UNICO I+D Project
6G-SORUS-RIS

SORUS-RIS-A2.2-E2

Clasificación final de UEs según el modelo de comportamiento

Abstract

This document presents the research activities carried out in the second period of SORUS-RIS-A2.2. These activities include the mapping of users' cognitive and emotional variables to multimodal psychophysiological signals, the analysis of the reliability and robustness of these signals, and, finally, the definition of the final version of the user clustering algorithm based on their response to latency, in the context of using existing services in the real world. The results demonstrate the usefulness of this approach and provide user profiling that can be considered in future applications to provide users with more personalized telecommunication services.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia

Document properties

Document number	SORUS-RIS-A2.2-E2
Document title	SORUS-RIS-A2.2-E2. Clasificación final de UEs según el modelo de comportamiento
Document responsible	Ioannis Arapakis, Mireia Masias
Document editor	Ioannis Arapakis, Mireia Masias
Editorial team	Ioannis Arapakis
Target dissemination level	Public
Status of the document	Final
Version	1.0
Delivery date	15-04-2024
Actual delivery date	15-04-2024

Production properties

Reviewers	Ioannis Arapakis

Disclaimer

This document has been produced in the context of the SORUS-RIS Project. The research leading to these results has received funding from the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D programme.

Content

List of Figures 4
List of Tables 5
List of Acronyms 6
Resumen Ejecutivo 7
Executive Summary 9
1. Introduction 11
2. Mapping of user behaviours into multimodal data channels 12
2.2 Video streaming dataset 13
2.1.1 Dataset 14
2.1.2 Analyses and results 14
2.2 Language Model Responses dataset 17
2.2.1 Dataset 17
2.2.2 Analyses and results 17
2.3 Online search dataset 19
2.3.1 Dataset 19
2.3.2 Analyses and results 19
3. Reliability and robustness of the multimodal signals 21
3.1 EEG signals 22
3.1.1 Video streaming dataset 22
3.1.2 Large Language Model Responses dataset 23
3.2 Peripheral physiology signals 24
3.2.1 EDA 24
3.2.1 EMG 25
3.2.2 HR 25
4. User profiles based on psychophysiological signals 27
4.1 User profiling with the video streaming dataset and the LLM interaction dataset 28
4.2 User profiling with the online search dataset 32
5. Conclusions 35
References 38

List of Figures

Figure 1. Correlation matrix for the multimodal signals in the video streaming dataset.....11

Figure 2. Correlation matrix for the multimodal signals in the LLM responses dataset.....14

Figure 3. Correlation matrix for the multimodal signals in the online search dataset.....17

Figure 4. Sample EDA signal, showing the original EDA signal (blue), the tonic component (green) and the phasic component (red).....21

Figure 5. Average phasic component of the EDA signal in the online search dataset, for seven seconds after the search.....22

Figure 6. Sample of EMG signal (red) showing large peaks due to noise.....22

Figure 7. HR signals from two participants, showing a good quality signal (up), and a very noisy signal (down).....23

Figure 8. Graphical representation of the internal validation methods for different numbers of clusters in the video streaming and LLM interaction dataset.....26

Figure 9. Graphical representation of the clustering solution: Dendogram (left) and visual representation in two dimensions (video streaming and LLM interaction dataset).....27

Figure 10. Correlations between clusters' components (video streaming and LLM interaction dataset).....27

Figure 11. Graphical representation of the internal validation methods for different numbers of clusters in the online search dataset.....29

Figure 12. Graphical representation of the clustering solution: Dendogram (left) and visual representation in two dimensions (online search dataset).....30

Figure 13. Correlations between clusters' components (online search dataset) 30

List of Tables

Table 1. Summary of the models for subjective quality.....12

Table 2. Summary of the models for attentional focus 13

Table 3. Summary of the models for enjoyment.....13

Table 4. Summary of the models for response quality.....15

Table 5. Summary of the models for response ranking.....15

Table 6. Summary of the models for the multimodal signals in the online search task 17

Table 8. Summary of the automatic detection of bad channels and epochs in the LLM responses dataset.....19

