Visual influence on auditory perception: Is speech special?

Christian Kroos, Katherine Hogan

MARCS Auditory Laboratories, University of Western Sydney, Australia
c.kroos@uws.edu.au

Abstract

Few studies have investigated whether visual perception can influence auditory perception outside the speech domain in a way comparable to the McGurk effect. Here we used common environmental events, a wooden and a metal spoon hitting a granite surface, to test for a change in auditory perception of the spoon's material induced by an incongruent video. To obtain a sensitive measure and avoid directing the participants' attention towards the stimulus manipulation, an AXB matching task was employed where in the target trials all three auditory stimuli were the same, while on the visual side the X stimulus was presented with no video and A and B were presented with a control video and a video either showing the wooden or the metal spoon. No visual influence was found. A second even more sensitive test investigated whether two neighbouring stimuli of an acoustic wood-to-metal continuum that were found to be auditorily non-discriminable became discriminable when one of them was enforced with the appropriate visual stimulus. Again, no effect was found. We interpreted the results as showing that the McGurk effect is limited to vocal-tract-like events.

Index Terms: cross-modal perception, McGurk effect, non-speech events

1 Introduction

This study is the first in a series of experiments investigating the longstanding yet unanswered question: Is the visual influence exerted on auditory perception that is well-documented in speech confined to the specific (human) activity of spoken language or is it the manifestation of a general mechanism of combining auditory and visual information? The question requires a more precise definition of visual influence for the purpose of the study. Here we were concerned with a change in identification of a perceived event through the addition of a visual signal as demonstrated by the the well-known McGurk effect [1, 2]. In other words, is there a McGurk-like effect for non-speech sounds?

There is indeed evidence for the existence of a more general perceptual mechanism of auditory-visual interference or even integration. For instance, in the so-called *ventriloquist* effect the detection of the spatial location of an auditory stimulus is displaced toward the location of a synchronously presented visual stimulus [3]. Similarly, the length of a musical note produced with a mallet on a marimba is perceived as longer as its objective acoustic duration when the visual stroke gesture indicates a longer note [4]. A direct comparison of speech and non-speech sounds was conducted by Rosenblum and Fowler [5], based on the speech-loudness-vocal-effort hypothesis in speech and hand clapping with different strengths as the complement in the non-speech domain. The results of the non-speech condition showed an overall significant effect of the visually perceived effort, implying that the visual information affected the auditory perception of loudness. However, the effect was more pronounced in speech.

The studies cited above showed that visual information is able to bias the auditory perception of *properties* of the investigated sounds, but they did not show this for the identification of the multi-modal events as such. In the non-speech domain the researcher is faced with the problem that in most languages relatively few labels for environmental sounds exist that do not refer to the whole multi-model event, but pertain only to the associated sound. As a consequence, it becomes necessary to explicitly ask the participant in a perception experiment to respond only according to what they hear while at the same time instruct them to still pay attention to the visual stimulus thereby making it likely that the participants become aware of the nature of the experiment. The problem could not be avoided in the ground-breaking study of Saldaña and Rosenblum [6] aimed at determining whether a complement to the McGurk effect could be found in the non-speech domain. They dubbed sounds from an artificially created continuum between a cello pluck and bow sound onto video clips showing a human arm and hand bowing or plucking a cello string. They found significant effects of the video display on the ratings of the sounds along the pluck-bow continuum, however, the effect was weaker and qualitatively different from the also tested speech McGurk effect. In contrast to the classic McGurk effect in speech, the participants were also able to detect the audiovisual incongruency of the non-speech stimuli. The pluck and bow stimuli exhibit a noticeable asymmetry: Though the participants were able to distinguish the two sounds in an auditory-only condition, participants without musical training can be assumed to be much less familiar with them compared with speech sounds. On the other hand, the visual stimuli was unmistakably clear, because of the presence or absence of the bow. Accordingly, the observed effect might have been the consequence of a post-perceptual cognitive process, that is, that despite being instructed to base their judgement on what they heard, the participants might have based their decision - when in doubt - on what they saw. Furthermore, the pluck and bow sounds lack ecological validity.

