Congruent audio-visual alarms for supervision tasks

Eliott Audry, Jérémie Garcia

CONGRUENT AUDIO-VISUAL ALARMS FOR SUPERVISION TASKS

Eliott Audry

Omnicontact-SafetyData,
2 allée Santos Dumont,
92150 Suresnes
eliott.audry@enac.fr

Jérémie Garcia

ENAC-Université de Toulouse,
7 Avenue Edouard Belin,
31400 Toulouse
jeremie.garcia@enac.fr

ABSTRACT

Operators in surveillance activities face cognitive overload due to the fragmentation of information on several screens, the dynamic nature of the task and the multiple visual or audible alarms. This paper presents our ongoing efforts to design efficient audio-visual alarms for surveillance activities such as traffic management or air traffic control. We motivate the use of congruent cross-modal animations to design alarms and describe audio-visual mappings based on this paradigm. We ran a preference experiments with 24 participants to assess our designs and found that specific polarities between visual and audio parameters were preferred. We conclude with future research directions to validate the efficiency of our alarms with different cognitive load levels.

1. INTRODUCTION

Maritime or aeronautical surveillance systems allow the recovery and fusion of information from ships and aircraft (type, position, speed, etc.) for traffic monitoring purposes via a display device. In both areas, the priority for operators is to guarantee safety through the prevention and resolution of potential conflicts (risk of collision, breakdowns, etc.). In addition, the detection of abnormal behavior and the early identification of associated threats (disaster, illegal or criminal activity, pollution, terrorist act, etc.) are major challenges for all surveillance operators.

To carry out their monitoring tasks, operators rely on complex systems, mainly graphical, to represent all traffic on a map and perform operations such as filtering certain information or selecting an element to obtain detailed information [17]. The systems also include visual or audible notifications and alarms when one or more algorithms integrated into the systems triggers an event [1,17,22].

As with most surveillance activities, a major problem concerns the cognitive overload and underload of operators [15,26]. This cognitive load problem is mainly due to the fragmentation of information on several screens but also to the dynamic nature of the task, visual and auditory distractions as well as interruptions. This overload can lead to blindness or unintentional deafness [4], [20] that prevents the perception of a visual notification or audible alarm when the user is overly solicited by the visual search for an element on the interface, for example. On the other hand, the phenomenon of cognitive underload, when traffic is calm, causes vigilance and attention maintenance problems that also have a negative impact on the quality of surveillance since operators can miss alarms.

Our goal is to rethink the design of audible alarms for surveillance by focusing on redundant modalities: instead of conceiving visual information and audible alarms as separate entities from monitoring systems, our approach consists in integrating several modalities in congruence with the sound to strengthen its perception and more effectively inform the monitoring operator even in cognitively complex situations.

2. BACKGROUND AND MOTIVATION

To support users reacting to dangerous or unpredicted events detected by algorithms, surveillance systems rely on audio or visual alarms. On one hand, visual animations are often used for helping users perceiving changes [24] or to shift their attentions [13], [18]. On the other hand, audible signals transmit important information or alert users through an item requiring immediate attention regardless of where users' current visual focus is.

The work by Gaver et al. highlights the ability of sound to provide useful information on processes and problems [10]. Several guides and experiments have been developed to guide sound interface designers to draw attention to and communicate the urgency of notification [14], [30], facilitate situational awareness of other operators [12] or for use in aircraft systems [22] or rail systems [23]. Texeira et al. [27] propose a gradual design of audible alarms allowing operators to distinguish the criticity level of alarms. The results of the implementation of such alarms suggest that more intelligible information reduces stress and the time spent verifying ambiguous cases or false alarms.

While sound interfaces offer potential benefits for monitoring activities, they are generally considered in isolation of visual components in current systems. Existing design guidelines rarely deal with their explicit combination, which would, among other advantages, improve situational awareness during change [24]. Our perception of the world takes advantage of all our senses and we constantly combine the different ways we understand and interact with our environment. One of the mechanisms we use to merge the inputs of these different channels is frequently defined as cross-modal interaction [25]. One of the main characteristics of a cross-modal interface is the transmission of information through two or more modalities, for example when oral comprehension is facilitated if the speaker's lip movements are visible.

