# Caution and Warning Alarm Design and Evaluation for NASA CEV Auditory Displays

***SHFE Information Presentation Directed Research Project (DRPP) report 12.07***

*Durand R. Begault*[1]
*Martine Godfroy*[1]
*Aniko Sandor*[2]
*Kritina Holden*[3]

[1]Human Systems Integration Division (Code TH)
NASA Ames Research Center
Moffett Field CA 94035
[2]LZ Technology [3]Lockheed Martin Corporation
Usability Testing and Analysis Facility (Code SF3)
NASA Johnson Space Center
Houston TX 77058

## Abstract

The design of caution-warning signals for NASA's Crew Exploration Vehicle (CEV) and other future spacecraft will be based on both best practices based on current research and evaluation of current alarms. A design approach is presented based upon cross-disciplinary examination of psychoacoustic research, human factors experience, aerospace practices, and acoustical engineering requirements. A listening test with thirteen participants was performed involving ranking and grading of current and newly developed caution-warning stimuli under three conditions: (1) alarm levels adjusted for compliance with ISO 7731, "Danger signals for work places – Auditory Danger Signals", (2) alarm levels adjusted to an overall 15 dBA s/n ratio and (3) simulated codec low-pass filtering. Questionnaire data yielded useful insights regarding cognitive associations with the sounds. *Funded by a directed research program of NASA's Space Human Factors Engineering project.*

## 1. Background

In this report, existing alarms currently in use with NASA flight deck displays are analyzed and then alternative designs are proposed that are compliant with ISO 7731, "Danger signals for work places – Auditory Danger Signals", and that correspond to suggested methods in the literature to insure discrimination and audibility. Listening tests are performed with thirteen subjects to evaluate the results.

Future development of auditory "sonification" techniques into the design of alarms will allow auditory signals to be extremely subtle, yet extremely useful for indicating trends or root causes of failures. A summary of 'best practice' engineering guidelines was given previously by the authors, along with results of an experiment involving subjective classification of alarms by ten subjects (Begault, D. R., Godfroy, M., Sandor, A. and

Holden, K. “Auditory alarm design for NASA CEV applications” *Proceedings of the 13th International Conference on Auditory Display, Montreal, CA, 26-29 June 2007*).

Based on the results of this previous study, a set of synthesized alternative alarms to the existing class 1-3 alarms was developed. It this study, the effects of level and bandwidth on the acceptability of the alarms was examined: (1) alarm levels adjusted for compliance with ISO 7731, "Danger signals for work places – Auditory Danger Signals", (2) alarm levels adjusted to an overall 15 dBA s/n ratio, and (3) simulated codec low-pass filtering. Rating and ranking data were obtained, along with questionnaire data. The results indicated no significant difference as a function of bandwidth, yielding useful implications for engineering implementation and test and verification. The questionnaire data yielded useful insights regarding cognitive associations with the sounds. Some comments seemed to be driven by an internal reference as to what an ‘urgent’ alarm should sound like, independent of preference.

An ‘auditory alarm’ for purposes of this report refers to any audio signal used for alerting or warning a user within a human-machine interface, while an ‘alarm’ refers generically to either audio or visual cues. The use of auditory alarms in current Shuttle applications is reviewed in technical documents. Auditory alarms are part of the collective caution and warning (c/w) system that consists primarily of visual cues (illuminated light displays and switches, an illuminated message on an dedicated matrix panel, or a text message on a CRT).

There are four classes of alarms used on shuttle, which can be prioritized in ascending order as follows. A “class 0” alarm visually indicates up and down arrows on the CRT display next to a specific parameter, indicating that it has exceeded its predefined upper or lower boundary limits. There is no auditory component for a class 0 alarm. A “class 3” alarm is technically an “alert” and generates a steady tone of 512 Hz for approximately 1 second (this can be changed by the crew to longer durations, up to 99 seconds), along with an illuminated button and fault message on the CRT. A “class 2” alarm generates an illuminated text message on a dedicated matrix panel (panel number F7), and illuminates parameter lights on another panel (number R13U). The alarm consists of an alternating tone between 375 and 1000 Hz. It is silenced (“killed”) by pressing a master alarm switch.

There are two types of class 1 “emergency” alarms that are highest priority: (1) smoke detection and (2) rapid cabin depressurization. The smoke detection alarm consists of a “siren” sound, i.e., a tone varied from 666 to 1,460 Hz and then back to 666 Hz over a 5 second interval. Smoke detection lights are indicated on a dedicated panel (number L1). The cabin depressurization alarm is indicated via a “klaxon” sound, consisting of two tones at 270 and 2500 Hz that are periodically iterated. Pressing the master alarm switch also silences these alarms. Under the current design, it is possible for all of the auditory alarms to sound simultaneously.

