Role of form and motion information in auditory-visual speech perception of McGurk combinations and fusions

Guillaume Gibert, Andrew Fordyce, Catherine J. Stevens

MARCS Auditory Laboratories, University of Western Sydney, Australia

g.gibert@uws.edu.au, a.j.fordyce@gmail.com, kj.stevens@uws.edu.au

Abstract

The perception of biological motion is influenced by motion and form information. Point-light technique has been used to capture the kinematic properties of biological motion. Integration of auditory-visual information in speech perception has been shown to be influenced by such degraded forms of display. The present experiment investigates the role of global shape information and motion in multimodal speech perception. Grayscale stimuli were created from video recordings. Point-lights and point-lights joined by lines formed the stimuli that were created from motion capture data. It was hypothesized that the addition of global shape information would improve the perception of biological motion leading to a higher number of perceptual illusions and that fusion and combination McGurk effects would be identical. Twenty four Australian English subjects were asked to discriminate congruent and incongruent stimuli consisting of non-words and displayed in grayscale Video, Point-light or joined Point-light displays. Results indicate that additional global form information provided by the joint lines compared to the Point-light condition does not influence speech perception for congruent and incongruent stimuli. Nevertheless, reaction times were slower in response to this additional shape information compared with Point-light stimuli. A difference in reaction time was observed for the Video stimuli between combination and fusion responses to McGurk stimuli with subjects responding faster when the stimulus auditory /ga/ and visual /ba/ elicited a combination response /bga/ compared to the reaction time when the incongruent stimulus auditory /ba/ and visual /ga/ elicited a fusion response /da/. Fusion and Combination McGurk effects may be generated by two different perceptual processes.

Index Terms: multimodal speech perception, McGurk effect, point-light display, motion capture

1. Introduction

Biological motion plays a special role in human visual perception. The perception of biological motion seems to rely on a specialized brain system as a 'motion blind' patient can still report human action stimuli [1, 2]. However this patient cannot report the spatial disposition of the actor. The contrary is also true; patients with normal motion coherence thresholds are sometimes unable to discriminate biological motion [2]. In fact, perceiving the motion of biological forms involves integrating form and motion information and also form from motion [3]. Additional explicit linking of joints does not change the overall integration of the audiovisual stimuli. This is likely because face and form activate primarily the ventral system while motion stimuli activate primarily the dorsal system. Recognition of biological movements may activate both systems as well as their confluence in the superior temporal sulcus (STS) [4]. The dorsal stream may be divided into at least 2 major substreams: one specialized for spatial and visuo-spatial functions and another one specialized for the analysis of complex motion. STS integrates motion information from the dorsal system and object information from the ventral system. Moreover, motion information arrives from the dorsal stream in the STS some 20 ms ahead of form information from the ventral stream. But only form and motion arising from the same biological object are integrated within 100 ms of the moving form becoming visible [3].

Kinematic properties of biological motion are isolated by blurring images [5] or more often by the use of the point-light (PL) technique. The first example of the PL technique was presented by Johansson [6]. He recorded an actor with PL on his major joints performing various actions in the dark. Whereas subjects could not identify static images, subjects were able to recognize accurately and quickly the underlying human performance. Local form information is not necessary for biological motion perception as this kind of PL display does not provide it. On the contrary, there is evidence that global form information plays an important role [7]. In fact, computational models of biological motion perception rely in general on template matching models. The templates tend to be global form templates (i.e., stick figures) and their temporal evolution [8] or PL templates [9].

