DOI: 10.61782/fa.2025.0205

FORUM ACUSTICUM EURONOISE 2025

# CLUSTERING INDIVIDUALS BASED ON PERFORMANCE IN AUDIO-VISUAL TESTING

**Timothy Van Renterghem**
Ghent University, Department of Information Technology, WAVES research group, Technologiepark 126, Gent-Zwijnaarde B 9052, Belgium

## ABSTRACT

Audio-visual interactions play a crucial role in shaping individuals' perception of their surroundings and have been proposed as an effective strategy for improving environmental noise perception. However, the relative influence of visual and auditory inputs, as well as the degree of audio-visual integration, will vary across individuals. To investigate these influences, a dataset of 196 valid participations in an audio-visual performance test was collected under controlled laboratory conditions. Using Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and k-means clustering, three distinct participant groups were identified based on six relative performance indicators. These clusters were categorized as *visual-first responders* (cluster A), *balanced integrators* (cluster B) and *visually dominated* individuals (cluster C). Visual-first responders and the visually dominated group show hardly any audio-visual integration in the audio-visual acuity test, while the balanced integrators clearly do. Notably, clusters A and C, accounting for 68% of the test population, are expected to benefit most from noise annoyance mitigation strategies that incorporate green window views.

**Keywords:** *environmental noise perception, psycho-acoustics, audio-visual integration.*

## 1. INTRODUCTION

The concept of enhancing environmental noise perception through audio-visual interactions is gaining increasing attention. A key example is the ability of visual vegetation in a window view to mitigate noise annoyance, a phenomenon convincingly demonstrated by multiple studies [1-5]. In these, contrasting vegetation views are typically compared at fixed sound exposure levels to ensure fair comparisons. While much of the focus has been on the quantity of greenery, recent controlled virtual reality research [5] suggests that the aesthetic quality of urban vegetation is a key factor in noise annoyance mitigation. Specifically, visually appealing greenery may encourage prolonged viewing, which in turn enhances cognitive restoration and stress reduction—two critical mechanisms counteracting the negative effects of environmental noise exposure.

Beyond exposure levels and green-related (or potentially non-green) contextual factors, individual characteristics also play a role. People process auditory and visual information differently, and the extent to which these inputs are integrated can vary widely. Research indicates that differences in audio-visual processing are shaped by cognitive abilities, personality traits, attentional capacities, and neural processing styles [6]. While green window views have shown a broad, positive effect across large-scale studies, it is crucial to investigate for whom these effects are most pronounced and whether certain subpopulations benefit less. This study takes an initial step in that direction by analyzing data from an audio-visual acuity test.

## 2. AUDIO-VISUAL PERFORMANCE TEST

The audio-visual dominance/acuity test used in this work is based on an object recognition task proposed in Ref. [7] and implemented in Ref. [8]. In front of a computer screen with headphones, participants were randomly presented with two

***Corresponding author***: timothy.vanrenterghem@ugent.be

objects, indicated as A and B (see Fig. 1), and were asked to correctly classify these objects as fast as possible by pressing the left or down arrow key, corresponding to object A and B, respectively.
Objects were defined by visual features alone, auditory features alone or in combination. The visual part of the object consisted of a circle deforming into an ellipse, either horizontally (object A) or vertically (object B). The auditory part was of a pure tone of 540Hz (object A) or 560Hz (object B).

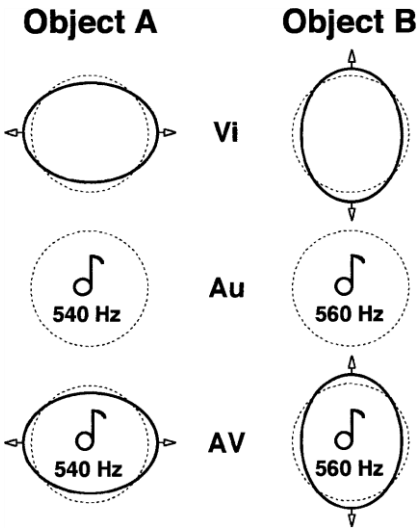


**Figure 1**. Illustration of the audio-visual objects used in the acuity test.

After each trial, the accuracy of the response was recorded. If correct, the reaction time was also saved. After initial practice trials, each participant completed 72 trials, consisting of 6 repetitions of audio-only (Au), visual only (Vi), and the audio-visual (AV) representations for both objects A and B. This enabled calculating the average accuracy and reaction times for audio-only, video-only, and audio-visual cues, finally leading to 6 parameters describing the audio-visual acuity of the test persons.