Research on multi-sensory experience often uses the term congruence or cross-modal correspondence to refer to non-arbitrary associations between different modalities and their consequences on the processing of human information. For example, studies have revealed cross-modal associations between high-pitched sounds and bright, small objects at upper spatial locations, and between low-pitched sounds and dark rounded objects at lower locations [21]. This cross-modal congruence was identified as relevant for interface design [8], [25], and exploited in particular by Hoggan et al. [16] who showed that the perceived quality of the buttons on a touch screen was correlated with the congruence between the visual

and audio/tactile feedback used to represent them. Other studies suggest that bimodal feedback can increase performance and reduce perceived mental workload [28].

To address the challenges faced by operators with notifications and audio or visual alarms, designing cross-modal signals seems like a promising way to improve both the quality and the quantity of information transmitted to the users even with cognitive load issues. In the context of air or maritime fleet control, operators are required to pass multiple medical checks, including vision and auditory tests, to be fit for the position. Thus, we do not consider issues related to color blindness or deafness in this paper.

3. CROSS-MODAL ANIMATIONS DESIGNS

Before designing new systems for surveillance activities, we first wanted to explore congruent audio-visual mappings for simple animations. We define an animation as a temporal evolution of one or more audio and visual parameters of a multimodal stimuli. The temporal evolution is driven by a modulation signal that will be mapped to one or many audio-visual parameters.

The stimulus is made of a circular colored shape and a sound produced with frequency modulation synthesis [5]. This stimulus is meant to be overlaid on any item that raises an alarm in a surveillance system. For instance, such an alarm can be triggered when an aircraft altitude is too low, or when two ships are not respecting the minimal distance between them.

3.1. An ecological approach

We follow an ecological approach to the design of the stimulus, i.e. relationships that exist in the world such as a bigger object produces lower resonances or the closer the louder when an object moves. This approach is inspired by the sonic finder [9], work in designing audio alarms for medical contexts [7] or background monitoring [6].

In our application domain, the congruence of visual and spatial position seems appropriate. Indeed, the spatial position of the items on the map is already used to represent their GPS coordinates so we can match the position of the sound source to the location on screen with sound spatialization techniques.

3.2. Criticity levels

Complex surveillance systems are likely to produce a multitude of alerts, with the possibility of them happening simultaneously. To resolve the conflicts, the operator needs to perceive the level of criticity and assign them the proper amount of cognitive charge to be efficient. Here we designed two criticity levels, low and high.

Sound and visual parameters offer several possibilities for creating appropriate warning scales. For instance fundamental frequencies, harmonic series, envelope shape and modulation speed can influence the perceived urgency of sounds [7], [11]. These results have several implications for our two criticity levels. First, high criticity sound use a higher inharmonicity ratio in the synthesis to produce more inharmonic spectrum. Second, we add distortion to produce higher frequencies that also enhance the perceived emergency [11]. Finally, the animation, i.e. the temporal changes, should be different so that the induced changes are slow and "round" for low criticity and fast and sharp for high criticity. We use two different modulation envelopes : a sine function for low criticity and a sawtooth function with doubled speed for high criticity. Figure 1 illustrates these two modulation settings.

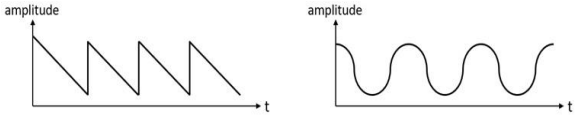

Figure 1: Modulation functions of animation parameters. Left: sawtooth. Right: sine

Regarding the visual parameters, we decided to mimic several existing systems by using the color hue to encode the criticity levels. We use yellow for low criticity and red for high criticity. This choice is intended for the lab experiment setup as a common design guideline but should not be interpreted as fixed rule. We are aware that for an end-user environment experiment, the designers will have to compose with the limitations of the panel of colors available to them.