In the case of speech perception, biological motion is part of multimodal information processing. The relative importance of coarse global facial information has been examined by blurring talking faces [10]. Even when visual details were severely reduced by blurring (until 8 cycles per face width), visual speech had a powerful influence on auditory speech. The PL paradigm has been applied to auditory-visual speech perception for congruent [11, 12] and incongruent stimuli [13]. Results showed that isolated kinematic displays provide enough information to increase speech intelligibility in noise for people with normal hearing [11, 12], people with cochlear implants [14], and to influence audiovisual speech integration in response to incongruent stimuli [13]. In this latter study, Rosenblum and Saldaña investigated the perception of congruent and incongruent audiovisual speech stimuli with two different kinds of display: fully illuminated video and PL. The PL stimuli were created by attaching retro-reflective dots to the speaker's face. Twenty-eight dots were placed on the tongue, incisors, lips, chin, cheeks and jaw. The speaker was then videotaped under low illumination. Two auditory-visual congruent (/ba/, /va/) stimuli and one incongruent (audio /ba/- visual /va/) were presented to subjects who had to report what they heard. Visual PL stimuli significantly influenced the heard speech even though the fully-illuminated video had greater visual influence generating a higher number of McGurk effects [15]; that is an automatic perceptual phenomenon appearing under incoherent multimodal information (e.g., when confronted with incongruent auditory and visual speech, subjects report hearing a percept different from the acoustic signal). The

incongruent stimulus auditory /ba/ and visual /va/ elicit mainly a 'visual' response /va/ but in fact there are other kinds of responses in the McGurk effect paradigm. For example, a 'fusion' response occurs when an auditory /ba/ is dubbed with a visual /ga/ and subjects perceive /da/ and a 'combination' response occurs when an auditory /ga/ is dubbed with a visual /ba/ and subjects perceive mainly /bga/. Incongruent stimuli have been shown to elicit longer reaction times for fusion stimuli [16] and for pooled fusion-combination stimuli [17]. Jordan and colleagues [18] extended the Rosenblum and Saldaña experiment by using a larger number of congruent and incongruent stimuli ('fusion' and 'combination'). They used auditory and visual combinations of /ba/, /bi/, /ga/, /gi/, /va/ and /vi/. The incongruent stimuli were constructed by dubbing auditory /ba/ with visual /ga/ and /va/ and auditory /bi/ with visual /gi/ and /di/ and also by dubbing auditory /ga/ and /gi/ with visual /ba/ and /bi/ respectively. Results showed that color and grayscale faces have identical visual influences on identification of the auditory components of congruent and incongruent stimuli whereas PL stimuli had a lower influence as already reported by [13]. No additional information regarding the number of fusion and combination responses induced by the incongruent stimuli was reported.

The setup used in the latter studies [13, 18] to create the PL stimuli was likely to induce additional 3D kinematic information. More specifically, an imperfect chromakey of natural video could leave traces of head and skin motion and so the apparent geometry of the dots change [19]. To avoid this issue, Odisio and colleagues [19] used true 2D PL displays to evaluate the synthesis of speech movements. In the evaluation of their PL rendering, the authors found poorer fusion responses with PL compared to natural faces at all Signal-to-Noise Ratios (SNR) whereas combinations were only significantly different for SNR greater than -18 dB. Due to the large number of points, the authors argued that their PL display was not a true one because it could provide cues on the underlying 3D structure in the absence of motion.

In the present paper, we are interested in, i) replicating and extending the previous results using true 2D PL stimuli with a number of points that do not allow the participants to identify the static display as a face; and ii) determining the role of global form information (by linking the PL by 'joints') for auditory-visual speech perception. The creation of the stimuli and more specifically the true 2D PL is described in the next section. Then, the results of a perception experiment with PL and 'joined' PL stimuli are reported in terms of perception accuracy and reaction time for the different kinds of displays. Results from video stimuli are also described as a baseline.

It is hypothesized that additional global information, provided by joined lines, will improve biological motion perception. Consequently, the number of perceptual illusions would be higher in Joined PL display compared with PL display. The second hypothesis is that the additional global form information will not modify reaction time because motion and form information arise from the same biological object. The third hypothesis is that fusion and combination responses to incongruent auditory-visual stimuli result from the same perceptual process. Consequently, no difference is expected either in terms of the number of illusions or in terms of reaction time.