## 3. RESULTS

### 3.1 Data set

The performance test was held as a part of other audio-visual experiments [5][9], operated in 3 batches. In total, 218 valid audio-visual acuity tests were taken. As an additional criterion, participants with accuracies of less than 33% on both the Au, Vi and AV challenges were removed as these participants might not have understood the task well or might have pressed the keys randomly. In total, 196 datapoints were finally retained for further analysis.
The age of the participants was inclined towards younger people and students, but not restricted to this group. The average ages were 32.9 years (SD=standard deviation=13.9), 27.6 years (SD=12.9) and 29.6 years (SD=11.9) over the 3 batches. Overall, there were slightly more female than male participants (56%, 61% and 45% in each of the batches, respectively). In Table 1, the overall performance on the acuity test is summarized.

**Table 1**. Overall performance of the test panel with relation to the audio-visual acuity test.

| | Mean | SD |
|---|---|---|
| N | 196 | |
| Acuity test : Correctness Audio only (%) | 74% | 24% |
| Acuity test : Correctness Audio-Visual (%) | 87% | 18% |
| Acuity test : Correctness Visual only (%) | 86% | 17% |
| Acuity test : Reaction time Audio only (ms) | 810 | 150 |
| Acuity test : Reaction time Audio-Visual (ms) | 680 | 140 |
| Acuity test : Reaction time Visual only (ms) | 710 | 130 |

### 3.2 Cluster analysis

#### *3.2.1 Relative acuity indicators*

In order to identify distinct groups based on the 6-dimensional parameter space, clustering analysis was performed. To rule out the fact that some persons will generally perform better in the tests, regardless of the presentation modality, both the correctness scores ("corr") and reaction times ("rt") are considered in a relative way per test person. Following parameters were therefore defined for subsequent analysis: "corr_AV-A", "corr_AV-V", "rt_A-AV", "rt_V-AV", "corr_V-A", and "rt_A-V". Based on the average responses, positive values are expected for these relative parameters.

#### *3.2.2 Optimized clustering*

Uniform Manifold Approximation and Projection (UMAP) [10] was applied for dimensionality reduction, preserving both local and global structure, followed by k-means clustering to identify patterns in the data. Key UMAP parameters, such as the number of neighbors and minimal distance, were tuned for optimal embedding.
The 2D representation obtained via UMAP (see Fig. 2) was clustered using k-means, with the optimal number of clusters determined iteratively through the elbow method and silhouette analysis. The best clustering performance was achieved with 3 clusters, yielding the highest silhouette score (0.68) and the lowest Davies-Bouldin Index (0.69). In

theory, silhouette scores range from -1 (poor clustering) to 1 (perfect clustering), while a Davies-Bouldin Index of 0 indicates ideal, well-separated clusters, with values above 1 suggesting poor clustering. Visual inspection of Fig. 2 already suggested the presence of 3 clusters. Without using UMAP, the silhouette score was 0.42, while the Davies-Bouldin Index increased to 1.20, highlighting UMAP's effectiveness in enhancing pattern discovery while reducing noise and redundancy, finally leading to a better clustering performance.

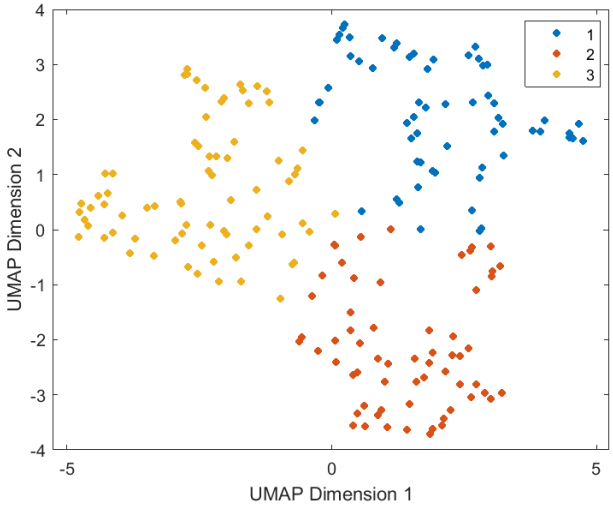


**Figure 2**. Clustering results in a 2-dimensional UMAP space, the latter being a reduction from a 6-dimensional parameter space.