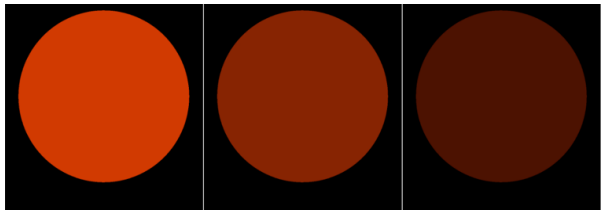

Figure 2: animating the size of the shape

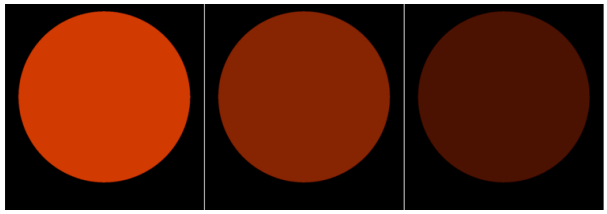

Figure 3: animating the brightness of the shape

The remaining visual, non-positional parameters that seem suitable to be animated are the size of the shape and its brightness as illustrated in Figure 2 and Figure 3. The size of the shape creates a motion that can guide the users' attention [13], [18]. The brightness has the advantage of preserving the shape which can be useful when the shape communicates the type of ships or other relevant information. Regarding audio parameters, we decided to animate the amplitude of the sound source, the pitch, and the dry/wet reverberation ratio and the lowpass filter cutoff frequency.

3.3. Congruent mappings

Based on the available parameters and our ecological approach, we propose four mappings between audio and visual parameters:

- M1 uses size as visual parameter and amplitude as sound parameter. It mimics a moving object going back and forth.
- M2 uses size as visual parameter and pitch as sound parameter. It mimics an object increasing or reducing its size which should respectively produce lower or higher sounds.
- M3 uses brightness as visual parameter and the dry/wet reverb ratio as sound parameter. It creates a diffuse sound stimulus similar to a temporal blur.
- M4 uses brightness as visual parameter and lowpass filter cutoff frequency as sound parameter. It mimics a fog that has a dampening effect on higher pitches [29].

4. PREFERENCE STUDY

Before evaluating the impact of cross-modal congruent alarms on surveillances tasks, we first need to validate our design approach. We conducted a preference study to better characterize subjective preferences on audio-visual mappings.

4.1. Hypothesis

We hypothesize that the ecological mappings should be preferred over non-ecological ones on both the associations between parameters and the polarity, i.e. whether an increase in the sound parameter should indicate an increase or decrease in the visual dimension [29]. For instance, size with pitch should be perceived as a better association than size with reverberation amount. Conversely, brightness with low pass filter cutoff frequency should be perceived as a better association than brightness with amplitude. Regarding polarity, we expect that the polarity suggested in M1, M2, M3 and M4 mappings to be preferred over opposites polarities.

4.2. Method

We ran an online preference test with 24 participants, 15 men and 8 women (M: 36 years; SD: 10,7 years) recruited with various research diffusion lists. One of them indicated being a professional in surveillance systems.
The first part of the online experiment introduces the tasks and indicates guidelines such as being in a quiet environment or wearing headphones before starting the experiment. The second part contains the tasks and the last part gathers information on the participants such as their age or their experience with surveillance systems and sound synthesis. The results were collected and anonymized before performing statistical analyses.

4.3. Task design

For each task, there is an animated visual (brightness or size) and a sound playing. The participant must rate the degree of harmony of the matching between the sound and the video.
We followed a [2×4×4×2] within-subject design with 4 primary factors: VISU ∈ [SIZE, BRIGH], AUDIO ∈ [AMP, PIT, REV, LPF], POLAR ∈ [NO, VR, AR, 2R], CON ∈ [CONT, DISC], as detailed below.
We tested two visual parameters (VISU): the size of the shape (SIZE) and its brightness (BRIGH). We tested four different audio parameters (AUDIO): the amplitude (AMP); the pitch (PIT); the reverberation ratio (REV); and the lowpass filter frequency (LPF).
Polarity (POLAR) is represented by the way one variable vary in association with another. There are two possible polarities: positive where both variables vary in the same direction, negative where variables vary in opposite directions. Based on those, we defined four orders of playing our audiovisual items: the visual variable is played forward and the audio variable is also played forward (NO), the visual is played forward and the sound in reverse (SR), the visual is played in reverse and the sound forward (VR), and both are played in reverse (2R).
We created two different conditions (CON) to challenge the robustness of the participants' preferences. A condition will consist in one of the two modulation curves, i.e. the function controlling the animation. The modulation curve is either a sawtooth function (SAW) or a sine function (SIN) as presented in Figure 1.

