Towards congruent cross-modal audio-visual alarms for supervision tasks

Eliott Audry, Jérémie Garcia

Towards congruent cross-modal audio-visual alarms for supervision tasks

ELIOTT AUDRY*[1,2], JEREMIE GARCIA†[2]

[1]SAFETY-DATA-CFH, TOULOUSE, FRANCE
[2]ENAC – UNIVERSITE DE TOULOUSE, TOULOUSE, FRANCE

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 conclude with the design of a study to validate our designs as well as future research directions.

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]. In addition to continuous visual representations of traffic, 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.

* eliott.audry@enac.fr
† jeremie.garcia@enac.fr

In this paper we describe our ongoing efforts to design congruent cross-modal audio-visual alarms. The paper first introduces the background and motivations for this work. We describe our design space and possible correlations between visual and audio parameters for simple animations. We then present the design of a study to assess these correlations. We conclude with perspectives for future research.

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 high-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 certain types of 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.

DESIGNING CROSS-MODAL ANIMATIONS

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.

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] as well as 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.

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. to Figure 1 illustrates these two modulation settings.

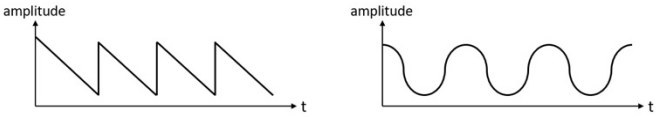

Figure 1: Modulation functions or animating parameters. Left: sawtooth. Right: sine

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 designer will have to compose with the limitations of the panel of colors available to him.

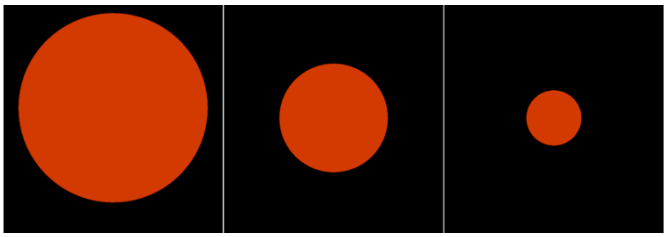

Figure 2: Size animation

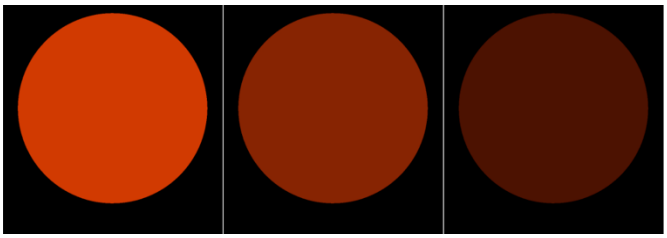

Figure 3: Brightness animation

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 distance of the sound source to the listener, the pitch, and the dry/wet reverberation ratio and the lowpass filter cutoff frequency. The distance of the sound source is similar to the amplitude of the sound but also considers the fact that higher frequencies are attenuated faster than lower frequencies while distance increases [2,19]. Also, the relation between size and sound distance is not linear, even when the object starts to be small the sound can still be heard at a reasonable level.

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 distance 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 [3].

PLANNED EXPERIMENTATION

Before evaluating the impact of cross-modal congruent alarms on surveillances tasks, we first need to validate our ecological design approach. We are currently designing a web-based experiment to better characterize subjective preferences on audio-visual mappings.

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 distance. Regarding polarity, we expect that the polarity suggested in M1,M2, M3 and M4 mappings to be preferred over opposites polarities.

PARTICIPANTS AND APPARATUS

We will contact participants with and without expertise in surveillance systems via email for a study about audiovisual cross-modal preferences. A first part of the web page will gather information on the participants such as their age or their experience with surveillance systems and sound synthesis. A second part will introduce the tasks and indicate guidelines such as being in a quiet environment or wearing headphones before starting the experiment. Finally, the last part will contain the tasks.

PREFERENCE TASKS

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 will follow a [2×4×4×2] within-subject design with 4 primary factors: VISU ∈ [SIZE, BRIGH], AUDIO ∈ [DST, PIT, REV, LOW], POLAR ∈ [NO, VR, AR, 2R], CON ∈ [CONT, DISC], as detailed below.

Four different audio parameters will be tested:

AUDIO	Abbr.
Distance	DST
Pitch	PIT
Reverberation ratio	REV
Lowpass filter frequency	LOW

Table 1: Audio variables tested with a size or brightness animation, and their abbreviations

Polarity is represented by the way one variable vary in association with another. There are 2 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 the experiment will be divided into 2 blocks of 32 items. The different parameters will be fairly divided between the 2 blocks, and each participant will be randomly affected to one of them. Participants will be presented all items in a randomized order and will have to rate each of them.

The rating of the harmonicity of the association between the audio and the visual will be done on a discontinuous scale, also known as Likert scale, from 0 to 4: The lowest rating corresponding to "Strongly disagree", then "Disagree", "Neutral", "Agree", and the highest rating "Strongly agree".

DATA COLLECTION

We will store the results in separated anonymized files. We will test preferences between each cross-modal combination with repeated measures ANOVA and compare the results with our assumptions.

DISCUSSION AND PERSPECTIVES

This paper addresses the perception efficiency of audio-visual alarms for fleet surveillance activities. Our work focuses on cross-modal congruent parameters interactions between sound alarms and visual animations, to improve operators' reaction time, ease of use and localization of alarms. Our contributions include new propositions on audiovisual congruence interactions, and the design of an experiment to validate these assumptions as a first step.

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

It should be noted that for this experiment we will test a subset of correlations between visual and sound that seemed pertinent to us, 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.

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.

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*.
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.