Published in final edited form as:
Curr Biol. 2009 May 12; 19(9): 735–739. doi:10.1016/j.cub.2009.03.019.

Audiovisual integration of speech in a bistable illusion

K.G. Munhall[1,2], **M.W. ten Hove**[3], **M. Brammer**[4], and **M. Paré**[1,5]

1 *Department of Psychology, Queen's University, Kingston, Canada*

2 *Department of Otolaryngology, Queen's University, Kingston, Canada*

3 *Department of Ophthalmology, Queen's University, Kingston, Canada*

4 *Centre for Neuroimaging Sciences, Institute of Psychiatry, London, UK*

5 *Department of Physiology, Queen's University, Kingston, Canada*

Summary

Visible speech enhances the intelligibility of auditory speech when listening conditions are poor [1], and can even modify the perception of otherwise perfectly audible utterances [2]. This audiovisual perception is our most natural form of communication and one of our most common multisensory phenomena. However, where and in what form the visual and auditory representations interact is still not completely understood. While there are longstanding proposals that multisensory integration occurs relatively late in the speech processing sequence [3], there is considerable neurophysiological evidence that audiovisual interactions can occur in the brain stem and primary auditory and visual cortices [4,5]. One of the difficulties testing such hypotheses is that when the degree of integration is manipulated experimentally, the visual and/or auditory stimulus conditions are drastically modified [6,7] and thus the perceptual processing within a modality and the corresponding processing loads are affected [8]. Here we used a novel bistable speech stimulus to examine the conditions under which there is a visual influence on auditory perception in speech. The results indicate that visual influences on auditory speech processing, at least for the McGurk illusion, necessitate the conscious perception of the visual speech gestures, thus supporting the hypothesis that multisensory speech integration is not completed in early processing stages.

Results and Discussion

In the present studies we held audiovisual stimulus conditions constant and allowed subjective organization of the percept to determine the extent of multisensory integration. This was achieved through the use of a dynamic version of Rubin's vase illusion [9]. In our stimulus an irregular vase rotated and its changing profile produced a talking face profile (see Figure 1). The face articulated the nonsense utterance /aba/ while the accompanying acoustic signal was a voice saying the nonsense utterance /aga/. Two visual and two auditory percepts occur with these stimuli. Visually, the faces appeared to be the figure and the vase was the background or vice versa. Auditorily, subjects heard either the recorded audio track, /aga/, or heard the so-called combination McGurk effect, /abga/ [3]. In this illusion, both consonants are 'heard' even though only the /g/ is present in the acoustic signal. This percept results from visual influences

Corresponding Author: Kevin Munhall E-mail: kevin.munhall@queensu.ca, Dept. of Psychology, 62 Arch St., Queen's University, Kingston, ON Canada K7L 3N6.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

on auditory perception. When subjects only heard the acoustic signal, /aga/, there was no phonetic influence of the visual information. Three experiments are presented here.

Experiment 1 looked at the association of the McGurk illusion and the perception of either the vase or face. Complete independence of these percepts would suggest that visual influences on auditory speech perception might occur at an early stage of processing, either subcortically or in primary sensory cortex. Recent work on figure-ground perception indicates that, beyond the simple competition between low-level processing units, figural assignment may involve widespread recurrent processing [e.g., 10] and biased competition between high-level shape perception units [11]. If audiovisual integration in speech is not sensitive to the suppression of face perception in the bistable stimulus it must precede or be independent of this process. Alternatively, complete association of face perception and perception of the McGurk illusion would suggest that audiovisual integration of speech depended on categorical stimulus representations for object perception. Two different stimuli were presented to subjects. In the first condition, the vase rotated and its shape produce a profile of an articulating face saying the utterance /aba/ (Figure 1A: moving face, moving vase). In the second condition, the vase rotated but the face profile remained constant (Figure 1B: still face, moving vase). This was achieved by subtle changes to the 3D vase in this condition such that its visible rotation did not produced any profile changes. Such a stimulus could only be produced using animation. Each of these stimuli were combined with a recording of /aga/. The control condition was not expected to produce the McGurk effect since there was no visual information for a consonant. Subjects watched single tokens and gave two responses. First they reported whether they perceived a vase or a face then they told the experimenter whether they heard /aga/ or /abga/.