2. Method

2.1. Material

2.1.1. Video and Motion Capture data

A native Australian English speaker, 25 year old male, was asked to produce twice the following Australian English consonants /b/, /d/, /g/, and /v/ in a Vowel-Consonant-Vowel (VCV) context where the initial and final vowels were /a/. The speaker was instructed to articulate naturally without artificial emphasis. In the first session, the speaker was video-taped with a Sony DV Cam Digital video camera (resolution: 960 x 540 pixels, frame rate: 25 Hz) and a Sennheiser EW 100 G2 lapel microphone was used to record the sound. In the second session, a motion capture device (Northern Digital Optotrak 3020) was used to track the 3D coordinates of 24 sensors glued on the speaker's face and 3 additional ones mounted on a crown while producing the same set of non-words (see Figure 1 for the location of the sensors). The 3D marker positions were captured at 60 Hz. In addition, sound was synchronously recorded using a Behringer C-2 condenser microphone connected to the Optotrak Acquisition Unit II.

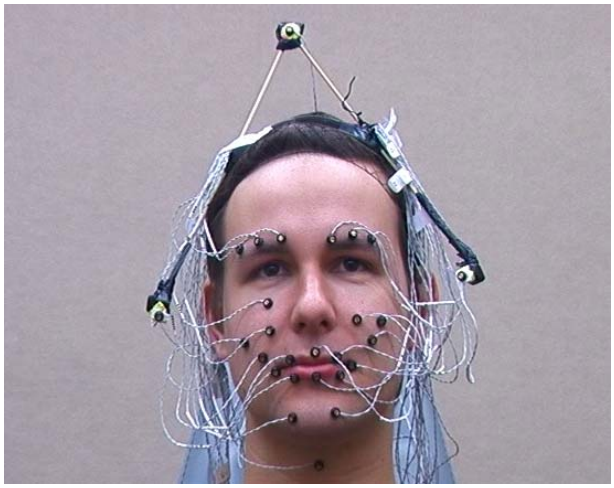

Figure 1: Location of the 27 active motion capture sensors on the speaker's face.

2.1.2. Video stimuli

Videos were segmented and labeled using Praat [20]. The images and the sound were extracted from the video using the software FFMPEG (http://ffmpeg.org/). The images were converted from color to grayscale (see Figure 2 a) using the Java Advanced Image toolbox. They were then recombined using the software MENCODER (http://www.mplayerhq.hu/) to create congruent and incongruent stimuli. There were four congruent stimuli /aba-aba/, /ada-ada/, /aga-aga/ and /ava-ava/ (the first non-word corresponds to the acoustics, the second one to the visual signal). The incongruent stimuli were constructed by dubbing audio /aba/ with visual /ada/, /aga/, and /ava/ and by dubbing visual /aba/ with audio /ada/, /aga/, and /ava/ leading to six incongruent stimuli: /aba-ada/, /aba-aga/, /aba-ava/, /ada-aba/, /aga-aba/ and /ava-aba/. The incongruent stimuli were synchronized on the acoustic consonantal burst onset except for the /v/ where the onset of the consonant was used. This ensured the synchronization to

be within the 200 ms duration asymmetric bimodal temporal integration window [21]. The video were cut to start 300 ms before the auditory onset of the first vowel /a/ and to end 300 ms after the offset of the second vowel /a/.

2.1.3. Point-light stimuli

Identically to the creation of the video stimuli, sound files provided by the Optotrak device were segmented and labeled using Praat [20]. Point-light images (see Figure 2 c) were created from the Optotrak data for each frame (60 frames/s) using MATLAB (The MathWorks, Inc.). They consisted of an orthogonal projection of the 3D sensors location facing the camera. Joined Point-light (JPL) images (see Figure 2 b) were created by joining the PL with lines. The points situated on each eyebrow, the outer lips, the cheekbones and the jaw line were joined successively. Whereas PL display provides only motion information, JPL provides additional global shape information. Videos were then created using MENCODER as described above. Congruent and incongruent stimuli were created using the same method as described previously leading to the same amount of stimuli.

(a) Grayscale Video (b) Joined point-light (c) Point-light

Figure 2: Visual displays used for the perception experiment (a) Grayscale Video (b) Joined Point-Light (JPL) and (c) Point-Light (PL)

2.2. Participants

Twenty four first year undergraduate psychology students from the University of Western Sydney participated in this experiment. They were all native Australian English speakers. They received course credit for their participation. All reported normal or corrected-to-normal vision and no hearing loss. This study was approved by the University of Western Sydney Human Research Ethics Committee.

