Feature-based audiovisual speech integration of multiple streams

JUAN CAMILO GIL-CARVAJAL[1,2,*], JEAN-LUC SCHWARTZ[3], TORSTEN DAU[2] AND TOBIAS SØREN ANDERSEN[1]

[1] *Cognitive Systems, DTU Compute, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Hearing Systems, DTU Health Tech, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[3] *GIPSA-lab, Univ, Grenoble Alpes, CNRS, Grenoble INP*

Speech perception often involves the integration of auditory and visual information. This is shown in the McGurk effect, in which a visual utterance, e.g., /ipi/, dubbed onto an acoustic utterance, e.g., /iki/, produces a combination percept, e.g., /ipki/. However, it is still unclear how phonetic features are integrated audiovisually. Here, we studied audiovisual speech perception by decomposing the auditory component of McGurk combinations into two streams. We show that auditory /i_i/, where the underscore indicates an intersyllabic silence, dubbed onto visual /ipi/ produce a strong illusion of hearing /ipi/. We also show that adding an acoustic release burst to /i_i/ creates a percept of /iki/. An auditory continuum was created with stepwise temporal alignments of the release burst and /i_i/. When dubbed onto /ipi/, this continuum was perceived mostly as a visually driven response /ipi/ when the burst overlapped with either acoustic vowel. Other temporal alignments frequently produced combination responses. Mostly /ikpi/ combinations were obtained when the burst was closer to the initial vowel, and reverse /ipki/ responses when it was closer to the final vowel. These results are indicative of feature-based audiovisual integration where burst and aspiration are sufficient cues for the consonant /k/, while the perception of /p/ depends on place information in the visual stream.

INTRODUCTION

The visible facial gestures accompanying the voice of the talker in face-to-face conversations facilitate speech perception (Sumby and Pollack, 1954). This is particularly advantageous in noisy listening situations (Binnie *et al.*, 1974). However, it is still unclear how phonetic information is integrated. The McGurk effect demonstrates phonetic integration for speech comprehension (McGurk and McDonald, 1976). Here we study the McGurk *combinations*, in which the audiovisual pairing of an auditory non-labial consonant (e.g., /iki/) and a visual labial consonant (e.g., /ipi/) leads to a cluster percept in which both consonants are represented (e.g., /ipki/ or /ikpi/). A typical finding in many studies that have reported the perceived consonant order of the combination response is that the labial consonant leads the non-

[*] Corresponding author: juac@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. © The Authors. ISSN: 2596-5522.

labial (e.g., Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009), although not always (Hampson *et al.*, 2003). It has also been suggested that the combination responses occurred more frequently with unvoiced consonants (Colin *et al.*, 2002). This could be due to the strong consonantal burst and aspiration that have been shown to increase the frequency of the combination responses (Green and Norrix, 1997). However, the role of these acoustic features on the perceived consonant order has remained unclear.

A few studies (e.g., Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009) varied the timing in combination illusions by altering the audiovisual stimulus onset asynchrony (SOA). This approach could be suboptimal for studying the perceived consonant order, since the cross-modal asynchrony could influence the strength of the integration in addition to the perceived consonant order. To minimize the effect of asynchrony on audiovisual integration, we chose to vary only the timing of the consonants such that the vowels were still synchronous across modalities. To do so, we isolated the burst and aspiration from the vowels of the auditory component. An auditory continuum was then created by pairing the vowels and the burst at nine SOAs. To test the effect of varying the timing of the visual articulatory gestures, we paired the auditory continuum with two visual contexts. This resulted in two audiovisual continua in which the visual consonant was pronounced either in the offset of the initial syllable or the onset of the final syllable in a vowel-consonant-vowel (VCV) context. Our hypothesis was that varying the timing of the consonantal burst and aspiration could systematically change the perceived consonant order of the combination response. If so, the combination responses could provide information about the temporal organization of the stimulus features integrated across modalities.

METHODS

Participants

The test subjects were 14 native French speakers (mean age 25, five female). All subjects reported to have normal hearing and normal or corrected-to-normal vision. Before the testing, all participants provided written consent and all experiments were approved by the Comité d'Ethique pour les Recherches Non Interventionnelles (CERNI), reference number IRB00010290.

Speech material

The speech material recorded consisted of the disyllables /i_i/, /ip_i/, /i_pi/, /i_ki/, /ikpi/, and /ipki/ articulated by a native French speaker. The underscore represents an intersyllabic silence, which corresponded to 300, 114, and 189 ms for /i_i/, /ip_i/ and /i_pi/, respectively. The speaker pronounced each syllable synchronously with two consecutive beats of a metronome, which was delivered through EarPods and set at 130 beats per minute. The idea was to provide an auditory reference that could enable the production of speech utterances with similar speaking rate and duration.