For the moving face, moving vase stimulus, the results show a strong association between consciously perceiving the face and perceiving the McGurk effect (Figure 2); 66 percent of the responses shared this perceptual pattern. Only 9 percent of the responses reported the McGurk effect when the vase was the percept. The control stimulus (still face, moving vase) produced a quite different pattern of responses. Approximately 90 percent of the speech responses were percepts of the auditory stimulus /aga/. These responses were split between the vase and face percepts with a slight bias toward perceiving the face. The /abga/ responses (~10%) were split between the face and vase percepts. This three-way interaction was reliable by Chi Square test ($p<.001$). When the 2×2 response contingencies tables were evaluated separately for each stimulus, the moving face, moving vase showed a reliable association between face perception and the perception of the McGurk combination (Fisher Exact Probability test, $p<.05$) while the stimulus with a still face and moving vase showed no association ($p>.5$).

The small number of /bg/ percepts in the moving face, moving vase condition when the vase was reported was approximately equal to the number of /bg/ percepts for the still face, moving vase condition (~10%). This common response rate suggests that this may be simply response bias or error. While motion in a suppressed image in binocular rivalry can still elicit motion aftereffects [12] and contribute to the perception of apparent motion [13], the moving face seems to require conscious perception in order to influence auditory speech.

The presence of vase motion alone produced a large number of face percepts. This is not associated with audiovisual integration as virtually no McGurk effects were observed for this condition. When the two motion conditions in Experiment 1 are contrasted we see strong evidence for the importance of dynamic facial information and its conscious perception as prerequisites for audiovisual speech perception. These findings are consistent with studies showing that awareness that an acoustic signal is speech is a prerequisite for audiovisual integration [14].

Two control experiments were carried out to help clarify the results. Experiment 2 tested further how motion influenced vase/face perception and in addition how sound influenced this percept. Three different levels of movement of the stimulus were shown with and without the speech soundtrack. In one condition, a static frame of the vase and face was shown for the duration of the dynamic stimuli. This frame was identical to the left most frame in Figure 1A. The other two conditions were identical to the visual conditions tested in Experiment 1.

Figure 3 shows the mean proportion of face percepts for the three movement conditions as a function of whether a speech utterance was played along with the visual stimuli. A robust effect of movement condition is evident (F(2,24) = 36.4, p<.001) while only a modest influence of the presence of sound can be seen (F(1, 24) = 3.8, p=.06) and no interaction. The presence of motion dramatically decreased the percentage of vase percepts from the high of 76 percent in the static image condition to a low of 28 percent in the moving face, moving vase stimulus. Each of the three motion conditions were reliably different from each other (p<.01). The presence of auditory speech increased the percentage of face percepts but by less than 10 percent on average.

From a pictorial viewpoint, the stimulus was biased toward perceiving the vase by the surface texture information and three-dimensional rendering of the vase [15]. The still image's high proportion of vase percepts reflects this. When auditory speech and any motion (either the vase alone or both vase and face) were presented, the proportion of vase percepts consistently decreased from the silent, still image-condition high water mark. The onset of motion in an image during binocular rivalry [16] or higher velocity [17] in an image tends to increase the likelihood of that image dominating perception. The reduction in vase percepts in the still face, moving vase condition is inconsistent with these findings. The independence of facial form and motion pathways [18] suggests a possible high-level associative account. Nevertheless, the moving face, moving vase condition is the only visual condition in which face percepts dominate (>50%). The influence of sound was modest and relatively consistent across the different visual conditions. If early audiovisual interactions were driving the visual percepts, the moving face condition would have been expected to show the strongest influence of the presence of sound. This was not the case. The results suggest that the conditions determining the perception of the unimodal stimulus (vision) are primarily determining multisensory integration [19].

