WERE THOSE COCONUTS OR HORSE HOOFS? VISUAL CONTEXT EFFECTS ON IDENTIFICATION AND PERCEIVED VERACITY OF EVERYDAY SOUNDS

Terri L. Bonebright

Department of Psychology
DePauw University
tbone@depauw.edu

ABSTRACT

This study examined whether visual context has an effect on the identification of everyday sounds. Scenes portraying actions that lead to everyday sounds were paired with the actual sounds, acoustically similar sounds and acoustically contrasting sounds Participants identified sounds, rated their confidence on the identifications, the veracity of the sounds and their familiarity with the sounds. Results showed that participants identified the actual and contrasting sounds correctly more often than the similar sounds, which were frequently incorrectly identified as the sound that occurred from the action in the visual scene. However, the confidence ratings for the identifications were lower for the similar sounds, and they rated them as less realistic than the actual sounds. Thus, even though similar sounds were frequently misidentified as the actual sound taking place in the scene, participants did recognize that such sounds were not quite correct for the visual action being portrayed.

1. INTRODUCTION

Watching movies is a favorite pastime for many people, most of whom readily accept the premise that the visual scene and the accompanying soundtrack, including the ambient sounds from the environment, were recorded simultaneously. In many cases, however, the visual tracks are recorded separately from the audio, and many of the sounds, especially the background noises, are recorded by producing sounds from objects other than the ones seen in the video [1]. Some of these sound effects are synthesized or sampled recordings while others are produced by Foley artists, who use a variety of different objects to produce sounds for the background sound track. The desired result is to produce a sound track that the movie viewer will perceive as realistic, regardless of what is actually used to produce a given sound. One example of a sound effect produced by Foley artists that movie watchers may be aware of is the use of halves of coconuts clapped together to create the sounds of horses galloping over the landscape. Foley artists routinely manipulate a number of objects to produce sounds for entirely different actions, such as crinkling cellophane for the sound of a fire crackling or breaking stalks of celery for the sound of bones breaking. This has led some filmmakers to argue that viewers have been conditioned by the media to expect "real" sounds that are not encountered in a natural environment [1].

Another newer application of sound effects to create a realistic experience is found in the development of virtual environments [2]. Researchers in this area have found that realistic 3-D sound environments can be produced using HRTF-constructed stimuli [3] and that synchrony between the sounds and visual stimuli is critical for realistically perceived sounds [4].

Since these examples suggest that listeners can be fooled into perceiving such sounds as realistic, it is important to determine whether people are able to correctly identify everyday sounds when they are presented without any accompanying visual stimuli. Researchers have shown that people are quite good at this in general [5, 6, 7, 8, 9], and that when they make misidentification errors, they are typically made with sounds that are acoustically similar.

Studies have also been performed to help determine if context can have an impact on everyday sound identification. Ballas and Mullins [10] and Gygi and Shafiro [11] showed that sounds embedded within a sequence help identification rates if they are semantically similar. For example, people are better at identifying the sound of a stapler if the preceding sound was a typewriter. Context has also been shown to provide enhancement for identification of visual objects within a scene [12,13,14]. However, the intermodal effects of sound and visual stimuli have not been investigated systematically in the same way. The exception to this are studies using speech that show that visual and auditory stimuli combine to produce interactive effects, such as the McGurk effect [15,16], the freezing effect [17], and the ventriloquist effect [18].

The purpose of the present study was to examine the effect of visual scenes with staged actions with objects that result in everyday sounds on the identification of those sounds. The scenes were paired with the actual sounds made by the objects, acoustically similar sounds to those made by the objects, and contrasting sounds that were acoustically dissimilar to those made by the objects. The responses collected from the participants after exposure to the sound/video combinations were identifications of the sounds, confidence ratings of those identifications, and ratings of veracity of the sounds. In addition, participants rated the familiarity of the sounds using a written list (see Table 1).

Four hypotheses were proposed based on the previously reviewed literature. First, it was expected that the visual context would affect the identifications of the sounds such that the actual and contrast sounds would be more likely to be correctly identified compared to the similar sounds. This would be the case if the acoustically similar sounds were confused with the actual sounds as suggested from previous research [5, 6, 7, 8, 9], and if the effect of the visual scene was not strong enough to override the perception of the acoustically contrasting sound.

