Audio Engineering Society

Convention Paper 8150

Presented at the 128th Convention
2010 May 22–25 London, UK



# Effect of Signal-to-Noise Ratio and Visual Context on Environmental Sound Identification

Tifanie Bouchara[1], Bruno L. Giordano[2], Ilja Frissen[2], Brian F.G. Katz[1] and Catherine Guastavino[2]

[1] LIMSI-CNRS, BP133, 91403 Orsay, France
tifanie.bouchara@limsi.fr, brian.katz@limsi.fr

[2] CIRMMT & McGill University, 555 Sherbrooke St. West, Montreal, QC, H3A 1E3, Canada
bruno.giordano@music.mcgill.ca, ilja.frissen@mail.mcgill.ca, catherine.guastavino@mcgill.ca

## ABSTRACT

The recognition of environmental sounds is of main interest for the perception of our environment. This study investigates whether visual context can counterbalance the impairing effect of signal degradation (signal-to-noise ratio, SNR) on the identification of environmental sounds. SNRs and semantic congruency between sensory modalities, i.e. auditory and visual information, were manipulated. Two categories of sound sources, living and nonliving were used. The participants' task was to indicate the category of the sound as fast as possible. Increasing SNRs and congruent audiovisual contexts enhanced identification accuracy and shortened reaction times. The results further indicated that living sound sources were recognized more accurately and faster than nonliving sound sources. A preliminary analysis of the acoustical factors mediating participants' responses revealed that the harmonic-to-noise ratio (HNR) sound signals was significantly associated with the probability of identifying a sound as living. Further, the extent to which participants' identifications were sensitive to the HNR appeared to be modulated by both SNR and audiovisual congruence.

## 1. INTRODUCTION

Environmental sounds, i.e., non-speech non-musical sounds, are a fundamental component of our auditory experiences, and play an important role in our everyday interaction with the multisensory environment. Compared to recent advances in our knowledge of the perception of environmental sounds [1][2], still very little is known about how their processing is affected by simultaneous visual information. Previous studies have shown that audiovisual semantic congruency improves identification performance and reduces identification time. This effect was observed in [3] within a priming paradigm where visual stimuli preceded the sound, in

[4] with synchronous and semantically congruent visual and auditory stimuli and in [5] with images that had a high conceptual relationship with the sound sources.

Several previous studies revealed crossmodal interactions between auditory and visual stimuli. For instance, the simultaneous presentation of a visual stimulus can modify the perceived position of a sound source (ventriloquist effect, [6]), or the phonemic analysis of speech sound (McGurk effect, [7]). Still in the domain of speech perception, other studies have shown that a congruent and simultaneous presentation of a visual stimulus could reduce the detection threshold of speech, and more important could enhance word identification by lip-reading, reducing reaction time while enhancing accuracy. Visual information thus improves the processing of speech in degraded listening conditions, typically a noisy environment, where identification times are reduced for lower degradation levels, i.e., for higher signal-to-noise ratios (SNR) ([8], [9]), or with a degraded signal due to transmission failure or signal treatments, such as cochlear implants degrading spectral content of signals. Speech perception has been studied extensively in degraded conditions, both for noisy environment and content degradation, in presence of visual stimuli or not. On the other hand, although some studies investigated environmental sound perception in degraded conditions like low spectral resolution [10], none of these studies investigated environmental sound perception in noise or in degraded condition with visual context.

Our study assessed the effect of visual context on the identification of environmental sounds presented at variable SNR. We tested sound identification at various levels of noise. The audio-visual semantic congruency varied with three types of pictures, i.e., congruent, incongruent and abstract (without semantic content). These bimodal conditions where visual stimuli were combined to sound presentation were compared to unimodal auditory condition in which no visual cues were provided. Sounds were selected from two different categories depending on whether they were produced by a living/animate agent or a nonliving object. Participants were asked to identify the category of the sound sources as fast as possible, irrespective of whether it was presented in a bimodal or unimodal condition.

Harmonic-to-noise ratio (HNR) is known as one of the acoustical features most likely used by listeners to differentiate between living and nonliving sounds [11]. We also investigated how the processing of acoustical information, especially HNR, is influenced by both SNR and visual context.

It was hypothesized that, as in the case of speech, a) a semantically congruent visual stimulus facilitates the sound identification in terms of both recognition time and accuracy and b) that the effect varies in proportion to the SNR.

## 2. EXPERIMENT 1 – AUDIOVISUAL STIMULI SELECTION

An initial experiment was carried out to select a set of highly identifiable auditory and visual stimuli, and which participants would consistently associate with each other. This set was then used in Experiment 2. Stimuli and results of experiment 1 are reported Table 2 (see Appendix)..

### 2.1. Participants

Ten individuals (3 females, 7 males) participated in this experiment. They reported normal hearing and normal or corrected-to-normal vision according to self-report.

### 2.2. Auditory stimuli

Fifty environmental sounds (sampling rate = 44.1 kHz, bit depth = 16; maximum duration = 3 sec) were selected from a larger set investigated in a previous study [12]. Based on the results reported in [12], all sounds were identified with a percent accurate score 93% on average (SD = 9%; range = 70-100%). Half of the stimulus set (living sounds) comprised sounds generated by animate agents (22 animal or human vocalizations and 3 bodily made sounds such as "*buzzing fly*" and "*blowing nose*"). The other half of the stimulus set (nonliving sounds) comprised sounds generated by inanimate sound sources (e.g., sounds of impacting solids or liquids, [13]). All sounds were selected so that they could be easily associated with a picture of the sound-generating event (e.g., wind sounds were not considered).

