by a shift in the N1 amplitude, can cause a change in phonetic classification. Subjects were presented with consonant vowel (CV) syllables /ba/ and /fa/ in Auditory-only, AV congruent (e.g., visual /ba/ paired with acoustic /ba/), or incongruent (visual /ba/ paired with acoustic /fa/ or vice versa) settings. They reported whether they heard /ba/ or /fa/. These CVs were used for two reasons. First, they primarily differ in the initial phone /f/; if /f/ is removed, the remainder of /fa/ is heard as /ba/ because the voiced portions of the two syllables have similar formant transitions. Thus, for the AV incongruent stimuli, visual information can lead to the strengthening or weakening of the /f/ representation in the auditory cortex, resulting in a /fa/ or /ba/ auditory perception, respectively. Second, /ba/ and /fa/ CVs evoke distinct N1 amplitudes when presented in Auditory-only or congruent AV settings; the N1 is larger (more negative) for /ba/ than /fa/. We hypothesized (Fig. 1) that, when subjects are presented with visual /fa/ and acoustic /ba/ and hear /fa/ (*illusion-fa*), the visual input strengthens the weighting of the phone /f/ representation, leading to a reduction in the N1 amplitude. When subjects are presented with visual /ba/ and acoustic /fa/ and hear /ba/ (*illusion-ba*), the visual input weakens the weighting of the phone /f/ representation, leading to an enhancement of the N1 amplitude.

Materials and Methods

Subjects

Twenty adults participated in this study. Data from one subject, who reported a language deficit, was excluded from the analyses. The remaining 19 subjects (8 female, 14 right handed, 1 ambidextrous) had a mean ± SD age of 20.9 ± 1.8 years and reported normal hearing, normal/corrected vision, and no history of language deficits or neurological disorders. Five subjects were non-native fluent English speakers. English fluency is defined here as having spoken English continuously for a minimum of 10 years before participation. All subjects provided written informed consent in accordance with the guidelines of the University of California, Davis Institutional Review Board, and they were monetarily compensated for their participation.

Stimuli

The visual and acoustic stimuli were extracted from a video of a female speaker (mean $f_0 = 210$ Hz) uttering CV syllables. The original video was recorded using a Panasonic digital camera AG-DVX100 (30 frames/s) and Adobe Premiere Pro 2.0 (Adobe Systems). Adobe Premiere was also used to edit the video. We selected four video clips of the speaker uttering /ba/ and four clips of the speaker uttering /fa/ from the original video. Each video clip lasted 3 s and contained a CV segment that began and ended with a still face (no mouth movements) and silence. For each CV, three of the four video clips were selected for acoustic stimuli, whereas the fourth clip was selected for the visual stimulus (silent video). The audios were then extracted from the videos, so that they could be paired with different silent videos. All acoustic /ba/ CVs lasted ~425 ms; however, the acoustic /fa/ CVs were always 50 ms longer due to the phone /f/ (Fig. 1*A*). Thus, the fricative /f/ was edited for all three /fa/ sounds to last exactly 50 ms because differences in voice onset time may affect the latency and amplitude of AEPs. The fourth previously selected video, per CV, was stripped of its audio to create a silent video file. The two silent videos (i.e., /ba/ and /fa/) were mixed and matched with the six acoustic tokens (3 /ba/ and 3 /fa/) to create congruent (*AV-congruent*) and incongruent (*AV-incongruent*) pairs of AV stimuli, in addition to visual-only (*V-only*) and auditory-only (*A-only*) conditions. The two silent videos were chosen for two reasons. (1) The lip closure in the /ba/ CV and lip

tuck of the /fa/ CV occurred at about the same time (frame 34 of the video clip, 1089 ms relative to the beginning of the videos). (2) The voiced portion of both the /ba/ and /fa/ CV sounds corresponding to the two chosen silent videos occurred at the same time (1270 ms relative to the beginning of each video). Hence, the 6 CV sounds were paired with the two silent videos in the *AV-congruent* and *AV-incongruent* conditions by aligning the voicing onset (e.g., /b/) of each CV to the 1270 ms time-point relative to the beginning of either of the two silent videos. For the auditory /ba/ CVs, this alignment is straightforward because /b/, thus CV onset, is voiced. However, for /fa/, the /f/ fricative onset occurred at 1220 ms relative to the start of the videos, but the voicing onset occurred at 1270 (same as /ba/) because /f/ was trimmed to exactly 50 ms in duration. The EEG triggers informing sound onsets (Fig. 1*A*) of all stimuli always occurred at the onset of the /fa/ sound (1220 ms after video onset); hence, the triggers for the /ba/ sounds always occurred 50 ms before the onset of /ba/. For this reason, the AEPs of /ba/ were always delayed by 50 ms.

The viseme /fa/ differed in prearticulatory mouth movements from the viseme /ba/, which is taken into account during the EEG analyses. All acoustic CVs were equalized to the same root mean square value. All three acoustic tokens per CV were used in each condition. Because acoustic features vary between utterances even by the same talker, we averaged the EEG data across the three acoustic tokens for each CV within a percept type (e.g., *A-only* /ba/ had all three /ba/ CVs). Thus, differences in the N1 AEP between acoustic /ba/ and /fa/ CVs may be less attributed to meaningless physical variations between the /ba/ and /fa/ utterances.

Procedure

Subjects sat ~85 cm in front of a 24-inch Dell monitor. EEG and behavioral responses were acquired while subjects watched and listened to the AV videos and made judgments on what they heard. The EEG was recorded with a 64-channel cap (BioSemi Active Two system, 10–20 Ag-AgCl electrode system, with Common Mode Sense and Driven Right Leg passive electrodes serving as grounds, A/D rate 1024 Hz). The stimuli were presented using Presentation Software (version 18.1, Neurobehavioral Systems). The sound was played though one loudspeaker (Vizio sound bar, model S2920W-C0) situated below the monitor, at a mean intensity level of 70 dBA sound pressure level. To ensure accurate timing for the EEG analyses, the sound onset triggers were embedded with the wave file metadata. The visual stimuli were cropped to show only the lower half of the talker's face (Fig. 1); the resulting dimensions of these cropped visual stimuli were 13.5 cm (width) × 12 cm (height) on the monitor.

The experiment consisted of six blocks that lasted just over 10 min each. One participant completed only five blocks. Each block consisted of 204 trials presented in an event-related mixed design and randomized among all stimulus types. Stimulus assignment within each block was as follows: 24 trials each for *A-only* /ba/, *A-only* /fa/, *V-only* /ba/, *V-only* /fa/, *AV-congruent* /ba/, and *AV-congruent* /fa/. For the *AV-incongruent* condition, there were 30 trials per CV combination in each block (i.e., 30 visual /ba/ and acoustic /fa/ trials, and 30 visual /fa/ and acoustic /ba/ trials). The larger number of trials in the *AV-incongruent* condition was intentional, in anticipation of illusion-failure trials. Trial duration was ~2700 ms plus a variable jitter of 1 to 500 ms. In the *A-only* condition, subjects listened to an acoustic token while watching a still image of the speaker with her mouth closed. In the *V-only* condition, subjects watched a silent video of the speaker uttering either /ba/ or /fa/. In the *AV-congruent* condition, subjects watched and listened to congruent video and audio files (e.g., visual /ba/ and acoustic /ba/). In the *AV-incongruent* condition, subjects watched and listened to incongruent video and audio files (i.e., visual /ba/ and acoustic /fa/ or vice versa). Subjects indicated whether they heard /ba/ or /fa/ by pressing a keyboard button using their left middle or index finger, respectively. They were instructed to make a quick decision even when unsure of the answer.

To distinguish between conditions and percept types, we henceforth use the following naming convention: the *A-only* condition produced *A-ba* and *A-fa* percept types; the *V-only* produced *V-ba* and *V-fa*; the *AV-congruent* condition produced *AV-congruent-ba* and *AV-congruent-fa*; the *AV-incongruent* condition produced *illusion-ba* (visual /ba/, acoustic /fa/, heard /ba/), *illusion-failure-ba* (visual /ba/, acoustic /fa/, heard /fa/), *illusion-fa* (visual /fa/, acoustic /ba/, heard /fa/), and *illusion-failure-fa* (visual /fa/, acoustic /ba/, heard /ba/). In a subsequent analysis, we also collapsed across the illusion and illusion-failure percepts, producing *illusion+failure-ba* (visual /ba/, acoustic /fa/, heard /ba/ or /fa/) and *illusion+failure-fa* (visual /fa/, acoustic /ba/, heard /ba/ or /fa/).

Data analysis

Behavior

Custom MATLAB code was used to parse the logfiles outputted from the Presentation software, to obtain participants' response type (/ba/ or /fa/) and response time (RT) on each trial. First, we obtained the response data for the *A-only*, *V-only*, and *AV-congruent* conditions, separately for each CV stimulus, /ba/ and /fa/. Because these control conditions have a correct answer, responses were analyzed in terms of accuracy (i.e., percent correct). For the *AV-incongruent* conditions, the response data were analyzed to determine how often the illusory percept was experienced. For example, for the visual /ba/ plus acoustic /fa/ trials, the occurrence of the *illusion-ba* percept was calculated as the number of trials with a /ba/ response divided by the total number of trials with a /ba/ or /fa/ response. Trials without a response (misses) were not included in any analysis. RTs were computed as the amount of time that elapsed between the onset of the audio signal and the response for each trial. The 50 ms delay between the acoustic onset of /ba/ and /fa/ was taken into account when computing RTs. For the *V-only* trials, a silent wave file, which contained a marker at the onset of the video's corresponding audio signal, was played; thus, the RT for *V-only* trials was computed in the same way as the other conditions relative to the start of the "acoustic" onset. For the control conditions (*A-only*, *V-only*, *AV-congruent*), the RTs were calculated for correct trials only. For the *AV-incongruent* conditions, the RTs were calculated for each percept type separately (e.g., *illusion-ba* and *illusion-failure-ba*).

AEPs

Preprocessing of EEG data was done using EEGLAB (Delorme and Makeig, 2004), ERPLAB (Lopez-Calderon and Luck, 2014), and in-house MATLAB code. Statistical analysis of the AEP peaks was done using the cluster-based permutation test implemented in FieldTrip toolbox (Maris and Oostenveld, 2007; Oostenveld et al., 2011).

Preprocessing. Each subject's continuous EEG file, which included the entire dataset, was initially down-sampled to 512 Hz and epoched from 0 to 2700 ms relative to the beginning of the trial, and the mean voltage of each epoch was removed. By "the beginning of the trial," we do not mean the beginning of mouth movements or sounds; rather, it refers to the instance when the trial begins with silence and still frames. As a reference, the sound of /ba/ occurred 1270 ms following the beginning of the trial. Next, we rejected trials with voltage shifts $>\pm 200$ μV between 1070 and 1270 ms (200 ms and 150 ms window before the onset of acoustic /ba/ and /fa/, respectively) at the frontal channels (FP1, FP2, and FPz). This was done before running independent component analysis (ICA) to remove epochs with ocular artifacts (e.g., blinks) occurring in this important period (beginning of articular mouth movements), as visual perception of mouth movements would not be possible while the eyes are closed. ICA was then performed on the epoched individual files that included all conditions and blocks, excluding bad channels. ICA components with topographies indicative of ocular artifacts were rejected (mean 2 components per subject). ICA correction was performed on the file that included all conditions, ensuring that common ICA components were removed from all conditions; this concatenated file was later reepoched and sorted according to condition/percept (see below). Following ICA correction, bad channels (maximum of 2, 4 subjects) were interpolated using EEGLAB's spherical interpolation method. Individual data were then average-referenced and band-passed filtered between 0.1 and 30 Hz using a zero-phase (fourth-order) Butterworth filter. Next, individual data were reepoched from −100 ms to 500 ms, relative to the acoustic stimulus onset (onset of /f/; Fig. 1), linearly detrended for each percept type separately, and rebaselined to the 100 ms preacoustic-stimulus period. Then, trials with amplitude shifts $>\pm 100$ μV in any channel were excluded from the data. Finally, for each subject, each

percept type file was averaged across trials in the time domain to produce AEP waveforms.

