Published in final edited form as:

Psychol Sci. 2014 September ; 25(9): 1835–1837. doi:10.1177/0956797614542132.

Lip reading without awareness

John Plass¹, Emmanuel Guzman-Martinez¹, Laura Ortega¹, Marcia Grabowecky^{1,2}, and Satoru Suzuki^{1,2}

¹Department of Psychology.. Northwestern University, Evanston, Illinois

²Interdepartmental Neuroscience Program. Northwestern University, Evanston, Illinois

Speech perception is vitally important, thus the visual system may possess exceptional sensitivity to track speech-related signals conveyed by lip movements even without awareness of the speaking face. We tested this possibility using a technique called continuous flash suppression (CFS) (e.g., Tsuchiya & Koch, 2005), in which a critical stimulus, here a speaking face, was presented to one eye and strong dynamic-masking noise was presented to the other eye (Figure 1A), rendering the speaking face invisible. We determined whether the visual system could still encode spatiotemporal patterns of lip movements when observers were aware of only the randomly flashing masking display.

We measured the encoding of invisible lip movements as crossmodal facilitation of spoken word categorization (e.g., Sumby & Pollack, 1954). Participants determined whether each spoken word was a target word (a tool name) or a non-target word (a name of a non-tool object) while they concurrently viewed a face that either spoke the same word—the congruent condition—or a different word—the incongruent condition.

Prior research suggests that spatial attention influences unaware as well as aware visual processing (e.g., Cohen et al., 2012), and that attention to the mouth region is necessary for lip movements to facilitate spoken word perception (e.g., Alsius et al., 2005; Driver & Spence, 1994). To help direct attention to the mouth region, on half of the trials, we presented the face without the dynamic mask (Figure 1B), and we instructed participants to localize a small probe briefly presented near the mouth (Supplementary Figure S1). These “attention-enforcement” trials were randomly intermixed with the critical masked-face (face invisible) trials. To further enforce attention to the mouth region on the masked-face trials, the probe also appeared on the masked face, and participants were instructed to report its location whenever the face became visible through the mask. Participants reported seeing the face on 5% of the masked-face trials, and the data from those trials were removed from the analyses. If the visual system automatically extracts lip movements even when they are invisible, spoken-word categorization should be facilitated by congruent lip movements even when the face is invisible on the masked-face trials.

Send correspondence to: John Plass or Satoru Suzuki, Northwestern University, ¹Department of Psychology and ²Interdepartmental Neuroscience program, 2029 Sheridan Rd., Evanston, IL 60208, Phone: (847) 467-1271, Fax: (847) 491-7859, johnnclass@gmail.com or satoru@northwestern.edu.

Indeed, responses to the spoken target words on the masked-face trials were significantly faster when the lip movements were congruent than incongruent, $t_{45}=2.66$, $p<0.01$ (Figure 1C). The accuracy of target-word classification was 92% with no evidence of a speed-accuracy trade-off. This congruency effect was absent, $t(45)=1.15$, $n.s.$, on the attention-enforcement trials (Figure 1E), likely because having to localize a briefly flashed probe interfered with the processing of lip movements. The congruency effect was also absent for responses to the spoken non-target words (1590ms [congruent] vs. 1589ms [incongruent], $t_{45}=0.13$, $n.s.$, on the masked-face trials, and 1647ms [congruent] vs. 1651ms [incongruent], $t_{45}=0.07$, $n.s.$ on the attention-enforcement trials). This lack of congruency effects on non-target trials may reflect that processes beyond word identification are necessary to decide that the target information is absent; consistent with this possibility, responses to non-targets were significantly slower than those to targets in this and the following experiments ($t_s>3.50$, $p_s<0.005$).

To control for the possibility that participants might have failed to report face visibility on some of the masked-face trials, we performed a second experiment incorporating a more stringent indicator of face visibility. A tinted translucent ellipse was placed over the mouth region of the face. On each trial, after responding to the spoken word, participants were asked to report the color of the ellipse (red, blue, green, or yellow); critically, they were required to guess if they thought that they had not seen the face. All masked-face trials on which participants correctly reported the color (30%) were removed from analysis.

The same pattern of results was obtained. On the masked-face (face invisible) trials, responses to the spoken target words were significantly faster when the lip movements were congruent than incongruent, $t_{23}=2.12$, $p<0.05$ (Figure 1D), with a mean accuracy of 92% with no evidence of a speed-accuracy trade-off. Also consistent with the original experiment, there was no congruency effect for target responses on the attention-enforcement trials, $t_{23}=0.80$, $n.s.$ (Figure 1F) or for non-target responses (1777ms [congruent] vs. 1743ms [incongruent], $t_{23}=1.06$, $n.s.$ on the masked-face trials, and 1708ms [congruent] vs. 1779ms [incongruent], $t_{23}=1.97$, $n.s.$ on the attention-enforcement trials.).

These results demonstrate that even when a speaking face is rendered invisible by a dynamic mask with strong motion signals, the visual system accurately encodes invisible lip movements to facilitate auditory perception of the corresponding spoken words. This crossmodal effect is likely to occur at the level of encoding words; it has been shown that invisible lip movements do not generate a McGurk effect (Palmer & Ramsey, 2012), suggesting that invisible lip movements do not influence auditory perception at the level of encoding syllables.