### 2.3. Visual stimuli

A set of 50 constant-size colour still pictures was selected either from online resources[1] or were taken by the first author. The pictures represented the agent or the object producing each of the 50 environmental sounds

[1] www.dreamstime.com

(25 living and 25 nonliving pictures). The agent/object was presented: 1) against a white background (20 pictures); 2) against a textured background representing part of the context (e.g., grass, beach; 20 pictures); 3) against a uniform color background (e.g., blue; 10 pictures). Ten additional pictures (5 living, 5 nonliving) were included to the set of pictures to make the task more challenging (i.e., participants could not use a process of elimination). These additional pictures were chosen so as not to represent any of the sources of the 50 sounds, and not to have any conceptual link with the other pictures. Part of these control stimuli had a white background with the object only, whereas another part had a coloured or textured background.

### 2.4. Procedure

Participants carried out two different tasks during two subsequent experimental sessions. The first session aimed to assess the identifiability of the visual stimuli. All the 60 pictures were simultaneously presented to the subjects arranged on a 6 rows-by-10 columns grid, with a reduced size of 2.8 x 2.1 $cm^2$ (corresponding to an approximate visual angle of 2° x 1.5°). Stimuli were presented in random order. Participants were asked to freely identify what was depicted in the picture by typing below each picture a maximum of two words (verbs or nouns).

The aim of the second session was to assess the correct association between sound stimuli and the corresponding visual stimuli. Participants carried out a forced choice task. On each trial, they were presented one of the 50 sounds, and were asked to choose which among the 60 pictures best matched the sound. Each of the 50 sounds was presented twice in random order for a total of 100 trials. No feedback was given. The entire experiment lasted ca. 20 minutes.

### 2.5. Results and discussion

Experiment 1 was functional to selecting audio-visual stimulus pairs for Experiment 2 that are well recognized in both modalities, and for which identification errors could not originate from inaccurate identifications of the visual stimulus, or from inaccurate audio-visual associations.

The first experimental session produced scores for identification accuracy for the visual stimuli. Accuracy was computed following a procedure similar to that described in [12]. The identification of two pictures of living sources ("*Gasping woman*" and "*Burping person*") and two pictures of nonliving sources ("*Splashing water*" and "*Honking bike horn*") were not adequate (percent correct < 75%), and were excluded for Experiment 2.

The second session produced measures of sound-to-picture associations (*pAssoc*), defined as the number of times a given sound was associated with the visual stimulus selected by the experimenter as the depiction of the sound-generating event. One nonliving sound source ("*boiling kettle*") was associated with a living picture and was exluded from further experimentation. Much confusion was observed between the different liquid sound sources. For this reason, we excluded the pair of audio-visual liquid stimuli with the lowest *pAssoc* score ("*lapping water*" , *pAssoc* = 0.75). The picture of the "*blowing balloon*" was never associated with the corresponding sound, because participants always chose the visual stimulus depicting an airplane. We finally excluded three more nonliving sources with the lowest *pAssoc* score in order to balance the number of living and nonliving stimuli: "*grunting pig*", *pAssoc*=0.8; "*quacking duck*", *pAssoc*=0.8; "*calling eagle*", *pAssoc*=0.85. The final set included 40 sound-picture pairs, each representing a different source (20 living, 20 nonliving).

## 3. EXPERIMENT 2 – RECOGNITION IN DEGRADED SNR

Experiment 2 addressed the effect of visual context on the identification of environmental sounds presented in various levels of acoustic noise. Four audio-visual (AV) conditions were tested: audio + congruent visual with a picture of the source (AVc), audio + incongruent visual with a picture of a source from the opposite category, i.e., living picture for nonliving sound and vice versa (AVi), audio + neutral visual with an abstract picture (AVn) and a control audio-only condition (A). The noise presented simultaneously with the sound was manipulated so as to achieve eight different levels of SNR. The experiment was an 8 (SNR) x 4 (AV condition) x 2 (category: living vs. nonliving) factorial design.

### 3.1. Participants

Nineteen individuals participated in the experiment (average age = 24 years, 9 males and 10 females). All except two of the participants were right-handed. All declared normal or corrected-to-normal sight and

normal hearing. None of them participated in Experiment 1.

### 3.2. Apparatus

Participants were tested individually in an isolated listening room. Stimuli were presented using a custom-built computer program. Participants sat in front of a computer screen, at a distance of ca. 60cm. Visual stimuli were presented in the middle of the screen, against a white background, with a size of 10.9 x 6.8 $cm^2$ (subtending a visual angle of 5.5°x3.4°). Auditory stimuli were presented through Sennheiser HD 280 headphones. A reference white noise signal with the same RMS level as the RMS-level-equalized sounds in absence of noise had a presentation level of 50dB SPL. Consequently the level of the same reference signal in the -18dB signal-to noise ratio condition was 68 dB SPL. Responses were collected from a computer keyboard placed on the table directly in front of the participant.