Number of trials and criterion for subject inclusion. The mean number and SD of trials of all subjects per percept type after artifact removal were as follows: *A-ba*, 112 ± 21; *A-fa*, 125 ± 17; *V-ba*, 118 ± 21; *V-fa*, 127 ± 17; *AV-congruent-ba*, 129 ± 16; *AV-congruent-fa*, 131 ± 15; *illusion-ba,* 53 ± 54; *illusion-fa,* 124 ± 51. As can be seen, *illusion-ba* had a small mean number of trials with a large SD, indicative of relatively infrequent illusory perception and large subject variability. Thus, to obtain a reliable AEP signal, initial AEP analysis was limited to subjects with a minimum of 40 trials per percept type. Hence, data from 19 of 19 subjects were included for the control percept types *A-ba*, *A-fa*, *V-ba*, *V-fa*, *AV-congruent-ba*, and *AV-congruent-fa*. Data from 9 of 19 subjects were included for *illusion-ba* (99 ± 45 trials); 17 of 19 subjects for *illusion-fa* (136 ± 39 trials); 15 of 19 subjects were included for *illusion-failure-ba* (122 ± 35 trials); and 6 of 19 subjects were included for *illusion-failure-fa* (99 ± 38 trials). In a subsequent analysis (see Interindividual variability), we examined the AEPs when they were averaged across the illusion and illusion-failure percepts. For this analysis, all 19 subjects were included with a mean trial number (± SD) for the combined percepts *illusion+failure-ba* of 155 ± 27 trials and *illusion+failure-fa* of 160 ± 23 trials.

Experimental design and statistical analysis

Recall that subjects were presented with the CVs /ba/ and /fa/ in *A-only*, *AV-congruent*, or *AV-incongruent* contexts, as well as a *V-only* control condition. They reported their auditory perception, whether they heard /ba/ or /fa/. These CVs were used for two reasons. First, they mainly differ in the initial phone /f/; if /f/ is edited out, the rest of the CV is heard as /ba/ because the formant transitions of the voiced portions of the two syllables are similar. Second, /ba/ and /fa/ are distinct by their N1 amplitudes when presented in the *A-only* condition (the N1 is larger for /ba/ than /fa/). The N1's main neural generators originate in the core and belt regions of the auditory cortex (Scherg et al., 1989; Zouridakis et al., 1998). However, AEPs represent the superposition of multiple generators, and contributions from other regions of the brain including the superior temporal sulcus/gyrus (STS/G) cannot be ruled out. Consistent with this attribute, the N1 has been linked to the encoding of simple acoustic features (e.g., sound onsets and pitch) (Jones et al., 1998), which are favorably processed in the core of the auditory cortex, as well as phonetic features (e.g., formants), which have representations in both low- and high-level auditory networks (Ostroff et al., 1998; Toscano et al., 2010; Carpenter and Shahin, 2013; Pereira and Toscano, 2016). We hypothesized (Fig. 1*A*) that, when subjects experience *illusion-fa* (visual /fa/, acoustic /ba/, heard /fa/), the visual context strengthens the weighting of the phone /f/ representation, leading to a reduction in the N1 amplitude. When subjects experience *illusion-ba* (visual /ba/, acoustic /fa/, heard /ba/), the visual context weakens the weighting of the phone /f/ representation, leading to an enhancement in the N1 amplitude. Only then we can claim that visual speech alters phonetic encoding at the auditory cortex.

Behavioral data statistics. Statistical analysis of behavioral data was based on repeated-measures ANOVAs. The variables inputted into each ANOVA are outlined in the results. *Post hoc* analyses used Tukey's honest significant difference tests, and contrasts with p values of 0.05 or less were considered significant. The Greenhouse–Geisser correction method was applied to p values from the main ANOVAs (main effects and interactions) if the sphericity assumption was violated. Effect sizes are denoted by partial η squared (η_p^2). Statistics were performed using Statistica version 13 (Dell Software).

EEG data statistics. We analyzed the AEP data in two ways. First, statistical analysis of AEPs was initially conducted the same way as the behavioral results. This was done using repeated-measures ANOVAs to test for significant differences between the different percept types' N1 amplitude (and latency) obtained at the frontocentral channels, FCz and Cz. However, to correct for the multiple-comparisons problem (due to multiple channels and time points), we also analyzed the data using cluster-based permutation tests (CBPTs) implemented in the FieldTrip toolbox (Maris and Oostenveld, 2007; Oostenveld et al., 2011). Both methods yielded qualitatively similar results, and thus we will focus on the results of the CBPTs.

We initially conducted the CBPTs on the entire waveforms (−100 to 500 ms), but a problem arose due to the latency shift between the N1s of /ba/ and /fa/ CVs (e.g., see Fig. 3*A*). Consequently, the CBPT may yield significant differences between the AEPs of these percept types simply due to this latency shift, rather than a real difference in N1 amplitude. Thus, to circumvent this problem, we developed a hybrid approach, in which we isolated the N1 peak for each participant and percept type, and submitted the amplitude values within a window around each individual's N1 peak (i.e., 10 ms before and after the peak, resulting in a 22 ms window including the peak) for all 64 channels to the CBPT. Peak analysis was performed as follows: (1) The N1 latency was obtained from the group-averaged AEP waveforms at 20 frontocentral channels (F3, F1, Fz, F2, F4, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, CP4) for each percept type. (2) The N1 peak latency (most negative point for the N1) within an 80 (±40) ms around the group peak from Step 1 was obtained for each individual and percept type using the mean AEP waveform of the 20 channels. (3) The amplitude values for each percept type within the 22 ms (11 sample points) window around each subject's N1 peak latency were extracted from the data for all channels and submitted to the CBPTs.

For each percept type contrast, we conducted CBPTs to determine whether and in which channels there were significant differences in N1 amplitude between percept types. First, two-tailed paired-samples t tests were performed on the amplitude values (i.e., the 11 samples of the N1 peak) of two percept types for each channel, to determine univariate effects at the sample level. Only data samples (i.e., time points within each channel) whose t value surpassed an α level of 0.05 (two-tailed) were considered for cluster formation, such that neighboring time points and channels with a univariate p value < 0.05 were grouped together. Neighboring channels were defined using FieldTrip's triangulation method. Finally, cluster-level test statistics were calculated as the sum of all the t values within each time-channel cluster. To evaluate the significance of these cluster-level statistics, a nonparametric null distribution (i.e., a Monte Carlo approximation) was created by repeating the above steps for each of 2000 random partitions (i.e., permutations) of the data, whereby the percept type labels of the data were randomly shuffled. For each permutation, the maximum of the cluster-level test statistics was recorded to form the null distribution. Monte Carlo significance probabilities (p values) were computed by comparing the real cluster-level test statistics with the null distribution of maximum cluster-level statistics. Cluster-based differences between percept types were considered significant if the cluster's Monte Carlo p value was <0.0167. This p value threshold was selected because most analyses involved three contrasts.

Statistical methods that take all channels into account often lead to significant differences between percept types over various scalp locations. This may complicate the interpretation of the results. Because we are explicitly interested in auditory activity, we only considered an effect as pertinent to our research question when it reached significance at channels FCz and/or Cz. Channels FCz and Cz were chosen because auditory activity is traditionally examined at these sites, as evidenced by a vast amount of auditory and AV speech research (Näätänen and Picton, 1987; Stekelenburg and Vroomen, 2007; Shahin et al., 2012; Herrmann et al., 2014). However, the frontocentral activity represents one pole of the auditory generators. Simultaneous activity in the lower temporo-occipital channels represents the opposite pole of the auditory generators that give rise to activity at FCz and Cz. This is evident in the topographies of the subsequent figures. We also chose the FCz/Cz sites as a determinant of a significant effect at the auditory cortex because these channels are least overlapped by visual evoked potentials (see Subtraction of *V-only* AEPs), which are usually largest at occipital sites and reverse at frontal sites. The majority of contrasts produced two significant channel clusters: one frontocentral and one temporo-occipital-parietal. We only discuss the significant results at the frontocentral cluster, which often included FCz/Cz, and only report the cluster-level p values of the significant contrasts ($p < 0.0167$).

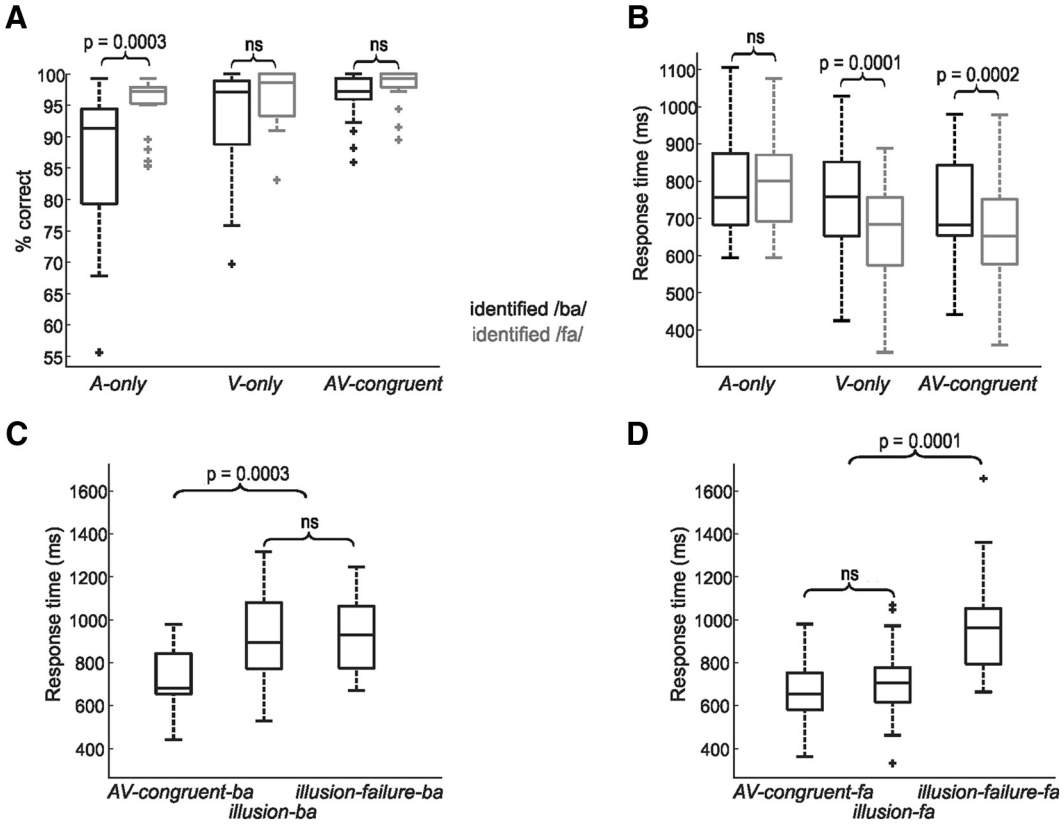

Figure 2. Accuracy and response time. ***A***, Boxplot depicting percent correct identification of /ba/ and /fa/ CVs for the *A-only*, *V-only*, and *AV-congruent* control conditions. Here and in subsequent figures, plus signs indicate outliers. ***B***, RTs for the /ba/ and /fa/ CVs for the three control conditions. ***C***, Boxplots depicting RTs for the *AV-congruent-ba*, *illusion-ba*, and *illusion-failure-ba* percept types. ***D***, Boxplots depicting RTs for the *AV-congruent-fa*, *illusion-fa*, and *illusion-failure-fa* percept types. ns, Nonsignificant ($p > 0.05$) effects.