Brancazio, Best and Fowler [7] examined the role of the phonological significance of the stimulus in the McGurk effect. They used an AXB matching task and included classic McGurk stimuli (cross-dubbed /pa/ and /ta/) as well as bilabial, dental and lateral clicks. Though common in some African languages, clicks are perceived by native English listeners as non-speech sounds. Although the observed effect was weaker for the clicks than for the English stop consonant syllables, the results still provided evidence that phonological significance of the auditory-visual stimulus might be dispensable to elicit the McGurk effect. Clicks are, however, still vocal-tract events (see section 4. Discussion).

Taking the earlier studies into consideration we listed the following essential requirements for potential auditory-visual stimuli to be used in this study:

- Ecological validity;
- Strong connection between visual and auditory part in natural environment;
- High familiarity;
- Acoustic properties that remotely resemble a speech syllable consisting of an initial consonant followed by vowel.

A wooden and metal object hitting a stone surface fulfilled all the criteria, even more so in the case of an everyday object like a spoon being struck by human hand against the surface.

To avoid inadvertently directing the participants' attention towards the discrepancy between auditory and visual stimulus we chose an AXB matching task similar to the one used in [7] and added a simple secondary visual-only detection task (a cross appearing for a brief moment toward the end of the video).

Despite the above outlined methodological criticism of Saldaña and Rosenblum's study [6] we hypothesised to find in accordance with their results and the results of Brancazio et al. [7] a clear visual influence in the non-speech domain.

2 Method

2.1 Stimuli

The target stimuli were recorded in an adequately lit sound proof booth with a plain off-white background using a Sony NP-F330 digital video camera. An 18mm thick granite slab was placed on a simple plastic table 2.5 meters away from the camera. The recorded action consisted of a hand-held wooden or metal spoon, approximately 30cm in length, held briefly at a height of about 25cm above the granite slab, then swinging a few centimetres upwards, reversing the movement direction and finally hitting the granite slab. The frame was chosen to show the actor's hand and part of the forearm to identify her as responsible for the striking action and thus making it obvious to the perceiver that the recorded action consisted of biological motion. The spoon was turned in an angle of approximately 45 degrees to allow the observer to see the spoon clearly and instantly recognise the material (wood or metal). A control video stimulus was recorded outdoors with the same camera. It displayed a tree swinging gently in a light breeze. See Figure 1 for still images extracted from the video clips.

The audio tracks were recorded together with the video using a Bruel & Kjaer 2671 microphone positioned on the left side just outside the visual field of the camera (except for the control video, where no sound was recorded). The acoustic signal from the microphone was fed directly to the camera's audio input and digitized with a sample rate of 48 kHz by the in-built AD-converter of the camera. The gain was fixed and not adjusted for the different material types to preserve any natural loudness differences due to material or strength of striking movement.

From the multiple recordings being made, the ones which showed the spoon most clearly and were similar in timing and perceived hitting strength were selected to become the final tokens. Their acoustic tracks became the endpoints of an artificial wood-to-metal continuum (see Figure 2 and Figure 3 for their respective spectrograms). A fine-grained 50-step wood-to-metal continuum was created by importing the sound tracks into Matlab (The MathWorks, Inc.), aligning the two signals via their impulse-shaped acoustic onsets (spoon hitting stone surface) and mixing them after modifying their respective average amplitudes. Mixing rather than morphing was chosen to maintain the real world significance of the stimuli as morphed sounds might have no real world complement. Almost identical durations of the two selected hitting sounds made any temporal manipulation before the mixing unnecessary. The resulting sounds were labelled according to how much metal sound they contained, i.e., the unmodified wood sound received the label 'WM000', the unmodified metal sound 'WM100', while e.g., 'WM020' denoted a mix of the wood sound at 0.8 of its average amplitude and the metal sound at 0.2 of its average amplitude.