These auditory alarms have three primary functions. First, they indicate that a specific condition exists that did not occur previously in time, and that now requires attention.

This may include the corollary function of waking a sleeping crewmember. Second, they have a rudimentary function of stating: "*look over here at this specific visual display*". This is a form of "directed attentional shift" that is significant in the larger context of the cognitive challenge of fault management. Third, their function is to relate the relative urgency of the alarm through the semantic content contained in the alarm type. The type of alarm indicates: "*where in the hierarchy of possible auditory alerts does this new alarm lie?"* and "*how quickly do I need to attend to this problem?"*

## 2. Experiment goals and conditions

This experiment sought to build on the results of the previous experiment and to determine the effect on ratings as a function of specific conditions related to level or limited bandwidth conditions.

## 3. Subjects

Thirteen subjects were run at the Usability Testing and Analysis Facility, located in Building 15 of NASA- Johnson Space Center. The subjects were not crewmembers or otherwise familiar with the current class 1, 2 or 3 alarms.

## 4. Experimental design

The goal of the experiment was three-fold: (1) to compare the rating of alarms regarding their suitability for representing a particular class of emergency situation; (2) to determine if there was an effect of calibration method on overall preference; (3) to determine if there was an effect of bandwidth on overall preference.

Regarding the first goal, a method was used based roughly on ITU-R recommendation BS.1534, which was developed to gain continuous scale sound quality ratings for stimuli with 'large scale' (obvious) differences. In this method, a graphic user interface (GUI) is used that allows continuous playback of the stimuli to be compared within a single trial. The GUI allows subjects to both rate and rank the various types of alarms. A difference from the BS.1534 recommendation was that a fixed reference was not used; i.e. subjects were not instructed to compare stimuli to any specific reference. Therefore, the stimuli ratings represent "absolute category ratings". Subjects were instructed to rate each of the alarms within a single trial on a continuous scale of "suitability" of the alarm for the particular warning that was to be conveyed. The suitability scale ranged from "very unsuitable" (1.0) to "very suitable" (5.0)

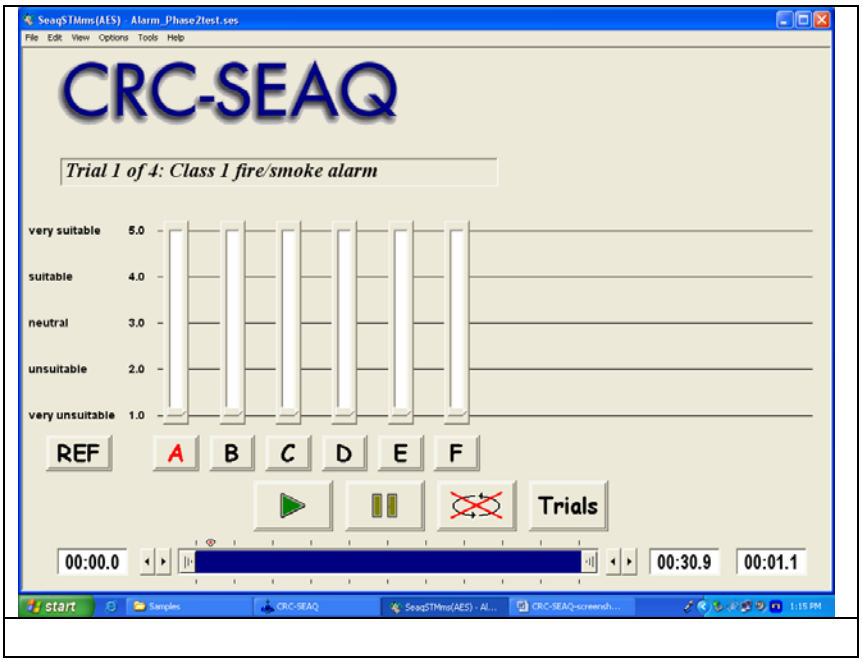


**Figure 1.** Screenshot of the listening test user interface, for a single trial.

There were a total of four trials, one for each alarm type, under three experimental conditions (twelve trials in total). The continuous rating scale appears at left (very unsuitable" = 1.0 – "very suitable" = 5.0); the individual stimulus files (alarms in a context of flight deck noise) were activated by the buttons with letter labels (A-F). The reference sound ("REF") was the shuttle's background noise. Subjects could repeat listening to each stimulus in order to make comparative ratings with the slider control located above each stimulus.

A set of four trials was developed, one for each category of alarm currently used on Shuttle: class 1 (fire/smoke); class 1 (depressurization); class 2 (warning); class 3 (caution). Within each trial there was one "hidden reference" representing the existing alarm used for each condition, and five "alternative" alarms based on results from a previous study on alarms conducted by the same authors.