### 3.3. Stimuli selection

#### 3.3.1. Auditory stimuli

Forty highly identifiable sound stimuli (20 living, 20 nonliving) were selected based on the results of Experiment 1 so that they could be unambiguously match with the corresponding visual stimulus. All stimuli were equated in overall RMS level. Then stimuli were processed to obtain 8 different levels of signal-to-noise ratio (-18 to 0 dB in 3 dB increments), by adding a 3 sec white noise, plus one control condition without noise.

As levels were equated in overall RMS, we noted a difference in peak level between the categories, with the peak of nonliving sounds higher than the peak level of living sounds by approximately 8 dB on average (t(38), $p < 0.001$). Central time was calculated as the centroid of each signal vector. It was shorter for nonliving sounds than for living sounds (mean = 1.28 and 1.54 s, respectively, $t(38) = 2.98$ , $p < 0.01$).

#### 3.3.2. Visual stimuli

Forty visual stimuli (20 living, 20 nonliving), each corresponding to one of the sound stimuli, were selected based on the results of Experiment 1. As is Experiment 1, some were sources presented in isolation against a white background, whereas in other pictures the source was presented against a textured background as part of the context for the sound source, or against a colored background.

Twenty-four abstract pictures were included in the set of visual stimuli for the audio + neutral visual condition. They were abstract drawings with different color traces selected so as not to be clearly associated with any of the sound sources. Some exemplars are shown in Figure 1.


Figure 1: Examples of pictures for each category of visual stimuli: living, nonliving and abstract.

### 3.4. Design and procedure

Participants were asked to listen to a sound, while watching the screen, and decide as quickly as possible, by pressing one of two keys, if the sound was a living sound generated by an animal/a human being or a nonliving sound generated by a nonliving object. Stickers, “living” and “nonliving”, were put on the response key (the ‘v’ and ‘n’ keys). The allocation of the category of the stimuli to the response key was counterbalanced across participants.

Participants were told that the sound could be masked by a noise or not and that pictures were of the source producing the sound or not. They were instructed to watch the screen attentively although they were informed that the task was to judge sounds only. As soon as participants pressed one of the response keys, the playback of the sound was stopped and the picture was removed. The next trial began after a 700ms pause. Before the main experiment, participants were familiarized with the task with a block of 16 training trials, during which they were presented 6 sound stimuli not selected for the main experiment.

During the main experiment, for each of the 8 SNR levels and the 4 audio-visual modality parameters tested in the experiment, 12 sources of each category were randomly picked among the 20 stimuli. This was a compromise between the number of trials participants could reasonably carry out during the experimental

session, and the maximum number of repeated presentations for each single sound that limited learning effects. Each participant carried out 768 trials (8 SNR conditions * 4 AV conditions * 2 categories * 12 sources) divided into 16 blocks of 48 trials. Participants were allowed to take breaks between blocks. The order of stimulus presentation was randomized. The entire experiment lasted approximately 45 minutes.

## 4. RESULTS

Figures 2 show the across-participants average of the reaction time and percent correct in the different experimental conditions. As expected, for both the living and nonliving categories and for each of the audiovisual condition, a decrease in SNR resulted in a decrease in accuracy and in an increase in reaction time.

a.

b.

