Legal and Criminological Psychology (2005), 10, 103–108
© 2005 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

Mixing sound and vision: The interaction of auditory and visual information for earwitnesses of a crime scene

Daniel B. Wright* and Gary Wareham
Psychology Department, University of Sussex, Brighton, UK

Purpose. Previous research has shown that visual information impairs the perception of the sound of individual syllables, often called the McGurk effect. In everyday life sounds are seldom heard as individual syllables, but are embedded in words, and these words within sentences. The purpose of this research is to see whether auditory and visual information interact in the perception of a contextually rich scene that is of forensic importance.

Methods. Participants were shown a video of a man following a woman. The man either says 'He's got your boot' or 'He's gonna shoot'. Half the participants saw the actor say the same phrase as they heard, and half saw a different phrase than they heard.

Results. When the visual and acoustic patterns did not match, people made mistakes. Many reported the fusion: 'He's got your shoe'.

Conclusions. This is the first demonstration of the interaction of auditory and visual information for complex scenes. The scene is one of forensic importance and therefore the findings are of importance within the emerging field of earwitness testimony.

When people perceive a sound, they are influenced not just by the physical sound, but also by visual and contextual information. For example, if you hear someone say the syllable *ba*, but you see them saying *ga*, you may perceive a fusion of these, the sound *da*. This is called the McGurk effect (McGurk & MacDonald, 1976). It is a robust and striking effect. The influence of contextual information on the perception of auditory information is also easily demonstrated by asking a classroom of students: 'How many animals of each kind did Moses take on the ark?' The typical response is two, the so-called Moses illusion (Erickson & Mattson, 1981). Similarly, the visual memories we have of complex scenes can be altered by verbal post-event information (Loftus, Miller, & Burns, 1978; Wright & Loftus, 1998). Here we report a study where participants view a contextually rich scene. For some, the visual and auditory cues are congruent, for others they are incongruent. By using complex scenes our findings are applicable to witness

*Correspondence should be addressed to Dr Daniel B. Wright, Psychology Department, University of Sussex, Falmer, Brighton BN1 9QH, UK (e-mail: DanW@Sussex.ac.uk).

DOI:10.1348/135532504X15240

testimony (see Wright & Davies, 1999, for a general review; see Yarmey, 1995, for a review specifically of earwitness testimony).

Recalling and interpreting what somebody said can be critical for criminal cases. Consider the case of Derek Bentley. In 1952 Derek Bentley and Chris Craig were involved in a robbery when they were confronted by police. According to police officers, Bentley said, 'Let him have it'. The police claimed that this incited Craig, then 16, to shoot and kill Police Constable Sydney Miles. According to this interpretation Bentley was also culpable. Craig was convicted of murder and as a juvenile served 10 years. Bentley, aged 19 but with a mental age of 11, was convicted and in 1953 was the last person to be hanged in the United Kingdom. After a long campaign, in 1998 he was posthumously pardoned. There has been much dispute about what, if anything, Bentley said, but if he had said, 'Let him have it', the meaning is ambiguous. His trial lawyers argued that he was urging Craig to hand the gun over to the police. This would have meant Bentley should not have been convicted.

There are two types of information that are relevant to earwitness testimony: who said the phrase and what was said. Sometimes the police are interested in the identity of speaker. In these cases, the police sometimes produce a voice identification parade. Cook and Wilding (1997a, b) have done much research on this. They find that visual information can interfere with voice identification, what they call the *face overshadowing effect*. The police are also often interested in what was said, which is the focus of this study. As in the Derek Bentley case, often the identity of the speaker is not an issue.

The aim of the current study is to see what occurs when visual and auditory are incongruent in a contextually rich, but ambiguous setting. This is similar to testing whether the McGurk effect, which is normally tested by presenting single syllables, applies to more complex stimuli. The McGurk effect can occur even when the sound and visual information are not well synchronized (180 ms out; Munhall, Gribble, Sacco, & Ward, 1996) and when the visual information is quite coarse (MacDonald, Andersen, & Bachmann, 2000). It has been shown to exist with prelinguistic infants (Rosenblum, Schmuckler, & Johnson, 1997) and in other cultures, although sometimes the effects are smaller (Sekiyama & Tohkura, 1991).