To determine if there was a significant effect of calibration method or of bandwidth on the suitability ratings, the four trials were presented under three conditions, for a total of 12 trials:

***Condition 1***

> Full bandwidth, alarm adjusted to +15dB (A-weighted level; "+15dB(A)") relative to the background noise level (reflecting common "rule of thumb" implementation method and equivalent to the broadband estimate method of ISO 7731-1986(E) "Danger Signals for Work Places- Auditory Danger Signals").

***Condition 2***

Full bandwidth, alarm adjusted to +13 dB in at least one 1/3-octave band re noise (reflecting the one-third-octave band method of ISO 7731-1986(E) "Danger Signals for Work Places- Auditory Danger Signals"). This generally results in somewhat lower signal levels compared to a +15 dBA rule (condition 1) since an individual spectral component can 'drive' the overall level.

***Condition 3***

Telephone bandwidth (signal low-pass filtered to 3 kHz), adjusted to +15 dBA re background noise. The low-pass filtering emulates the characteristic of some low-bit rate codecs (e.g., G.729) and is of interest since higher frequency components of alarms or signal sweeps will be inaudible or distorted compared to a full-bandwidth version. The level will be somewhat higher compared to condition 1 or condition 2 since high frequency information is absent from the signal.

## 5. Stimuli

The core set of stimuli for this experiment were derived from alarms chosen in a previous study by at least 75% of participants as being 'strongly identified' as a class 1, 2 or 3 alarm (reported in Begault, D. R., Godfroy, M., Sándor, A. and Holden, K. "Auditory alarm design for NASA CEV applications" *Proceedings of the 13th International Conference on Auditory Display, Montreal, CA, 26-29 June 2007*). In that study, subjects sorted a set of 49 candidate alarms that included current Shuttle caution and warning signals, synthesized variants on these signals, and an ad hoc collection of various alerts heard in everyday life (e.g., the 'flight attendant bell' from a commercial airliner).

For each of four trial types, six stimuli, one of which included the current alarm used on shuttle, were presented: class 1 (fire/smoke); class 1 (depressurization); class 2 (warning); class 3 (caution). See Table I.

Table I indicates the arrangement of trials. For all trials, the sounds were randomized and conditions were counterbalanced.

TABLE 1. Stimuli within each trial. ("y16", etc. represent stimuli index code)

| Stimuli (randomized) | Trial 1 class 1 fire/smoke | Trial 2 class 1: depressurization | Trial 3 class 2 - warning | Trial 4 class 3 - caution |
|---|---|---|---|---|
| A | y16 | y23 | y10 | y5 |
| B | y17 | y24 | y11 | y6 |
| C | y18 | y25 | y12 | y7 |
| D | y19 | y21 | y13 | y8 |
| E | y20 | y22 | y14 | y9 |
| F | current class 1 F/S | current class 1 depr. | current class 2 | current class 2 |

Stimuli were processed digitally and analyzed using an Agilent HP 35670A dynamic signal analyzer and a Bruel & Kjaer Head and Torso simulator (4100D). BeyerDynamic DT-990 headphones were used for calibration and playback to subjects.

Each alarm was presented multiple times at 1 - 2 s intervals, within a background noise spectrum designed to simulate realistic noise conditions for space operations. The background noise spectrum was taken from an actual recording of U.S. Lab made 6 in. from the forward Audio Terminal Unit (ATU), and was binaurally processed using NASA SLAB software so that, over headphones, the noise seemed 'externalized' and not completely inside of the head. The calibrated noise level was approximately 60 dBA (Leq, slow 15s time weighting); this level is equivalent to the maximum noise level permitted by current HSIR requirement of NC 52 (NC = "noise criteria level", a set of curves based on an average noise spectrum). The alarms were set to either +13 dBA (Lmax fast) or +15 dBA relative to the measured background noise. The overall levels were likely to have varied as much as -/+ 2 dB due to headphone donning. This variability is not considered to influence the overall results.

## 6. Result I: Effect of condition

Repeated Measures Analysis of Variance (R-ANOVA) indicated a significant main effect of condition ($F(2,12) = 28.7$, $p < .0001$). Further analyses indicated no significant effect of condition caused by bandwidth (condition 1 versus condition 3), and a significant effect of level showing a slight preference for the higher sound level (15 dBA) warning condition. Refer to Appendix A and Figures 1A- 4A (appendix A) for a detailed presentation of these data. Despite the indication of a significant difference among conditions, the overall ratings were similar enough among conditions to allow for pooling of data to study overall ratings and rankings of alarms.