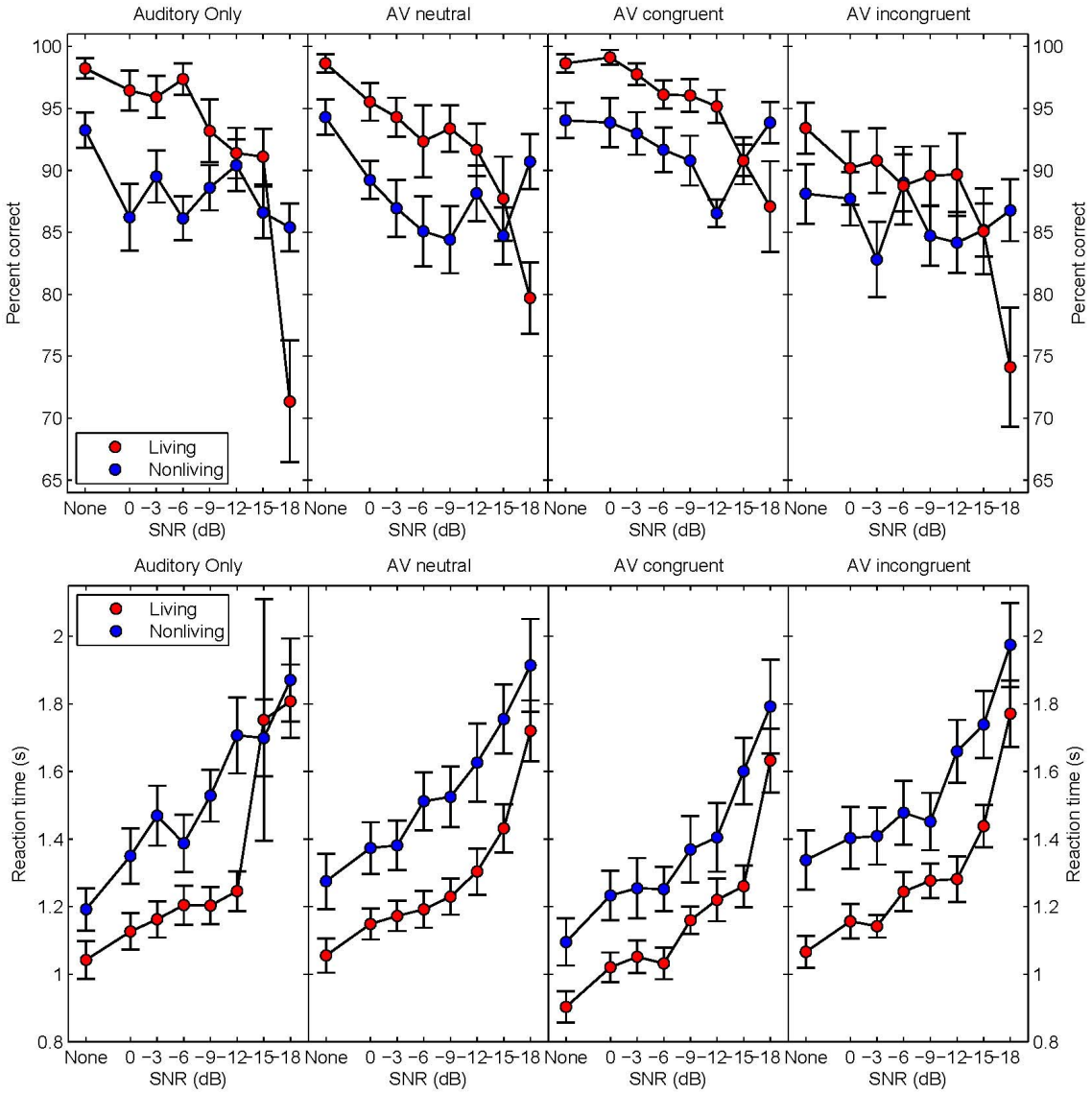


Figure 2: Results across-participants in the different experimental conditions as a function of the signal-to-noise ratio (SNR). None = no noise; error bar = ± 1 SE. a. Average of the percentage of correct answers. b. Average reaction time for the correct answers

Another expected result concerns the difference between the AVc and A or AVi conditions, with reaction times being shorter for the congruent bimodal condition than for the incongruent bimodal and unimodal conditions. Interestingly, living sounds (red) were identified faster than nonliving sounds (blue). Surprisingly, one can observe in Figure 2 an inversion between living and nonliving sounds, with the living sounds being recognized more accurately for small degradation (SNR > -15dB) but not for high SNR (i.e., -15dB and -18dB). This effect is more pronounced for A and AVi than for AVc or AVn. However, the reaction time increases significantly between -12dB, and -15dB and -18dB, signifying that participants needed more time to think before answering.

### 4.1. Accuracy

Table 1 reports the average percentage of errors for the different experimental conditions.

For both living and nonliving categories accuracy is highest for bimodal congruent AVc condition and lowest for bimodal incongruent AVi condition. No difference between other audio-visual conditions can be noted. A difference of average accuracy occurs between living sounds identified more accurately than nonliving sounds.

| | | A | AVn | AVc | AVi |
|---|---|---|---|---|---|
| Accuracy (%) | Living | 91.9 (1.7) | 91.7 (1.6) | 95.1 (1.0) | 87.7 (2.5) |
| | Nonliving | 88.3 (1.0) | 87.9 (1.1) | 91.8 (0.9) | 86.1 (1.7) |
| | Global | 90.1 (1.2) | 89.8 (1.2) | 93.5 (0.8) | 86.9 (2.0) |
| RT (ms) | Living | 1318 (81) | 1282 (51) | 1160 (48) | 1297 (52) |
| | Nonliving | 1525 (80) | 1545 (85) | 1375 (81) | 1557 (85) |
| | Global | 1422 (78) | 1413 (66) | 1267 (63) | 1427 (66) |

Table 1. Average accuracy and reaction times and in the different conditions of audio-visual congruence. Values between brackets are Standard Errors of the Mean.

Accuracy data were analyzed with a repeated measures ANOVA with condition (A, AVn, AVc, Avi), SNR (no noise and 7 SNR values) and category (living, nonliving) as within-subject factors. The interaction between SNR and condition and that between SNR and category were significant, $F(2,378) = 2.64$ and $F(7,126) = 11.59$, respectively, $p < 0.001$, indicating that the effect of SNR differed across experimental conditions, and across categories of environmental sounds. The effect of category was significant, $F(1,18) = 6.36$, $p = 0.021$: participants made fewer errors with living sounds (8.41% of answers) than with nonliving sounds (11.45 % of answers). The effect of SNR was also significant, $F(7,126) = 26.76$, $p < 0.001$, originating from a decrease in error rate with an increase of SNR. The effect of condition was significant, $F(3,54) = 13.48$, $p < 0.001$: participants made fewest errors in the congruent condition (AVc, 6.54% errors) and more errors for incongruent condition (AVi, 13.12% errors) than for the audio-only condition (A, 9.93% errors) and neutral condition (AVn, 10.20%). Planned comparisons confirmed that performance in AVc was significantly better than in any of the other conditions, $F$'s > 18.8, $p$'s < 0.001. None of the other interactions was significant, $F \leq 1.36$, $p \geq 0.07$.