However, there are limitations to the McGurk effect. For example, Walker, Bruce, and O'Malley (1995) have found that the effect is much less pronounced with familiar faces than non-familiar faces. This may be due to there being more meaning in faces of people who we know. Easton and Basala (1982) reported that it was more difficult to detect the effect with meaningful stimuli. However, Deckle, Fowler, and Funnell (1992) did find the effect for words. More recently, Sams, Manninen, Surakka, Helin, and Katto (1998) found the effect with their Finnish sample for words embedded within three word sentences. The current study investigates whether the McGurk effect can occur for a sentence within a complex scene.

Method

Eighty participants, half of whom were male, volunteered for this study. The ages ranged from 9 to 66 years old, with a mean of 26 years. They were recruited from a supermarket and the University of Sussex, and were tested individually. No reliable differences were found by gender, age, or test location.

Four versions of a video were created for this study (see left side of Table 1). They were all identical until the final scene, and all lasted approximately 1 minute. In the early evening, a woman is shown walking along a street. A man is behind her. Thinking that he might be following her she begins to walk faster. He is still behind her. She starts to run and her boot falls off. The man picks it up and runs after her. A male bystander sees this chase and then says either, 'He's gonna shoot'. or 'He's got your boot'. These phrases were designed to be acoustically similar. Each is four syllables in length:

He's gon na shoot
He's got your boot

However, the two have very different meanings. Like Bentley's, 'Let him have it', one interpretation is that the man is committing a crime and one is not. Further, the final words 'shoot' and 'boot', could be fused into 'shoe'. It is acoustically similar to shoot and semantically similar to boot.

Table 1. Responses for what participants thought the bystander said

	Video		Responds with. . .				
Cond.	Heard	Saw	Heard	Saw	Fuse	Other	Total
1	He's gonna shoot	He's gonna shoot	18		2	0	20
2	He's got your boot	He's got your boot	20		0	0	20
3	He's got your boot	He's gonna shoot	15	0	1	4	20
4	He's gonna shoot	He's got your boot	9	0	10	1	20
		Total	–	–	13	5	80

This final scene was shot from approximately one metre away and clearly showed the actor saying the critical sentence. The scene was shot twice. In one the bystander said, 'He's gonna shoot', and in the other he said, 'He's got your boot'. The actor was trained to say these at the same pace to allow them to be synchronized. Four versions were created: the two original ones and two where the visual and acoustic materials were incongruent. Care was taken to synchronize the lip movement and sounds for each syllable.

Participants were randomly allocated to one of four conditions of a 2 × 2 design, with the restriction that each film was seen by 20 people. All participants were tested individually. They were told to watch the film carefully. The very last part of the film was when the bystander said the critical phrase. Immediately (i.e. within a couple of seconds of hearing the phrase) after this, participants were asked, 'Can you tell me what the man at the end of the film said?' Finally, participants were debriefed.

Results

Participants either heard and saw different phrases or the same phrase. If they heard and saw different phrases their responses were coded as one of the following four options: what was heard, what was seen, a *fused response*, or other incorrect response. We defined a fused response as recalling the word 'shoe' as the final word. If the participants heard and saw the same phrase their responses were coded as one of three options: heard and saw, a fused response, or other incorrect response. Fused responses were determined in the same way. Of course these were not fused because the participants heard and saw the same phrase, but the scoring was kept the same for comparison purposes. The right side of Table 1 shows the responses for each condition. Other coding strategies were also used and led to the same conclusions.

Two observations are clear from Table 1. First, the people in the congruent conditions, where they heard and saw the same phrase, were very accurate (95% correct). Second, when the people in the incongruent conditions made mistakes, they never reported what the lips 'said', but mostly reported a fuse, stating that 'shoe' was the last word. Logistic regressions were run to examine these effects further. There were two independent variables (congruent vs. incongruent, heard shoot vs. heard boot), and the dependent variable was correct versus incorrect. The interaction was non-significant ($\chi^2(1) = 0.95, p = .33$) and was removed from the model. Both main effects were significant. People who heard 'shoot' were more likely to make an error ($\chi^2(1) = 5.75, p = .02$, conditional odds ratio = 4.41, with a 95% confidence interval of 1.23–15.80). More importantly, those participants given the incongruent stimuli were much more likely to make an errant response ($\chi^2(1) = 16.61$, $p < .001$, conditional odds ratio = 15.10, 95% CI 3.02–75.54). Thus, visual information interferes with voice perception.[1]

Eighteen people responded incorrectly. Some of these people gave errant phrases for a few parts of the phrase, but every participant who made an error incorrectly produced the final word. This is the word that determines whether the man is attacking the woman or simply returning something of hers. Most of the errors had the word 'shoe' being recalled. However, there were other errors. Three errors were acoustically similar to the correct responses (e.g. 'booze' instead of 'boot'), and two errors were not ('He's got your food' and 'He's got your glue'). 'He's got your glue' was interesting because although 'glue' shares few acoustic or semantic features with 'boot' and 'shoot', it does rhyme with the most popular error, the fusion 'shoe'. This is just a single case and while it deserves mention, further examples are necessary before more is made of it.