Experiment 3 was carried out to test whether perceptual alternations could be accounted for simply by an alternation between eccentric (face) and central (vase) fixations. The distributions of gaze fixation positions associated with either of the two reported percepts were compared using an analysis derived from signal detection theory. For each subject, we found that the distribution of fixation positions associated with each percept overlapped extensively and only in 0.1% of the cases (22/23506) could the gaze distributions be considered as significantly different. This finding is consistent with the report that the changes in perception of the Rubin's face-vase stimulus are not associated with changes in eye positions [20] and with work showing that the McGurk effect is not dependent on whether the visual speech is viewed using central (foveal) or paracentral vision [e.g., 21].

Recent evidence indicates that the attentional state of the subject influences audiovisual integration of speech [22,23]. The McGurk effect is reduced under high attention demands. Further, subjects appear to have perceptual access to the individual sensory components as well as the unified multisensory percept [24]. Findings such as these contradict the view that multisensory integration is pre-attentive and thus automatic and mandatory [25] and are consistent with the involvement of higher order processes in phonetic decisions. The evidence that auditory processing is influenced by visual information subcortically as early as 11 ms following acoustic onset for speech stimuli [4] or cortically in less than 50 ms [5] for tone stimuli is, at first look, difficult to reconcile with such findings.

One possible solution is that multisensory speech processing involves interaction between auditory and visual information at many levels of perception yet the final phonetic categorization, and ultimately audiovisual integration, takes place quite late. Multisensory processing may involve rapid attentional mechanisms that modulate early auditory or visual activity [26], promote spatial orienting [27] or provide contextual modulation of activity [28]. Yet, the dynamic structure of speech may require integration over longer timescales than the speed at which vision and audition can initially interact. The production of human speech is quite slow with the modal syllable rate being approximately 3-6 Hz [29]. It has long been recognized that information for speech sounds does not reside at any instant in time but rather is extended over the syllable [30]. Thus, even within a modality the temporal context of information determines its phonetic identity. For audiovisual speech of the kind presented here, the information for consonant identity is extended in time [31] and perception requires extended processing to integrate this perceptual information.

It remains to be seen whether this conclusion extends to all audiovisual speech phenomena. Vision can influence auditory speech perception in at least two distinct ways [32]. The first involves correlational modulation. Visible speech strongly correlates with some parts of the acoustic speech signal [33]. The acoustic amplitude envelope and even the detailed acoustic spectrum can be predicted by the visible speech articulation. This redundancy may permit early modulation of audition by vision, for example, by the visual signal amplifying correlated auditory inputs [34].

The second way in which visible speech influences auditory speech is by providing complementary information. In this case, vision provides stronger cues than the auditory signal or even information missing from the auditory signal. This latter case is the situation that best describes the perception of speech in noise and the combination McGurk effect. In both of the examples the correlation between auditory and visual channels is broken because of the loss of information in the auditory channel. For the combination McGurk, a /b/ could be plausibly produced during the intervocalic closure in /aga/ with minimal or without any auditory cues. The strong cue of a visible bilabial closure provides independent information to the speech system. It is possible that such complementary visual information can only be combined with the auditory signal late during phonetic decision making after both modalities carry out considerable processing.

In experimental settings, the natural correlation between auditory and visual speech can also be broken by having the visible speech provide contradictory cues for the auditory signal. This is the case for the standard fusion McGurk effect [2] where an auditory /b/ is combined with a visual /g/ and /d/ is heard. Both modalities yield sufficient but contradictory cues for consonant perception though for the strongest effect the auditory /b/ must be a weak percept. Whether the perceptual system also makes a late phonetic decision under these conditions is unclear. The evidence from attention studies suggests that this is the case [22,23].

Bistable phenomena in vision, audition, and multisensory processing are well accounted for by ideas of distributed competition involving different neural substrates and perceptual processes [35,36,37]. Audiovisual speech perception may share this form of distributed processing. However, the data presented here indicate that multisensory decision making in speech perception requires high-level phonetic processes including the conscious perception of facial movements. The unique stimuli used in these experiments will be an important tool in further characterizing the network of processes involved in this multisensory perception.

Experimental Procedures

Subjects

The studies were approved by the Queen's University General Research Ethics Board and all subjects gave informed consent before participating in the research.