## 7. Results II: Ranking and sorting of alarms

For each trial (alarm type in each of the conditions), the six alarms were rated on the 1-5 "suitability" scale, and also ranked. One of the alarms included the currently used alarm, allowing an analysis of its position relative to other newly synthesized, ISO-compliant alarms. Some alarms were also found to be 'unsuitable' by nearly all subjects. Refer to Appendix A, and Figure 5A, for a graphic representation of these data.

Table II below indicates the mean value ratings and rankings for each stimulus type across subjects. For each condition, the mean rating and rankings are shown, top to bottom. The current Shuttle alarm is indicated in cross hatching. Averaging across conditions, the current alarms are ranked as follows, with 2.0 = "unsuitable", 3.0 = "neutral", and 4.0 = "suitable:

| | |
|---|---|
| Class 3: | 2.0 |
| Class 2: | 3.3 |
| Class 1 Depressurization: | 3.6 |
| Class 1 Fire-Smoke: | 3.4 |

TABLE II. Ratings and rankings for each condition. Cross-hatching indicates current shuttle alarm. Y = condition 1; Z = condition 2; YY = condition 3.

| | | | | | | |
|---|---|---|---|---|---|---|
| Class 1 | | | | | | |
| Emergency depressurization. | *y2* | 4.0 | *z24* | 3.3 | *yy2* | 4.0 |
| | *y24* | 3.8 | *z25* | 3.0 | *yy25* | 3.8 |
| | *y25* | 3.7 | *z23* | 2.9 | *yy24* | 3.6 |
| | *y23* | 3.4 | *z2* | 2.7 | *yy23* | 3.2 |
| | *y21* | 2.3 | *z21* | 2.6 | *yy21* | 2.4 |
| | *y22* | 2.1 | *z22* | 2.1 | *yy22* | 2.2 |
| Class 1 | | | | | | |
| Emergency fire-smoke | *y1* | 3.7 | *z16* | 3.3 | *yy16* | 3.9 |
| | *y20* | 3.6 | *z17* | 3.2 | *yy17* | 3.7 |
| | *y16* | 3.5 | *z1* | 3.2 | *yy18* | 3.6 |
| | *y17* | 3.4 | *z18* | 3.1 | *yy1* | 3.4 |
| | *y18* | 3.2 | *z20* | 2.2 | *yy20* | 3.3 |
| | *y19* | 2.0 | *z19* | 1.5 | *yy19* | 2.0 |
| Class 2 | | | | | | |
| Warning | *y14* | 3.7 | *z14* | 3.2 | ***yy3*** | 3.8 |
| | *y13* | 3.7 | *z13* | 2.9 | *yy13* | 3.7 |
| | *y3* | 3.4 | *z3* | 2.7 | *yy14* | 3.7 |
| | *y10* | 3.3 | *z10* | 2.4 | *yy10* | 3.4 |
| | *y12* | 1.7 | *z12* | 1.8 | *yy12* | 1.9 |
| | *y11* | 1.6 | *z11* | 1.7 | *yy11* | 1.7 |
| Class 3 | | | | | | |
| Caution | *y5* | 3.0 | *z9* | 3.0 | *yy7* | 3.4 |
| | *y7* | 3.0 | *z7* | 2.9 | *yy9* | 3.1 |
| | *y9* | 2.9 | *z8* | 2.5 | *yy5* | 2.9 |
| | *y6* | 2.8 | *z5* | 2.4 | *yy8* | 2.8 |
| | *y8* | 2.5 | *z4* | 2.4 | *yy6* | 2.6 |
| | ***y4*** | 1.6 | *z6* | 2.0 | *yy4* | 2.1 |

The following observations can be made from Table II across conditions, with regards to the current alarm and the new alarms with the highest average rating:

• For class 3 alarms, the current alarm (y4, z4, yy4) is ranked last or near last. Alarm stimulus 7 is ranked 2nd or 1st and has the highest average rating (3.1), slightly above "neutral".

• For class 2 alarms, the current alarm (y3, z3, yy3) is ranked 3rd under two conditions and 1st under the limited bandwidth condition. Alarm stimuli type 13 are consistently ranked 2nd under all conditions, while alarm stimuli type 14 are ranked 1st under two conditions and 3rd under the limited bandwidth condition. Alarm stimuli 13 and 14 have the highest average rating of 3.5, midway between "neutral" and "appropriate".

• For class 1 depressurization, the current alarm is ranked 1st under two conditions and 4th under one condition. Alarm stimuli type 24 are ranked 2nd, 1st and 3rd under the different conditions. The average ratings for these two alarms is 3.6, slightly above the midpoint between "neutral" and "appropriate".

• For class 1 smoke-fire, the current alarm is ranked 1st, 3rd and 4th under the different conditions. Alarm stimuli type 16 are ranked 3rd under condition 1 and 1st under conditions 2 and 3. The average rating for the current alarm is 3.4, and the average rating for alarm stimuli type 16 is 3.5.