For example, it would be expected that a person would incorrectly identify Velcro ripping as paper being torn while watching a person tearing paper. However, it would *not* be expected that someone hearing a foghorn would mistakenly identify this sound as a telephone ringing, even if the visual scene displayed a person answering a telephone. Second and third, confidence ratings of the identifications and the veracity ratings of the sounds were expected to be highest for the actual and similar sounds compared to the contrast sounds. Such results would occur if the visual context impacted and biased the perception of the listener [15, 16, 17, 18]. For example, if listeners are swayed by the visual context and use it help identify the actual and similar sounds, their confidence in their identification and perception of realism should be high. However, if the sounds perceptually mismatch with the visual scene, there should be an impact on the confidence and assessment of the overall realism resulting in lower ratings for both, even though the sound may be correctly identified. Finally, the familiarity ratings for the sounds were expected to be correlated with the number of correct identifications since actual experience with sounds should assist the ability to label them.

2. METHOD

2.1 Participants

There were 45 undergraduate students (31 female and 14 male) who participated in the study for extra credit for psychology courses. The mean age was 20.71 years, and the range was 18 to 22 years and the majority (95%) of them were Caucasian. All participants reported normal hearing and either normal or corrected-to-normal vision. Thirty-five participants completed the sound/video condition; 10 completed the sounds-only control condition.

2.2 Apparatus

The scenes were filmed using a Canon GL1 digital video camcorder. An Audio-technica MB 4000C microphone was used to record the auditory stimuli that were recorded by the experimenters. The video and audio recordings were edited using FinalCut Pro 4.0. The final videos were presented to participants using PowerPoint on Apple Powerbooks with Sony MDR-CD850 stereo headphones.

2.3 Auditory and Visual Stimuli

There were 36 everyday sounds made by objects chosen for use in the experiment (see Table 1) based on data from a previous study [8]. Thirteen of the sounds were the sounds made by the objects in the videos (actual sounds); 10 of the sounds were acoustically similar sounds to those made by the objects in the videos that had been misidentified as the actual sounds (similar sounds); and 13 of the sounds were acoustically dissimilar and had not been confused with the sounds in the videos (contrast sounds). Three of the actual sounds (book closing, stapler stapling, and paper ripping) were also used as similar sounds for the videos.

Table 1: Sound stimuli and their relationship with the videotaped scenes

Actual	Similar	Contrast
3-Ring Binder (closed)	Purse (snapped shut)	Hair Dryer (turned on)
Book (shut)	Balloon (popped)	Vinyl Record (scratched)
Soda Can (crushed)	Book (shut)	Vacuum Cleaner (turned on)
Soda Can (opened)	Stapler (stapling)	Touch-tone Phone (dialed)
Chalkboard (erased)	Eraser (erasing in paper)	Rattle (shaken)
Keys (jingled)	Chains (clinked)	Chalkboard (written on)
Hammer (pounding)	Basketball (bounced)	Tires (Screeching)
Paper (ripping)	Tape (pulled off roll)	Sword (taken out of sheath)
Telephone (ringing)	Alarm Clock (ringing)	Foghorn (blown)
Scissors (snipped)	Whip (cracked)	Baseball (hit with bat)
Spoon (dropped)	Nails (dropped)	Ratchet (turned)
Stapler (stapling)	Cigarette Lighter (flicked)	Glass (breaking)
Velcro (pulled apart)	Paper (ripping)	Saw (sawing wood)

The scenes were the action on the object that produced the actual sound and were staged with a single person in a context where such an action might normally happen. They were videotaped with the target action and sound repeated 3 times. During the recording, the audio was also recorded so that there were other minor ambient sounds available in the soundtrack. After recording was completed, the videos were edited and the sounds were synchronized with the actions for all three types of sounds. The resulting 39 videos were distributed across 3 sets of 13 videos so only one of the scenes was represented in each set, and the sound conditions (actual, similar, and contrast) were counterbalanced across the sets. Due to the small number of trials per individual, the conditions were unequally distributed for each set, such that there were no fewer than 3 and no more than 6 from each condition. This was done to prevent participant bias based on expectations for answers on given trials. Each of the 3 sets of videos were placed in PowerPoint slides in 2 random orders resulting in 6 sets of PowerPoint slides for the sound/video condition procedure.