Stimuli

Audiovisual stimuli were created using a dynamic version of the Rubin Vase illusion [9]. Experiments 1 and 2, used an animated version (Figure 1) of a vase created with Maya (Autodesk) with the face profile determined by the same video sequence as Experiment 1. In both stimuli, the vase was irregular and as it rotated, its edge would produce a different face profile. Figure 1a shows three frames from the movie in which the vase rotates and its changing shape produced a face profile that articulates the utterance /aba/. The face profile matches the original movie exactly on a frame-by-frame basis. Figure 1b shows 3 frames from the control movie in which a slightly different vase rotates but its changing shape produces no change in the face profile. The difference in profile changes between 1a and 1b is due to subtle differences in the animated vase 3D shape between the two conditions. In Experiment 3, a video of a rotating, custom-constructed vase was edited using the profile of a female speaker saying the utterance /aba/.

Procedure

Experiment 1—12 subjects were presented with 2 types of stimuli in single trials. The stimuli were the two dynamic stimuli from Experiment 2 (rotating vase that produced an articulating face, rotating vase that produced a still face) both presented with the audio track, /aga/. Subjects were asked to indicate whether they saw a face or vase. Only a single response was permitted for each trial. After reporting this, they were instructed to record whether the sound they perceived was most like /aga/ or /abga/. Following 10 warm up trials, the subjects were presented with 60 experimental trials, 30 of each condition in randomized order.

Experiment 2—14 subjects were presented with 6 types of stimuli in single trials. Three visual stimuli (still frame, moving face, moving vase: rotating vase that produced an articulating face, and still face, moving vase: rotating vase that produced a still face) were presented with either the audio track, /aga/, or silence. Each trial was composed of a single rotation of the vase or in the case of the still frame a period equaling the duration of the dynamic trials. The subjects' task was to indicate whether they saw the face or the vase first, then indicate each time it changed within a trial. Following 12 warm up trials, each of the stimuli was presented five times with order randomized across stimulus type making 30 experimental trials in total. When subjects reported more than one state in a single trial both responses were included in the analyses as separate responses for that condition. Subjects generally reported only one perceptual state for the bistable stimulus in each trial with the overall average number of states reported being 1.05 states per trial. There were no differences in the number of states seen across the conditions.

Experiment 3—We tested 7 subjects on a behavioural task in which the stimulus was displayed in loops of 10 continuous utterances. Subjects responded after each loop with a keypress indicating whether they heard a /b/ sound or not. In addition to the behavioural task, we examined whether the varying audiovisual percept of the bistable stimulus depended on the subject's gaze fixation positions by monitoring horizontal and vertical eye position of the subjects while they view the stimulus during repeated trials and reported when their percept changed.

Horizontal and vertical eye position were sampled at a rate of 1 kHz using the search-coil-in-magnetic-field technique [38] with an induction coil that consisted of a light coil of wire embedded in a flexible ring of silicone rubber (Skalar) that adheres to the limbus of the human eye, concentric with the cornea [39]. The search coil was positioned in the dominant eye of the subjects and only after the surface of the eye had been anesthetized with a few drops of anesthetic (Tetracaine Hcl, 0.5%). Details of this method were described previously [21].

Analysis –Experiment 3—The distributions of fixations for the signal detection analysis were computed in the following manner. At each millisecond in each type of utterance (i.e., ones perceived either as /bg/ or /g/), we calculated separately the probabilities that the positions of horizontal and vertical gaze fixation were greater than a position criterion, which was incremented in 1-deg steps across the image from either the left margin of the image or its bottom margin. The ensuing fixation position probabilities (for each percept) were then plotted against each other in a receiver operating characteristic (ROC) curve, and the area under each curve (AUROC) computed to capture the amount of separation between the two distributions of fixation positions. This quantitative measure gives the general probability that, given one draw from each distribution, the fixation positions from the distributions associated with the two percepts would be distinct.

Acknowledgments

The National Institute on Deafness and other Communication Disorders (grant DC-00594), the Natural Sciences and Engineering Research Council of Canada. Bryan Burt assisted in data collection. The authors acknowledge the artistic work of Patty Petkovich, Rob Loree, and particularly Mike Watters for creation of the vase stimuli.