2.3. Procedure

The experiment was conducted in a sound proof experimental booth. Visual stimuli were displayed on an 18'' computer screen (refresh rate 60Hz) and audio stimuli were presented through 2 loudspeakers. Participants (seated 0.5 m from a computer screen) were instructed to listen to each stimulus and to identify the non-word by clicking on the corresponding labeled button of a graphic user interface. The labeled buttons consisted in a list of 5 items (e.g., for /aba-aga/, the items were /aBa/, /aGa/, /aDa/, /aBGa/ and /aTHa/). Prior to the actual experiment, a pilot study with 4 subjects was conducted to determine all the potential responses for each stimulus and to limit the effect of a 'multiple choice' condition leading to a higher number of illusions compared to a 'free choice' condition [22]. The choice positions were randomly assigned for each stimulus. All stimuli were presented in a random order by a Java program using Java Media Framework. No upper limit of time was defined but participants were instructed to respond quickly and to report their first percept. The practice block consisted of 3 stimuli. The experiment comprised 10 blocks of 30 stimuli ((4 congruent + 6 incongruent) x 3 displays). Participants could rest in between blocks. The stimuli were played once. After choosing an item, the next stimulus was presented. Responses and reaction time were recorded by the program. The reference time for the reaction time was at the beginning of each video.

3. Results

In the following section, a 'correct' response refers to the acoustic stimuli. In the case of incongruent stimuli, a visual influence implies a lower correct response rate than for congruent stimuli. For each subject, responses with a reaction time shorter than 200 ms and greater than 3 standard deviations were rejected. The results recorded in response to the video stimuli are presented as a baseline and will not be compared statistically with the results of the PL and JPL stimuli.

3.1. McGurk effects

Given the non normality of the distributions (Shapiro-Wilk parametric test, $p<0.05$), the median percentage of 'correct' responses for each congruent and incongruent stimulus is presented in Table 1 instead of the mean percentage. The impoverished PL and JPL displays do not affect the perception of congruent stimuli but elicit equal or greater accuracy than the grayscale Video display (except for /ava-ava/). For incongruent stimuli, the effect of PL and JPL displays are weaker than the Video display as already reported [13, 19]. The number of McGurk illusions is low for incongruent stimuli with visual alveolar /d/ and velar /g/. Stronger McGurk effects are found for incongruent stimuli with visual bilabial consonant /b/ and labio-dental consonant /v/.

Table 1. *Median percentage of correct responses (identical to the audio consonant) for each congruent and incongruent stimulus. Results for congruent stimuli are highlighted.*

Stimuli	Video	Joined Point-light	Point-light
/aba-aba/	90	100	100
/aba-ada/	0	84	80
/aba-aga/	10	90	90
/aba-ava/	0	68	80
/ada-ada/	100	100	100
/ada-aba/	0	85	70
/aga-aga/	100	100	100
/aga-aba/	0	60	50
/ava-ava/	100	100	100
/ava-aba/	72	100	100

A step-down non parametrical Dunn test [23] ($\alpha=0.05$) was performed on the congruent and incongruent audio /aba/ stimuli for the Video display. The number of illusions was significantly higher for all the incongruent stimuli compared to the congruent one (Critical Q-value: 2.394, Q-value: 5.95 for /aba-ada/, 4.99 for /aba-aga/, 7.03 for /aba-ava/). Then, the other incongruent stimuli were compared. The number of illusions for the incongruent /ada-aba/ stimulus was statistically different (Q-value: 5.40, Critical Q-value: 1.96) compared to the congruent one. Second, for /aga-aba/, Video display was influencing the perception of the audiovisual

stimuli (Q-value: 5.76, Critical value: 1.96). Third, for the /ava-aba/ stimuli a difference was found for the Video display (Q-value: 2.52, Critical value: 1.96).

In order to compare the JPL and PL displays, a two-way nonparametric Friedman test for identical treatment effects was applied. No statistical difference in terms of number of illusions between PL and JPL displays was found ($\chi^2(1,5)=0.047$, $p=0.827$).