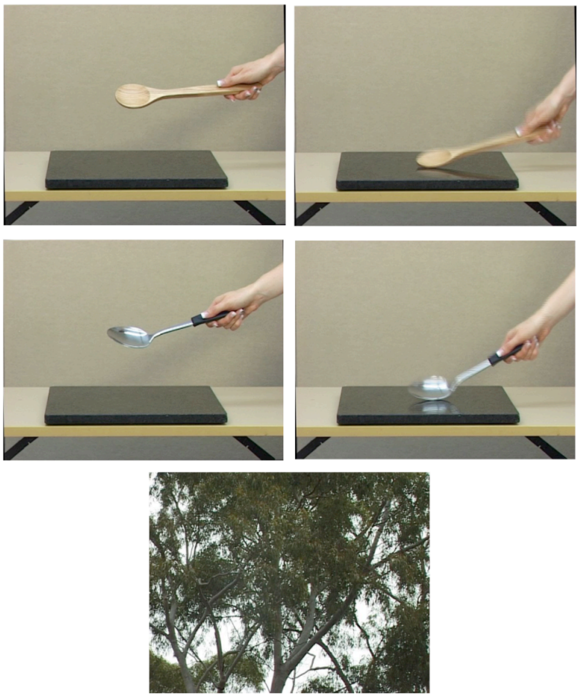

Figure 1: *Still images from the video clips. Top row: wooden spoon (beginning and impact). Centre row: metal spoon (beginning and impact). Bottom row: control video showing a tree.*

With two pre-tests the perceptual 25%, 50%, and 75% points on the continuum were determined. In an AXB matching task, four participants were required to rate whether the target sound X was more similar to sound A or B, which consisted of the unmodified endpoints of the continuum. The target sound was presented 20 times in a fully counterbalanced design. In the first pre-test the wood-metal continuum spanned 11 steps (including the endpoints) to narrow down the region where a change of perception occurred. In the second pre-test 21 steps were employed ranging from WM020 to WM060. The results showed a steep curve with the participant changing from selecting wood to metal within a tight interval. The participants performed at chance level (50%) for sound WM038 and rated the target sound more similar to the metal sound 25% of the time at sound WM032 and 75% of the time at sound WM044. Note that the absolute numbers have lit-

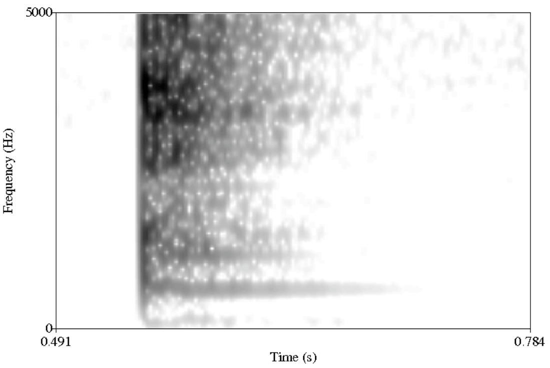

Figure 2: *Spectrogram of the sound produced by the wooden spoon hitting the granite surface.*

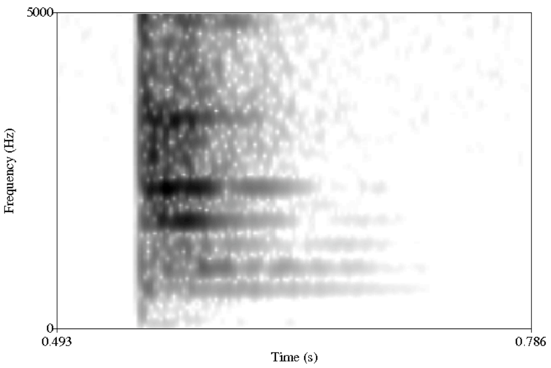

Figure 3: *Spectrogram of the sound produced by the metal spoon hitting the granite surface.*

tle meaning as the average loudness of the two sounds was left unadjusted to maintain the inherent relation to perceived visual hitting strength and material properties. Furthermore, for the purpose of the main experiment the perceptual points needed to be determined only approximately.