### 4.2. Reaction Time

Analyses focused on the reaction time (RT) measures for the correct answers. Table 1 reports the average RT for the different experimental conditions.

For both living and nonliving categories the RTs are shorter for bimodal congruent AVc condition. No difference between other audio-visual conditions can be noted. A main difference of RT occurs between categories with living sounds being recognized faster than nonliving sounds.

RT data were analyzed with a repeated measures ANOVA with condition (A, AVn, AVc, Avi), SNR (no noise and 7 SNR values) and category (living, nonliving) as within-subject factors. The effect of category was significant, $F(1,18) = 41.10$, $p < 0.001$, with participants responding more rapidly for living than for nonliving sounds, mean RT = 1264 and 1500 ms, respectively. The effect of noise level was also significant, $F(7,126) = 51.60$, $p < 0.001$, confirming decreased RTs for higher SNRs. The effect of condition was significant, $F(3.54) = 30.48$, $p < 0.001$, with shorter RT for AVc than to either A, AVn or AVi conditions (see Table 1). Planned comparisons confirmed that performance in AVc was significantly faster than in any of the other conditions, $F$'s > 36.4, $p$'s < 0.001. None of the interactions between the considered factors was significant, $F \leq 1.46$, $p \geq 0.09$, suggesting, for instance,

that the enhancement of reaction time with congruent visual stimuli is not modulated by noise level, as was hypothesized.

### 4.3. Effect of the Harmonic-to-Noise Ratio

We carried out a preliminary analysis to estimate the extent to which the perceptual processing of acoustical information for making the living vs. nonliving distinction was influenced by the experimental manipulation of SNR and visual context. To this purpose, we assessed the influence of sound harmonicity on the probability that participants identified a sound stimulus as living. Analyses focused on sound harmonicity because: a) previous studies of environmental sounds showed harmonicity is a central factor in perceptual processes [11]; b) the harmonicity of a signal is likely the acoustical property that better distinguishes between the living sounds and nonliving sounds in this experiment, where the vast majority of living sounds was a vocalization.

The software system described in [14] was used to extract the maximum of the time-varying Harmonic-to-Noise Ratio (HNR) for each of the sounds without noise. HNR measures the ratio between the energy of the harmonic and nonharmonic components of a sound signal, where HNR increases with increasing harmonic energy (e.g., HNR for a pure tone is higher than for a random signal), and HNR equals zero when the harmonic and nonharmonic components have the same energy. On average, HNR was significantly higher for living than nonliving sounds, HNR = 21.45 and 6.24 dB, respectively, unpaired samples $t(38) = 5.18$, $p < 0.001$. Also note that 8 of the living sounds had an HNR lower than the maximum nonliving HNR, and 4 nonliving sounds had an HNR higher than the minimum living HNR, i.e., participants could have achieved a performance of at least 70% correct by answering based on HNR alone.

A probit regression model was fit using the across-participants probability that a given sound was identified as living as dependent variable, and with the sound HNR as independent variable. One different probit model was fitted to each of the 32 datasets from the 4 AV conditions combined with the 8 SNR conditions. Note that the HNR values used to predict the probability of living identifications were always computed from the sounds without noise, even when the dependent variable was collected from a noise condition. As such, the working assumption was made that participants were always capable to separate the sound from the background noise, and that they were able to estimate the harmonicity of the target sound independently of the level of the background noise. It should be noted that the same analysis was attempted by extracting HNR from the sound-in-noise stimuli. In this case, the HNR of the mixture appeared not to be significantly associated with participants' responses. For the sake of brevity, these results are not reported here.

Subsequent analyses focused on the slope of the probit models, measuring how rapidly the probability of answering "living" changed as a function of HNR. Figure 3 shows two of the probit models fitted to two of the considered datasets (A condition in absence of noise in blue and A condition with -18 dB SNR in red). In this figure, the model in red has a smaller slope than the model in blue, i.e., the answers of participants changed less slowly as a function of HNR. In the following, the slope term is taken as a measure of the extent to which the identification of participants were sensitive to variations in the HNR of a sound, steeper slopes measuring a higher sensitivity.

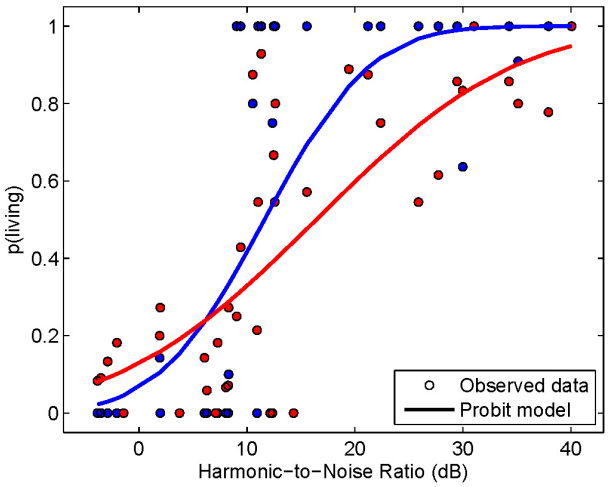


Figure 3: Observed and modeled probability of answering "living" as a function of the maximum of the time-varying HNR measured in the sound stimuli in absence of noise. All data from the A condition with no noise (blue) and with SNR = -18 (red).