3.2. Reaction time

Given the non normality of the distributions (Shapiro-Wilk parametric test, $p<0.05$), the median reaction times of correct responses for congruent stimuli and incorrect responses for incongruent stimuli instead of the mean values are reported in Table 2 and Table 3, respectively. Reaction times for the congruent stimuli are in general shorter for PL and JPL displays compared to the Video display. An exception was found for the /aga-aga/ stimulus. For incongruent stimuli, the latter pattern is not verified for all kinds of stimuli. Reaction times for the Video stimuli were longer for the fusion responses to the stimulus /aba-aga/ than for the combination responses to the stimulus /aga-aba/.

Table 2. *Median reaction times (in seconds) of correct responses (i.e., identical to the audio consonant) for each congruent stimulus.*

Stimuli	Video	Joined Point-light	Point-light
/aba-aba/	2.389	2.149	2.130
/ada-ada/	2.037	2.032	1.994
/aga-aga/	1.997	2.086	2.037
/ava-ava/	2.067	1.878	1.877

A step-down non parametrical Dunn test [23] ($\alpha=0.05$) was performed on the congruent and incongruent audio /aba/ stimuli for the Video display. The reaction times were not significantly different for the incongruent stimuli /aba-ada/ and /aba-ava/ compared to the congruent one (Critical Q-value: 2.394, Q-value: 0.021 for /aba-ada/, 0.286 for /aba-ava/). On the contrary, the reaction time for the incongruent /aba-aga/ stimulus was significantly slower than for the congruent /aba-aba/ (Q-value: 3.140, Critical Q-value: 2.394). For the /ada-aba/ stimulus compared to the /ada-ada/ stimulus, no difference was found for Video display (Q-value: 1.495, Critical Q-value: 1.960). For incongruent /aga-aba/ stimulus, no difference was found for Video display (Critical Q-value: 1.960, Q-value: 0.132). Finally, for /ava-aba/ incongruent stimulus a difference was found for Video display (Q-value: 3.118, Critical Q-value: 1.960).

Table 3. *Median reaction times (in seconds) of incorrect responses (i.e., identical to the audio consonant) for each incongruent stimulus.*

Stimuli	Video	Joined Point-light	Point-light
/aba-ada/	2.098	2.727	2.110
/aba-aga/	3.049	2.539	2.838
/aba-ava/	2.195	2.604	2.302
/ada-aba/	2.224	2.167	2.411
/aga-aba/	2.027	1.938	2.068
/ava-aba/	2.531	2.336	2.000

A two-way nonparametric Friedman test for identical treatment effects was applied to assess if the PL and JPL displays were statistically different in terms of reaction time. JPL display elicits slower reaction time than PL display ($\chi^2(1,5)=6.034$, $p<0.05$).

3.3. Confusion matrices

The confusion matrices for responses to incongruent stimuli with auditory /aba/ and visual /aba/ are provided for descriptive purposes in Table 4 and in Table 5, respectively.

Table 4: Confusion matrices for incongruent stimuli with auditory /aba/ and for each kind of display.

/aba-ada/	aBa	aDa	aBDa	aTHa	aVa
Video	31	43	13	134	19
JPL	191	0	10	22	17
PL	189	2	11	17	21
/aba-aga/	aBa	aGa	aBGa	aDa	aVa
Video	41	10	14	118	57
JPL	204	2	10	9	15
PL	209	0	8	3	20
/aba-ava/	aBa	aVa	aBVa	aTHa	aDVa
Video	10	105	44	35	46
JPL	156	25	34	13	12
PL	168	17	34	9	12

Table 5: Confusion matrices for incongruent stimuli with visual /aba/ and for each kind of display.