These conditions create a set of 64 possible mappings. To avoid fatigue and concentration biases we created two sets of 32 items. The different parameters are fairly divided between the 2 blocks, and each participant will be randomly affected to one of them. Participants were presented all items in a randomized order and had to rate each of them as illustrated in Figure 4.

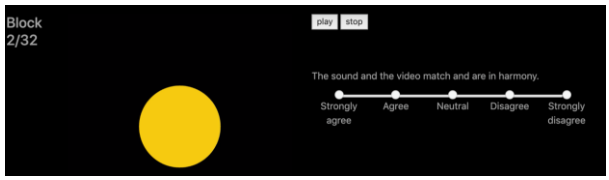

Figure 4: Example of an animated audio-visual item to be rated by the subjects

The rating of the harmonicity of the association between the audio and the visual is done on a Likert scale, from 1 to 5: The lowest rating corresponding to "Strongly disagree", then "Disagree", "Neutral", "Agree", and the highest rating "Strongly agree".

4.4. Results

We proceed with a statistical analysis of the results, first for the global audio-visual mappings, then on more detailed variables with polarities or modulations and compared the results with our assumptions.

We first ran test preferences between each audio-visual possible combination with repeated measures ANOVA. The aggregated results of the possible audio-visual association without considering the modulation nor the polarity resulted in a neutral score for each mapping and none is standing out as statistically significant.

	Amplitude	Pitch	Reverberation	LPF
Size	**m+ = 3,83** (M1) m- = 2,73 p < 0,001	**m+ = 4,00** m- = 2,94 (M2) p < 0,001	m+ = 2,85 **m- = 3,81** p < 0,001	**m+ = 3,94** m- = 2,88 p < 0,001
Brightness	**m+ = 3,81** m- = 2,83 p < 0,001	**m+ = 3,38** m- = 2,46 p < 0,001	m+ = 2,52 **m- = 3,83** (M3) p < 0,001	**m+ = 3,73** (M4) m- = 2,9 p = 0,001

Table 1: Significance of the effect of polarities for each mapping (bold = favorite polarity), and their mean score value.

We then studied preferences between polarities with dependent Student's t-test. The results indicate a significant effect of polarities on the users' preferences. Table 1 shows the mean score of each polarity within each mapping, and the result of the dependent samples t-test between them. Every single test returned a significative result ($p < 0.05$) on the effect of polarity on the preference score.

The distribution of the score only between the preferred polarity of each mapping is presented in Figure 5. No visual variables are significantly preferred in association with Amplitude, Reverberation, and Low-pass filer. However, the pitch is significantly preferred ($p < 0.001$) when associated with size Size (M = 4.0) than with Brightness (M=3.38).

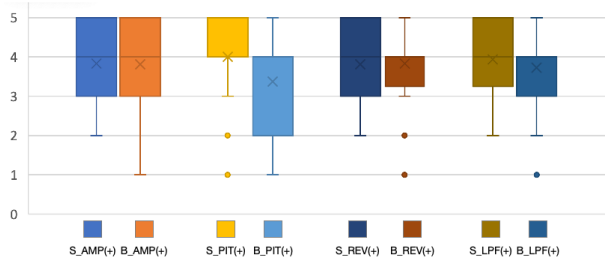

Figure 5: Box plots of score distribution for the preferred polarity of each association

We investigated the effect of the continuous or discontinuous modulation type with dependent Student's t-test. In six out of height cases the mapping preferences were robust to the variation of modulation. However, the preference for polarities did vary depending on the modulation function for two of the mappings, both in the positive polarity setup. The Size and AMP association is preferred with the discontinuous modulation over the continuous one ($M = 3.1$ vs $M = 2.3$, $p = 0.015$). The Size and LPF association is also preferred with the discontinuous modulation over the continuous one ($M = 3.3$ vs $M = 2.5$, $p = 0.015$).