References

1. Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. J Acoust Soc Am 1954;26:212–215.
2. McGurk H, MacDonald J. Hearing lips and seeing speech. Nature 1976;264:746–748. [PubMed: 1012311]
3. Massaro, DW. Perceiving talking faces: from speech perception to a behavioral principle. MIT Press; Cambridge, MA: 1998.
4. Musacchia G, Sams M, Nicol T, Kraus N. Seeing speech affects information processing in the human brainstem. Exp Brain Res 2006;168:1–10. [PubMed: 16217645]
5. Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ. Multisensory auditory–visual interactions during early sensory processing in humans: a highdensity electrical mapping study. Brain Res Cogn Brain Res 2002;14:115–128. [PubMed: 12063135]
6. Ross LA, Saint-Amour D, Leavitt VN, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex 2007;17:1147–1153. [PubMed: 16785256]
7. Callan D, Jones JA, Munhall KG, Kroos C, Callan A, Vatikiotis-Bateson E. Multisensory-integration sites identified by perception of spatial wavelet filtered visual speech gesture information. Journal of Cognitive Neuroscience 2004;16:805–816. [PubMed: 15200708]
8. Alsius A, Navarra J, Campbell R, Soto-Faraco SS. Audiovisual integration of speech falters under high attention demands. Current Biology 2005;15:839–843. [PubMed: 15886102]
9. Rubin, E. Synoplevde Figurer. Kopenhagen: Gyldendalske; 1915.
10. Domijan D, Setic M. A feedback model of figure-ground assignment. J Vision 2008;8:1–27.
11. Peterson M, Skow E. Inhibitory competition between shape properties in figure-ground perception. Journal of Experimental Psychology: Human Perception and Performance 2008;34:251–267. [PubMed: 18377169]
12. Lehmkuhle S, Fox R. Effect of binocular rivalry suppression on the motion aftereffect. Vision Res 1976;1:5, 855–859.

13. Wiesenfelder H, Blake R. Apparent motion can survive binocular rivalry suppression. Vision Res 1991;31:1589–1600. [PubMed: 1949627]

14. Tuomainen J, Andersen T, Tiippana K, Sams M. Audio-visual speech perception is special. Cognition 2005;96:B13–B22. [PubMed: 15833302]

15. Hasson U, Hendler T, Bashat DB, Malach R. Vase or face? A neural correlate of shape-selective grouping processes in the human brain. J Cog Neurosci 2001;13:744–753.

16. Fox R, Check R. Detection of motion during binocular rivalry suppression. J Exp Psychol 1968;78:388–395. [PubMed: 5705853]

17. Blake R, Yu K, Lokey M, Norman H. Binocular rivalry and visual motion. J Cogn Neurosci 1998;10:46–60. [PubMed: 9526082]

18. Alais D, Parker A. Independent binocular rivalry processes for form and motion. Neuron 2006;52:911–920. [PubMed: 17145510]

19. Sanabria D, Soto-Faraco S, Chan J, Spence C. Intramodal perceptual grouping modulates multisensory inegration: evidence from the crossmodal dynamic capture task. Neuroscience Letters 2005;377:59–64. [PubMed: 15722188]

20. Andrews TJ, Schluppeck D, Homfray D, Matthews P, Blakemore C. Activity in the fusiform gyrus predicts conscious perception of Rubin's vase-face illusion. Neuroimage 2002;17:890–901. [PubMed: 12377163]

21. Paré M, Richler R, ten Hove M, Munhall KG. Gaze Behavior in Audiovisual Speech Perception: The Influence of Ocular Fixations on the McGurk Effect. Perception and Psychophysics 2003;65:553–567. [PubMed: 12812278]

22. Alsius A, Navarra J, Campbell R, Soto-Faraco SS. Audiovisual integration of speech falters under high attention demands. Current Biology 2005;15:839–843. [PubMed: 15886102]

23. Tiippana K, Andersen TS, Sams M. Visual attention modulates audiovisual speech perception. Eur J Cogn Psychol 2004;16:457–472.