Results

Behavior

Accuracy and illusion efficacy

We first examined the accuracy of CV recognition for the control conditions, *A-only*, *V-only*, and *AV-congruent* (Fig. 2*A*). A 3 × 2 ANOVA with the variables condition and percept type (/ba/, /fa/) revealed main effects of condition ($F_{(2,36)} = 12.2$, $p = 0.0001$; $\eta_p^2 = 0.4$) and percept type ($F_{(1,18)} = 11.3$, $p = 0.003$; $\eta_p^2 = 0.38$), and an interaction between both variables ($F_{(2,36)} = 4.1$, $p = 0.045$; $\eta_p^2 = 0.19$). The main effect of condition was due to more accurate identification of the CVs for the *AV-congruent* condition versus the *A-only* condition ($p = 0.0002$; Tukey's) and for the *V-only* condition than the *A-only* condition ($p = 0.013$), with no difference in accuracy between the *AV-congruent* and *V-only* conditions ($p = 0.15$). The main effect of percept type was attributed to more accurate identification of /fa/ than /ba/ across all conditions. The interaction, however, revealed that more accurate identification of /fa/ than /ba/ was only significant for the *A-only* condition ($p = 0.0003$), and the differences in accuracy between conditions only occurred for the /ba/ CV (*AV-congruent* > *A-only*, $p = 0.0001$; *V-only* > *A-only*, $p = 0.01$).

Compared with the control conditions, perception in the *AV-incongruent* condition exhibited instability. *Illusion-ba* (/fa/ heard as /ba/) was experienced on average 33% of the time, and the *Illusion-fa* (/ba/ heard as /fa/) was experienced on average 76% of the time (*Illusion-fa* > *Illusion-ba*, $t_{(18)} = 5.3$, $p = 0.00005$). This asymmetry is not surprising, as the occurrence of visually mediated auditory illusion varies across phonemes (McGurk and MacDonald, 1976; MacDonald and McGurk, 1978).

RT

RT relative to the onset of the sound can reflect the ease or difficulty experienced during CV identification. We begin by reporting the RT results for the control conditions (*A-only*, *V-only*, *AV-congruent*), followed by the illusion and illusion-failure percepts. One subject had zero trials for *illusion-failure-fa*; thus, the subject was excluded from RT analyses that included this percept type.

For the control conditions (Fig. 2*B*), an ANOVA with the variables condition and percept type revealed a main effect of condition ($F_{(2,36)} = 30.1$, $p < 0.00001$; $\eta_p^2 = 0.62$) that was due to slower RTs occurring in the *A-only* condition than the *V-only* ($p = 0.0001$; Tukey's) and the *AV-congruent* ($p < 0.0001$) conditions, with no difference between the RTs of the *V-only* and *AV-congruent* conditions ($p = 0.36$). There was also a main effect of percept type ($F_{(1,18)} = 8.3$, $p = 0.01$; $\eta_p^2 = 0.31$); participants responded slower to the /ba/ than /fa/ CVs. Finally, there was an interaction between condition and percept type ($F_{(2,36)} = 18.5$, $p = 0.00003$; $\eta_p^2 = 0.51$), whereby the RT difference between percept types (RT /ba/ > RT /fa/) was found only on the *V-only* ($p = 0.0001$) and *AV-congruent* ($p = 0.0002$) trials, but not on the *A-only* trials ($p = 0.99$). This pattern of results suggests that the visual information for /fa/ was more rapidly identifiable than that for /ba/.

Next, we examined the RTs for the percepts *AV-congruent-ba*, *illusion-ba*, and *illusion-failure-ba* (Fig. 2*C*). We included the *AV-congruent-ba* condition in this analysis because *AV-congruent-ba* and *illusion-ba* have the same perceptual outcome but different acoustic stimuli. Thus, comparing RTs across these two percepts

may indicate the strength of the illusion. If *AV-congruent-ba* and *illusion-ba* have similar RTs, then this would suggest that the /fa/ versus /ba/ decision on illusion trials (*illusion-ba*) is perceptually equal in difficulty as that on *ba-congruent* trials. However, if the RT on *AV-congruent-ba* trials is faster than that on *illusion-ba* trials, then this would indicate that participants may be struggling with the /fa/ versus /ba/ decision on illusion trials, perhaps due to their ambiguity. A one-way ANOVA revealed a main effect of percept type ($F_{(2,36)} = 16.1$, $p = 0.00001$; $\eta_p^2 = 0.47$). Participants responded significantly faster on *AV-congruent-ba* compared with *illusion-ba* ($p = 0.0003$; Tukey's) and *illusion-failure-ba* trials ($p = 0.0001$), with no significant RT difference between the latter two percept types ($p = 0.8$). In other words, compared with /ba/ congruent trials, participants took longer to respond on visual /ba/ plus acoustic /fa/ trials, regardless of whether or not the illusion was perceived. This suggests that the AV incongruent stimulus led to at least some perceptual ambiguity (even when the illusion was perceived) and, consequently, greater difficulty in deciding between /ba/ and /fa/ than the *AV-congruent-ba* trials.

Finally, we examined the RTs for the percepts *AV-congruent-fa*, *illusion-fa*, and *illusion-failure-fa* (Fig. 2*D*). The *AV-congruent-fa* trials were included for the same rationale as the *AV-congruent-ba* trials in the previous analysis. A one-way ANOVA revealed a main effect of percept type ($F_{(2,34)} = 35.7$; $p < 0.00001$; $\eta_p^2 = 0.68$). Participants responded significantly faster on *AV-congruent-fa* and *illusion-fa* trials than *illusion-failure-fa* trials ($p = 0.0001$ for both contrasts, Tukey's). There was no significant difference in RT between *AV-congruent-fa* and *illusion-fa* trials ($p = 0.5$). Thus, a different pattern of results was observed than the previous analysis. Specifically, participants took a similar amount of time to respond on *AV-congruent-fa* and *illusion-fa* trials (unlike *AV-congruent-ba* and *illusion-ba*), suggesting that perceiving /fa/ in visual /fa/ plus acoustic /ba/ trials was relatively unambiguous. However, participants responded slower only when the /fa/ illusion failed, indicating that there was more ambiguity on *illusion-failure-fa* trials than *illusion-fa* and *AV-congruent-fa* trials.

A caveat of the above RT analyses is that some percept types were experienced on very few trials in some subjects, especially for the *illusion-failure-fa* percept. Reanalysis of the previous ANOVA (*AV-congruent-fa*, *illusion-fa*, *illusion-failure-fa*) with 8 subjects that had at least 20 trials per percept type yielded qualitatively similar results.

Summary of behavioral results

Correct identification of CVs in the control conditions (Fig. 2) was on average >90%, suggesting that participants were generally paying attention to the stimuli. Correct identification of the CV /ba/ occurred more frequently for the *V-only* and *AV-congruent* conditions than the *A-only* condition; furthermore, participants were more accurate at identifying /fa/ than /ba/ on *A-only* trials. This suggests that the acoustic /ba/ CV was more difficult to identify than the acoustic /fa/ CV. Participants also responded faster to /fa/ than /ba/ on *V-only* and *AV-congruent* trials, suggesting that visual information was stronger for the phone /f/ than /b/. Overall, more subjects experienced *illusion-fa* (acoustic /ba/, heard /fa/) than *illusion-ba* (acoustic /fa/, heard /ba/). Perhaps the combination of a less identifiable auditory /ba/ and more rapidly identifiable visual information for /fa/ contributed to a stronger /fa/ than /ba/ illusion. Subjects responded equally slower on *illusion-ba* and *illusion-failure-ba* trials compared with *AV-congruent-ba* trials, suggesting that they found these percepts during *AV-incongruent* trials equally ambiguous. However, subjects responded equally fast on *illusion-fa* trials compared with *AV-congruent-fa* trials. Together, this pattern of results suggests that the *illusion-fa* percept was less ambiguous than the *illusion-ba* percept, in line with the previous point: a stronger /fa/ than /ba/ illusion. Subsequent results of the N1 AEP amplitude dynamics concur with this conclusion. That is, the N1 shift due to illusory perception is more robust for *illusion-fa* than *illusion-ba*.

AEPs

From the outset, we confirm that we found no effects of illusion on the N1 latency (obtained at FCz/Cz, using ANOVAs). For example, the N1 latency for *illusion-ba* (/fa/ heard as /ba/) exhibited a similar latency as the N1 for *AV-congruent-fa*. We conclude that the auditory stimulus, not perception, drives N1 latency. However, illusory perception affected N1 amplitude, which we detail below using the CBPTs.

The AEP data of the AV percepts were examined both with and without subtraction of the *V-only* evoked potentials. Examining AEPs after subtraction of the evoked potentials of the silent videos is a way to rule out contributions from visual stimuli, leaving only contributions from the auditory cortex (for drawbacks of this method, see Subtraction of *V-only* AEPs). Because the results of the two analyses concurred, we discuss the raw AEP results (i.e., without subtraction of the *V-only* potentials) in detail and briefly discuss the normalized (with *V-only* subtraction) AEP results.

Control conditions (*A-only*, *V-only* and *AV-congruent*, $n = 19$)

The behavior of the N1 AEP in the control conditions served as a reference to that of the experimental conditions. Figure 3*A*, *B*, *C* (top panels) depicts group AEP waveforms (mean across FCz and Cz) for the /ba/ and /fa/ percept types during the *A-only, V-only*, and *AV-congruent* conditions. Figure 3*A*, *B*, *C* (middle panels) depicts the N1 topographies (mean of 22 ms around the peak) for the two percept types and the three conditions. Figure 3*A*, *B*, *C* (bottom panels) depicts the cluster of channels where a significant difference between the /ba/ and /fa/ percept types occurred for the three conditions. Finally, Figure 3*D* depicts a boxplot of the N1 amplitude averaged across channels FCz and Cz evoked by /ba/ and /fa/ in the three control conditions. The CBPT showed that the N1 was larger for /ba/ than /fa/ for all three conditions at frontocentral sites (*A-only*, $p = 0.001$; *V-only*, $p = 0.007$; *AV-congruent*, $p = 0.001$). However, the difference between the N1 amplitudes of /ba/ and /fa/ percept types at FCz/Cz only reached significance for the *A-only* and *AV-congruent* conditions. This, we believe, indicates a more robust percept type effect at the auditory cortex than the one observed for the *V-only* condition. Upon examination of individual subject data averaged across FCz/Cz, larger (more negative) N1s for /ba/ vs /fa/ were exhibited in 15 of 19 subjects in the *A-only* condition, 13 of 19 for the *V-only* condition, and 18 of 19 in the *AV-congruent* condition. Finally, even when the *V-only* AEP waveforms were subtracted from *AV-congruent* AEP waveforms, the /ba/ vs /fa/ effect remained qualitatively similar ($p = 0.001$).

It is worth noting that all control conditions exhibited significant differences between the two percept types at temporo-occipital-parietal sites. These differences are consistent with stronger auditory generators (opposite poles of the vertex N1), but also the focal and bilateral hot spots (red colored) at these sites (Fig. 3) are indicative of contributions from visual potentials.