We will denote the final five sounds as WM-0% (pure wood sound), WM-25%, WM-50%, WM-75%, and WM-100% (pure metal sound). They were dubbed on all three video stimuli (wooden spoon, metal spoon, tree) using Adobe Premiere Pro 2.0. For the target stimuli they were precisely aligned using the synchronous original sound track as reference, for the control video the time for the impact sound was chosen to match the one of the target stimuli relative to the beginning of the clip.

The whole stimulus set was copied and a red cross sign was added to the video clips using Adobe Premiere's title generator appearing for the duration of one frame (40ms) in the spatial centre of the frame towards the end of the clip. This was put in place to ensure that the participants would attend to the visual signal despite the primary matching task consisting of an auditory comparison.

2.2 Participants

Thirty participants (27 female, mean age 21.2, range 18 to 40 years of age) took part in this study. All were undergraduate students from the University of Western Sydney who participated for course credit. They were required to have normal or corrected-to-normal vision and no hearing impairments.

2.3 Procedure

Participants were tested in a cross-modal, two-alternative, forced-choice matching task. On each trial, the participants viewed a set of three stimuli sequentially as part of a counter-balanced AXB design. The A and B stimuli were presented audiovisually while the X stimulus consisted of the acoustic signal only. The participants were required to judge whether the A or B stimulus auditorily more closely resembled the X stimulus. In the AXB triads the acoustic signals were either the same in all three stimuli, or one of the A or B stimuli was different by being one step apart from the other two on the continuum axis, e.g., if the A and X stimuli consisted of WM-50% then the B was WM-75% or WM-25%. The former triads were to be used in the final analysis while the latter were made part of the experiment to introduce small but real (that is not visually induced) auditory differences in order to avoid transparency of the research aim to the participants. On the visual side, either A or B always consisted of the control video (tree). An additional set of AXB triads consisted of the same auditory stimuli combinations as above but both videos, that of the A and the B stimulus, displayed the control video.

Note that a typical non-control trial of the current experiment would have the following complement in an analogous experiment in the speech domain (only one example given): /ba/ as the auditory-only X stimulus, auditory /ba/ dubbed on an unrelated video as the A stimulus and auditory /ba/ dubbed on the video of a speaking face uttering /ga/ as the B stimulus. As in this case the B stimulus would be perceived as /da/ due to the McGurk effect, the A stimulus unaffected from the unrelated video would consistently be rated as more similar to the X stimulus.

In a secondary task the participants had to indicate whether or not they had perceived the cross appearing in one third of the trials.

Participants were individually tested in a sound proof booth, seated on a chair with an approximate viewing distance of 50cm to a cathode-ray tube monitor displaying the stimuli. Two loudspeakers placed behind the monitor converted the stored acoustic signals to audible sound waves. The participants responded by pressing the appropriate button on a button box and proceeded to the next trial by pressing a foot pedal. The experiment was run using the experiment control software DMDX [8] which was responsible for all stimulus presentations as well as registering the participants' responses. Three repetitions of the stimulus set described above were presented (one of them containing the trials with the cross) in fully randomised trial order.