24. Soto-Faraco S, Alsius A. Conscious access to the unisensory components of a cross-modal illusion. Neuroreport 2007;18:347–350. [PubMed: 17435600]

25. Bertelson P, Vroomen J, de Gelder B, Driver J. The ventriloquist effect does not depend on the direction of deliberate visual attention. Perception and Psychophysics 2000;62:321–332. [PubMed: 10723211]

26. Foxe JJ, Simpson GV, Ahlfors SP. Cued shifts of intermodal attention: parieto-occipital |10 Hz activity reflects anticipatory state of visual attention mechanisms. Neuroreport 1998;9:3929–3933. [PubMed: 9875731]

27. Spence C, Driver J. Audiovisual links in exogenous covert spatial orienting. Perception and Psychophysics 1997;59:1–22. [PubMed: 9038403]

28. Kayser C, Petkov CI, Logothetis NK. Visual modulation of neurons in auditory cortex. Cerebral Cortex 2008;18:1560–1574. [PubMed: 18180245]

29. Greenberg S. Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. Speech Communication 1999;29:159–176.

30. Liberman A, Cooper F, Shankweiler D, Studdert-Kennedy M. Perception of the speech code. Psychol Rev 1967;74:431–461. [PubMed: 4170865]

31. Munhall KG, Tohkura Y. Audiovisual gating and the time course of speech perception. J Acoust Soc Am 1998;104:530–539. [PubMed: 9670544]

32. Campbell R. The processing of audio-visual speech: empirical and neural bases. Phil Trans R Soc B 2008;363:1001–1010. [PubMed: 17827105]

33. Yehia HC, Kuratate T, Vatikiotis-Bateson E. Linking facial animation, headmotion and speech acoustics. J Phonetics 2002;30:555–568.

34. Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A. Neuronal oscillations and visual amplifications of speech. Trends Cogn Sci 2008;12:106–113. [PubMed: 18280772]

35. Pressnitzer D, Hupé JM. Temporal dynamics of auditory and visual bistability reveal principles of perceptual organization. Current Biology 2006;16:1351–1357. [PubMed: 16824924]

36. Blake R, Logothetis NK. Visual competition. Nat Rev Neurosci 2002;3:13–21. [PubMed: 11823801]

37. Shams L, Kamitani Y, Shimojo S. Visual illusion induced by sound. Brain Res Cogn Brain Res 2002;14:147–152. [PubMed: 12063138]

38. Robinson DA. A method of measuring eye movements using a scleral search coil in a magnetic field. IEEE Transactions in Biomedical Engineering 1963;10:137–145.

39. Collewijn H, van der Mark F, Jansen TC. Precise recording of human eye movements. Vision Research 1975;15:447–450. [PubMed: 1136166]