To create stimuli in which only the timing of the burst and aspiration was affected, we used two separate auditory streams. One stream contained only the vowels and consisted of the recorded sound of /i_i/. The second stream contained only the consonantal burst and aspiration with a duration of 100 ms, which were extracted from the recorded articulation of /i_ki/. An auditory continuum was then generated offline by pairing the two streams at nine different SOAs with a step size of 50 ms. At one end (–200 ms), the waveform of the initial vowel fully overlapped with the burst, and at the other end (200 ms), the final vowel fully overlapped with the burst. At 0 ms (arbitrarily defined), the burst was in the middle of the two vowels. Two audiovisual continua were also created by pairing the auditory continuum with the visual articulatory gestures for /ip_i/ (consonant offset) or /i_pi/ (consonant onset).

In addition to the continua, we tested the recorded cluster articulations /ikpi/ and /ipki/, and the articulation of the vowels /i_i/, which were presented to the subjects in the auditory, visual and congruent audiovisual conditions. The articulations /ip_i/ and /i_pi/ were only tested in the visual and the congruent audiovisual conditions. In total, 40 stimuli were tested in the experiment. The audio of the speech material had a sampling rate of 48 kHz, and a resolution of 24 bits. The video had a resolution of 720 x 576 and frame rate of 25 Hz. The total duration of each stimulus was one second. All visual articulations started and ended with a neutral expression of the speaker with the mouth closed, which lasted at least two video frames.

Experiments

The experiments were conducted in a sound-proof booth. The subjects were seated 80 cm in front of a 21.5-inch Dell monitor, from which the videos were reproduced. The sound was played back monaurally at 65 dBA from a loudspeaker positioned above the computer monitor. Prior to the experiment, the subjects received written instructions in French and performed a training session that lasted one trial. The experiment was conducted in two separate blocks of ten trials each. Within a trial, all stimuli were presented in random order. The subjects were asked to report what they heard in each trial, or what they saw in the case of the visual trials. The response options were labelled on a computer keyboard and corresponded to /k/, /p/, /kp/, /pk/, and "no consonant" (n.c.). The subjects took a five-minute break between the experimental blocks. The total duration of the experiment session was 60 minutes.

Data analysis

Data analysis of responses was carried out in three steps: (1) Assessment of the auditory consonant identification by contrasting all percepts containing /k/ with pure /p/ percepts; (2) assessment of the visual influence by contrasting all percepts containing /p/ with "no consonant" or pure /k/ percepts; and (3) assessment of the perceived consonant order by defining the PK-index. The index was estimated as the proportion of /pk/ responses out of the total combination responses, and hence, represents the conditional probability of obtaining a /pk/ response given the occurrence of a combination response. For all analyses, the effects were evaluated using linear-mixed model ANOVAs with subject as a random factor and SOA and stimulus type as fixed factors. Post-hoc multiple comparisons were performed with Tukey's HSD test. Analyses involved the mean proportions of percepts with /k/ (step

1), the mean proportion of percepts with /p/ (step 2) or the mean PK-index (step 3). A significance level of 0.05 was considered in all analyses.

RESULTS

Assessment of the auditory consonant identification

To study the effect of SOA and articulatory gestures on the perception of the auditory stream, we analyzed the responses containing /k/, which indicate that the burst was indeed perceived as a consonant. Figure 1 shows the response percentages obtained for the auditory and audiovisual continua with the two visual contexts as a function of SOA. A linear-mixed model ANOVA with fixed factors for visual context (three levels: no visual, /ip_i/, and /i_pi/) and SOA, (nine levels) and subject as a random factor was performed on the sum of /k/, /kp/ and /pk/ response percentages. The outcome of the test revealed a significant main effect of SOA [$F(8,338) = 35.62$, $p < 0.0001$] and visual context [$F(2,338) = 58.10$, $p < 0.0001$] on the perception of /k/, whereas their interaction was not significant [$F(16,338) = 1.25$, $p = 0.228$].