*3.2.3 Cluster centroids*

Table 2 shows the cluster centroids after transforming back to the original 6-parameter space. This table should be jointly analyzed with Table 3, indicating which cluster centroids give rise to statistically significantly different means of the relative acuity parameters.

The clustering resulted in groups of roughly equal size, reinforcing the robustness of the classification. Combined with strong clustering performance metrics (see Section 3.2.2), statistically significant differences were observed across the six indicators, further supporting the existence of three distinct groups in how people process audio-visual information.

**Table 2**. Average values for the relative parameters in the identified clusters.

| | cluster A | cluster B | cluster C |
|---|---|---|---|
| n | 61 | 62 | 73 |
| corr_AV-A (%) | 4.1% | 3.3% | 28.5% |
| corr_AV-V (%) | 3.0% | 0.7% | -0.1% |
| rt_A-AV (ms) | 173 | 63 | 174 |
| rt_V-AV (ms) | 9 | 97 | 15 |
| corr_V-A (%) | 1.1% | 2.6% | 28.6% |
| rt_A-V (ms) | 164 | -34 | 159 |

**Table 3**. p-values of two-sample t-tests analyzing the means between each cluster pair for the relative performance parameters. "*" means statistically significant at the 5% level, "**" at the 1% level, and "***" at the 0.1% level.

| | cluster A-B | cluster A-C | cluster B-C |
|---|---|---|---|
| corr_AV-A (%) | 0.483 | <1E-10 *** | <1E-10 *** |
| corr_AV-V (%) | 0.030 * | 0.017 * | 0.522 |
| rt_A-AV (ms) | <1E-10 *** | 0.965 | <1E-10 *** |
| rt_V-AV (ms) | <1E-10 *** | 0.487 | <1E-10 *** |
| corr_V-A (%) | 0.302 | <1E-10 *** | <1E-10 *** |
| rt_A-V (ms) | <1E-10 *** | 0.724 | <1E-10 *** |

## 4. DISCUSSION

### 4.1 Single modality analysis

Clusters A and C process visual information much faster than cluster B. In Cluster C, in contrast to cluster A, this also leads to a strong accuracy gain. Notably, Cluster B even leads to faster reaction times for audio only relative to visual only, but also here, similar to cluster A, accuracy gains are very minor.

Persons belonging to cluster C can be called *visually dominated*, as the correctness is much higher (29%) and responses are much faster (159 ms) for visual only compared to audio only inputs. Cluster B could be called the *balanced cluster*. Unimodal audio-input even leads to slightly faster reactions (-34 ms) to the challenges in the performance test but with nearly no accuracy gains (3%). So both modalities have a similar reliability and reaction time. When comparing to the other clusters, this group seems better in processing audio information and its description might tend towards audio dominance. However, audio dominance is clearly not reached, as the speeding up in case of audio only inputs is very limited, while visual

input is still slightly more accurate. So this means that he visual dimension is still important.
Cluster A positions somewhere in between clusters B and C. Similar to cluster C, reactions to audio only are much faster (164 ms), and similar to cluster A, unimodal audio and visual inputs lead to more or less the same correctness (1%). This group is therefore called *visual-first responders*.

### 4.2 Audio-visual integration analysis

In cluster A and C, audio-visual integration is limited. Audio-visual inputs, relative to visual only, hardly lead to a decrease in reaction time (only 9 ms and 15 ms difference on average), while correctness gains are very limited or non-existent (3% and -0.1%, respectively). Although the correctness gain is limited, in cluster A it is nevertheless statistically significantly different from the other clusters. These characteristics align with the concept of modality appropriateness [11], which posits that the dominant modality (in this case, visual information) becomes the primary source of information for decision making.
Note that for cluster A, visual information only increases processing speed but not accuracy relative to audio only. In cluster C, both speeding up and a strong increase in accuracy is obtained when comparing unimodal cues, leading to true visual dominance. Regarding audio-visual integration, both seem to follow the race model [12] but with a very weak redundant multi-sensory gain.
In cluster B there is a significant reaction time decrease for combined audio-visual cues relative to visual only information, but without accuracy gains. Audio-visual exposure significantly enhances visual processing. This can be explained since both audio and visual information are reliable inputs, with audio slightly faster and visually slightly more accurate. But also relative to audio-only, reaction to the audio-visual stimulus is significantly faster. Strong audio-visual integration is thus observed in this cluster, suggesting co-activation [13].