5. DISCUSSION

Our results show that for each possible audio-visual association, there is a preferred polarity. These preferences are consistent with our ecological mappings M1, M3 and M4 but not for M2, that associates a size increase with a pitch decrease. In fact, the opposite preference was observed. We believe that this might be due to the fact that size is a physical invariant in everyday life, thus making it unlikely to change dynamically. Cases involving size change might imply transformations such as stretching the object which might produce a higher pitch.

We assumed that size would be preferred with amplitude (M1) or pitch (M2) and brightness with the reverberation ratio (M3) or lowpass filter frequency (M4). While our ecological approach seems appropriate, the results of the study does not show preferences for specific associations between visual parameters and audio parameters expect for the pitch. Even if the polarity is not the one we hypothesized, users seem to favor an association of pitch with size rather than brightness.

Regarding the other associations, it is possible that the brightness and the amplitude can be related via an intensity metaphor. Similarly, the dry/wet reverberation ratio can also be perceived with an object moving further away in a reverberating room and because there is also a attenuation of the high frequencies with the distance.

Regarding the effect of modulation on polarities, we observed an effect of the discontinuous over the continuous one with the negative polarity for Size and AMP and Size and LPF. This might be due to the fact that synchronization perception is facilitated with a discontinuity.

6. CONCLUSION AND PERSPECTIVES

Our goal is to design efficient audio-visual alarms to support fleet surveillance activities. We motivated the use of cross-modal congruent parameters interactions between sound alarms and visual animations, to improve operators' reaction time, ease of use and localization of alarms. We proposed audiovisual congruence interactions based on an ecological approach and conducted an experiment to assess user preferences of the possible associations.

While the results do not suggest preferences for specific associations, we found that for each possible association, a polarity is significantly preferred. These particular polarities can be used by designers to combine audio and visual stimuli.

To better characterize our design, we also need to validate our criticity level guidelines and to investigate the effect of congruency on attention-related tasks in a surveillance context. We are currently setting up another study to assess the effect of these new interactions on operators' reaction time and error rate against the existing alarm designs.

In our study, we only tested a subset of correlations between visual and sound that seemed the most relevant in our application domain but we are not excluding other cross-modal correlations to be promising and will further investigate these in future work. We are also concerned by the difference between an abstract warning signal designed in a lab, and an alarm signal in a professional environment associated with a strong mental representation [12]. For this reason, future work will focus on conducting field studies with maritime fleet centers and air traffic controllers.

7. ACKNOWLEDGEMENTS

We would like to thank Stephane Conversy and Jean-Luc Marini for their help and support in this project. This project has received funding from ANRT.

8. REFERENCES

[1] Sylvie Athènes, Stéphane Chatty, and Alexandre Bustico. 2000. Human factors in ATC alarms and notifications design: an experimental evaluation. *Proceedings of the USA/Europe Air Traffic Management R&D Seminar*.A. Bee, C. Player, and X. Lastname, "A correct citiation," in *Proc. of the 1st Int. Conf. (IC)*, London, UK, 2001, pp. 1119-1134.

[2] Michel Beaudouin-Lafon and William W. Gaver. 1994. ENO: Synthesizing Structured Sound Spaces. *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, ACM, 49–57.

[3] Tifanie Bouchara, Christian Jacquemin, and Brian F. G. Katz. 2013. Cueing Multimedia Search with Audiovisual Blur. *ACM Trans. Appl. Percept.* 10, 2: 7:1–7:21.

[4] Mickaël Causse, Jean-Paul Imbert, Louise Giraudet, Christophe Jouffrais, and Sébastien Tremblay. 2016. The role of cognitive and perceptual loads in inattentional deafness. *Frontiers in human neuroscience* 10: 344.

[5] John M. Chowning. 1973. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the audio engineering society* 21, 7: 526–534.

[6] Stephane Conversy. 1998. Ad-hoc synthesis of auditory icons. Georgia Institute of Technology.

[7] Judy Edworthy, Sarah Loxley, and Ian Dennis. 1991. Improving Auditory Warning Design: Relationship

between Warning Sound Parameters and Perceived Urgency. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 33, 2: 205–231.

[8] Thomas K. Ferris and Nadine B. Sarter. 2008. Cross-modal links among vision, audition, and touch in complex environments. *Human Factors* 50, 1: 17–26.