/ada-aba/	aBa	aDa	aBDa	aVa	aBTa
Video	3	56	171	0	10
JPL	3	183	52	0	2
PL	4	162	68	2	4
/aga-aba/	aBa	aGa	aBGa	aPa	aDa
Video	2	46	192	0	0
JPL	5	143	91	0	1
PL	3	133	104	0	0
/ava-aba/	aBa	aVa	aVBa	aPa	aTHa
Video	10	169	43	0	18
JPL	1	227	7	0	5
PL	0	226	4	0	10

In the case of incongruent stimuli composed by auditory /aba/, Video display generated more fusion responses than combination responses. For example, /aba-ada/ Video stimuli generated mainly /aTHa/ fusion than combination or acoustics response. Identically, for /aba-aga/ more fusion responses /aDa/ were chosen than combination or acoustics responses. The incongruent /aba-ava/ elicited mainly a visual response /ava/. Because of the weakness of the effects for PL and JPL displays, the 'incorrect' responses are larger and no real preference for fusion, combination, acoustic or visual responses seemed to be revealed from the data. On the contrary, incongruent stimuli with visual /aba/ generated a large number of illusions for the PL and JPL displays. The same combination responses were chosen for the /aga-aba/ stimulus for all kinds of display. For example /aBGa/ was the most likely response for all displays (after the acoustics for PL and JPL). Identically, for /ada-aba/, a /aBDa/ response was preferred. The incongruent stimulus /ava-aba/ did not elicit a McGurk effect for the PL and JPL displays.

4. Discussion

This study investigated the role of form and motion in auditory-visual speech perception. Motion capture data was used to create true PL displays providing essentially motion information and JPL displays providing motion and global form information. It was hypothesized that the additional shape information would improve the perception of biological motion compared to degraded PL display and without damaging the global reaction time because motion and form information arise from the same biological object. Finally, we hypothesized that fusion and combination percepts come from the same perceptual processing. The results from a perception experiment where 24 subjects were asked to listen to congruent and incongruent auditory-visual stimuli show that true 2D PL display can generate auditory-visual illusions. Contrary to the hypothesis, there was no effect of the additional global form information in terms of number of perceived illusions and there was impairment in terms of reaction time. Finally, fusion and combination percepts differ in terms of reaction time.

The technique used in this study to generate PL was different from the classical one. PL displays were not derived from video recordings but from a motion capture device. Contrary to video-based techniques, the use of a dedicated device provide the accurate positions (<1mm) of sensors glued on the face. Normalized PL (same size) can then be created from this set of positions information. No additional 3D kinematics information (traces of head and skin motion, different apparent geometry of the PL) was added and these PL could be considered as true 2D points moving on the screen. Moreover, contrary to Odisio and colleagues [19], the small number of points was less likely to provide cues to the underlying 3D structure in the absence of motion. Yet, this PL display generated McGurk effects and the number of illusions was smaller for the PL display than for the Video display as already reported. McGurk effects have been shown to be robust even in adverse conditions such as filtering of facial information [24, 25]. The lack of key morphological features in this kind of display affects the number of illusions but does not suppress it. Our results for the /aba-ava/ incongruent stimuli are different from [13]. In fact, fewer occurrences of McGurk effects were elicited in our experiment. Several differences in the protocols could explain the differences in the results. In our experiment no PL was placed on the teeth or the tongue and the PL were derived from different kinds of data (motion capture vs. video). An additional experiment using additional sensors (e.g., from a Wave system (Northern Digital Inc.)) glued on the tongue and the teeth could demonstrate if the difference comes from the lack of sensors in the inter-oral region or from the technology used to derive the PL.

Regarding the addition of global form information, results contradict the hypothesis. No facilitation of biological motion perception was demonstrated either in terms of the number of perceptual illusions or in terms of reaction time. On the contrary, reaction times were significantly slower for the JPL display compared to the PL one. Even though motion and form information arise from the same biological object, the 'linear' relation between the points is virtual and a rough approximation. The perceptual system may not be considering these two channels of information arising from the same biological object. This effect may have been amplified by the various shapes (eyebrows, cheekbones). The presentation of the lips only may have elicited a higher number of illusions. An additional experiment with more sensors glued on the lips (outer and inner contours) defining a more realistic contour should demonstrate that this non effect was due to the non-realistic shape displayed.