Experimental condition (*AV-incongruent*)

The purpose of the analyses described in this section is to test our hypothesis that visual context alters phonetic representations, as

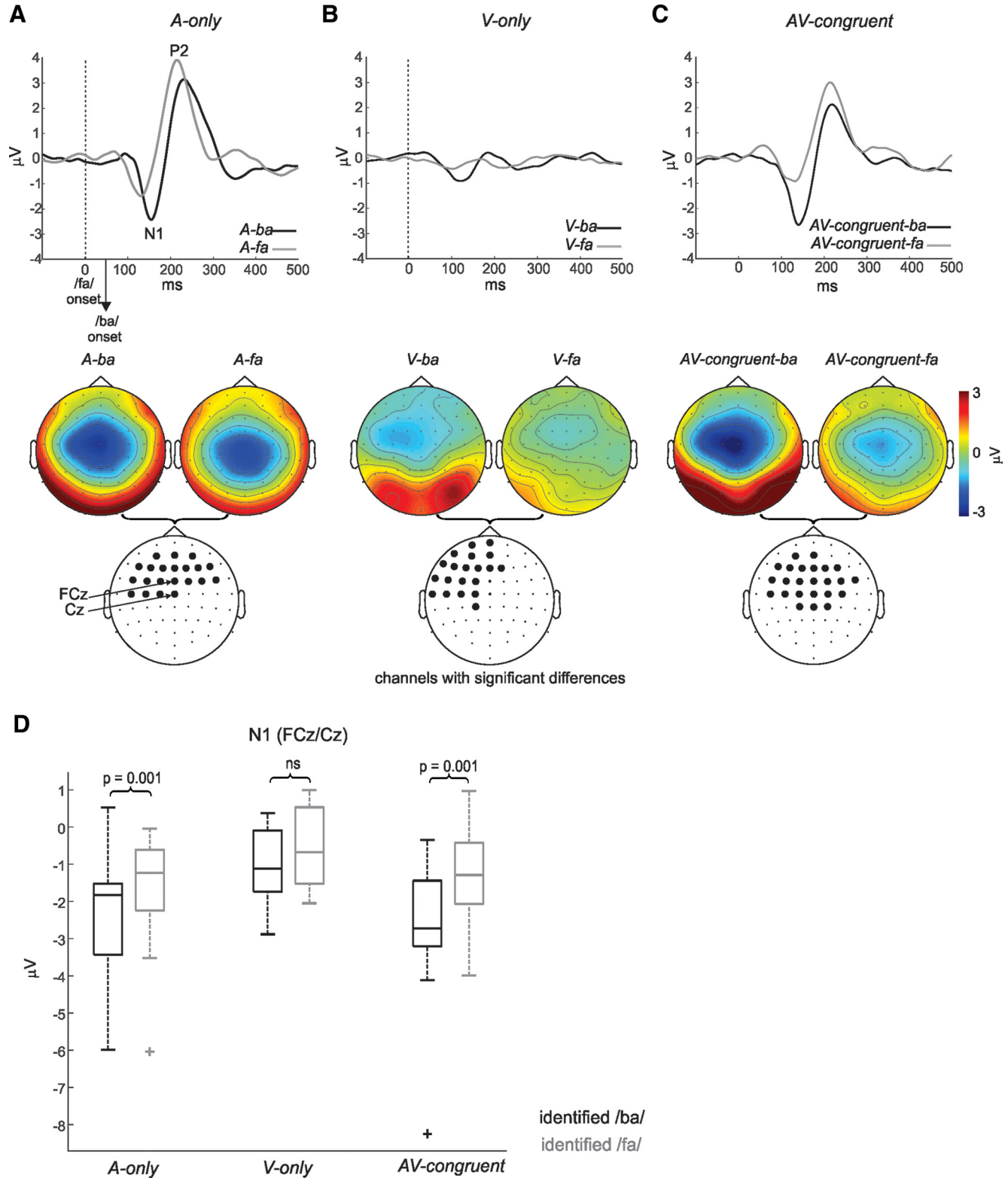

Figure 3. AEPs of control conditions ($n = 19$). ***A–C***, Group AEP waveforms (top panels) and N1 topographies (middle panels) evoked by the CVs /ba/ and /fa/ of *A-only*, *V-only*, and *AV-congruent* conditions. Bottom, Head plots of the cluster of channels where the contrast distinguishing the N1s of /ba/ and /fa/ percepts reached significance for each of the three conditions. ***D***, Boxplot depicting the N1 amplitude data at channels FCz/Cz for the percept types /ba/ and /fa/ for the three control conditions. Here, and in subsequent figures: (1) Time 0 ms indicates the onset of acoustic /fa/. Acoustic /ba/ commenced 50 ms later; this shift should be kept in mind when examining all AEP waveforms. (2) AEP waveforms reflect the mean evoked potential across FCz and Cz. (3) The N1 topographies reflect the group average of each subject's mean amplitude across the 22 ms surrounding the individual's N1 peak latency. (4) The *p* values displayed are those of the cluster-level statistics. (5) ns, Nonsignificant difference at either channel FCz or Cz, or both.

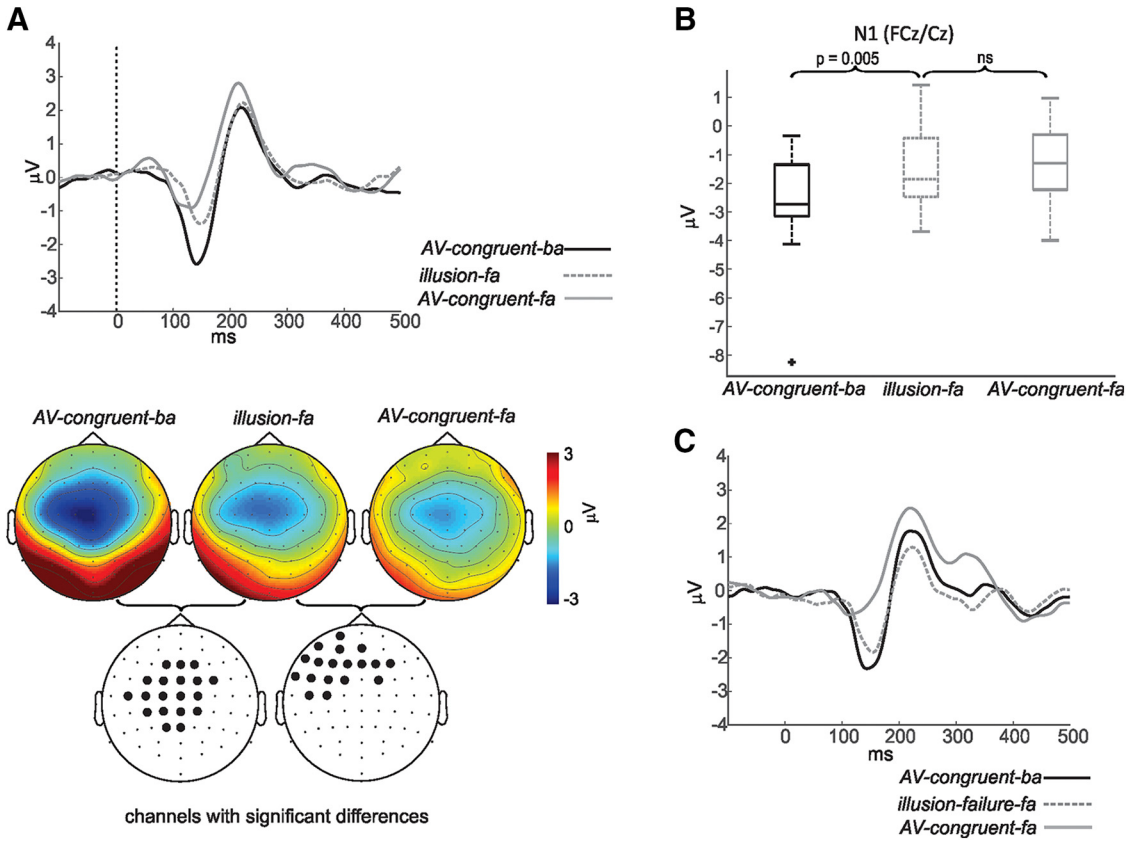

Figure 4. AEPs of illusory /fa/. ***A***, Group ($n = 17$) AEP waveforms (top, channels FCz/Cz) and N1 topographies (middle) of the *AV-congruent-ba*, *AV-congruent-fa*, and *Illusion-fa* (/ba/ heard as /fa/) percept types. Bottom, Head plots of the cluster of channels where significant differences in N1 amplitude were observed between *illusion-fa* versus *AV-congruent-ba* (left) and *illusion-fa* versus *AV-congruent-fa* (right). ***B***, Boxplot depicting the N1 amplitude data for the same percept types averaged across channels FCz/Cz. ***C***, Group ($n = 6$) AEP waveforms for the *AV-congruent-ba*, *AV-congruent-fa*, and *Illusion-failure-fa* percept types.

indexed by changes in the N1 amplitude. That is, the N1 of acoustic /fa/ when heard as /ba/ (*illusion-ba*) due to pairing with visual /ba/ should increase in amplitude, becoming more negative and thus more similar to the N1 amplitude evoked by the *AV-congruent-ba* percept. Conversely, the N1 of acoustic /ba/ when heard as /fa/ (*illusion-fa*) due to pairing with visual /fa/ should decrease in amplitude, becoming less negative and thus more similar to the N1 amplitude evoked the *AV-congruent-fa* percept. We first present the results of the illusion and illusion-failure percepts separately. This limited the number of subjects per analysis because each subject had to attain at least 40 artifact-free trials per percept type to be included in any given analysis (see Number of trials and criterion for subject inclusion). In a subsequent analysis (see Interindividual variability), we collapsed across the illusion and illusion-failure percepts, which allowed us to examine the effects across all subjects.

Illusion-fa (visual /fa/, acoustic /ba/, heard as /fa/; $n = 17$)

On *illusion-fa* trials, individuals listened to visual /fa/ paired with acoustic /ba/ and heard /fa/. According to our hypothesis and the N1 behavior observed in the control conditions (see Control conditions (*A-only*, *V-only* and *AV-congruent*, $n = 19$)), the N1 amplitude of *illusion-fa* should exhibit a smaller amplitude than that of *AV-congruent-ba* and a similar amplitude as that of *AV-congruent-fa*. Figure 4*A* (top) depicts group AEP waveforms (means across FCz/Cz) for the percept types *Illusion-fa*, *AV-congruent-ba*, and *AV-congruent-fa* for the 17 subjects that experienced the illusion (mean illusion 83%) and met the inclusion criteria. Figure 4*A* (middle) shows the N1 topographies of the three percept types, and the bottom shows the clusters of channels where the N1 difference between two percept types (*AV-congruent-ba* vs *illusion-fa* or *illusion-fa* vs *AV-congruent-fa*) reached significance. Figure 4*B* shows a boxplot depicting the group's N1 amplitudes averaged across channels FCz and Cz for the three percept types. The CBPTs revealed that smaller N1 amplitudes occurred for *illusion-fa* than *AV-congruent-ba* ($p = 0.005$); indeed, 14 of 17 subjects exhibited an N1 AEP that was smaller for *illusion-fa* than *AV-congruent-ba* at FCz/Cz. Also, the CBPT revealed a larger left lateralized frontocentral activity for the N1 of *illusion-fa* than *AV-congruent-fa* ($p = 0.004$); but because this effect was not significant at channels FCz and Cz, we consider it a weak effect. This pattern of results is consistent with our hypothesis that the N1 amplitude reflects illusory perception rather than the acoustic characteristics distinguishing the phonemes. Finally, the contrast between the N1 amplitudes of the congruent percepts yielded a significant effect ($p = 0.001$) at channels FCz/Cz, due to larger N1s for *AV-congruent-ba* versus *AV-congruent-fa*.

In a subsequent step, we conducted the same analysis as above but with the evoked potentials of the silent condition subtracted from the AEPs of the three percept types (data not shown). The results remained qualitatively similar, except that the N1 amplitude of *illusion-fa* was intermediate to the N1 amplitudes of the congruent percepts at channels FCz/Cz (N1 *illusion-fa* < N1 *AV-congruent-ba*, $p = 0.001$; N1 *illusion-fa* > N1 *AV-congruent-fa*, $p = 0.007$); 14 of 17 subjects exhibited an N1 AEP that was smaller for *illusion-fa* than *AV-congruent-ba*.