## 8. Results III: Questionnaire data

At the end of each trial, subjects were asked the following questions regarding specific areas of interest:

1. Have you recognized any of the sounds?
2. Describe the basis of your ratings.
3. Why did you rate "this" sound higher than "that" sound? What is the difference that you noticed?
4. Based on the sounds that you heard and the ratings that you gave them, what do you think the best sound would be for this category of alarms? Try to describe it.

The analysis of these verbal subjective impressions will be presented in a future paper. Overall, there was a tendency for subjects to be influenced by their cognitive association with a sound (e.g., "airliner fasten seat belt chime"). Many of the comments made statements to the effect that "this is what this sort of alarm should sound like". In other words, their ratings of suitability were probably as much influenced by an internal model of 'suitability', based on a cognitive reference to ecological experience of alarms, as they were by an internal scale of personal preference. The authors have commented previously in the context of 'best practices' that associative influences are important to recognize in the design of an auditory display.

## 9. Overall summary

No significant effect was found as a function of low-pass filtering that might occur with implementation of a low bit-rate codec, suggesting that this factor can be ignored when considering the design of alarms. The significant but small effect of level ("15 dB rule" versus the "13 dBA rule" of conditions 1 and 2) may be explained by a preference for a higher signal-noise ratio; there is no specific impact of this result. The rating and ranking data show that particular attention should be paid to the design of improved class 2 and 3 warnings. However, none of the alarms tested were rated any higher than 3.6 on average, suggesting that more work needs to be paid to the means of their synthesis and that new possibilities for alarms ought to be explored.

A future study will examine a subset of the conditions using crew members as subjects. Their data compared to 'layperson' data will be of interest, particularly as regards ratings, rankings, and questionnaire data results.

## 10. Acknowledgments

The authors thank Dr. Efrem Reeves for his assistance with calibration verification.

## APPENDIX A: Analysis of variance of the experimental conditions

Significant Effect of condition:

**ANOVA Table for Condition**

| | DF | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 311 | 902.514 | 2.902 | | |
| Category for Condition | 2 | 36.869 | 18.434 | 28.734 | <.0001 |
| Category for Condition * S... | 622 | 399.045 | .642 | | |

Reliability Estimates - All Treatments: .759; Single Treatment: .513

Significant effect of level (condition 1 (C1) versus condition 2 (C2)), but no effect of bandwidth (condition 1 (C1) versus condition 3 (C3)).

C1, C2: $p<0.0001$ S; C2, C3: $p<0.0001$ S, C1, C3: $p=0.14$, NS

The reduction of the level leads to a decrease in the rating (mean C1=2.99; mean C2=2.62, mean C3= 3.08).

A significant interaction is observed between subjects and conditions, essentially due to subjects 2 and 6, who express a very low rating in condition 2.

**Tableau ANOVA pour Variable compacte sans titre #1**

| | ddl | Somme des carrés | Carré moyen | Valeur de F | Valeur de p | Lambda | Puissance |
|---|---|---|---|---|---|---|---|
| Sujet | 12 | 104.428 | 8.702 | 3.260 | .0002 | 39.124 | .997 |
| Sujet(Groupe) | 299 | 798.085 | 2.669 | | | | |
| Catégorie pour Variable compacte sans t... | 2 | 36.869 | 18.434 | 31.634 | <.0001 | 63.268 | 1.000 |
| Catégorie pour Variable compacte sans t... | 24 | 50.567 | 2.107 | 3.616 | <.0001 | 86.775 | 1.000 |
| Catégorie pour Variable compacte sans t... | 598 | 348.478 | .583 | | | | |

When comparing exclusively C1 and C3, the subject effect is no longer significant and the interaction effect disappears. This observation confirms that the absence of effect of the frequency filtering is common to all subjects.