Two random orders of all 36 sounds were produced and placed in PowerPoint slides for presentation to the participants in the sounds-only condition. The slide used to designate each sound had the number displayed in the middle of the screen that corresponded to the trial on the response sheet.

Finally, 2 random orders of a written list of all 36 sounds were produced that were used for participants in both the sound/video condition and the sounds-only control for the familiarity ratings.

2.4 Procedure

For the sound/video condition participants were randomly assigned to one of the six sets of PowerPoint slides. The experimenter read a set of instructions to the participants while they read along. Participants were told that they would be viewing videotapes of people in 13 everyday situations. They were also told that after each scene, they would be asked to identity the sound the object made and to rate their confidence in their identification (1, not confident, to 7, very confident) and the veracity of the sound (1, not realistic, to 7, very realistic).

Participants completed a practice trial and were allowed to ask questions about the procedure. After they completed the 13 video trials, they were given a written list of all 36 sounds and rated each of them on familiarity (1, not familiar, to 7, very familiar). To finish the procedure, participants completed a brief follow-up questionnaire after which they were fully debriefed.

For the sounds-only condition, participants were randomly assigned to one of the two orders of the 36 sounds. After each sound trial, they made an identification of the sound and rated their confidence in this identification as well as a rating of the sound's veracity. After these trials were completed, they rated the written list of sounds for familiarity. These tasks were the same as those performed by the sound/video condition group, except that this group was not exposed to the videos.

3. RESULTS & DISCUSSION

For the sound/video condition there was one within-subjects independent variable, the sound and video pairings, which had three conditions, actual, similar, and contrast. The dependent variables were the number of correct sound identifications, the ratings of confidence for the sound identifications on a 7-point scale (1, not confident, to 7, very confident), the ratings of veracity of the sound (1, not realistic, to 7, very realistic), and the rating of the familiarity of each sound (1, not familiar, to 7, very familiar). The sounds-only control condition had data for all three dependent variables.[1]

3.1. Sound identifications

For the number of correct identifications for the sounds, a repeated measures ANOVA and follow-up analytical comparisons revealed that the actual sounds (M = 4.06, SD = 1.04) had the highest mean number of correct identifications, followed by the contrast sounds (M = 2.56, SD = 1.05) with the similar sounds (M = .62, SD = .82) showing the lowest mean number of correct identifications, $F(2,66) = 93.31$, $p < .001$, η^2_p =.74. These results provide partial support for the hypothesis since it was expected that the actual and contrasting sounds would have higher identification rates than the similar sounds. Contrary to the hypothesis, it was found that the actual sounds had a higher identification rate than the contrasting sounds. Considering these data as percentages clearly shows the difference in identification rates with actual sounds identified 95%, contrast sounds 61%, and similar sounds 14% of the cases (see Figure 1). Further examination of the incorrect identifications of the similar sounds showed they were misidentified 60% of the time as the sound made by the object in the video; however, the contrast sounds were never identified in this matter. The control group, who only heard the sounds, had an identification rate of only 49%.

[1] The control group for this design was not included in the statistical analyses with the experimental groups due to the different number of stimuli in the control versus experimental conditions. However, the control group data were included in the results to give an *indication* of how people perform these auditory tasks when they have no contextual visual information.

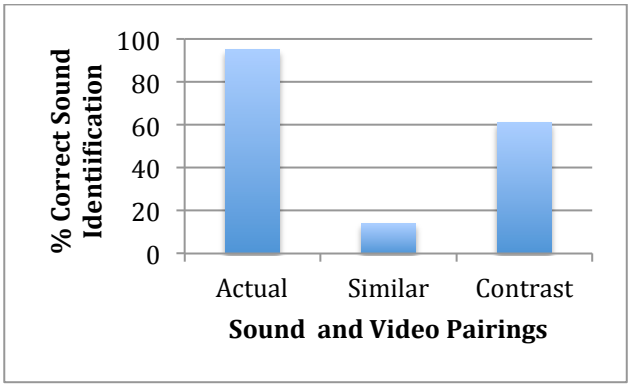

Figure 1: Percentage of correct sound identifications for the sound and video pairings.