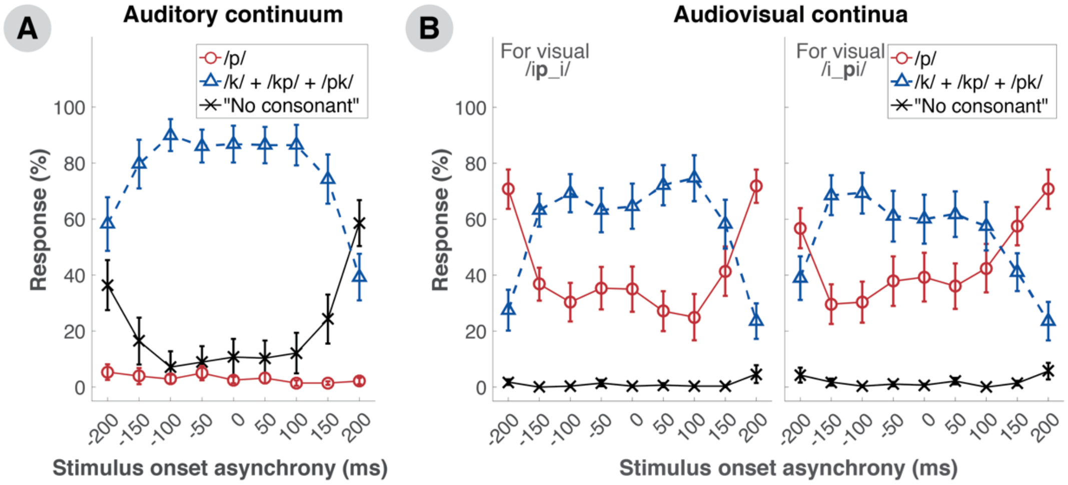

Figure 1: Mean response percentages obtained for the auditory continuum **(A)** and the audiovisual continua **(B)** for the two visual articulatory contexts as a function of SOA. The error bars show the standard error of the mean across subjects.

The main effect of the visual context on the perception of the auditory stream is reflected in the lower percentage of responses containing /k/ in the audiovisual continua than in the auditory continuum. Post-hoc multiple comparisons using Tukey's HSD test confirmed the significant differences in the responses containing /k/ between the auditory continuum and each of the two audiovisual continua ($p < 0.0001$), but not between the two audiovisual continua ($p = 0.211$). This indicates that both audiovisual continua decreased the perception of /k/ across SOA in a similar way, which could be due to a "visual capture" effect reflected in the larger proportion of /p/ responses obtained in the audiovisual continua than in the auditory continuum.

Varying SOA also affected the perception of /k/ responses. The effect was more pronounced at the extreme SOAs for which the percentage of responses with /k/ decreased for all continua. The post-hoc multiple comparisons using Tukey's HSD test showed significant differences between –200 ms and all other SOAs ($p < 0.001$), except for 200 ms ($p = 0.04$), and 200 ms contrasted with all other SOAs ($p < 0.001$). Interestingly, for all comparisons between –150 ms and 100 ms, no significant differences were found ($p = 0.84$), revealing a "plateau" region in which /k/ was similarly perceived across SOAs.

The responses obtained for the auditory continuum indicate that the two auditory streams in the continuum (burst and vowels) were perceived as /iki/ across most SOAs despite the lack of formant transitions. Importantly, /k/ was not clearly perceived for all of the SOAs, since at –200 and 200 ms the "no consonant" responses increased at the expense of the responses containing /k/. This suggests that it is not the burst alone that is perceived as a VCV, which is further supported by the fact that the subjects correctly perceived the auditory stimulus /i_i/ (mean response percentage of 95%).

Assessment of the visual influence

The effect of SOA and visual articulatory gestures on the visual influence was studied by analyzing the sum of /p/, /kp/ and /pk/ responses, in which the visual /p/ influenced speech perception. Figure 2 shows the response percentages obtained for the audiovisual continua in the two visual articulatory contexts as a function of SOA, and for the five unimodal visual articulations. A linear-mixed model ANOVA was fitted to the sum of responses containing /p/. The subject was treated as a random factor, whereas the visual context (two levels: /ip_i/ and /i_pi/) and SOA (nine levels) were treated as fixed factors. The test revealed a significant effect of SOA [$F(8,221) = 8.88$, $p < 0.0001$], while the visual context [$F(1,221) = 2.05$, $p = 0.153$], and the interaction of visual context and SOA [$F(8,221) = 0.73$, $p = 0.66$] were insignificant.

The main effect of SOA on the perception of the visual stream is reflected by the decreased responses containing /p/ for the median SOAs (between –50 and 50 ms). Post-hoc multiple comparisons using Tukey's HSD test confirmed the significant differences between –50 ms contrasted with all other SOAs ($p < 0.01$), except for 0 and 50 ms ($p = 0.78$), and for 50 ms compared to all other SOAs ($p < 0.05$), except for –50 and 50 ms ($p = 0.78$). In contrast, no significant differences were found for all comparisons in the range from –200 to –100 ms ($p = 0.99$), nor in the range from 100 to 200 ms ($p = 0.99$). These results suggest that audiovisual integration occurred more frequently when the burst was closer to the vowels.