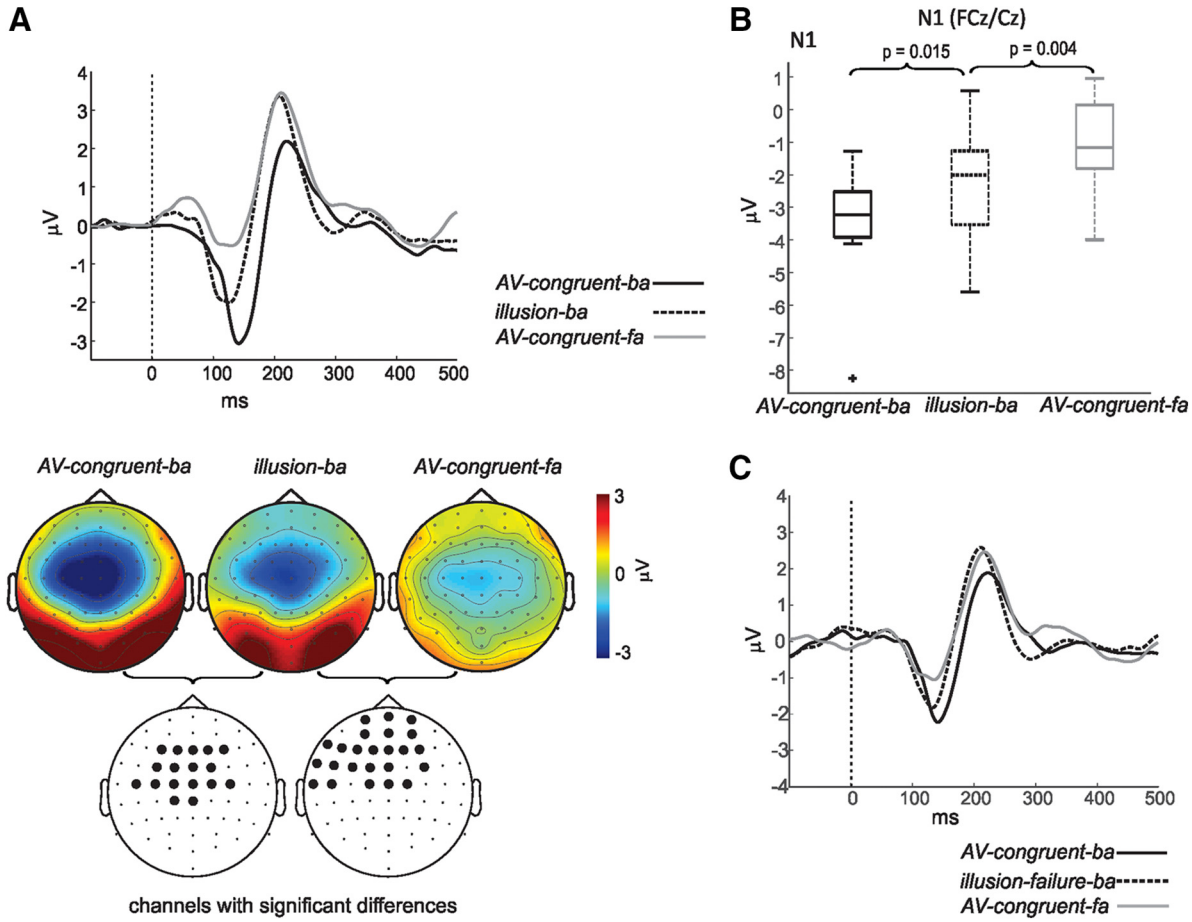

Figure 5. AEPs of illusory /ba/. ***A***, Group ($n = 9$) AEP waveforms (top, channels FCz/Cz) and N1 topographies (middle) of the *AV-congruent-ba*, *AV-congruent-fa* and *Illusion-ba* (/fa/ heard as /ba/) percept types. Bottom, Head plots of the cluster of channels where significant differences in N1 amplitude were observed between *illusion-ba* versus *AV-congruent-ba* (left) and *illusion-ba* versus *AV-congruent-fa* (right). ***B***, Boxplot depicting the N1 amplitude data for the same percept types averaged across channels FCz/Cz. ***C***, Group ($n = 15$) AEP waveforms for the *AV-congruent-ba*, *AV-congruent-fa*, and *Illusion-failure-ba* percept types.

Illusion-failure-fa **(visual /fa/, acoustic /ba/, heard as /ba/; $n = 6$)**
Figure 4*C* depicts the same analysis as in Figure 4*A*, except that, instead of using the N1 data for *illusion-fa*, we used the N1 data for *illusion-failure-fa*. However, this analysis was limited to 6 subjects who met the inclusion criteria. The low number of subjects experiencing *illusion-failure-fa* reflects the robust illusion perception mediated by visual /fa/. In this contrast, we expected the N1 of the *illusion-failure-fa* (/ba/ heard as /ba/) to be similar in amplitude to that of *AV-congruent-ba* and significantly larger than that of *AV-congruent-fa*, that is, opposite the expectations for *illusion-fa.* Even though all 6 subjects exhibited an N1 amplitude of *illusion-failure-fa* that was larger than the N1 amplitude of *AV-congruent-fa*, this effect did not reach significance using the CBPT, most likely due to the small number of subjects. Because of the small number of subjects, we also used a less conservative statistical approach, in which we obtained the N1 amplitude averaged across time (i.e., the 22 ms window around the N1) and channels (i.e., FCz and Cz) for each of these percept types within each subject. These values were submitted to paired-samples *t* tests, which yielded a significant effect ($t_{(5)} = 2.99$, $p = 0.03$) of larger N1 amplitudes occurring for *illusion-failure-fa* than *AV-congruent-fa*, and no difference between the N1 amplitudes of *AV-congruent-ba* and *illusion-failure-fa* ($t_{(5)} = 0.58$, $p = 0.58$). These effects were observed even when the evoked potentials of the silent condition were subtracted from the percepts' AEP waveforms. Despite the low number of subjects, the results support our hypothesis, when the illusion fails (individuals hear /ba/ as /ba/; Fig. 4*C*), the N1 does not shrink in amplitude, like it does during illusory perception (individuals hear /ba/ as /fa/; Fig. 4*A*). This N1 result is consistent with the RT results; the RT for *illusion-fa* was equal to the RT of *AV-congruent-fa*, whereas the RT of *illusion-failure-fa* was delayed relative to the RTs of *illusion-fa* and *AV-congruent-fa*.

Illusion-ba **(visual /ba/, acoustic /fa/, heard as /ba/; $n = 9$)**
On *illusion-ba* trials, individuals listened to a visual /ba/ paired with acoustic /fa/ and heard /ba/. According to our hypothesis, the N1 amplitude of *illusion-ba* should be larger than that of *AV-congruent-fa* and similar to that of *AV-congruent-ba*. Figure 5*A* (top) depicts group AEP waveforms averaged across channels FCz and Cz for the percept types *Illusion-ba*, *AV-congruent-ba*, and *AV-congruent-fa* for the 9 subjects that experienced the illusion (mean illusion 59%) and met the inclusion criterion. Figure 5*A* (middle) shows the N1 topographies of the three percept types, and the bottom shows the clusters of channels where the N1 difference between two percept types (*AV-congruent-ba* vs *illusion-ba* or *illusion-ba* vs *AV-congruent-fa*) reached significance. Figure 5*B* shows a boxplot depicting the group's N1 amplitudes averaged across FCz and Cz for the three percept types. The CBPTs revealed a systematic increase in N1 amplitude differentiating the percepts (N1 *illusion-ba* $<$ N1 *AV-congruent-ba* $> p = 0.015$; N1 *illusion-ba* $>$ *AV-congruent-fa*, $p = 0.004$).

Eight of 9 subjects exhibited an N1 that was larger for *illusion-ba* than *AV-congruent-fa* at channels FCz/Cz. This result is partially consistent with our hypothesis because the N1 of *illusion-ba* was larger than the N1 of the *AV-congruent-fa* percept, but intermediate in amplitude to those of the congruent percepts, rather than exclusively similar to the N1 of the *AV-congruent-ba* percept. This is unlike the results of the *illusion-fa* and *illusion-failure-fa* percepts, whose N1 amplitudes clearly reflected auditory perception. This N1 behavior likely reflects the ambiguity of the *illusion-ba* percept (experienced 59% of the time) and the greater stability of the *illusion-fa* percept (experienced 83% of the time). Finally, the contrast between the N1 amplitudes of the congruent percepts in this group of subjects yielded a significant effect ($p = 0.001$) due to larger N1s for *AV-congruent-ba* versus *AV-congruent-fa*.

In a subsequent step, we conducted the same analysis as above but with the evoked potentials of the silent condition subtracted from the AEPs of the three percept types. This was done for the same rationale as the *Illusion-fa* analysis. However, unlike the *illusion-fa* analysis, which revealed similar results with and without subtraction of the *V-only* AEPs, the *illusion-ba* results did not hold after the subtraction. That is, the N1 amplitude at FCz/Cz for *illusion-ba* was not different from that of *AV-congruent-fa*: it did not change with perception. However, the N1 of *illusion-ba* at FCz/Cz was significantly smaller than that of *AV-congruent-ba* ($p = 0.001$).

***Illusion-failure-ba* (visual /ba/, acoustic /fa/, heard as /fa/; $n = 15$)**

Figure 5*C* depicts the same analysis as in Figure 5*A*, except instead of using the N1 data for *illusion-ba*, we used the N1 data for *illusion-failure-ba* for the 15 subjects that met the inclusion criterion. The larger sample size for this analysis, compared with the *illusion-ba* analysis, is due to a weak illusion and strong illusion-failure experienced by the subjects for the visual /ba/ and acoustic /fa/ pairing. In this contrast, we expected the N1 of the *illusion-failure-ba* to be similar in amplitude to that of *AV-congruent-fa* and significantly smaller than that of *AV-congruent-ba* opposite the expectations for *illusion-ba* (Fig. 5*A*,*B*). However, this was not the case. Larger N1 amplitudes occurred for *illusion-failure-ba* than *AV-congruent-fa* ($p = 0.003$), with no difference between the N1 amplitudes of *AV-congruent-ba* and *illusion-failure-ba*. Eleven of 15 subjects exhibited a larger N1 AEP for *illusion-failure-ba* than *AV-congruent-fa*. These effects were observed even when the evoked potentials of the silent condition were subtracted from the percepts' AEP waveforms, except that the significant effect (N1 *illusion-failure-ba* > N1 *AV-congruent-fa*) was observed at Cz but not FCz. In short, despite the failure to experience the /ba/ illusion (individuals heard /fa/ as /fa/), the N1 behaved similarly as to when individuals perceived the illusion. The N1 of *AV-congruent-fa* was smaller than the N1s of both *illusion-ba* and *illusion-failure-ba*. Thus, the N1 did not robustly reflect perception, for either percept type in this case. Based on our previous analysis (Fig. 5*A*,*B*) and the RT results, which showed equal delays for *illusion-ba* and *illusion-failure-ba* relative to *AV-congruent-ba* (Fig. 2*D*), this N1 behavior is likely due to the perceptual ambiguity of *illusion-ba*.

Interindividual variability

This analysis examined interindividual variability by splitting the subjects into stronger and weaker illusion-perceivers. To obtain adequate power for this analysis, we used all 19 subjects, which was possible by collapsing across illusion and illusion-failure trials. Mixing the two percepts is practical: First, the influence of the illusion can still be gauged because the group that experienced the illusion more robustly will have more illusion trials and fewer illusion-failure trials than the weaker group. Second, the preceding analyses showed that, regardless of illusion success or failure, the N1 amplitude shifts in the same direction. This analogous neurophysiological behavior observed during the illusion and illusion-failure (especially during *illusion-ba* and *illusion-failure-ba*) is not surprising. Several reports on the McGurk illusion have shown that activity at STS is larger for incongruent than congruent stimuli, regardless of whether individuals perceive or fail to perceive the illusion (Benoit et al., 2010; Nath and Beauchamp, 2012). Nath and Beauchamp (2012) further showed that there was no difference in activity when individuals perceived or failed to perceive the illusion. This is consistent with our results that both percepts represent ambiguous perception, with differing degrees of ambiguity across stimulus combinations.