3.2 Confidence ratings for sound identifications

For the confidence ratings for the identifications, a repeated measures ANOVA with post hoc analytical comparisons revealed that actual sounds (M = 6.43, SD = .60) showed the highest ratings while there was no difference between the similar (M = 4.64, SD = 1.19) and the contrast (M = 5.00, SD = 1.31) sound ratings, $F(2,66) = 30.11$, $p < .001$, η^2_p = .48 (see Figure 2). These results show partial support for the hypothesis since the actual sounds were given higher confidence ratings than the contrast sounds as predicted, but contrary to the hypothesis, the similar sounds were not rated higher than the contrast sounds and had lower ratings than the actual sounds. The control group's confidence ratings for all sounds showed a base rate that fell within the means of the experimental conditions (M = 5.13, SD = 1.03).

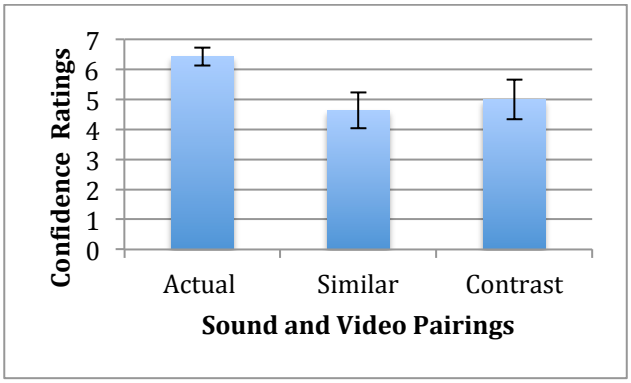

Figure 2: Confidence ratings for sound identifications for the sound and video pairings.

3.3 Veracity ratings for sounds

For the veracity ratings, a repeated measures ANOVA with post hoc analytical comparisons showed that actual sounds were viewed as most realistic (M = 6.39, SD = .54), followed by similar sounds (M = 3.86, SD = 1.13) with contrast sounds having the lowest veracity rating (M = 2.26, SD = 1.65), $F(2,66)$ = 116.50, $p < .001$, η^2_p =.78 (see Figure 3). These results provided partial support for the hypothesis since it was expected that the actual and similar sounds would have higher veracity ratings than the contrast sounds, but it was not

expected the similar sounds would be perceived as less realistic than the actual sounds. The means for the control group indicate the rated realism of sounds only was closest to the actual sound condition ($M = 5.96$, $SD = 1.24$).

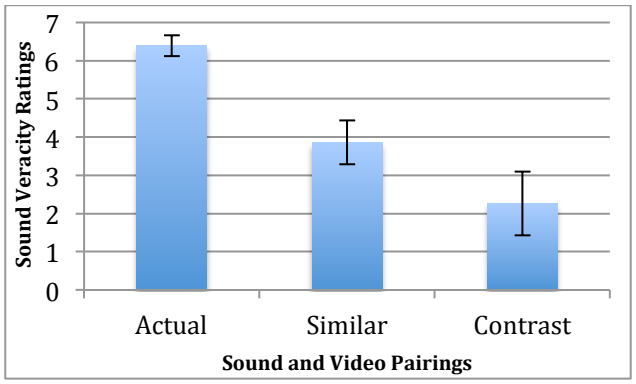

Figure 3: Sound veracity ratings for the sound and video pairings.

3.4 Familiarity ratings for sounds

Finally, ratings of familiarity for the sounds showed that the more familiar the sound was, the higher the number of correct identifications for the actual sounds and for the sounds in the control condition, $r(43) = .44$, $p < .001$. However, the familiarity ratings for the contrast, $r(33) = .23$, $p > .05$ and similar sounds, $r(33) = .17$, $p > 05$, provided no predictive value.