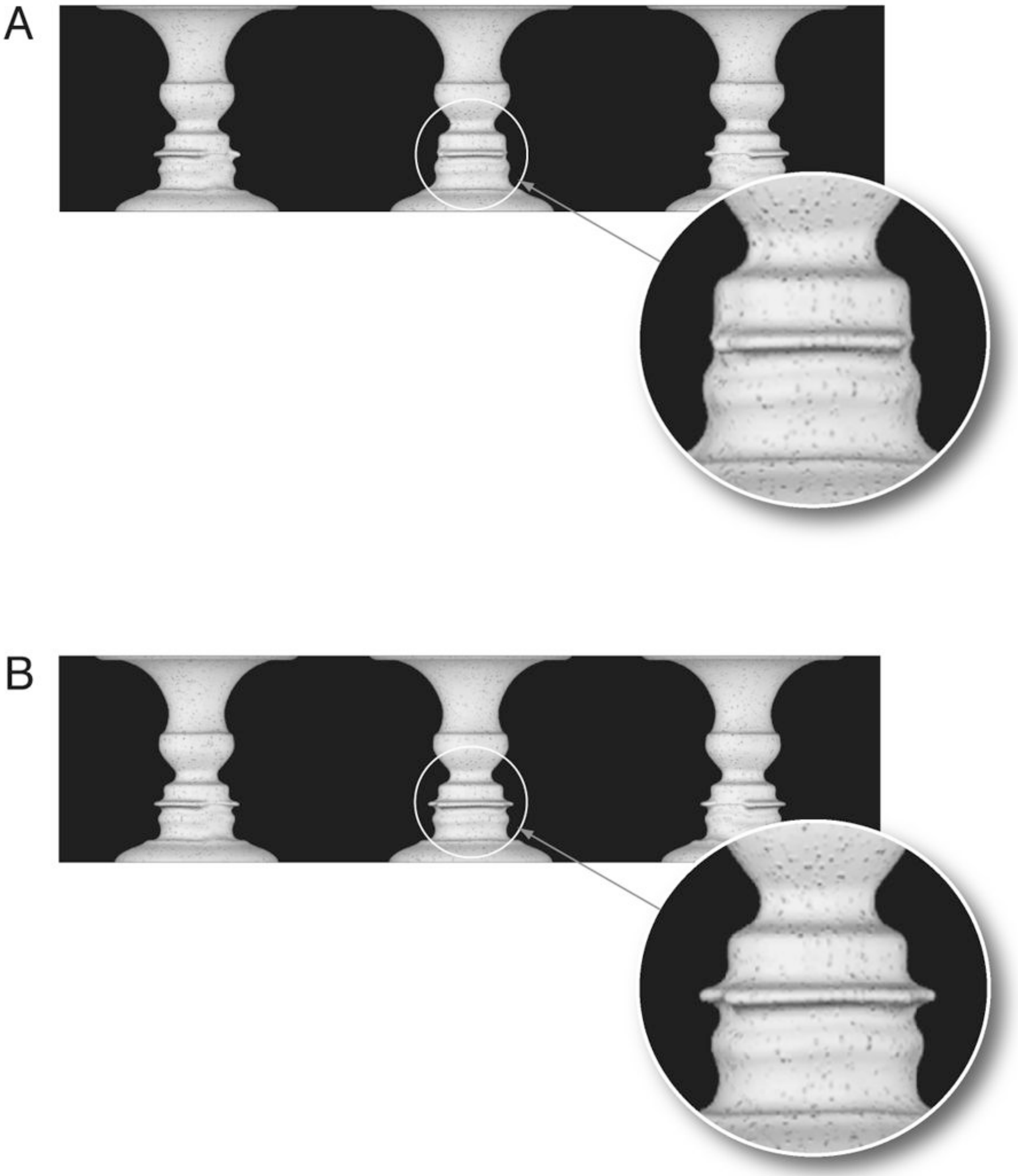

Figure 1.

Individual frames from rotating vase movie used in the dynamic vase conditions in Experiments 2 and 3. A. Moving face, moving vase: The face profile changes shape as if the face is articulating the utterance /aba/ as the vase rotates. The three frames correspond to the points in time during the first vowel, during the closure for the /b/, and during the second vowel. The circle shows the detail of the lip closure in the facial profile. B. Still face, moving vase: The face profile does not change shape as the vase rotates. The three frames correspond to the same points in time as the sequence shown in A. The circle shows the detail of the open lips in this condition in contrast to that shown in A.