3 Results

One participant had to be excluded from the analysis as her responses suggested that she had not paid attention to the visual stimuli (she had answered to have seen the cross in all trials). The responses of the remaining 29 participants were for each participant individually converted to percentage scores of how many times the target video (wooden spoon, metal spoon) was chosen

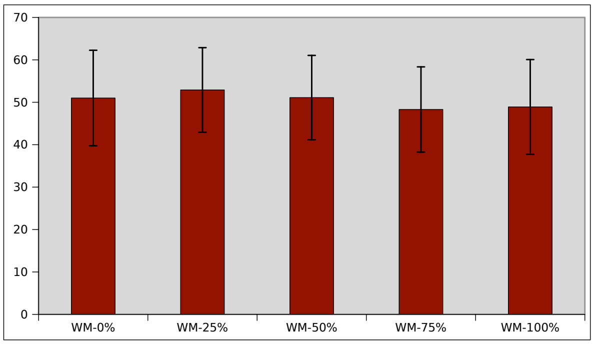

Figure 4: *AXB matching tasks with same auditory stimulus (along X axis) in A, B and X, but different visual stimuli: Means and standard deviations (shown as error bars) of the frequency (in percent) with which the sound presented with the WOOD video was chosen to be more similar to the auditory-only X stimuli than the sound presented with the control video (tree). Chance level: 50%.*

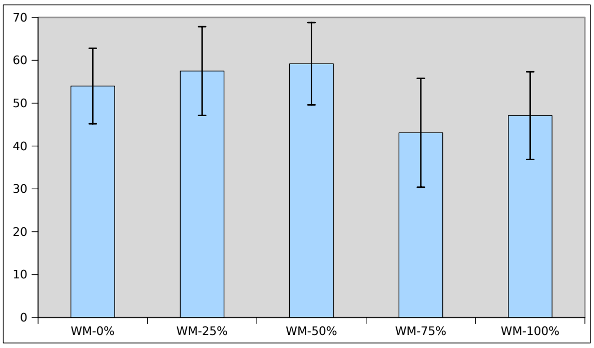

Figure 5: *AXB matching tasks with same auditory stimulus (along X axis) in A, B and X, but different visual stimuli: Means and standard deviations (shown as error bars) of the frequency (in percent) with which the sound presented with the METAL video was chosen to be more similar to the auditory-only X stimuli than the sound presented with the control video (tree). Chance level: 50%.*

as opposed to the control video (tree) across the tested six repetitions (three identical repetitions plus counterbalanced trials). Note that in the trials included in this analysis, the A, B, and X stimuli always consisted of the same auditory stimulus. Therefore, if a visual influence indeed exists, changing the perception of the auditory stimulus, the auditory stimulus presented with the target video should have been significantly less often chosen as being similar to the X stimulus than the stimulus presented with the control video.

Means and standard deviations across participants are shown in Figure 4 for the target videos showing the wooden spoon (WOOD) and in Figure 5 for the target videos showing the metal spoon (METAL), each dubbed with all five auditory stimuli.

As a first step, it was investigated whether there had been a response bias caused by the fact that the target visual stimulus was in an obvious way related to the auditory stimulus while the control stimulus was not, i.e., whether the participants had chosen the WOOD or METAL video as being more similar to the auditory-only X stimulus, solely because it displayed a hitting action. For this purpose, the results from trials, where the pure wood or metal sounds were presented combined with their original video as target video, were compared to chance level (50%) employing a one-sample t-test. With α set to 0.05 a statistically non-significant t-value for choosing the target video over the control video was found in these trials both for WOOD ($t(28) = 0.25$, $p = 0.81$) and METAL ($t(28) = -0.76$, $p = 0.46$). The difference between the WOOD mean and chance level was 1.03 and the 95% confidence interval extended from -7.53 to 9.60; the difference between the METAL mean and chance level was -2.78 and the 95% confidence interval extended from -10.65 to 4.90. Taken together the results show that no response bias existed that could have modified the participants' judgement based on the video alone without exerting any influence on the auditory perception of the stimulus. Nevertheless, in order to follow a maximally strict procedure the means and variances of these trials were used as baseline to which the remaining trials were compared instead of simply comparing them to chance level.