Across the 32 datasets, HNR explained an average of 55% of the variance of the observed probability to identify a stimulus as living, STD = 0.04. Figure 4 shows the slope of the probit models created for the different SNR values in the different experimental conditions. We analyzed the slopes within a two-way

ANOVA with SNR and AV condition (A, AVn, AVc and AVi) as factors. SNR significantly affected the slope of the probit models, $F(7,21) = 5.84$, $p < 0.001$. We further quantified this effect by creating a linear regression model with the 32 slopes from each of the datasets as dependent variable, and SNR as predictor. Within this model, the linear association between SNR and slope was significant, $p < 0.001$, showing that, overall, the HNR slope increased with increasing SNR values. The AV condition also had a significant effect on the sensitivity to HNR, $F(3,21) = 3.62$, $p = 0.03$. In particular, slope increased from the AVi condition (0.091) to the AVn condition (0.103) to the A condition (0.106) to the AVc condition (0.113). After post-hoc pairwise contrasts, the only significant difference between AV conditions was between the AVi and AVc conditions, with sensitivity to HNR higher for congruent visual stimuli than for incongruent visual stimuli, $t(7) = 5.94$, $p < 0.001$. Notably, although both SNR and AV conditions influenced the sensitivity for HNR, SNR had a stronger effect on this measure than AV condition, as revealed by the partial eta squared measure of effect size for both SNR and AV factors = 0.66 and 0.34, respectively.

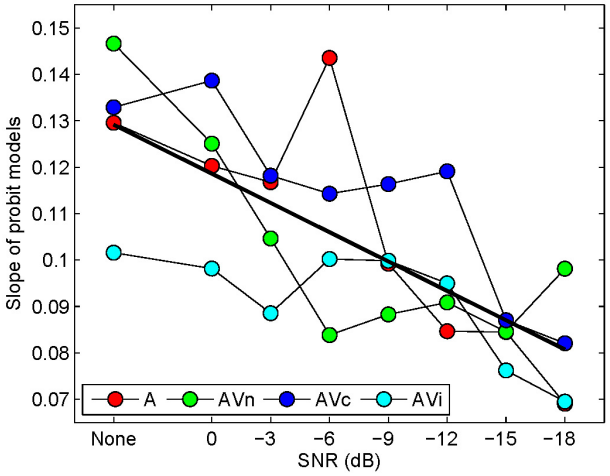


Figure 4: Slope of the probit model fitted to the probability of answering "living" in different experimental conditions. Higher slopes indicate a higher sensitivity to the Harmonic-to-Noise Ratio of the sound signals. The solid black line shows the linear regression model fitted with all slope measures as dependent variable, and SNR as predictor. The linear regression analysis confirmed a decrease in slope for decreasing values of SNR.

Finally, it should be noted that the linear relationship between SNR and the slope of the probit models does not follow an exactly linear relationship. Possible explanations for the deviation from linearity include: a) noise in the behavioral responses; b) the fact that participants did not focus on HNR but on a correlated acoustical measure; c) effects of eventual outliers on the slope estimates for the probit regression models. Further work will be necessary to take apart these factors, and to improve the quality of the prediction of participants' responses based on acoustical factors.

## 5. DISCUSSION

The effect of a visual context and signal-to-noise ratio on environmental sound identification was evaluated with a living/nonliving categorization task. Reaction times were shorter and accuracy was better for AVc condition than for the others, while accuracy was lower for AVi. These results are in agreement with previous crossmodal studies on sound identification in the presence of a visually congruent context ([4][5][15]) as semantic congruency improves identification performances. Moreover, participants benefit more from congruent visual information when the audio degradation is high. This is consistent with the "maximal efficiency theory" on crossmodal integration of visual and auditory information where modalities are combined in such a way to ensure a reliable percept [16].

Two different categories of sounds were chosen as sound stimuli for this experiment: nonliving and living sounds. This choice was motivated by previous classification of environmental sound sources and the fact that these two categories are easily differentiable. The results revealed differences in reaction times and accuracy between categories, with living being faster and more accurately recognized than nonliving sounds. These results are consistent with previous studies on natural sound identification [17]. Analysis of peak level and central time on reference stimuli revealed significant differences between nonliving and living sounds which could have explained the differences in accuracy and reaction times. As the accuracy decreases and reaction times increase with decreasing SNR, we could have supposed that louder sounds (i.e. with higher peak level as nonliving sounds), would have been identified faster and more accurately. However results show the inverse tendency with nonliving sounds being processed more slowly than living sounds. Differences

of peak level can also not be responsible for differences between categories.

The differences in accuracy may be biased by the chosen task. This could explain the observed interaction between category and SNR. Indeed, for living sounds the number of correct answers falls suddenly for low SNR while for nonliving sounds the number of correct answers increases. We believe that tasks with no reference to the category of source production could reduce this bias. For example, [3] suggested another "two-way" task where objects are categorized according to the size of the source, e.g. 'seagull' and 'elephant' would have been classed in two separate categories even produced by two living agents.