Dorsal motion processing mechanisms (e.g., V3a, V5) would have predominantly responded to the strong and visible flashing mask (e.g., Moutoussis et al., 2005). The invisible lip movements would thus likely have been processed through the ventral visual pathway including the superior temporal sulcus (STS), an area that selectively responds to biological motion and movements of facial features (e.g., Allison et al., 2000; Calvert & Campbell, 2003; Grossman et al., 2000), and facilitated spoken word perception via multimodal portions of the STS (e.g., Calvert et al., 2000). Sophisticated unconscious processing of

static images (e.g., words, faces, sex of human body, and contextual congruence; Jiang et al., 2006; Jiang et al., 2007; Mudrik et al., 2011; Yang et al., 2007) has been demonstrated. Our results extend these prior findings to the processing of dynamic information. Static information can theoretically be extricated from a dynamic mask by temporal averaging. However, unconscious extrication of the subtle dynamics of lip movements from the overwhelming random dynamics of the mask requires sophisticated tuning of the ventral visual system to the behaviorally relevant dynamics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The first and second authors made equal contributions. This study was supported by NIH grants R01EY021184 & R01EY018197 and NSF grant BCS 0643191.

References

- Allison T, Puce A, McCarthy G. Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences*. 2000; 4:267–278. [PubMed: 10859571]
- Alsius A, Navarra J, R Campbell R, Soto-Faraco S. Audiovisual integration of speech falters under high attention demands. *Current Biology*. 2005; 15:839–843. [PubMed: 15886102]
- Calvert GA, Campbell R. Reading speech from still and moving faces: the neural substrates of visible speech. *Journal of Cognitive Neuroscience*. 2003; 15:57–70. [PubMed: 12590843]
- Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*. 2000; 10:649–657. [PubMed: 10837246]
- Cohen MA, Cavanagh P, Chun MM, Nakayama K. The attentional requirements of consciousness. *Trends in Cognitive Sciences*. 2012; 16:411–417. [PubMed: 22795561]
- Driver, J.; Spence, CJ. Spatial synergies between auditory and visual attention. In: Umiltà, C.; Moscovitch, M., editors. *Attention and Performance XV: Conscious and Nonconscious Information Processing*. MIT Press; Cambridge, MA: 1994. p. 311–331.
- Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, Blake R. Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience*. 2000; 12:711–720. [PubMed: 11054914]
- Jiang Y, Costello P, Fang F, Huang M, He S. A gender- and sexual orientation-dependent spatial attentional effect of invisible images. *Proceedings of the National Academy of Sciences, USA*. 2006; 103:17048–17052.
- Jiang Y, Costello P, He S. Processing of invisible stimuli: Advantage of upright faces and recognizable words in overcoming interocular suppression. *Psychological Science*. 2007; 18:349–355. [PubMed: 17470261]
- Moutoussis K, Keliris G, Kourtzi Z, Logothetis N. A binocular rivalry study of motion perception in the human brain. *Vision Research*. 2005; 45:2231–2243. [PubMed: 15924938]
- Mudrik L, Breska A, Lamy D, Deouell LY. Integration without awareness: Expanding the limits of unconscious processing. *Psychological Science*. 2011; 22:764–770. [PubMed: 21555524]
- Palmer TD, Ramsey AK. The function of unconsciousness in multisensory integration. *Cognition*. 2012; 125:353–364. [PubMed: 22989871]
- Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*. 1954; 26:212–215.
- Tsuchiya N, Koch C. Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*. 2005; 8:1096–110.

Yang E, Zald D, Blake R. Fearful expressions gain preferential access to awareness during continuous flash suppression. *Emotion*. 2007; 7:882–886. [PubMed: 18039058]

A. Trial sequence of a masked-face (face invisible) trial. Each trial began with a fixation cross ($0.36^\circ \times 0.36^\circ$) lasting 3000 ms. While the face was presented to one eye, a strong dynamic mask, consisting of a random array of contrast-enhanced features taken from the face and updated at 10 Hz, was simultaneously presented to the other eye. The face was faded in over a 1000 ms period to reduce the likelihood that it would break through the mask to become visible. The face articulated either the spoken word—congruent condition—or a different word—incongruent condition. Participants indicated whether the spoken word was a target word or a non-target word as quickly and as accurately as possible by pressing the designated button. A secondary task was used to enforce participants' attention to the mouth region. In Experiment 1, a small circular probe (0.1° -diameter) was presented in one of six possible locations in the mouth region, and after each trial participants reported whether the face became visible and if so also reported the location of the probe (see Supplementary Figure SI). In Experiment 2, a tinted translucent ellipse (red, blue, green, or yellow) was placed over the mouth region, and after each trial participants reported the color (guessing if necessary). **B.** Trial sequence of an attention-enforcement trial. These trials were identical to the masked-face trials except that the face was fully visible (presented to both eyes with no dynamic mask) and that the attention-enforcing probe (Experiment 1) was briefly (200 ms) and unpredictably presented on each trial. **C & D.** Average response times to the spoken target words on the masked-face (face invisible) trials in the congruent and incongruent conditions for Experiments 1 (C) and 2 (D). **E & F.** Average response times to the spoken target words on the attention-enforcement trials in the congruent and incongruent conditions for Experiments 1 (E) and 2 (F). The error bars represent ± 1 SEM adjusted for within-participants comparisons. * $p < 0.05$, ** $p < 0.01$.