PERCEPTUAL EFFECTS OF VISUAL EVIDENCE OF THE AIRSTREAM

Connor Mayer[1], Bryan Gick[1,2], Tamra Weigel[1], and Douglas H. Whalen[2]
[1]Dept. of Linguistics, University of British Columbia, 2613 West Mall, British Columbia, Canada, V6T 1Z4
[2]Haskins Laboratories, 300 George Street, Suite 900, New Haven, Connecticut, USA, 06511

1. INTRODUCTION

It has been long understood that perceivers of speech integrate visual and auditory information from articulator movements, resulting in both interference (e.g., McGurk and MacDonald 1976) and enhancement (e.g., Sumby and Pollack 1954) of auditory perception. Few studies have investigated integration of other types of information in speech perception. Gick and Derrick (2009) found that during auditory speech perception, perceivers integrated tactile information in the form of light air puffs. These puffs, delivered cutaneously on the hand or neck, were designed to resemble speech aspiration. When puffs were present, aspirated stops were more often correctly identified as being aspirated, and unaspirated stops were more often misidentified as aspirated, showing that listeners integrate tactile information in auditory perception in much the same way as visual information.

The goal of the present study was to examine whether speech aspiration can influence perception when it is not felt on the skin, but is rather recoverable indirectly from the visual signal. We predict that when visual evidence of aspiration is present, it will be integrated in speech perception in much the same way as first hand tactile information is. An additional question that arises is whether perceivers automatically make use of this kind of ambient information, or whether they need to be consciously attentive to it.

2. METHODS

Participants were seated in a sound proof room and shown short video clips of a speaker producing the sequence "pom" and "bomb" in a noisy bar setting with multi-talker babble. The babble was set to such a volume that correct auditory-only identification of the sounds was about 70%. There were a total of nine conditions in the experiment: conditions 1 and 2 had a candle placed in front of the speaker: in 1, the speaker said "pom", visibly perturbing the candle by the aspiration of the /p/, while in 2 the speaker said "bomb", and the candle was not perturbed because of the lack of aspiration of /b/. Conditions 3 and 4 were identical to 1 and 2 except that the candle was placed to the side of the speaker, and thus was not perturbed. Conditions 5 and 6 used the same video as conditions 1 and 2, but with mismatched audio: in condition 5, perceivers saw a video "pom" accompanied by an auditory "bomb", while in condition 6 they saw the opposite. Conditions 7 and 8 used the video from conditions 3 and 4, but with ambiguous audio between "pom" and "bomb" created by morphing audio of the two words from conditions 3 and 4 using the program STRAIGHT (Kawahara 2003). Because morphing resulted in half the original sound files, both the "pom" videos and "bomb" videos in this condition used the same audio. This condition was intended to factor out the possibility of facial cues disambiguating the sounds. Condition 9 featured the candle to the side as in conditions 3 and 4, but with perturbation of the candle flame occurring at times not corresponding to the effects of the airstream. This condition was designed primarily for training purposes: perceivers were shown 10 tokens of it at the beginning of the experiment to downplay the significance of the flickering candle. Aside from condition nine, all conditions had 20 repetitions, resulting in a total of 170 tokens. Subjects were given a forced-choice task to identify whether they heard "pom" or "bomb" in each video clip by pressing the left and right arrows on a keyboard. Half the subjects pushed left for "pom", the other half pushed right. Stimuli were presented and input recorded using Psyscope B53 on an iMac. When the experiment was completed, subjects were asked what they had observed, whether they had been consciously aware of the candle flickering and whether they had used it in any conscious strategy to disambiguate the sounds. Subjects' data was separated into those who reported being consciously aware of the candle as offering perceptual cues and those who were not. A total of 19 native English speakers participated: 6 reported being aware of the candle while 13 did not. No subjects had any training in linguistics. All statistical analysis was done using R 2.9.1. Repeated measures analyses of variance were conducted with three audio conditions (aspirated, unaspirated, and ambiguous) and four video conditions ("pom" with candle, "pom" without candle, "bomb" with candle, and "bomb" without candle) for both groups of subjects with post-hoc Tukey's Honestly Significant Difference tests. Data from the training condition was not included in the analysis.