Further preliminary analyses investigated the acoustical factors involved in the identification of participants. To this purpose, analyses focused on the harmonic-to-noise ratio (HNR). It was discovered that the probability of categorizing a sound as "living" was significantly associated with the HNR of a sound signal. More precisely, high HNR sounds appeared to be more likely to be identified as living sounds. Further analyses quantified the extent to which participants' responses were sensitive to variations in the HNR of the sounds. Higher SNR levels appeared to be associated with a reduction in sensitivity to HNR. Sensitivity to HNR was also higher in the congruent audiovisual context than in the incongruent audiovisual context, suggesting an influence of visual context on the processing of acoustical information.

In this paper, we defined a context as everything surrounding the sound. According to this definition, stimuli presented in another modality are contextual. However the level of abstraction of the contextual cues can be manipulated. For example, [5] compared the presentation of a contextual object, semantically related to sound source, to the presentation of a contextual scene, representing the environment in which the sound event occurs. In a similar vein, our study could be extended to investigate the effect of visual context at two different levels of abstraction, namely comparing the effect of representing the sound source vs. the environment in which it is typically encountered (e.g. barn for domesticated animal sounds).

## 6. CONCLUSION

The effect of visual context congruency and signal-to-noise ratio on environmental sound identification was evaluated through a living/nonliving task. Identification of environmental sound sources was found to be more accurate and accelerated in the presence of a semantically congruent visual context representing the source. Simultaneously, degradation of the signal-to-noise ratio reduces identification accuracy and increases reaction time. The study also revealed differences in the perception of living versus nonliving sound sources. The analysis of harmonic-to-noise ratio is one of the main acoustical factors used to segregate living from nonliving sound sources. It was shown that both signal-to-noise ratio and visual context influenced the way acoustical information is processed to identify sound sources.

## 7. REFERENCES

[1] Gygi, B. and Shafiro, V. "Environmental Sound Research As It Stands Today." Proc. of Meetings on Acoustics,vol. 1, pp. 1-18 (2007)

[2] Schulte-Fortkamp, B. and Dubois, D. "Recent Advances in Soundscape Research." Acta Acustica united with Acustica, vol. 92, pp v-viii (2006)

[3] Schneider, T. R., Engel, A. K. and Debener, S. "Multisensory identification of natural objects in a two way crossmodal priming paradigm." J. Exp. Psychol. Human, vol. 55, pp. 121-132 (2008)

[4] Suied, C., Bonneel, N. and Viaud-Delmon, I. "Integration of auditory and visual information in the recognition of realistic objects." Exp. Brain Res., vol. 194, pp. 91-102 (2009)

[5] Özcan, E. and van Egmond, R. "The effect of visual context on the identification of ambiguous environmental sounds." Acta Psychologica, vol. 131, pp. 110-119 (2009)

[6] Howard, I.P. and Templeton, W.B. Human Spatial Orientation. Wiley, London. 1966.

[7] McGurk, H. and MacDonald, J. "Hearing lips and seeing voices." Nature, vol. 264, pp.746-748 (1976)

[8] Sumby, W. H. and Pollack, I. "Visual contribution to speech intelligibility in noise." J. Acoust. Soc. Am., vol. 26, pp. 212-215 (1954)

[9] Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. and Foxe, J. J. "Do you see what I'm saying?

Exploring Visual Enhancement of Speech Comprehension in Noisy Environments." Cereb. Cortex, vol. 17, pp. 1147-1153 (2007)

[10] Shafiro, V. "Development of a Large-Item Environmental Sound Test and the Effects of Short-Term Training with Spectrally-Degraded Stimuli." Ear & Hearing, vol. 29, pp. 775-790 (2008)

[11] Gygi, B., Kidd, G.R. and Watson, C. S. "Similarity and categorization of environmental sounds." Percep. Psychophys., vol. 69 (6), pp. 839-855 (2007)

[12] Giordano, B. L., McDonnell, J. and McAdams, S. "Hearing living symbols and nonliving icons: Category specificities in the cognitive processing of environmental sounds." Brain & Cognition, doi: 10.1016/j.bandc.2010.01.005 (2010)

[13] Gaver, W. W. "What in the world do we hear? An Ecological Approach to Auditory Event Perception." Ecol. Psychol., vol. 5, pp. 1-29 (1993)

[14] Boersma, P. "Praat, a system for doing phonetics by computer." Glot Int., vol. 5 (9/10), pp. 341-345 (2001)

[15] Laurienti, P.J., Kraft, R.A., Maldjian, J.A., Burdette, J.H. and Wallace M.T. "Semantic congruence is a critical factor in multisensory behavioral performance." Exp. Brain Res., vol.158, pp. 405-414 (2004)

[16] Burr, D. & Alais, D. "Combining visual and auditory information," special issue on Visual Perception, Prog. Brain Res., vol. 155, pp. 243-258 (2006)

[17] Suied, C. and Viaud-Delmon, I. "Auditory-visual Object Recognition Time Suggests Specific Processing for Animal Sounds." PlosOne, vol. 4(4), (2009 April)

[18] Ballas, J. A. and Mullins, T. "Effects of context on the identification of everyday sounds." J. Exp. Psychol. Human, vol. 4 (3), pp. 199-219 (1991)