Using a GLM, repeated measures contrasts were implemented comparing the baseline with the results from the other AXB trials to determine whether the video type changed the perception of the auditory tokens. Specifically, the AXB trials, which were composed of the auditory tokens WM-25%, WM-50%, WM-75%, and WM-100% (pure metal) and the WOOD video were compared to the WOOD baseline. Similarly, the AXB trials which were composed of the auditory tokens WM-0% (pure wood), WM-25%, WM-50%, WM-75% metal) and the METAL video were compared to the METAL baseline. Table 1 shows the resulting statistics. Albeit the implemented contrasts constituted planned comparisons, they were not orthogonal and therefore α was Bonferoni-adjusted to 0.0125 to account for multiple comparisons (four per video type). None of the contrasts reached statistical significance. Note, however, that one contrast (METAL video, WM-50% vs. WM-100%) would have reached statistical significance with an unadjusted α of 0.05. We therefore decided to cross-check the observed null result with a more sensitive test.

The AXB trials in which both videos consisted of the control video but the auditory stimulus differed in A and B allowed the assessment whether neighbours on the auditory continuum, e.g., WM-50% and WM-75%, could actually be distinguished by the participants. One-sample t-tests comparing the response percentages to chance level (50%) showed that the participants were able to discriminate between WM-0% (pure wood) and WM-25% (X was WM-0%: $t(28) = -6.81$, $p = 0.00$; X was WM-25%: $t(28) = -3.05$, $p = 0.05$) as well as between WM-75% and WM-100% (pure metal)(X was WM-75%: $t(28) = -3.63$, $p = 0.01$; X was WM-100%: $t(28) = 8.94$, $p = 0.00$), but not within the two remaining neighbouring steps, WM-25% vs WM-50% (X was WM-25%: $t(28) = -0.89$, $p = 0.38$; X was WM-25%: $t(28) = 0.37$, $p = 0.71$) and WM-50% vs WM-75% (X was WM-25%: $t(28) = 0.84$, $p = 0.41$; X was WM-25%: $t(28) = 0.74$, $p = 0.47$). Accordingly the latter two auditory differences were selected as the base for a more sensitive test as described in the following.

Assuming that there had been a very weak influence of the video on the auditory perception (leading to the results in the main test where with the METAL video WM-50% would have become

Table 1: *AXB matching tasks with the auditory stimulus being the same for A, B and X: Repeated measures contrasts (implemented through a GLM) comparing frequencies with which the sound presented with the WOOD or METAL video was chosen to be more similar to the auditory-only X stimuli than the sound presented with the control video (tree). For the WOOD video WM-0% constitutes the baseline (pure wood sound), for the METAL video WM-100% (pure metal sound). The performance on the baseline is compared with the performance on the other auditory stimuli. The degrees of freedom for the given F* value are $F(1, 28)$.

WOOD video				
Auditory token	F	p	Conf. interval	
			Lower	Upper
WM-25% vs. WM-0%	0.19	0.67	-10.44	6.76
WM-50% vs. WM-0%	0.00	0.98	-9.58	9.35
WM-75% vs. WM-0%	0.32	0.58	-7.23	12.74
WM-100% vs. WM-0%	0.14	0.71	-9.60	13.97
METAL video				
Auditory token	F	p	Conf. interval	
			Lower	Upper
WM-0% vs. WM-100%	2.07	0.16	-16.71	2.91
WM-25% vs. WM-100%	3.04	0.09	-22.49	1.80
WM-50% vs. WM-100%	5.34	**0.03**	-22.77	-1.37
WM-75% vs. WM-100%	0.58	0.45	-6.78	14.83

statistically significant with an unadjusted α), it stands to reason that the auditory stimuli must have sounded more ‘metallic’ due to the visual stimulus. Indeed, looking at the mean confirms that in these cases the auditory stimulus of the target video was less often perceived to be similar to the X stimulus than the the auditory stimulus of the control video, i.e., it was perceived as being different. Thus in the situation where the A/B and X auditory stimuli are comprised of the non-discriminable neighbouring sounds they should become discriminable if the one auditory stimulus that contains more metal sound in its mix is combined with the METAL video. All that is required is that the video emphasises the prominence of the congruent part in the sound mix. For the sake of completeness we tested both directions, that is, whether the METAL or the WOOD video would facilitate discrimination if combined with the auditory stimulus that contained more of its respective accompanying sound.