Twelve of the erroneous participants also made errors in other parts of the phrase. Three people said 'got your', one said 'got my', two said 'got a', and six said 'got her' in response to having her boot/shoe. There was only one participant who did not get the initial word: 'he's'. This individual said, 'Is that your shoe?'.

Discussion

Auditory and visual information can interact within a complex scene. While this has been shown for individual syllables in the McGurk effect, in this study we demonstrated that visual information interferes with the perception of complex phrases embedded in a forensically relevant situation. It was not just that the visual information interfered with perception, but that the majority of errors were a fusion of two words. Interestingly, the fusion was between the acoustic characteristics of one word, 'shoot', and the semantic characteristics of another word, 'boot'. This fusion tended to occur when the participant heard 'shoot' but saw 'boot'. Calvert *et al.* (1997) have shown that lip

[1] *The analyses were repeated without the non-fused errors. The conclusions are the same. The conditional odds ratio for which word they heard is 19.24 with a 95% confidence interval of 2.20–168.36. The conditional odds ratio for being presented with congruent or incongruent stimuli is 10.88 with a 95% confidence interval of 2.02–58.67.*

reading and auditory perception are processed by some of the same brain regions, but that each evokes activity in other regions. It may be that the visual material activates the semantic meaning of the word more than the acoustic material does. This is worth further investigation.

As stated earlier, one of our reasons for conducting this research was because we wanted to see the extent that non-acoustic information could influence the perception of speech. This is critical for many criminal cases. In the condition where the effect was most pronounced, participants heard the word 'shoot', but they saw the lips say 'boot' and onscreen, the male actor was shown carrying a boot. The visual and contextual information led to half of the participants responding 'shoe'. These participants were not only using contextual information. If that were the case, the participants in the other three conditions would have also frequently recalled 'shoe'.

The choice of stimuli was purposeful. We knew that the word 'shoe' related to both of the critical words, but in different ways. The particular phrases were chosen to show that visual and auditory information can interact. With other phrases the interaction may fuse in less predictable ways, depending on the particular characteristics of each phrase.

This study has important implications for assessing the accuracy of earwitness testimony. It is important to realize that perception of complex sounds is influenced both by other sensory input and by the situation within which the utterance is heard. More research is necessary in trying to understand earwitness testimony (Wilding, Cook, & Davis, 2000). In this study, the focus was on what was said rather than who said it. Subtle differences between words can cause ambiguity, and this could potentially be the difference between whether an individual is guilty or not guilty of a crime. This research shows that errors that people hear can be easily created by visual information.

References

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*, 593–596.

Cook, S., & Wilding, J. (1997a). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology, 11*, 95–111.

Cook, S., & Wilding, J. (1997b). Earwitness testimony. 2.: Voices, faces, and context. *Applied Cognitive Psychology, 11*, 527–541.

Deckle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception and Psychophysics*, *51*, 355–362.

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception and Psychophysics*, *32*, 562–570.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*, 540–551.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19–31.

MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: How much spatial degradation can be tolerated? *Perception*, *29*, 1155–1168.

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

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception and Psychophysics*, *58*, 351–362.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception and Psychophysics*, *59*, 347–357.

Sams, M., Manninen, P., Surakka, V., Helin, P., & Katto, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, *26*, 75–87.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, *90*, 1797–1805.

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception and Psychophysics*, *57*, 1124–1133.

Wilding, J., Cook, S., & Davis, J. (2000). Sound familiar? *Psychologist*, *13*, 558–562.

Wright, D. B., & Davies, G. M. (1999). Eyewitness testimony. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of applied cognition* (pp. 789–818). Chichester, UK: John Wiley & Sons.

Wright, D. B., & Loftus, E. F. (1998). How misinformation alters memories. *Journal of Experimental Child Psychology*, *71*, 155–164.

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law*, *4*, 792–816.

Received 13 October 2003; revised version received 25 February 2004