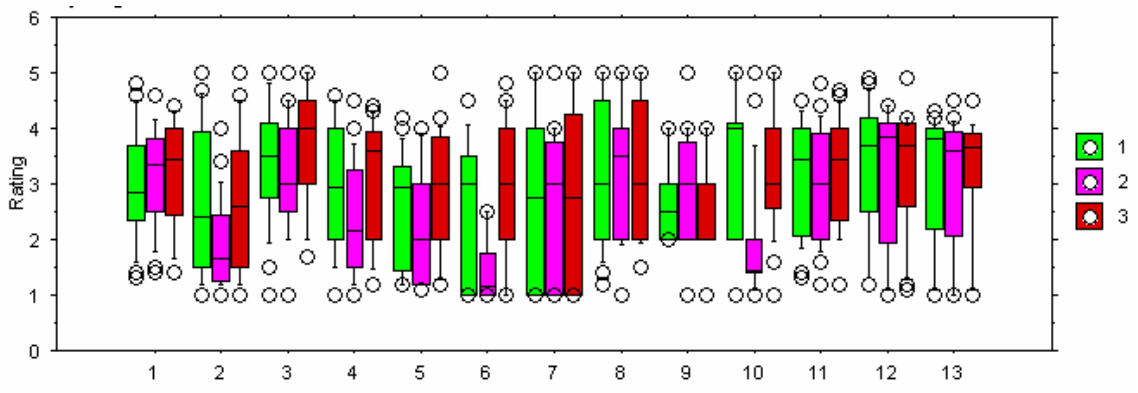


**Figure 1A:** ***Rating for the 13 subjects as a function of the condition***

There is no effect of interaction between the type of sound and the condition of presentation of the alarm, i.e. whatever the type of sound, the C2 condition leads to a reduction in the rating. It is not type-specific.

**Tableau ANOVA pour Variable compacte sans titre #1**

| | ddl | Somme des carrés | Carré moyen | Valeur de F | Valeur de p | Lambda | Puissance |
|---|---|---|---|---|---|---|---|
| Type | 3 | 31.224 | 10.408 | 3.679 | .0125 | 11.038 | .805 |
| Sujet(Groupe) | 308 | 871.289 | 2.829 | | | | |
| Catégorie pour Variable compacte sans t... | 2 | 36.869 | 18.434 | 28.773 | <.0001 | 57.545 | 1.000 |
| Catégorie pour Variable compacte sans t... | 6 | 4.379 | .730 | 1.139 | .3379 | 6.835 | .446 |
| Catégorie pour Variable compacte sans t... | 616 | 394.666 | .641 | | | | |

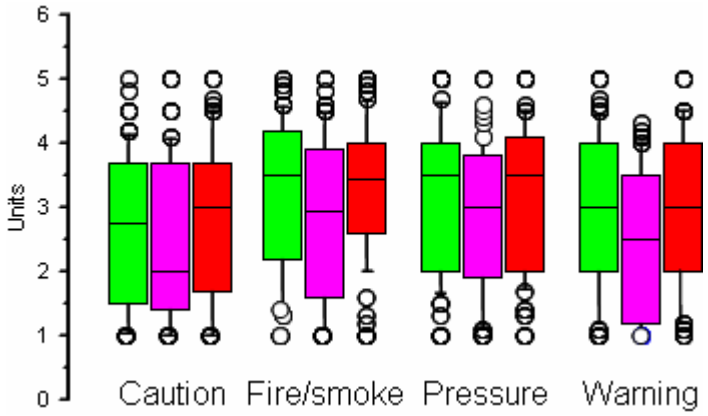


**Figure 2A:** ***Rating for the 4 Types of Alarms as a function of the condition***

Nevertheless a significant Condition* Type*Subject is observed that persists for the paired comparison of conditions C1 and C3 (second table). Observation of Figures 3A-4A suggest that some subjects might be sensitive to a reduction of the bandwidth for some specific type of sounds.

**Tableau ANOVA pour Condition**

| | ddl | Somme des carrés | Carré moyen | Valeur de F | Valeur de p | Lambda | Puissance |
|---|---|---|---|---|---|---|---|
| Type | 3 | 31.224 | 10.408 | 3.879 | .0097 | 11.638 | .828 |
| Sujet | 12 | 104.428 | 8.702 | 3.243 | .0002 | 38.921 | .997 |
| Type * Sujet | 36 | 69.265 | 1.924 | .717 | .8851 | 25.816 | .733 |
| Sujet(Groupe) | 260 | 697.596 | 2.683 | | | | |
| Catégorie pour Condition | 2 | 36.869 | 18.434 | 34.092 | <.0001 | 68.185 | 1.000 |
| Catégorie pour Condition * Type | 6 | 4.379 | .730 | 1.350 | .2333 | 8.098 | .524 |
| Catégorie pour Condition * Sujet | 24 | 50.567 | 2.107 | 3.897 | <.0001 | 93.519 | 1.000 |
| Catégorie pour Condition * Type * Sujet | 72 | 62.926 | .874 | 1.616 | .0018 | 116.376 | 1.000 |
| Catégorie pour Condition * Sujet(Groupe) | 520 | 281.172 | .541 | | | | |