Figure 6 shows means and standard deviations for the four possible cases and Table 2 gives the statistics for the four one-sample t-tests comparing the results against chance level. As can be seen none of the comparisons yields a statistically significant difference. Furthermore, a look at the confidence intervals for the these test confirms that the mean in the population is with 95% certainty between -10.9 and 11.53 indicating that humans would in the ‘worst’ case (if the true mean would be close to the boundary of the confidence interval) on average deviate from an equal distribution in their answers (chance level) only once over ten repetitions. Consequently, even in the strongest case their discrimination ability would be very poor, implying that even if there was a visual influence on the auditory percept, it would be of no behavioural relevance.

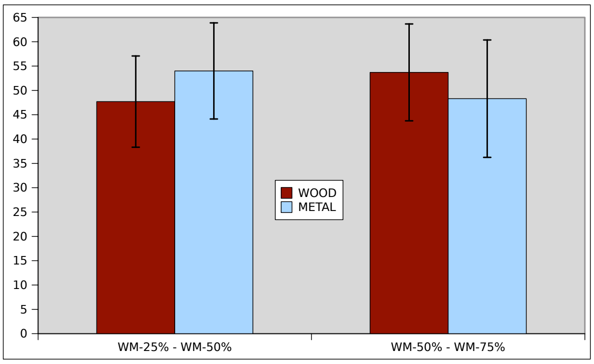

Figure 6: *Discrimination of two neighbouring sounds from the wood-to-metal continuum as revealed by the AXB matching task: Means and standard deviations (shown as error bars) of the frequency (in percent) with which the sound presented with the METAL or WOOD video was chosen to be more similar to the auditory-only X stimuli than the sound presented with the control video (tree). For the WOOD video (dark-red bars) the auditory stimuli consisted of WM-25% (presented with WOOD video) and WM-50% (presented with the control video), shown on the left hand side, and WM-50% (presented with WOOD video) and WM-75% (presented with the control video), shown on the right hand side. For the METAL video (light-blue bars) the auditory stimuli consisted of WM-25% (presented with the control video) and WM-50% (presented with the METAL video), shown on the left hand side, and WM-50% (presented with the control video) and WM-75% (presented with the METAL video), shown on the right hand side. Chance level: 50%.*

4 Discussion

Taken together the results provide strong evidence against the existence of McGurk-like effect with regard to non-speech environmental events. Obviously, generalisations of statements about the non-existence of a phenomenon have to be considered tentative, as it is not feasible to test all possible cases and conditions. We argue, however, that the results of this study can be generalised to a large degree due to the careful selection of the stimuli and the sensitivity of the employed task. Our findings are in contrast with [6], adding weight to the critique based purely on methodological considerations that we brought forward in the Introduction. The findings also seem to be at odds with the results from [7], despite that we were using a similar methodological approach. Closer examination, however, shows that this is not the case. Though perceived as non-speech by native English listeners, clicks are still vocal-tract events and that might be the crucial difference with respect to a McGurk-like effect.

We propose the following distinction to be made in addressing the generalisibility of the McGurk effect:

- Speech events;
- Vocal-tract events (e.g., throat clearing, laughter, speech events that are not perceived as such);
- Events involving sound production and modification in a way closely or remotely resembling a vocal-tract (e.g., frog croaking, steam whistle);
- Other events.