***Illusion+failure-fa* (visual /fa/, acoustic /ba/, heard as /ba/ or /fa/)**

We contrasted the N1 amplitudes for *Illusion+failure-fa, AV-congruent-ba*, and *AV-congruent-fa* for all 19 subjects (Fig. 6*A*). This is a similar contrast as the one depicted in Figure 4*A*, with the exception that we replaced the N1 data of *illusion-fa* with that of *illusion+failure-fa*. Figure 6*B* shows the same contrast but separated into two groups (split in the middle, top 9 vs bottom 10 illusion-perceivers). One group consisted of stronger /fa/ illusion-perceivers ($n = 9$, 97% illusion), and the other group comprised of moderate (weaker) /fa/ illusion-perceivers ($n = 10$, 57% illusion). When collapsing across all subjects, the CBPT revealed an effect that was consistent with, but less robust than, that observed for *illusion-fa* (compare Fig. 6*A* with Fig. 4*A*). That is, the N1 at channels FCz/Cz for *illusion+failure-fa* was intermediate in amplitude to the N1s of the congruent percepts (N1 *illusion-fa* < N1 *AV-congruent-ba*, $p = 0.001$; N1 *illusion-fa* > N1 *AV-congruent-fa*, $p = 0.006$, effect observed at FCz but not Cz,). So despite that the N1 reflected the combined illusion and illusion-failure percepts, the effect remained qualitatively the same as when the N1 reflected the illusion percept only (Fig. 4*A*,*B*). This suggests that, even when the illusion fails, there remains some degree of ambiguity resulting in some shift in the N1 amplitude. These results held even after the evoked potentials of the silent condition were subtracted from the current percepts' AEPs.

To further probe the ambiguity factor, we examined the N1 dynamics in the stronger (less ambiguous) and weaker (more ambiguous) illusion-perceivers. The N1 effect associated with a decrease in N1 amplitude when hearing /ba/ as /fa/ (*illusion-fa*) was most robustly exhibited in stronger illusion-perceivers. In strong /fa/ illusion-perceivers, the N1 amplitude evoked by *illusion+failure-fa* was similar to the N1 of *AV-congruent-fa* (i.e., no significant difference at FCz/Cz) but significantly smaller than the N1 of *AV-congruent-ba* ($p = 0.01$). The moderate (weaker) /fa/ illusion-perceivers did not exhibit a significantly smaller N1 for *illusion+failure-fa* relative to *AV-congruent-ba* (Fig. 6*B*), as was observed for stronger illusion-perceivers. However, like the stronger /fa/ illusion-perceivers, the weaker illusion-perceivers did exhibit larger N1s for *AV-congruent-ba* versus *AV-congruent-fa* (stronger illusion-perceivers: $p = 0.009$; weaker illusion-perceivers: $p = 0.01$, reaching significance at Cz but not FCz). These effects were reproduced when the evoked potentials of the silent condition were subtracted from the current percepts' AEPs, except that, for the stronger illusion-perceivers, the N1 of *illusion-fa* became intermediate relative to the N1s of the congruent percepts.

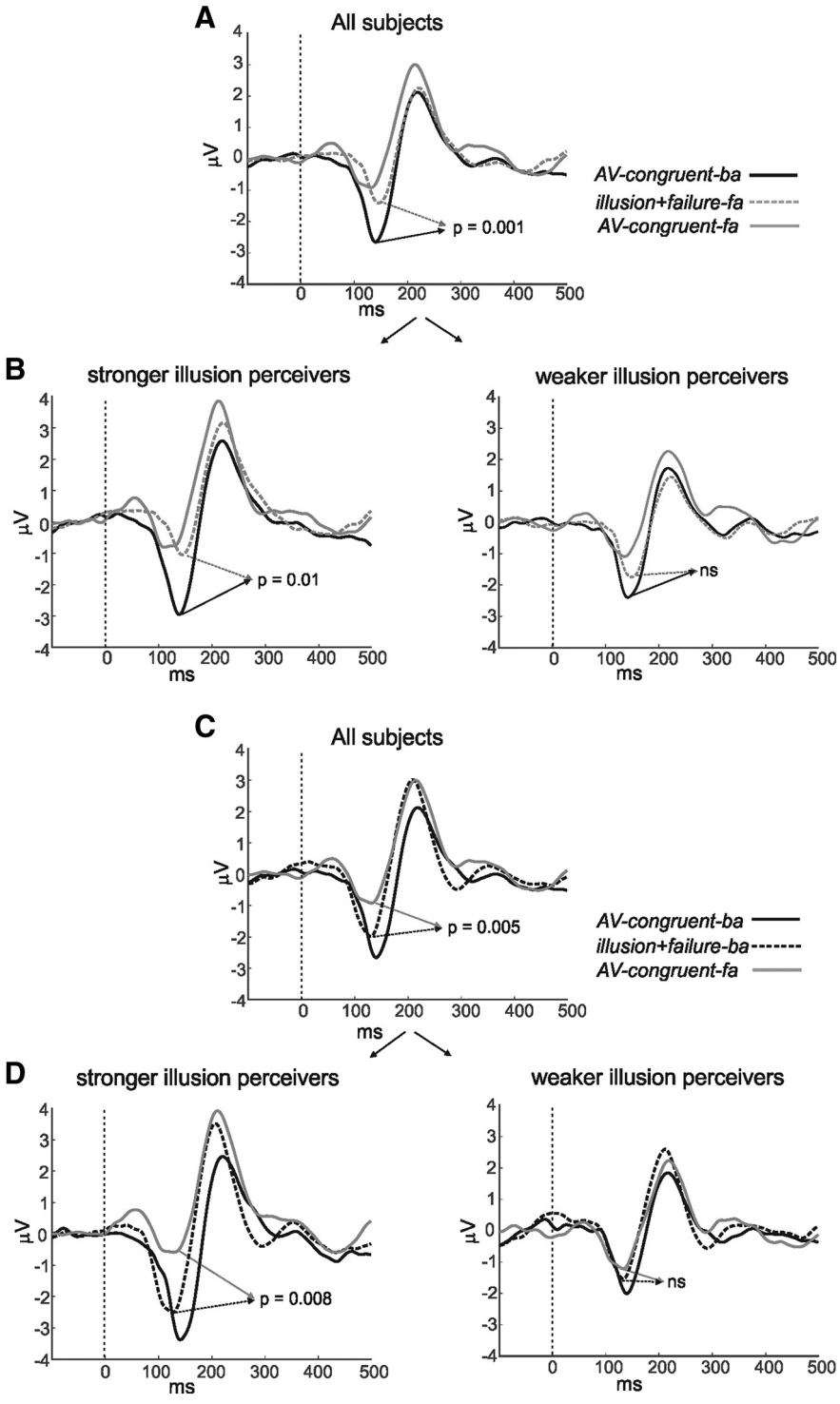

Figure 6. Interindividual variability. ***A***, AEP waveforms for the same illusory /fa/ contrast as in Figure 4*A*, except that the AEP waveforms represent the combination of AEPs of the illusion and illusion-failure percepts, as opposed to illusion percept only in Figure 4*A*. ***B***, The same contrast as in ***A***, except that the AEPs are split into strong ($n = 9$, mean illusion of 97%) and moderate ($n = 10$, mean illusion of 57%) /fa/ illusion-perceivers. ***C***, AEP waveforms for the same illusory /ba/ contrast as in Figure 5*A*, except that the AEP waveforms represent the combination of AEPs of the illusion and illusion-failure percepts. ***D***, The same contrast as in ***C***, except that the AEPs are split into moderate ($n = 9$, mean illusion of 59%) and weak ($n = 10$, mean illusion of 10%) /ba/ illusion-perceivers.

Illusion+failure-ba (visual /ba/, acoustic /fa/, heard as /ba/ or /fa/)

Similar to the *illusion+failure-fa analysis*, we contrasted the N1 amplitudes for *Illusion+failure-ba*, *AV-congruent-ba*, and *AV-congruent-fa* for all 19 subjects (Fig. 6*C*). This is a similar contrast as the one depicted in Figure 5*A*, with exception that we replaced the N1 data for *illusion-ba* with that of *illusion+failure-ba*. Figure 6*D* shows the same contrast but separated into two groups (top 9 vs bottom 10 illusion-perceivers). One group consisted of moderate (stronger) /ba/ illusion-perceivers ($n = 9$, 59% illusion) and the other group comprised of weaker /ba/ illusion-perceivers ($n = 10$, 10% illusion). When collapsing across all subjects, the CBPTs revealed that the percept type effect mirrored the effect for *illusion-ba* (compare Fig. 6*C* with Fig. 5*A*). That is, the N1 at channels FCz/Cz for *illusion+failure-ba* was intermediate in amplitude to the N1s of *AV-congruent-ba* ($p = 0.005$) and *AV-congruent-fa* ($p = 0.001$). So despite that the N1 in the current analysis reflected the illusion and illusion-failure percepts, the effect remained the same as when the N1 reflected the illusion percept only (Fig. 5*A*,*B*). Again, this suggests that, even when the illusion fails, there is ambiguity, which causes the N1 to shift. However, this effect was most robustly exhibited in stronger /ba/ illusion-perceivers. In this group, the N1 evoked by *illusion+failure-ba* was smaller than the N1 of *AV-congruent-ba* ($p = 0.008$) and significantly larger than the N1 of *AV-congruent-fa* ($p = 0.001$). The stronger illusion-perceivers also showed a significant difference between the N1 amplitudes of the congruent percepts (*AV-congruent-ba* > *AV-congruent-fa*, $p = 0.001$). In contrast, the weaker /ba/ illusion-perceivers did not exhibit differences between the N1 amplitudes of the *illusion+failure-ba* relative to those of the congruent percepts. They did not even exhibit a difference between the N1 amplitudes of the congruent percepts. Finally, these effects were reproduced when the evoked potentials of the silent condition were subtracted from the current percepts' AEPs.

Summary of AEP results

The N1 amplitude shift for illusory /fa/ was more robust than that for illusory /ba/. This is consistent with the behavioral results, which showed more robust illusory perception mediated by visual /fa/ than visual /ba/ (see Behavior). The N1 AEP was larger for /ba/ than /fa/ in the *A-only* and *AV-congruent* control conditions, but not in the *V-only* control condition. In the experimental condition (*AV-incongruent*), several analyses revealed that the N1 of the illusory percepts (*illusion-ba* and *illusion-fa*) shifted in amplitude with perception. The aforementioned shift mirrored the N1 ampli-

tude in response to *A-only* and *AV-congruent* /ba/ and /fa/. That is, when auditory /fa/ was perceived as /ba/ (*illusion-ba*), the N1 to auditory /fa/ increased in amplitude, becoming similar to the N1 amplitude of *AV-congruent-ba*. When auditory /ba/ was perceived as /fa/ (*illusion-fa*), the N1 to auditory /ba/ decreased in amplitude (i.e., shifted toward the N1 amplitude of *AV-congruent-fa*). In the last set of analyses, the subjects were divided into subgroups, based on how often they experienced each illusion. This analysis revealed that the above N1 effects were significantly more pronounced in individuals that experienced the illusion more robustly.