4. CONCLUSION

The results from this study clearly show that people watching videos of actions in which objects are "sounded" impact their perception of the sound. When the sound is the actual sound made or is an acoustically contrasting sound, their ability to make correct identifications is much better than when the sound is acoustically similar. These results even suggest that there is a facilitative effect for seeing the action and hearing the sound at the same time rather than just hearing the sound alone. The inaccurate identifications of the similar sounds show what would be expected from the Foley representations of sounds – people accept the sound as that portrayed by the video. However, it is important to note that in contrast to expectations that similar sounds would be *completely* perceived as real, listeners' confidence in such identifications and their assessment of the realistic nature of the sounds show that they do indeed recognize that the sound is not quite right. Since the stimuli in this study have only one sound that was actively portrayed, it is reasonable to predict that adding more background sound effects and more visual action would lead to people not noticing the discrepancy between the visual scene and an accompanying acoustically similar sound that is not the actual sound made by the object. In such cases, the coconuts banged together would indeed be perceived as horse hoofs galloping across the prairie.

5. ACKNOWLEDGMENT

I would like to thank Tanja Gazibara, Philip Schuman, Natalie Piltz, and J. Allen Lynch for their assistance with this project.

6. REFERENCES

[1] H. Mantell (Ed.) The complete guide to the creation and use of sound effect for film, T.V. and dramatic productions: And for exercising the mind, the ear, the imagination and the pen. Princeton: Films for the Humanities, Inc., 1983.

[2] S. Namba, Y. Hayashi, S. and Wako, "On the synchrony between figure movement and sound change," *Empirical Studies of the Arts*, vol. 21, 177-184, 1998.

[3] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *J. Acoustical Soc. Am.,* vol. 107, 582-527, 2000.

[4] N. F. Dixon and L. Spitz, "The detection of auditory visual desynchrony," *Perception*, vol. 9, 719-721, 1980.

[5] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 19, 250-267, 1993.

[6] W. H. Warren, Jr. and R. R. Verbrugge, "Auditory perception of breaking and bouncing events: A case study in ecological acoustics, *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, 704-712, 1984.

[7] N. J. Lass, S. K. Eastman, W. C. Parrish, K. A. Scherbick, D. M. Ralph, "Listeners' identification of environmental sounds, Perceptual and Motor Skills, vol. 55, 75-78.

[8] T. L. Bonebright, "Perceptual structure of everyday sounds: A multidimensional scaling approach," *Proceedings of 2001 ICAD.*

[9] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds," *Perception and Psychophysics*, vol. 69, 839-855, 2007.

[10] J. A. Ballas and T. Mullins, "Effects of context on the identification of everyday sounds," *Human Perception*, vol. 5, 199-219, 1993.

[11] B. Gygi and V. Shafiro, "The incongruency advantage for environmental sounds presented in natural auditory scenes," vol. 37, 551-565, 2011.

[12] S. J. Boyce and A. Pollatsek, "Identification of objects in scenes: The role of scene background object naming," *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 18, 531-543, 1992.

[13] I. Biederman, "Perceiving real-world scenes," *Science*, vol. 177, 77-80, 1972.

[14] J. Davenport and M. C. Potter, "Scene consistency in object and background perception," *Psychological Science*, vol. 15, 559-564, 2004.

[15] H. McGurk and J. W. MacDonald, "Hearing lips and seeing voices, *Nature*, vol. 264, 746-748, 1976.

[16] K. P. Green and A. Gerdeman, "Cross-modal discrepencies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels," *Journal of Experimental Psychology*, vol. 21, 1409-1426, 1995.

[17] J. Vroomen and B. de Gelder, "Sound enhances visual perception: Cross-modal effects of auditory organization on vision, *Journal of Experimental Psychology*, vol. 26, 1583-1590, 2000.

[18] J. Vroomen and B. de Gelder, "Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon," In G. Calvert, C. Spence, and B. E. Stein (Eds.), *Handbook of Multisensory Processes*, pp. 141-152, 2004.