**Tableau ANOVA pour Condition**

| | ddl | Somme des carrés | Carré moyen | Valeur de F | Valeur de p | Lambda | Puissance |
|---|---|---|---|---|---|---|---|
| Type | 3 | 29.312 | 9.771 | 4.222 | .0062 | 12.666 | .864 |
| Sujet | 12 | 47.050 | 3.921 | 1.694 | .0681 | 20.330 | .859 |
| Type * Sujet | 36 | 58.778 | 1.633 | .705 | .8962 | 25.397 | .724 |
| Sujet(Groupe) | 260 | 601.728 | 2.314 | | | | |
| Catégorie pour Condition | 1 | 1.338 | 1.338 | 3.401 | .0663 | 3.401 | .435 |
| Catégorie pour Condition * Type | 3 | .814 | .271 | .689 | .5594 | 2.067 | .190 |
| Catégorie pour Condition * Sujet | 12 | 3.724 | .310 | .789 | .6622 | 9.463 | .455 |
| Catégorie pour Condition * Type * Sujet | 36 | 26.312 | .731 | 1.857 | .0033 | 66.861 | .999 |
| Catégorie pour Condition * Sujet(Groupe) | 260 | 102.318 | .394 | | | | |

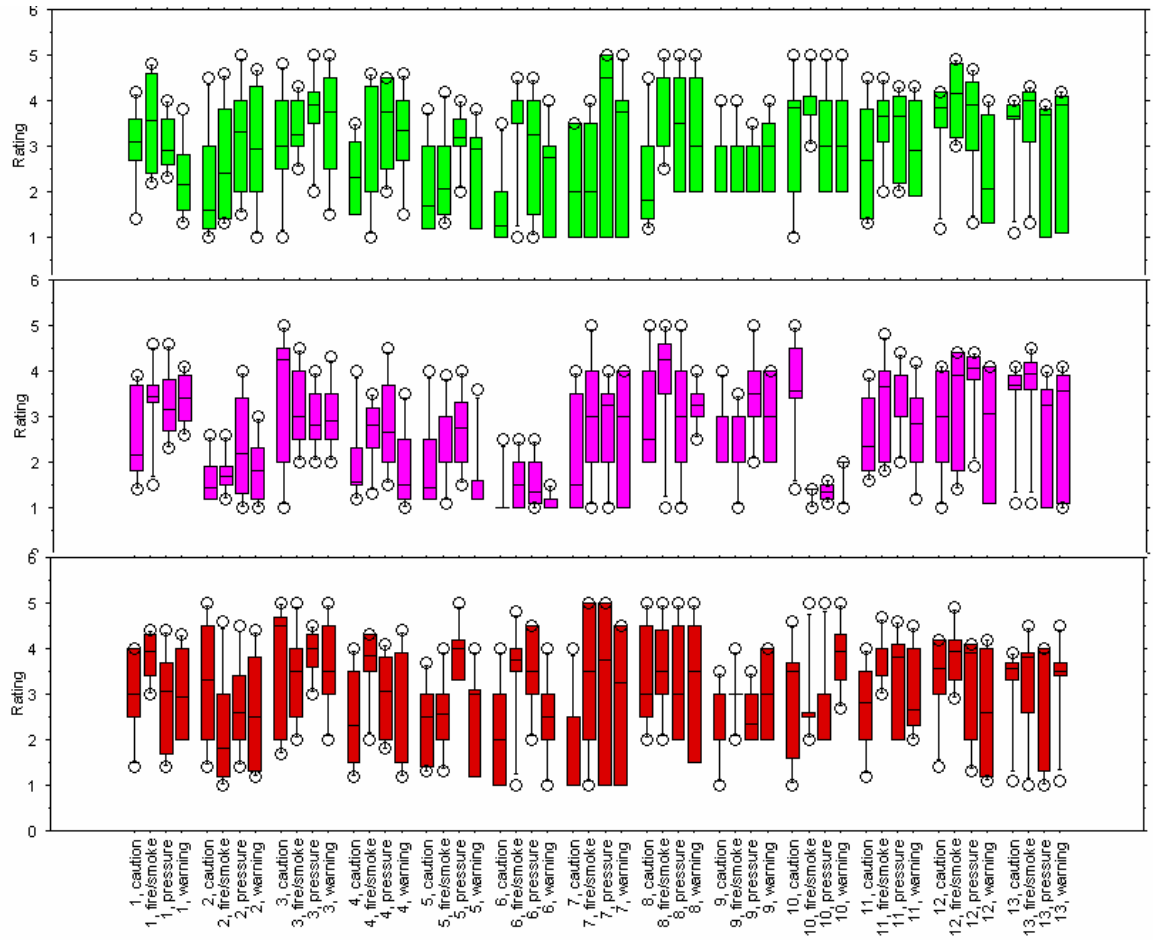


**Figure 3A:** ***Rating for the 13 subjects as a function of the condition (from top to bottom, C1, C2 and C3) and the type of sound (Caution, Fire/Smoke, Pressure, Warning)***