[19] Gygi, B. and Shafiro, V. "The incongruency advantage for sounds in natural scenes." AES 125th convention, San Francisco, USA, 2008

## 8. APPENDIX

| Category | Identification label | HNR (dB) | p(correct) Audio-only | p(correct) Visual-only | pAssoc AV association | Experiment 2 |
|---|---|---|---|---|---|---|
| living | Screaming woman | 34.26 | 0.95 | 0.90 | 1.00 | x |
| living | Buzzing fly | 12.56 | 0.95 | 1.00 | 1.00 | x |
| living | Crowing rooster | 40.07 | 1.00 | 1.00 | 1.00 | x |
| living | Neighing horse | 12.46 | 1.00 | 1.00 | 1.00 | x |
| living | Grunting pig | | 0.95 | 1.00 | 0.80 | training |
| living | Calling seagull | 29.46 | 0.95 | 1.00 | 1.00 | x |
| living | Barking seal | 9.41 | 0.80 | 1.00 | 0.90 | x |
| living | Croaking frog | 10.52 | 1.00 | 1.00 | 0.90 | x |
| living | Quacking duck | | 0.95 | 1.00 | 0.80 | training |
| living | Roaring lion | 9.04 | 1.00 | 1.00 | 1.00 | x |
| living | Gasping woman | | 0.85 | 0.40 | 0.80 | no |
| living | Whining dog | 35.11 | 0.95 | 1.00 | 0.90 | x |
| living | Blowing nose | 12.61 | 1.00 | 1.00 | 1.00 | x |
| living | Bleating sheep | 15.54 | 1.00 | 1.00 | 0.90 | x |
| living | Chirping cricket | 22.38 | 0.85 | 1.00 | 1.00 | x |
| living | Cawing crow | 11.31 | 0.80 | 1.00 | 0.90 | x |
| living | Howling wolf | 37.91 | 1.00 | 1.00 | 1.00 | x |
| living | Crying baby | 25.87 | 1.00 | 1.00 | 1.00 | x |
| living | Burping person | | 1.00 | 0.00 | 0.95 | no |
| living | Meowing cat | 31.02 | 1.00 | 1.00 | 0.95 | x |
| living | Mooing cow | 27.72 | 1.00 | 1.00 | 1.00 | x |
| living | Calling eagle | | 0.75 | 1.00 | 0.85 | training |
| living | Trumpeting elephant | 19.43 | 1.00 | 1.00 | 0.95 | x |
| living | Coughing man | 11.02 | 0.70 | 1.00 | 0.95 | x |
| living | Laughing woman | 21.2 | 0.90 | 1.00 | 1.00 | x |
| nonliving | Blowing balloon | | 0.90 | 1.00 | 0.45 | no |
| nonliving | Bubbling water | 6.28 | 1.00 | 0.90 | 0.95 | x |
| nonliving | Ringing bell | 10.94 | 1.00 | 1.00 | 1.00 | x |
| nonliving | Ringing bike bell | 14.32 | 0.90 | 1.00 | 1.00 | x |
| nonliving | Sawing wood | 8.29 | 1.00 | 1.00 | 1.00 | x |
| nonliving | Crackling fire | -3.85 | 0.75 | 1.00 | 0.85 | x |
| nonliving | Jingling keys | 6.07 | 1.00 | 1.00 | 0.95 | x |
| nonliving | Rolling dice | 3.75 | 0.75 | 1.00 | 1.00 | x |
| nonliving | Honking bike horn | | 0.75 | 0.70 | 1.00 | training |
| nonliving | Pouring water | 8.25 | 0.95 | 1.00 | 1.00 | x |
| nonliving | Crumpling paper | -2.91 | 0.95 | 1.00 | 0.95 | x |
| nonliving | Swinging racket | 7.28 | 0.85 | 1.00 | 1.00 | x |
| nonliving | Boiling kettle | | 0.95 | 1.00 | 0.90 | no |
| nonliving | Dripping water | 12.36 | 0.90 | 1.00 | 0.95 | x |
| nonliving | Running water | -3.54 | 1.00 | 1.00 | 0.95 | x |
| nonliving | Flushing toilet | 8.04 | 1.00 | 1.00 | 1.00 | x |
| nonliving | Sharpening knife | 12.14 | 0.90 | 0.90 | 1.00 | x |
| nonliving | Typing keyboard | 1.97 | 1.00 | 1.00 | 1.00 | x |
| nonliving | Dropping change | -2.05 | 0.95 | 1.00 | 1.00 | x |
| nonliving | Bouncing ping pong ball | 7.1 | 1.00 | 0.90 | 1.00 | x |
| nonliving | Blowing party whistle | 29.97 | 0.90 | 1.00 | 0.95 | x |
| nonliving | Splashing water | | 0.90 | 0.60 | 1.00 | no |
| nonliving | Flowing water | -1.44 | 0.90 | 1.00 | 0.85 | x |
| nonliving | Lapping water | | 0.80 | 1.00 | 0.75 | training |
| nonliving | Thundering thunder | 1.91 | 1.00 | 1.00 | 1.00 | x |

Table 2. Experimental stimuli investigated in Experiment 1. The column “experiment 2” indicates which stimuli where also investigated in Experiment 2. Columns p(correct) for audio-only indicates values from [12].