In their original article, McGurk and MacDonald [15] presented two kinds of incongruent stimuli leading to two kinds of responses: by dubbing an auditory /baba/ onto a visual /gaga/, a majority of adults reported a fused response /dada/ and by using the reverse dubbing, the majority reported combination responses /bagba/ or /gaba/. Incongruent auditory-visual speech stimuli have been shown to generate longer reaction times for fusion stimuli [16] and for pooled fusion-combination stimuli [17]. In the present study, several incongruent stimuli were tested by dubbing auditory /da/, /ga/ and /va/ to visual /ba/ more likely to generate 'combination' responses and by dubbing visual /da/, /ga/ and /va/ to auditory /ba/ more likely to generate 'fusion' or 'visual' responses. For the Video display, 'fusion' responses evoked by /aba-aga/ stimulus have a slower reaction time than combination responses due to the /aga-aba/ stimulus. In fact, incongruent stimuli eliciting 'combination' responses such as /ada-aba/ and /aga-aba/ are not significantly different in terms of reaction time compared to congruent stimuli. Due to the small amount of auditory-visual illusions elicited by the incongruent stimulus /ava-aba/, the slower reaction time may be due to a post-perceptual decision process. Incongruent stimuli generating combination responses seem to be processed as normal congruent speech. On the contrary, the 'fusion' stimulus /aba-aga/ generates a significantly slower reaction time, leading to the hypothesis that this kind of stimulus is not processed in the same way. In an electrophysiological experiment, Colin and colleagues [26] used an oddball paradigm where rare incongruent auditory-visual stimuli (deviants) were inter-mixed with high-probability congruent auditory-visual stimuli (standards). Although 'combination' and 'fusion' responses elicited electrophysiological Mismatch negativity (MMN) responses, two distinct patterns emerged from the data. 'Combination' responses elicited a MMN with two components: an early one covering the N1 exogenous component and a later one after the P2 component whereas 'fusion' responses generated a MMN containing three components: the first two components were similar to the MMN evoked by the 'combination' responses (the early component was larger covering P1 and N1 evoked responses) and a very late component starting 400 ms after the acoustic onset. These McGurk effects may be considered as two different perceptual processes. Additional experiments investigating separately 'combination', 'fusion' and 'visual' responses are needed to characterize the different perceptual processes involved in each case and to verify if the present results are not due to the different dynamics of the auditory and visual information. Moreover, reaction time information should be integrated in computational models of speech perception. In general, the only information used in building these models is the number of illusions and not the reaction time. For example, Omata and Mogi developed a computational model of auditory-visual speech perception [27]. Their results show that the asymmetric effect (fusion, combination) could be the 'distance' relationship between audio or visual information with the neural networks. The integration of the reaction time differences obtained here could lead to a better understanding of the perceptual phenomenon.

5. Acknowledgements

We would like to thank Dr. Jeesun Kim for providing the video and the Optotrak recordings used in this study. This work was supported by the Thinking Head project a *Special Initiative* scheme of the Australian Research Council (ARC) and the National Health and Medical Research Council (NH&MRC) [28].

6. References

[1] P. McLeod, W. Dittrich, J. Driver, D. Perrett, and J. Zihl, "Preserved and impaired detection of structure from motion by a "motion-blind" patient," *Visual Cognition,* vol. 3, pp. 363-391, 1996.

[2] T. Schenk and J. Zihl, "Visual motion perception after brain damage: II. Deficits in form-from-motion perception," *Neuropsychologia,* vol. 35, pp. 1299-1310, 1997.

[3] M. W. Oram and D. I. Perrett, "Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey," *Journal of Neurophysiology,* vol. 76, pp. 109-129, 1996.

[4] L. M. Vaina, J. Solomon, S. Chowdhury, P. Sinha, and J. W. Belliveau, "Functional neuroanatomy of biological motion perception in humans," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 98, pp. 11656-11661, Sep 25 2001.

[5] S. Kuhlmann and M. Lappe, "Recognition of biological motion from blurred natural scenes," *Perception,* vol. 35, pp. 1495-1506, 2006.

[6] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics,* vol. 14, pp. 201-211, 1973.

[7] B. I. Bertenthal and J. Pinto, "Global processing of biological motions," *Psychological science,* vol. 5, pp. 221-225, 1994.