Table 2: *Results from one-sample t-tests comparing the performance in the AXB matching task with regard to two neighbouring auditory stimuli from the wood-to-metal continuum to chance level performance (50%). For details see text and Figure 6.*

WOOD video				
Auditory tokens	$t(28)$	p	Confidence interval	
			Lower	Upper
WM-25% - WM-50%	-0.66	0.52	-9.43	4.83
WM-50% - WM-75%	1.00	0.33	-3.89	11.25
METAL video				
Auditory tokens	$t(28)$	p	Confidence interval	
			Lower	Upper
WM-25% - WM-50%	1.01	0.28	-3.48	11.53
WM-50% - WM-75%	-0.39	0.70	-10.90	7.45

Any of these categories might provide different results in experiments investigating auditory-visual interactions. In fact, the question whether speech is special with regard to visual influence on auditory perception might be much more complex to answer and might be tied in with the problem of defining, explaining and categorising speech in general. Even on the perceiver side it appears to make a difference whether the perceiver considers the *same* stimuli as speech or non-speech: [9] showed that auditory-visual integration of sine wave speech stimuli dubbed on a speaking face depended on whether or not the participants interpreted it as speech.

5 Conclusion

An AXB matching task was used to investigate whether visual perception is able to influence auditory perception to change the quality of the heard sound of an incongruent auditory-visual stimulus. Ecologically valid and highly familiar common environmental events, a wooden and a metal spoon hitting a granite surface, were selected to create this stimuli for the study. In the target trials all three auditory stimuli were the same, while on the visual side the X stimulus was presented with no video and A and B were presented with a control video and a video either showing the wooden or the metal spoon. No visual influence could be detected. A second test, considered to be more sensitive, was based on two pairs of neighbouring stimuli of an acoustic wood-to-metal continuum that were found by examining a control condition to be non-discriminable for the participants. Even when either the A or B stimulus was combined with a congruent video while the other was combined with the control video, the visual target stimulus did not shift auditory perception towards the event it was showing and the two sounds remained non-discriminable. The results are interpreted as evidence that a McGurk-like effect is limited to vocal-tract events or events that at their source resemble a vocal-tract.

In future studies we will further narrow down the essential conditions for a McGurk-like effect to happen outside its known domain of speech syllables.

6 Acknowledgements

The authors would like to thank Chris Davis and Jeesun Kim for many helpful comments. The first author wishes to acknowledge the support of NHMRC/ARC grant TS0669874.

References

[1] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, 1976.

[2] J. MacDonald and McGurk, “Visual influences on speech perception processes,” *Perception and Psychophysics*, vol. 24, no. 3, pp. 253–257, 1978.

[3] P. Bertelson and G. Aschersleben, “Automatic visual bias of perceived auditory location,” *Psychonomic Bulletin & Review*, vol. 5, no. 3, pp. 482–489, 1998.

[4] M. Schutz and S. Lipscomb, “Hearing gestures, seeing music: Vision influences perceived tone duration,” *Perception*, vol. 36, no. 6, p. 888 897, 2007.

[5] L. D. Rosenblum and C. A. Fowler, “Audiovisual investigation of the loudness-effort effect for speech and nonspeech events,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 17, no. 4, pp. 976–985, 1991.

[6] H. M. Saldaña and L. D. Rosenblum, “Visual influences on auditory pluck and bow judgments,” *Perception and Psychophysics*, vol. 54, no. 3, pp. 406–416, 1993.

[7] L. Brancazio, C. T. Best, and C. A. Fowler, “Visual influences on perception of speech and nonspeech vocal-tract events,” *Language and Speech*, vol. 49, no. 1, pp. 21–53, 2006.

[8] K. L. Forster and J. C. Forster, “A Windows display program with millisecond accuracy,” *Behavior Research Methods, Instruments, & Computers*, vol. 35, pp. 116–124, 2003.

[9] J. Tuomainen, T. S. Andersen, K. Tiippana, and M. Sams, “Audio-visual speech perception is special,” *Cognition*, vol. 96, no. 1, pp. B13–B22, 2005.