### 4.3 Relation to environmental noise perception

The cluster descriptions allow hypothesizing how audio-visual interactions influence environmental noise perception, particularly in relation to the green window view concept discussed in the Introduction.
Since vision dominates environmental processing in Clusters A and C, individuals in these groups are likely to benefit most from an attractive visual element, such as vegetation, which shifts auditory cues—and their associated disturbance—into the background. Auditory information integration remains minimal when both auditory and visual inputs are present. In Cluster C, the visually dominated group, the strongest effects could be expected, as visual input not only enhances the speed of scene evaluation but also contributes to a more accurate mental representation of the surroundings.
Together, Clusters A and C account for 68% of the test population in this work, aligning with the findings in Refs. [1-5], which demonstrated that the positive effects of green window view holds overall for the populations surveyed.
In contrast, the balanced audio-visual integration group (Cluster B) is expected to benefit less from an attractive window view. Here, environmental sounds are likely to carry equal weight to the visual scene and may even elicit faster responses. As a result, auditory information is less likely to be suppressed in favor of appealing visuals, leading to a weaker mitigation effect on noise annoyance.

## 5. CONCLUSIONS

Performance indicators from an audio-visual acuity test suggest that individual responses to green window views as a noise annoyance mitigation strategy are likely to be influenced by personal characteristics, forming three distinct clusters. It is hypothesized that the *visually dominated* group (Cluster C) and the *visual-first* responders (Cluster A) - together comprising 68% of the test population - stand to gain the most from this strategy. In contrast, Cluster B, consisting of more proficient listeners who integrate both auditory and visual information to a similar extent (the *balanced* group), may experience less benefits, if any.

## 6. ACKNOWLEDGMENTS

The author is grateful to master thesis students Elin Vermandere, Maarten Lauwereys and Amber Lippens for guiding the test panels through the audio-visual acuity tests that complemented other audio-visual perception experiments.

## 7. REFERENCES

[1] H. Li, C. Chau, and S. Tang: “Can surrounding greenery reduce noise annoyance at home?” *Science of the Total Environment*, vol. 408, pp. 4376–4384, 2010.

[2] T. Van Renterghem and D. Botteldooren: “View on outdoor vegetation reduces noise annoyance for dwellers near busy roads,” *Landscape and Urban Planning*, vol. 148, pp. 203–215, 2016.

[3] T. Leung, J. Xu, C. Chau, S. Tang, and P.-C. L. Pun-Cheng: "The effects of neighborhood views containing multiple environmental features on road traffic noise perception at dwellings," *Journal of the Acoustical Society of America*, vol. 141, pp. 2399–2407, 2017.

[4] B. Schäffer, M. Brink, F. Schlatter, D. Vienneau, and J.-M. Wunderli: "Residential green is associated with reduced annoyance to road traffic and railway noise but increased annoyance to aircraft noise exposure," *Environment International*, vol. 143, 105885, 2020.

[5] T. Van Renterghem, E. Vermandere, and M. Lauwereys: "Road traffic noise annoyance mitigation by green window view: optimizing green quantity and quality," *Urban Forestry and Urban Greening*, vol. 88, 128072, 2023.

[6] R. A. Stevenson and M. T. Wallace: "Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, pp. 884–898, 2013.

[7] M. Giard and F. Peronnet: "Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study," *Journal of Cognitive Neuroscience*, vol. 11, pp. 473–490, 1999.

[8] J. De Winne, P. Devos, M. Leman, and D. Botteldooren: "With no attention specifically directed to it, rhythmic sound does not automatically facilitate visual task performance," *Frontiers in Psychology*, vol. 13, 894366, 2022.

[9] T. Van Renterghem and A. Lippens: "The audio-visual incongruency asymmetry. Natural sounds in an urban visual setting are more relaxing than urban sounds in visual nature," *Urban Forestry and Urban Greening*, vol. 101, 128514, 2024.

[10] L. McInnes, J. Healy, N. Saul, and L. Großberger: "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, vol. 3, 861, 2018.

[11] R. B. Welch and D. H. Warren: "Immediate perceptual response to intersensory discrepancy," *Psychological Bulletin*, vol. 88, pp. 638–667, 1980.

[12] D. H. Raab: "Statistical facilitation of simple reaction times," *Transactions of the New York Academy of Sciences*, vol. 24, pp. 574–590, 1962.

[13] J. Miller: "Divided attention: Evidence for coactivation with redundant signals," *Cognitive Psychology*, vol. 14, pp. 247–279, 1982.