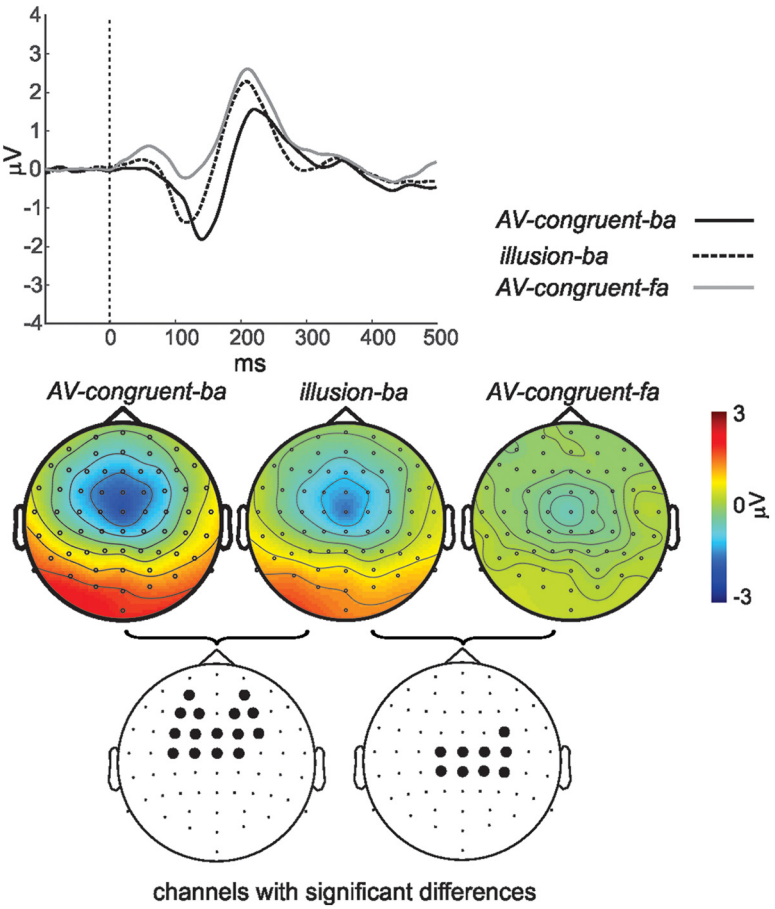

Figure 7. The same analysis depicted in Figure 5*A*, except the data were based solely on ICA components that represented auditory topographies.

Caveats and considerations

Subtraction of V-only AEPs

Subtraction of evoked potentials of the silent condition (*V-only*) from AEPs of the AV conditions is a practice commonly used in AV research (Stekelenburg and Vroomen, 2007; Alsius et al., 2014; Baart and Samuel, 2015). Our results except for one (*illusion-ba* vs *AV-congruent-fa*) largely held with this subtraction. However, a caveat of this approach is that, if the silent condition evokes auditory activity (Calvert et al., 1997; Pekkola et al., 2005; Besle et al., 2008), as can safely be deduced from the topographies of the *V-only* condition (Fig. 3*B*; frontocentral negativity), then by removing the evoked potentials of the silent condition, a critical effect (visual modulation of auditory cortex) is also removed. Also, the assumption of additivity (AV = A + V) is not well supported (Besle et al., 2004; van Wassenhove et al., 2005; Pilling, 2009). Nonetheless, in the present study, visual evoked potentials superimpose with the AEPs as observed in Figure 3*B*, *C* (left topographies) and Figure 5*A* (left and middle topographies). However, it is not clear to what extent, if any, this superimposition affects the AEPs at frontocentral sites (channels FCz/Cz), where auditory activity is prominent. Figures 3*B*, *C* and 5*A* show a distinct focal activity at parieto-occipital sites (red) that is consistent with visual potentials. However, they also show negative frontal and vertex activity (blue). The vertex activity is consistent with a negative auditory potential (N1), although the frontal activity most likely reflects the negative pole of the generators of the visual evoked potentials over parieto-occipital sites. We have no way of knowing whether this visual activity is related to viseme representations (a meaningful visual response coinciding with the auditory N1) or due to nonmeaningful mouth movements (i.e., a confound).

To provide reassurance that the visual activity did not influence the results, we conducted a supplementary analysis. We reanalyzed the contrast depicted in Figure 5*A*, with only ICA components (maximum of 3 per subject) that represented auditory activity. We discarded all other components, including those representing visual evoked potentials. We stress that the rejected ICA components were common across all conditions. If the N1 AEP effect observed in Figure 5*A* is an artifact of visual evoked potentials, then it should diminish following this operation. The results of the new analysis using the CBPTs are shown in Figure 7 (compare with Fig. 5*A*, AEP waveforms and topographies). Two striking differences distinguish the two figures: (1) the waveforms are substantially smaller in the new analysis; and (2) the posterior focal positivities (visual evoked potentials) have disappeared. More importantly, despite a reduction in amplitude and the absence of visual potentials, the outcome was qualitatively the same as the original analysis. The CBPTs revealed that the N1 amplitude of *illusion-ba* was smaller than that of *AV-congruent-ba* ($p = 0.013$), and larger (more negative) than the N1 of *AV-congruent-fa* ($p = 0.001$ reaching significance at Cz but not FCz; 8 of 9 subjects showed this effect). In other words, perception of *illusion-ba* increased the N1 amplitude of /fa/ to be intermediate to the congruent percepts, as was found in the original analysis, which included ICA associated with visual potentials (Fig. 5*A*). This supplementary analysis reduces the likelihood that the effects observed in this study are attributed to visual evoked potentials contaminating auditory activity.

Baseline choice

We baselined the poststimulus activity to the 100 ms preacoustic stimulus period. However, this period may be contaminated by activity attributed to the preceding visual stimulus. To alleviate our concern, we conducted a reanalysis of the data in which we baselined to an earlier period (−600 to −500 ms prestimulus). This earlier period occurred before any prearticulatory mouth movements. All effects reported in the original analyses were qualitatively replicated. The decision to use the 100 ms preacoustic period for the baseline was motivated by a strategy to maxi-

mize the number of trials because more trials are rejected (due to artifacts) with a longer prestimulus interval (and thus a longer epoch).

Discussion

Brief summary and theory

We show that the N1 amplitude is augmented when subjects perceive *illusion-ba* (visual /ba/, acoustic /fa/, heard /ba/) and weakened when they perceive *illusion-fa* (visual /fa/, acoustic /ba/, heard /fa/). These effects mirror the direction of the N1 amplitude observed for these CVs in ecological listening situations (e.g., *A-only* speech). Also, these N1 effects were more robust in stronger than weaker illusion-perceivers. The results support a mechanism by which the visual modality influences encoding of phonetic features at the auditory cortex, and that this influence commences at early stages of processing (van Wassenhove et al., 2005; Alsius and Munhall, 2013).

The robustness of AV integration is related to the speeding and suppression of N1-P2 AEPs (Besle et al., 2004; van Wassenhove et al., 2005; Roa Romero et al., 2015). The current data also showed this outcome. In a separate analysis (not presented in Results), we found that the N1-P2 latencies of the auditory /ba/ CVs occurred earlier during the *AV-congruent* condition versus the *A-only* condition ($p < 0.05$; paired *t* tests). Moreover, the P2 amplitudes for both /ba/ and /fa/ CVs were significantly smaller during the *AV-congruent* condition versus the *A-only* condition ($p < 0.05$). The dynamic reweighting model (Bhat et al., 2015), which builds on the predictive coding model (van Wassenhove et al., 2005), attributes the suppressive effect to a visually driven shift in neural processing from low- to high-level auditory networks. This shift causes inhibition of activity in the core and surrounding regions of the auditory cortex (hence, the reduced AEPs), and simultaneously excites networks along the STS/G and middle temporal gyrus that encode phonetic and linguistic features, allowing the visual system to engage these representations.

Although suppression of the N1-P2 may index a general process during AV integration, our results suggest an additional, more specific role for the N1 AEP. Consistent with previous accounts (Ghazanfar et al., 2005; Kayser et al., 2008, 2010) and the dynamic reweighting model, the N1 may also reflect visually driven changes to phonetic representations at the auditory cortex, via cross-modal inhibitory and excitatory mechanisms. For example, in the case of *illusion-ba*, the /b/ viseme excites representations of the /b/ phoneme while weakening representations of other phonemes, including the /f/ phoneme. However, the incoming acoustic signal /fa/ activates representations of the /f/ phoneme. The cumulative outcome of the two processes leads to an intermediate (ambiguous) neurophysiological state whereby perception can go either way: illusion (biased toward the visual cue) or illusion-failure (biased toward the acoustic cue). This neutral state is reflected in the N1 behavior, in which both the illusion and illusion-failure evoke an N1 amplitude that is intermediate to those of the congruent percepts (Fig. 5*A*). The RT data support this account, as subjects' RTs during *illusion-ba* and *illusion-failure-ba* were equally delayed relative to the *AV-congruent-ba* percept, suggesting that individuals struggled to identify the sounds as /ba/ or /fa/ regardless of whether they perceived the illusion or not. This intermediate state hypothesis applies to varying degrees depending on the viseme-phoneme pairs, with the *illusion-fa* being less ambiguous than the *illusion-ba*, for example. Indeed, for both the *illusion-fa* and *illusion-failure-fa* trials, the behavioral and N1 results more robustly reflected auditory perception: the N1 and RT shifted with the illusion, but not with the illusion-failure. Along these lines, we posit that, because phonetic representations overlap to varying degrees within the auditory cortex (Mesgarani et al., 2014), sometimes the intermediate state causes perception of a third phoneme (e.g., classical McGurk: visual /ga/, acoustic /ba/, heard /da/).

Evidence of visual priming of the auditory cortex

Prior AV and visual-only studies demonstrated that vision activates low- and high-level regions of the auditory cortex (Calvert et al., 1997; but Bernstein et al., 2002; Ghazanfar et al., 2005; Kayser et al., 2008, 2010; Okada et al., 2013). Ghazanfar et al. (2005) showed that species-specific face and voice integration takes place in the core and lateral belt of the auditory cortex, the same regions that give rise to the N1 AEP in humans (Scherg et al., 1989; Zouridakis et al., 1998). This influence also extends to higher-level networks. A recent fMRI study (Zhu and Beauchamp, 2017) showed that voxels in the posterior STS that respond to mouth movements also respond to speech sounds. This effect was not observed for voxels that respond to eye movements. More pertinent are the studies by Skipper et al. (2005) and Smith et al. (2013, discussed in the Introduction). Skipper et al. (2005) showed that activations of auditory networks during a successful McGurk illusion initially reflect representations of the acoustic stimulus, but with time transform to reflect the visually driven auditory representations. These accounts validate that vision influences phonetic representations at the auditory cortex, and this influence spans both low- and high-level auditory networks.

The role of multisensory networks

Previous imaging work has identified several brain regions that could act as hubs for fusion of AV percepts, including but not limited to the following: the posterior STS/G (Calvert et al., 2000; Beauchamp et al., 2004, 2010; Erickson et al., 2014), middle STS (Miller and D'Esposito, 2005; Venezia et al., 2017), middle temporal gyrus (Beauchamp et al., 2004), and superior parietal lobule (Molholm, 2006). However, the findings of these studies do not categorically link these networks with fusion of AV percepts per se. First, most aforementioned brain regions are within networks associated with phonological processing (Binder et al., 2000; Hickok and Poeppel, 2007; Hocking and Price, 2008; Mesgarani et al., 2014; Arsenault and Buchsbaum, 2015). These regions receive inputs from auditory and visual centers (Venezia et al., 2017; Zhu and Beauchamp, 2017). Hence, changes observed at these networks, and manifested in the N1 behavior, may also entail recruitment of phonetic representations along low- and high-level phonological networks to varying levels in response to incongruent and congruent stimuli. Second, the posterior STS has been shown to behave similarly in response to intermodal versus intramodal pairing of stimuli (Hocking and Price, 2008). This led some investigators to propose that activity within the posterior STS may reflect a comparison process to determine whether concurrent stimuli match one another (Hocking and Price, 2008). In short, there is tangible evidence that several high-level networks (e.g., STS/G and middle temporal gyrus) are engaged during AV integration. Their involvement may reflect a direct role (i.e., fusing of percepts) or a supportive top-down role, whereby following evaluation of a mismatch among incoming AV percepts (Hocking and Price, 2008), they communicate the outcome to low-level speech areas (Arnal et al., 2009; Blank and von Kriegstein, 2013; Bhat et al., 2015; Venezia et al., 2017) to regulate phonetic encoding. Alternatively, differences in activity

within these high-level networks may also reflect visually mediated changes in phonetic encoding.

Interindividual variability

We also observed individual differences for each AV stimulus pair, whereby stronger and weaker illusion-perceivers exhibited different patterns of N1 amplitudes. Interindividual variability in the frequency of illusory perception is well known in the McGurk effect literature (Gurler et al., 2015; Proverbio et al., 2016). One's age (Pearl et al., 2009; Setti et al., 2013), native language background (Hazan et al., 2010), and musical experience (Proverbio et al., 2016) can influence McGurk effect susceptibility, as can schizophrenia (White et al., 2014) and autism spectrum disorder (Stevenson et al., 2014). Differences in methodology, including the talker and segments used for the stimuli, also influence the illusion's strength (Hazan et al., 2010).