Across SOAs, the response percentages containing /p/ were independent of the visual context, as both visual /ip_i/ and /i_pi/ produced similar responses. For these two visual stimuli, the response almost always contained /p/, as reflected in 94% and 97% of responses for /ip_i/ and /i_pi/, respectively. However, cluster percepts were also frequently obtained due to the difficulty in detecting whether the visual articulation contained /k/ in addition to /p/. The two visual cluster articulations presented perceptual confusions and were perceived correctly in 63% and 54% of the trials for visual /ikpi/ and /ipki/, respectively. Also, the subjects were remarkably successful in recognizing when the visual articulation did not contain a consonant, as indicated by 99% correct identifications of /i_i/.

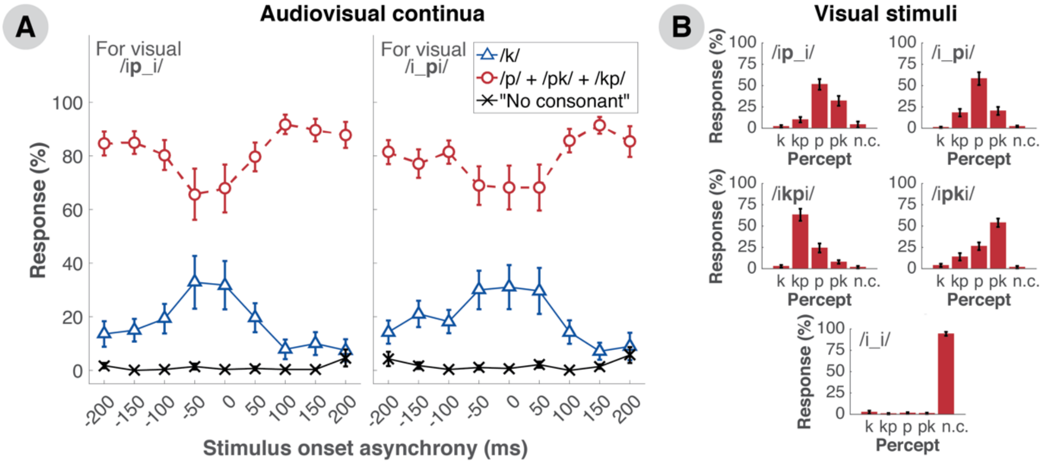

Figure 2: Mean response percentages obtained for the audiovisual continua in the two visual articulatory contexts as a function of SOA **(A)** and for the five visual stimuli tested **(B)**. The error bars show the standard error of the mean across subjects.

Assessment of the perceived consonant order

The PK-index was used to assess the perceived consonant order of the combination responses. Figure 3 shows the PK-index estimated for the two visual contexts as a function of SOA. A linear mixed-effects model was fitted to the mean PK-index. The visual context and SOA were taken as fixed factors and the subjects as a random factor. The outcome of the test revealed a significant main effect of visual context [$F(1,181.05) = 16.05$ $p < 0.0001$] and SOA [$F(8,180.92) = 60.71$, $p < 0.0001$], and an insignificant two-way interaction [$F(8,178.96) = 1.21$, $p = 0.29$].

The effect of SOA on the PK-index for the two audiovisual continua can be seen in the two distinct regions found with different combination responses. One region with a small PK-index, from –200 to –100 ms, for which the asynchrony produced mostly /kp/ responses, and another region with a high PK-index, from 50 to 200 ms, where mostly /pk/ responses were perceived. Tukey's HSD test showed significant differences in the PK-index for all possible pairwise comparisons across regions ($p < 0.0001$). No significant differences were found in the PK-index for any of the pairwise comparisons within the region with small PK-index ($p > 0.96$), nor within the region with high PK-index ($p > 0.89$). These results suggest that, for the two audiovisual continua the subjects perceived one order of consonants when the burst was closer to the initial vowel, and the reverse consonant order when the burst was closer to the final vowel.

For each audiovisual continuum, the consonant order reversal occurred at different SOAs. For the continuum paired with visual /ip_i/ the reversal occurred (earlier) at

–50 ms, and for the continuum paired with visual /i_pi/ the reverse percept occurred (later) at 0 ms. This is consistent with the fact that, in the case of the articulation of /ip_i/, the lips are closed earlier to produce the bilabial consonant than in the case of the articulation of /i_pi/. Post hoc Tukey's HSD test confirmed the significant differences between the two visual contexts on the PK-index ($p < 0.0001$). This indicates that the perceived consonant order depends on the timing of the acoustic burst and the visual articulatory gestures.