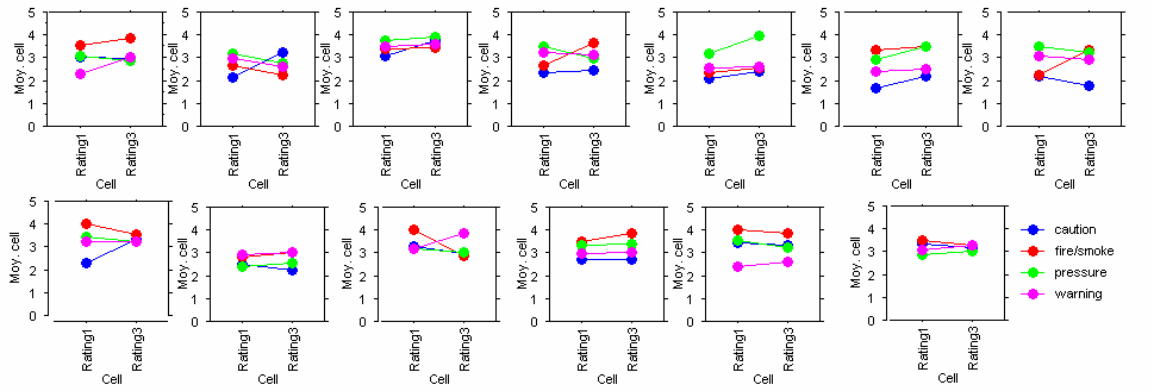


**Figure 4A:** ***Rating for the 13 subjects as a function of the condition (C1 top, C3 bottom) and the type of sound.***

Some sounds jump out as being rated particularly inappropriate for a given category: see Figure 5A. Specifically:
Caution: 1
Pressure: 21, 22
Warning: 11, 12
Fire/Smoke: 19

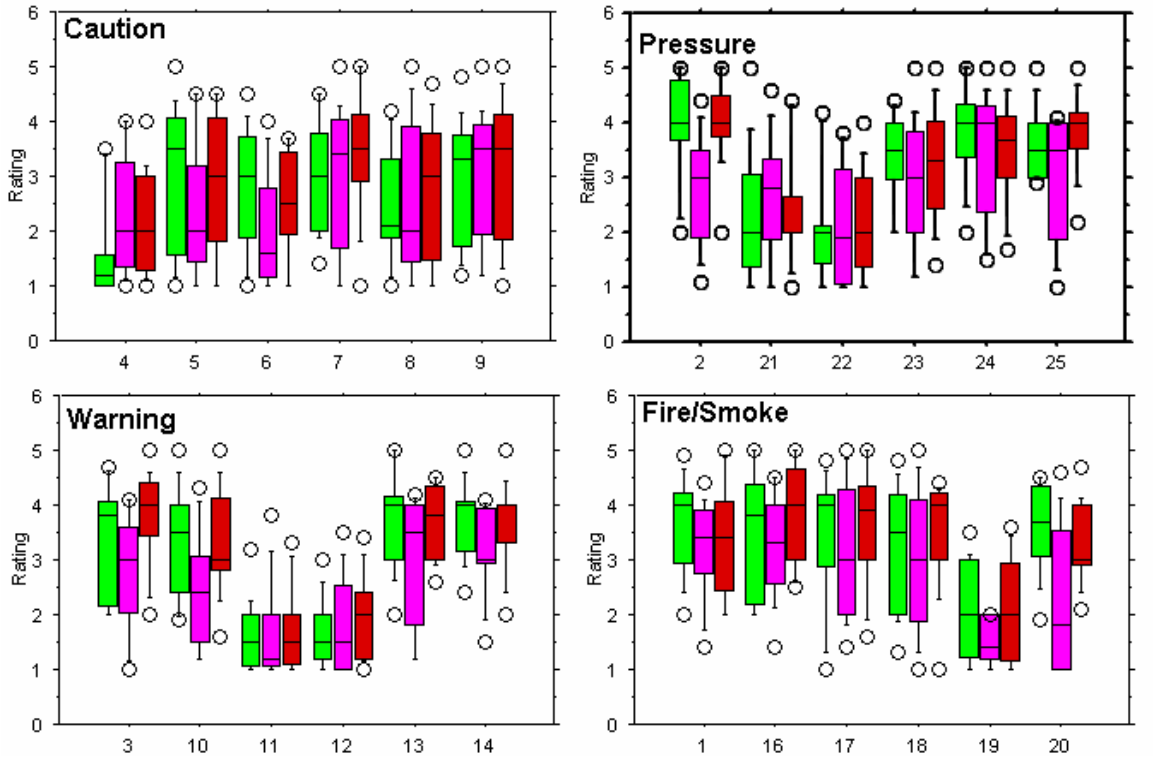


**Figure 5A: Effect of the condition as a function of the type of stimulus and the different sounds within a type**