3. RESULTS

Both groups of subjects displayed significant differences in responses between "pom" with and without candle and "bomb" with and without candle, indicating that they were generally able to perceive a difference between aspirated and unaspirated stops. Overall results for subjects who reported not being aware of the candle showed significant main effects only for audio [$F(2, 82) = 8.0845$; $p < 0.001$]. There was no significant effect for video [$F(3,82) = 0.7703$, $p = 0.51$], nor any interaction between audio and video [$F(2,82) = 0.0203$, $p = 0.89$]. Post-hoc tests showed that tokens with an auditory "pom" were more likely to be

identified as such. There were no significant differences between any video factors. For interactions, there was no significant difference in response between "pom" with and without candle (conditions 1 and 3; $p = 1$), nor between "bomb" with and without candle (conditions 2 and 4; $p = 1$). Conditions with mismatched audio and video (conditions 4 and 5), both resulted in lower percentages of correct answers than in conditions with natural audio, but these differences were not significant. Most interestingly, condition 5, auditory "bomb" with an accompanying candle flicker, did not show significant differences from conditions with natural audio "bomb" with candle (condition 2; $p = 0.10$) and without candle (condition 4; $p = 0.17$). There were also no significant differences between conditions where videos for "pom" and "bomb" were paired with identical ambiguous audio (conditions 7 and 8; $p = 1$).

Overall results for subjects who reported being consciously aware of the candle flickering showed a significant main effect of video [$F(3, 36) = 27.0888$; $p < 0.0001$]. There was no significant effect for audio [$F(2, 36) = 0.6778$, $p = 0.68$] nor any interaction between audio and video [$F(2, 36) = 1.3413$, $p = 0.25$]. Post-hoc tests showed that tokens with a candle flicker were more likely to be identified as "pom". However, there was no significant difference between video "pom" with candle to the side and video "pom" with candle in front ($p = 0.99$). As well, there were significant differences between audio "pom" and "bomb", with "pom" being more likely to be identified as such ($p < 0.05$): the lack of a main effect likely comes from the lack of a difference between the ambiguous and "pom" audio ($p = 0.7$). For interactions, there was no significant difference between "pom" with candle and "pom" without candle (conditions 1 and 3; $p = 0.85$). There was also no significant difference in responses between "bomb" with and without candle (conditions 2 and 4; $p = 0.99$). "Bomb" audio with "pom" video (condition 5) did show significant differences with "bomb" with candle (condition 2; $p < 0.05$) but not with "bomb" without candle (condition 4; $p = 0.28$). "Pom" audio with "bomb" video (condition 6) did not show a significant difference for "pom" with an accompanying candle flicker (condition 1; $p = 0.14$), nor did it show a difference with "pom" without candle (condition 2; $p = 0.97$), although in both cases the tokens without a flicker were more often identified as "bomb" . Subjects who noticed the candle did not show any difference in responses between conditions where videos for "pom" and "bomb" were paired with ambiguous audio (conditions 7 and 8; $p = 0.99$).

4. DISCUSSION

Depending on whether subjects reported being consciously aware of it, the presence or absence of the candle flickering had different effects on their responses. The subjects who were not consciously aware of the candle did not show evidence of any integration or interference effects: an accompanying perturbation of the candle had no bearing on correct identifications of "pom", and a flicker accompanying auditory "bomb" did not produce interference effects. Rather, these subjects appeared to make their choices based solely on the audio.

Subjects who were aware of the candle, however, clearly showed effects of it on their responses. When the flame was perturbed, regardless of the accompanying audio, "pom" responses increased. Significant differences in responses between 'pom' and 'bomb' audio, however, suggest that audio did play a role in disambiguation for this group, mainly in cases where the candle was absent. Neither group showed a difference between visual "pom" and "bomb" coupled with identical ambiguous audio, suggesting that subjects were not able to use facial cues in differentiation.

Although both direct and indirect consequences of articulation, whether auditory, visual, or tactile, clearly influence perception, these results indicate that for certain types of information it is not enough that they are merely present: to be used in perception they require listeners' active attention. It is difficult to say where the boundary lies between information that can be unconsciously integrated and information that cannot. Tactile stimuli as in Gick and Derrick (2009) are a relatively indirect consequence of speech articulation and one with which speakers presumably have less experience (eg. puffs on the back of the neck), yet these can be unconsciously integrated. When these stimuli are present only in the visual modality, however, they can only influence perception when perceivers are consciously aware of them: this suggests that there is a wide range of ambient information that can be used in speech perception, but not all of it can be unconsciously integrated.

REFERENCES

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. Nature, 462, 502-504.

Kawahara, H., & Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. ICASSP, V.1, Hong Kong. 256–259.

McGurk, H., & MacDonald, J.W. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212-215.

ACKNOWLEDGEMENTS

Thanks to Donald Derrick, Mark Scott, Molly Babel, and the members of the UBC ISRL. This project was funded by an NSERC Discovery Grant to the second author and by NIH Grant DC-02717 to Haskins Laboratories.