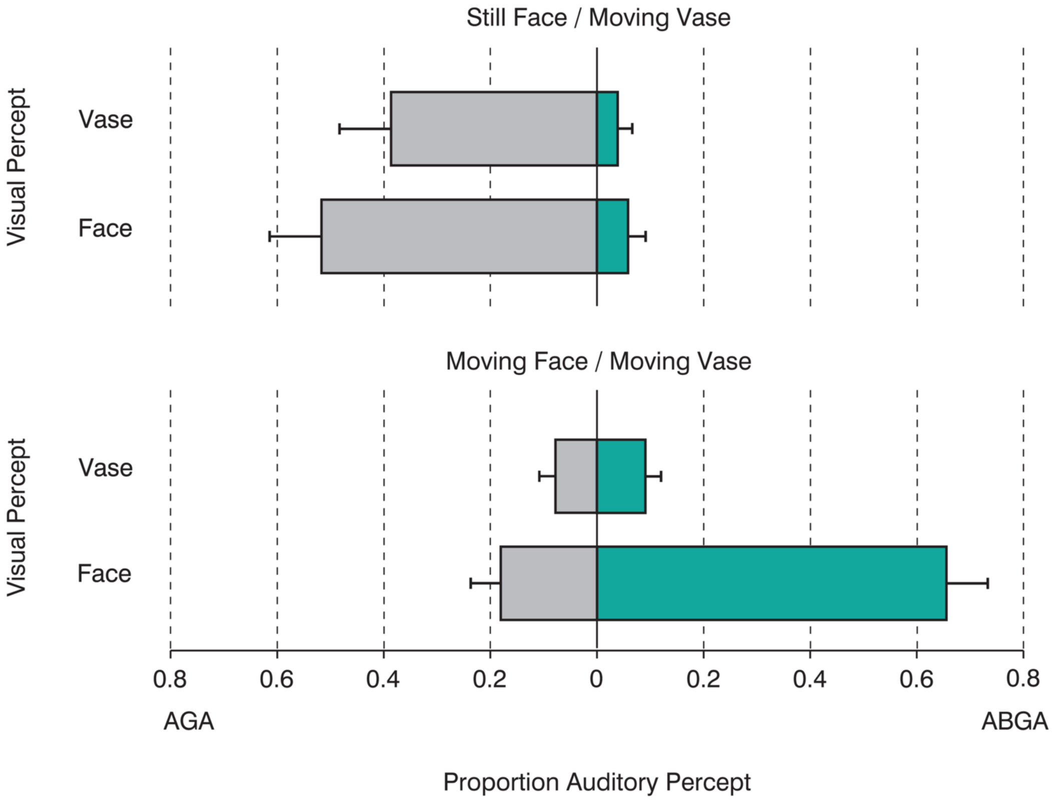

Figure 2.
Proportion of different speech percepts as a function of whether the face or vase was perceived for the still face, moving vase (upper panel) and moving face, moving vase (lower panel) conditions. The proportions are computed separately for the two stimulus conditions and sum to 1.0 for each panel. AGA responses correspond to perception of the sound track while ABGA responses indicate the McGurk combination percept.

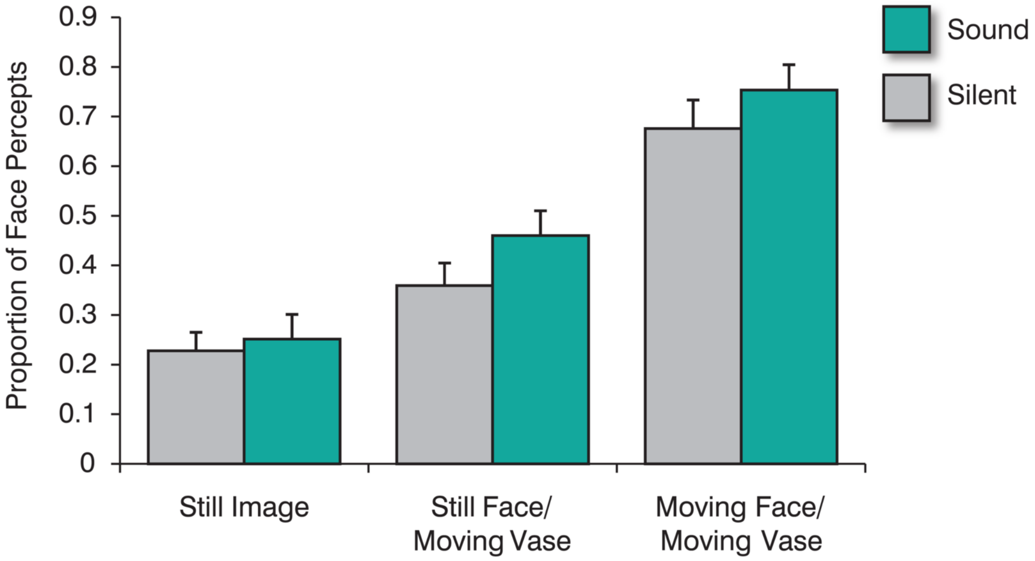

Figure 3.

Proportion of vase percepts reported as a function of visual motion conditions and presence or absence of auditory speech.