[8] J. Lange and M. Lappe, "A model of biological motion perception from configural form cues," *Journal of Neuroscience,* vol. 26, pp. 2894-2906, Mar 15 2006.

[9] J. Lee and W. Wong, "A stochastic model for the detection of coherent motion," *Biological Cybernetics,* vol. 91, pp. 306-314, 2004.

[10] S. M. Thomas and T. R. Jordan, "Determining the influence of Gaussian blurring on inversion effects with talking faces," *Perception & Psychophysics,* vol. 64, pp. 932-944, 2002.

[11] L. Lachs and D. B. Pisoni, "Specification of cross-modal source information in isolated kinematic displays of speech," *Journal of the Acoustical Society of America,* vol. 116, pp. 507-518, Jul 2004.

[12] L. D. Rosenblum, J. A. Johnson, and H. M. Saldaña, "Point-light facial displays enhance comprehension of speech in noise," *Journal of Speech & Hearing Research,* vol. 39, pp. 1159-1170, 1996.

[13] L. D. Rosenblum and H. M. Saldaña, "An audiovisual test of kinematic primitives for visual," *Journal of Experimental Psychology / Human Perception & Performance,* vol. 22, p. 318, 1996.

[14] T. R. Bergeson, D. B. Pisoni, and J. T. Reynolds, "Perception of Point Light Displays of Speech by Normal-Hearing Adults and Deaf Adults with Cochlear Implants," in *International Conference on Audio-Visual Speech Processing* St Jorioz, France, 2003, pp. 55-60.

[15] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature,* vol. 264, pp. 746-8, Dec 23-30 1976.

[16] K. P. Green and P. K. Kuhl, "Integral Processing of Visual Place and Auditory Voicing Information During Phonetic Perception," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 17, pp. 278-288, 1991.

[17] K. Sekiyama and D. Burnham, "Impact of language on development of auditory-visual speech perception," *Developmental Science,* vol. 11, pp. 306-320, 2008.

[18] T. R. Jordan, M. V. McCotter, and S. M. Thomas, "Visual and audiovisual speech perception with color and gray-scale facial images," *Perception and Psychophysics,* vol. 62, pp. 1394-1404, 2000.

[19] M. Odisio and G. Bailly, "Audiovisual Perceptual Evaluation of Resynthesised Speech Movements," in *International Conference on Spoken Language Processing* Jeju Island, Korea 2004.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 5.1.31 ed, 2010.

[21] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia,* vol. 45, pp. 598-607, 2007.

[22] C. Colin, M. Radeau, and P. Deltenre, "Top-down and bottom-up modulation of audiovisual integration in speech," *European Journal of Cognitive Psychology,* vol. 17, pp. 541-560, 2005.

[23] G. Cardillo, "Dunn Test: a procedure for multiple, not parametric, comparisons.," 2006, http://www.mathworks.com/matlabcentral/fileexchange/12827.

[24] C. S. Campbell and D. W. Massaro, "Perception of visible speech: Influence of spatial quantization," *Perception,* vol. 26, pp. 627-644, 1997.

[25] J. MacDonald, S. Andersen, and T. Bachmann, "Hearing by eye: how much spatial degradation can be tolerated?," *Perception,* vol. 29, pp. 1155-1168, 2000.

[26] C. Colin, M. Radeau, A. Soquet, D. Demolin, F. Colin, and P. Deltenre, "Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory," *Clin Neurophysiol,* vol. 113, pp. 495-506, Apr 2002.

[27] K. Omata and K. Mogi, "Fusion and combination in audio-visual integration," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 464, pp. 319-340, 2008.

[28] D. Burnham, R. Dale, K. Stevens, D. Powers, C. Davis, J. Buchholz, K. Kuratate, J. Kim, G. Paine, C. Kitamura, M. Wagner, S. Möller, A. Black, T. Schultz, and H. Bothe, "From Talking Heads to Thinking Heads: A Research Platform for Human Communication Science," ARC/NH&MRC Special Initiatives, TS0669874, 2006-2011.