Prior research has found that interindividual differences of the McGurk illusion are associated with activity in the left STS/G; specifically, individuals who experienced the McGurk illusion more often exhibited a stronger STS/G response (Hall et al., 2005; Nath and Beauchamp, 2012). The current results add to the literature on the neurophysiological differences associated with susceptibility to the McGurk effect. As stated, subjects who more strongly experienced *illusion-ba* showed a significant N1 amplitude distinction between the congruent percepts of /ba/ and /fa/, relative to the illusory percept and relative to one another (Fig. 6*D*, left). The weaker /ba/ illusion-perceivers did not show this N1 distinction, not even between the congruent /ba/ and /fa/ percepts. This was also the tendency for illusory /fa/ (Fig. 6*B*). Based on this observation, a mechanism for this interindividual variability could be that, compared with weaker illusion-perceivers, strong illusion-perceivers have less overlap within the auditory cortex between the neural representations of the two phonemes. In other words, the strong illusion-perceivers have neural networks (representing the phonemes) that are more distinct and have more specialized properties, which allow the visual input to more strongly impact these networks and consequently impact auditory perception.

In conclusion, the current findings provide evidence that visual speech influences phonetic encoding at the auditory cortex. We assert that this is one mechanism in which visual input shapes auditory perception, likely with support from high-level multisensory networks, such as the posterior STS. Further research should examine the generalizability of these findings to other AV phoneme combinations.

References

Alsius A, Munhall KG (2013) Detection of audiovisual speech correspondences without visual awareness. Psychol Sci 24:423–431. CrossRef Medline

Alsius A, Möttönen R, Sams ME, Soto-Faraco S, Tiippana K (2014) Effect of attentional load on audiovisual speech perception: evidence from ERPs. Front Psychol 5:727. CrossRef Medline

Arnal LH, Morillon B, Kell CA, Giraud AL (2009) Dual neural routing of visual facilitation in speech processing. J Neurosci 29:13445–13453. CrossRef Medline

Arsenault JS, Buchsbaum BR (2015) Distributed neural representations of phonological features during speech perception. J Neurosci 35:634–642. CrossRef Medline

Baart M, Samuel AG (2015) Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. J Mem Lang 85:42–59. CrossRef

Beauchamp MS, Lee KE, Argall BD, Martin A (2004) Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41:809–823. CrossRef Medline

Beauchamp MS, Nath AR, Pasalar S (2010) fMRI-guided TMS reveals that the STS is a cortical locus of the McGurk effect. J Neurosci 30:2414–2417. CrossRef Medline

Benoit MM, Raij T, Lin FH, Jääskeläinen IP, Stufflebeam S (2010) Primary and multisensory cortical activity is correlated with audiovisual percepts. Hum Brain Mapp 31:526–538. CrossRef Medline

Bernstein LE, Auer ET Jr, Moore JK, Ponton CW, Don M, Singh M (2002) Visual speech perception without primary auditory cortex activation. Neuroreport 13:311–315. CrossRef Medline

Besle J, Fort A, Delpuech C, Giard MH (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. Eur J Neurosci 20:2225–2234. CrossRef Medline

Besle J, Fischer C, Bidet-Caulet A, Lecaignard F, Bertrand O, Giard MH (2008) Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. J Neurosci 28:14301–14310. CrossRef Medline

Bhat J, Miller LM, Pitt MA, Shahin AJ (2015) Putative mechanisms mediating tolerance for audiovisual stimulus onset asynchrony. J Neurophysiol 113:1437–1450. CrossRef Medline

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10:512–528. CrossRef Medline

Blank H, von Kriegstein K (2013) Mechanisms of enhancing visual-speech recognition by prior auditory information. Neuroimage 65:109–118. CrossRef Medline

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997) Activation of auditory cortex during silent lipreading. Science 276:593–596. CrossRef Medline

Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 10:649–657. CrossRef Medline

Carpenter AL, Shahin AJ (2013) Development of the N1–P2 auditory evoked response to amplitude rise time and rate of formant transition of speech sounds. Neurosci Lett 544:56–61. CrossRef Medline

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21. CrossRef Medline

Erickson LC, Zielinski BA, Zielinski JE, Liu G, Turkeltaub PE, Leaver AM, Rauschecker JP (2014) Distinct cortical locations for integration of audiovisual speech and the McGurk effect. Front Psychol 5:534. CrossRef Medline

Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J Neurosci 25:5004–5012. CrossRef Medline

Gurler D, Doyle N, Walker E, Magnotti J, Beauchamp M (2015) A link between individual differences in multisensory speech perception and eye movements. Atten Percept Psychophys 77:1333–1341. CrossRef Medline

Hall DA, Fussell C, Summerfield AQ (2005) Reading fluent speech from talking faces: typical brain networks and individual differences. J Cogn Neurosci 17:939–953. CrossRef Medline

Hazan V, Kim J, Chen Y (2010) Audiovisual perception in adverse conditions: language, speaker and listener effects. Speech Commun 52:996–1009. CrossRef

Herrmann B, Schlichting N, Obleser J (2014) Dynamic range adaptation to spectral stimulus statistics in human auditory cortex. J Neurosci 34:327–331. CrossRef Medline

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402. CrossRef Medline

Hocking J, Price CJ (2008) The role of the posterior superior temporal sulcus in audiovisual processing. Cereb Cortex 18:2439–2449. CrossRef Medline

Jones SJ, Longe O, Vaz Pato M (1998) Auditory evoked potentials to abrupt pitch and timbre change of complex tones: electrophysiological evidence of "streaming"? Electroencephalogr Clin Neurophysiol 108:131–142. CrossRef Medline

Kayser C, Petkov CI, Logothetis NK (2008) Visual modulation of neurons in auditory cortex. Cereb Cortex 18:1560–1574. CrossRef Medline

Kayser C, Logothetis NK, Panzeri S (2010) Visual enhancement of the information representation in auditory cortex. Curr Biol 20:19–24. CrossRef Medline

Lopez-Calderon J, Luck SJ (2014) ERPLAB: an open-source toolbox for the

analysis of event-related potentials. Front Hum Neurosci 8:213. CrossRef Medline
MacDonald J, McGurk H (1978) Visual influences on speech perception processes. Percept Psychophys 24:253–257. CrossRef Medline
Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164:177–190. CrossRef Medline
McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:691–811. CrossRef Medline
Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343:1006–1010. CrossRef Medline
Miller LM, D'Esposito M (2005) Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. J Neurosci 25:5884–5893. CrossRef Medline
Molholm S, Sehatpour P, Mehta AD, Shpaner M, Gomez-Ramirez M, Ortigue S, Dyke JP, Schwartz TH, Foxe JJ (2006) Audio-visual multisensory integration in superior parietal lobule revealed by human intracranial recordings. J Neurophysiol 96:721–729. CrossRef Medline
Näätänen R, Picton T (1987) The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology 24:375–425. CrossRef Medline
Nath AR, Beauchamp MS (2012) A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. Neuroimage 59:781–787. CrossRef Medline
Okada K, Venezia JH, Matchin W, Saberi K, Hickok G (2013) An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. PLoS One 8:1–8. CrossRef Medline
Oostenveld R, Fries P, Maris E, Schoffelen J (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011:1–8. CrossRef Medline
Ostroff JM, Martin BA, Boothroyd A (1998) Cortical evoked response to acoustic change within a syllable. Ear Hear 19:290–297. CrossRef Medline
Pearl D, Yodashkin-Porat D, Katz N, Valevski A, Aizenberg D, Sigler M, Weizman A, Kikinzon L (2009) Differences in audiovisual integration, as measured by McGurk phenomenon, among adult and adolescent patients with schizophrenia and age-matched healthy control groups. Compr Psychiatry 50:186–192. CrossRef Medline
Pekkola J, Ojanen V, Autti T, Jääskeläinen IP, Möttönen R, Tarkiainen A, Sams M (2005) Primary auditory cortex activation by visual speech: an fMRI study at 3 T. Neuroreport 16:125–128. CrossRef Medline
Pereira O, Toscano J (2016) The N1 event-related potential component as an index of speech sound encoding for multiple phonetic contrasts. J Acoust Soc Am 140:3217. CrossRef
Pilling M (2009) Auditory event-related potentials (ERPs) in audiovisual speech perception. J Speech Lang Hear Res 52:1073–1081. CrossRef Medline
Proverbio AM, Massetti G, Rizzi E, Zani A (2016) Skilled musicians are not subject to the McGurk effect Skilled musicians are not subject to the McGurk effect. Sci Rep 6:30423. CrossRef Medline
Roa Romero Y, Senkowski D, Keil J (2015) Early and late beta band power reflects audiovisual perception in the McGurk illusion. J Neurophysiol 113:2342–2350. CrossRef Medline
Saint-Amour D, De Sanctis P, Molholm S, Ritter W, Foxe JJ (2007) Seeing voices: high-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. Neuropsychologia 45:587–597. CrossRef Medline
Sams M, Aulanko R, Hämäläinen M, Hari R, Lounasmaa OV, Lu ST, Simola J (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 127:141–145. CrossRef Medline
Scherg M, Vajsar J, Picton TW (1989) A source analysis of the late human auditory evoked potentials. J Cogn Neurosci 1:336–355. CrossRef Medline
Schroeder CE, Foxe J (2005) Multisensory contributions to low-level, "unisensory" processing. Curr Opin Neurobiol 15:454–458. CrossRef Medline
Setti A, Burke KE, Kenny R, Newell FN, Gunter TC, Plank M, Sommers M (2013) Susceptibility to a multisensory speech illusion in older persons is driven by perceptual processes. Front Psychol 4:575. Medline
Shahin AJ, Kerlin JR, Bhat J, Miller LM (2012) Neural restoration of degraded audiovisual speech. Neuroimage 60:530–538. CrossRef Medline
Skipper JI, Nusbaum HC, Small SL (2005) Listening to talking faces: motor cortical activation during speech perception. Neuroimage 25:76–89. CrossRef Medline
Smith E, Duede S, Hanrahan S, Davis T, House P, Greger B (2013) Seeing is believing: neural representations of visual stimuli in human auditory cortex correlate with illusory auditory perceptions. PLoS One 8:e73148. CrossRef Medline
Stekelenburg JJ, Vroomen J (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. J Cogn Neurosci 19:1964–1973. CrossRef Medline
Stevenson RA, Segers M, Ferber S, Barense MD, Wallace MT (2014) The impact of multisensory integration deficits on speech perception in children with autism spectrum disorders. Front Psychol 5:379. CrossRef Medline
Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215. CrossRef
Toscano JC, McMurray B, Dennhardt J, Luck SJ (2010) Continuous perception and graded categorization: electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. Psychol Sci 21:1532–1540. CrossRef Medline
van Wassenhove V, Grant KW, Poeppel D, Halle M (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A 102:1181–1186. Medline
Venezia JH, Vaden KI Jr, Rong F, Maddox D, Saberi K, Hickok G (2017) Auditory, visual and audiovisual speech processing streams in superior temporal sulcus. Front Hum Neurosci 11:174. CrossRef Medline
White TP, Wigton RL, Joyce DW, Bobin T, Ferragamo C, Wasim N, Lisk S, Shergill SS (2014) Eluding the illusion? Schizophrenia, dopamine and the McGurk effect. Front Hum Neurosci 8:565. CrossRef Medline
Zhu LL, Beauchamp MS (2017) Mouth and voice: a relationship between visual and auditory preference in the human superior temporal sulcus. J Neurosci 37:2697–2708. CrossRef Medline
Zouridakis G, Simos PG, Papanicolaou AC (1998) Multiple bilaterally asymmetric cortical sources account for the auditory N1m component. Brain Topogr 10:183–189. CrossRef Medline