[9] William W. Gaver. 1989. The SonicFinder: An interface that uses auditory icons. *Human–Computer Interaction* 4, 1: 67–94.

[10] William W. Gaver, Randall B. Smith, and Tim O'Shea. 1991. Effective sounds in complex systems: The ARKola simulation. *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, ACM, 85–90.

[11] A. Guillaume, C. Drake, M. Rivenez, L. Pellieux, and V. Chastres. 2002. Perception of urgency and alarm design. Georgia Institute of Technology.

[12] Carl Gutwin, Oliver Schneider, Robert Xiao, and Stephen Brewster. 2011. Chalk sounds: the effects of dynamic synthesized audio on workspace awareness in distributed groupware. *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, ACM, 85–94.

[13] Johanna Haider, Margit Pohl, and Peter Frohlich. 2013. Defining Visual User Interface Design Recommendations for Highway Traffic Management Centres. *2013 17th International Conference on Information Visualisation*, IEEE, 204–209.

[14] Elizabeth J. Hellier, Judy Edworthy, and I. A. N. Dennis. 1993. Improving auditory warning design: Quantifying and predicting the effects of different warning parameters on perceived urgency. *Human factors* 35, 4: 693–706.

[15] Helen M. Hodgetts, François Vachon, Cindy Chamberland, and Sébastien Tremblay. 2017. See no evil: Cognitive challenges of security surveillance and monitoring. *Journal of applied research in memory and cognition* 6, 3: 230–243.

[16] Eve Hoggan, Topi Kaaresoja, Pauli Laitinen, and Stephen Brewster. 2008. Crossmodal congruence: the look, feel and sound of touchscreen widgets. *Proceedings of the 10th international conference on Multimodal interfaces*, ACM, 157–164.

[17] Anne R. Isaac and Bert Ruitenberg. 2017. *Air traffic control: human performance factors*. Routledge.

[18] Björn B. de Koning, Huib K. Tabbers, Remy M. J. P. Rikers, and Fred Paas. 2009. Towards a Framework for Attention Cueing in Instructional Animations: Guidelines for Research and Design. *Educational Psychology Review* 21, 2: 113–140.

[19] Lester F. Ludwig, Natalio Pincever, and Michael Cohen. 1990. Extending the notion of a window system to audio. *Computer* 23, 8: 66–72.

[20] Arien Mack and Irvin Rock. 1998. *Inattentional blindness*. MIT press Cambridge, MA.

[21] Geoffrey R. Patching and Philip T. Quinlan. 2002. Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance* 28, 4: 755.

[22] Roy D. Patterson. 1982. *Guidelines for auditory warning systems on civil aircraft*. Civil Aviation Authority.

[23] UK Rail Safety. *Standards Board. Alarms and alerts guidance and evaluation tool*.

[24] Céline Schlienger, Stéphane Conversy, Stéphane Chatty, Magali Anquetil, and Christophe Mertz. 2007. Improving Users' Comprehension of Changes with Animation and Sound: An Empirical Assessment. In C. Baranauskas, P. Palanque, J. Abascal, and S.D.J. Barbosa, eds., *Human-Computer Interaction – INTERACT 2007*. Springer Berlin Heidelberg, Berlin, Heidelberg, 207–220.

[25] Charles Spence and Jon Driver. 1997. Cross-modal links in attention between audition, vision, and touch: Implications for interface design. *International Journal of Cognitive Ergonomics*.

[26] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*. Elsevier, 37–76.

[27] Bruno Teixeira De Sousa, Alessandro Donati, Elif Özcan, et al. 2016. Designing and deploying meaningful audio alarms for control systems. *14th International Conference on Space Operations*, 2616.

[28] Holly S. Vitense, Julie A. Jacko, and V. Kathlene Emery. 2003. Multimodal feedback: an assessment of performance and mental workload. *Ergonomics* 46, 1–3: 68–87.

[29] Bruce N. Walker and Gregory Kramer. 2004. Ecological Psychoacoustics and Auditory Displays: Hearing, Grouping, and Meaning Making. *Ecological psychoacoustics*: 150–175.

[30] Marcus O. Watson and Penelope M. Sanderson. 2007. Designing for attention with sound: challenges and extensions to ecological interface design. *Human Factors* 49, 2: 331–346.