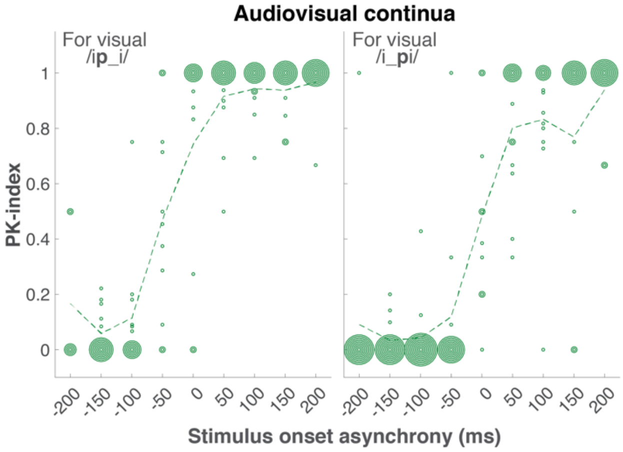

Figure 3: PK-index obtained for the audiovisual continua in the two visual articulatory contexts as a function of SOA. The concentric circles indicate the individual indices and the dotted lines the estimated mean across subjects.

DISCUSSION

The main result of the current study is that the perceived consonant order in McGurk combinations can be reversed consistently by varying the timing of the burst and aspiration of the auditory component. Importantly, we show that the timing at which the reversal in the perceived consonant order occurs seems to depend on the temporal alignment of the burst relative to the articulatory mouth gestures of the speaker. These results support the hypothesis that the burst and aspiration are important cues for audiovisual speech perception, affecting the perceived consonant order for McGurk combinations, and not only the strength of the integration of the cluster percept as has been shown earlier (Green and Norrix, 1997).

Our results are surprising in light of previous findings in which mostly the combination response with the visual labial consonant leading was found (e.g., Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009). While in previous studies the researchers varied the cross-modal timing of consonant-vowel (CV) stimuli, in the present study only the timing of the burst and aspiration was altered, while the vowels were kept synchronous. Our results are consistent with the prior reports in that we mostly found cluster responses with the labial consonant leading when the burst was closer to the onset of the vowel, as would be the case for CV stimuli. Also, since we found reverse combination responses when the acoustic burst was closer to the offset of the vowel, one could expect cluster responses with the acoustic non-labial consonant leading for vowel-consonant (VC) stimuli. This is partly what Hampson and colleagues reported when testing VC combinations (Hampson *et al.*, 2003). They found several responses with the reverse order of consonants, although these percepts only exceeded the most common order of consonants (with the labial leading) when the visual component lagged the auditory component. It thus appears that the position of the consonant within the stimulus, either consonant offset or onset, influences the perceived consonant order, and that such effect seems to be driven by the timing of the burst and aspiration relative to the articulatory gestures of the mouth, as seen in our results.

Our findings also suggest that the cluster percept in McGurk combination provides information on how the individual stimulus features are integrated. While the burst and aspiration seem to be sufficient cues for the perception of the consonant /k/ at most SOAs, the perception of the bilabial /p/ seems to depend on the place information in the visual stream. A combination response then arises when both /k/ and /p/ are perceived, whereas the order of the consonants in the cluster percept depends on the temporal organization of these acoustic features (burst and aspiration) and the mouth closing gestures. Finally, the experimental paradigm in this study further revealed the robustness of audiovisual speech perception, as the phonetic features were split into different streams across a range of temporal asynchronies and were yet integrated.

REFERENCES

Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (**1974**). “Auditory and visual contributions to the perception of consonants.” J. Speech, Lang. Hear. R., **17**, 619-630.

Colin, C., Radeau, M., Deltenre, P., Demolin, D., and Soquet, A. (**2002**). “The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations.” Eur. J. Cogn. Psychol., **14**, 475-491.

Green, K. P., and Norrix, L. W. (**1997**). “Acoustic cues to place of articulation and the McGurk effect: the role of release bursts, aspiration, and formant transitions.” J. Speech Lang. Hear. R., **40**, 646-665.

Hampson, M., Guenther, F. H., Cohen, M. A., and Nieto-Castanon, A. (**2003**). “Changes in the McGurk Effect across phonetic contexts.” Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems.

Massaro, D. W., and Cohen, M. M. (**1993**). “Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables.” Speech Commun., **13**, 127-134.

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

Soto-Faraco, S., and Alsius, A. (**2009**). “Deconstructing the McGurk–MacDonald illusion.” J. Exp. Psychol. Hum. Percept. Perform., **35(2)**, 580.

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