Table 9. Results of the cluster validation analysis (video streaming and LLM interaction datasets.....25

List of Acronyms

ECG - Electrocardiography

EDA – Electrodermal Activity

EEG – Electroencephalogram

EMG – Electromyography

FAA – Frontal Alpha Asymmetry

HR – Heart rate

LLM – Large Language Model

LPA – Latent Profile Analysis

PPG - Photoplethysmography

PSD - Power spectral density

QoE – Quality of Experience

SCL – Skin Conductance Level

SCR – Skin Conductance Response

Resumen Ejecutivo

Este documento presenta las actividades de investigación realizadas durante la segunda fase de SORUS-RIS-A2.2.

Durante la primera fase, se llevaron a cabo las siguientes acciones: identificar los conceptos psicológicos relevantes relacionados con la percepción de latencia y su impacto en los usuarios; revisar las metodologías disponibles para su medición; y mapear la sensibilidad subjetiva de los usuarios a la latencia utilizando conjuntos de datos abiertos, junto con una exploración preliminar de la asociación entre las señales psicofisiológicas y la latencia a través de un estudio piloto. Los resultados de estas actividades sugirieron que un enfoque multimodal, que combine métricas autorreportadas con métricas psicofisiológicas, es esencial para lograr un perfilado de usuarios más preciso.

Sobre esta base, la segunda fase de SORUS-RIS-A2.2 se centró en la construcción y el análisis sistemático de conjuntos de datos psicofisiológicos, que fueron fundamentales para completar la tarea 3 presentada en este entregable. Estos conjuntos de datos capturaron las respuestas de los usuarios en tres escenarios distintos que reflejan interacciones del mundo real: ver videos en YouTube, realizar búsquedas en línea utilizando un motor de búsqueda e interactuar con un Modelo de Lenguaje Grande (LLM). Al recolectar y analizar estos conjuntos de datos, se mapearon las variables cognitivas, emocionales y conductuales de los usuarios a sus señales psicofisiológicas multimodales. Los resultados sugieren que, en contextos de uso más pasivos, como ver videos en streaming o leer respuestas generadas por modelos de lenguaje, el impacto de la latencia se refleja principalmente en variables relacionadas con la motivación del usuario (señal de asimetría alfa frontal en EEG), la atención visual (alfa occipital), la emoción negativa (beta parietal) y la excitación emocional (EDA fásica). Sin embargo, cuando las variaciones de latencia son mínimas, las relaciones entre los indicadores multimodales y las respuestas autorreportadas de los usuarios no son fuertes. En contraste, en contextos más interactivos, como las búsquedas en línea, se observa una correspondencia más clara entre los niveles de latencia y las señales psicofisiológicas, particularmente EDA, EMG y HR.

Para asegurar la robustez de los datos recolectados, se evaluó la fiabilidad de las señales psicofisiológicas multimodales mediante una combinación de métodos automáticos y análisis experto. Los resultados indican que las señales de EEG y EDA son suficientemente robustas, mientras que las señales de EMG a menudo mostraron segmentos ruidosos. En el caso de las señales de HR, se encontró que el sistema basado en ECG es más fiable que el basado en PPG.

Finalmente, se definieron los perfiles de usuario determinando primero la sensibilidad individual a distintos niveles de latencia en los casos de uso analizados, y luego agrupando a los usuarios según sus respuestas. Los resultados indican dos grupos de usuarios distintos: aquellos que muestran una clara sensibilidad a la latencia y aquellos que no expresan respuestas cognitivas o emocionales evidentes. Esta sensibilidad se manifiesta de manera diferente dependiendo del

contexto: como respuestas fásicas de EDA (excitación emocional) en casos de uso más pasivos y como respuestas de EDA, EMG y HR en casos de uso más interactivos.

En conclusión, a través del entregable SORUS-RIS-A2.2-E2, se ha demostrado la utilidad de utilizar señales multimodales para evaluar las respuestas individuales a la latencia. Un aspecto central de este esfuerzo fue la construcción de los conjuntos de datos que permitieron un análisis detallado de las respuestas de los usuarios en diversos contextos. Estos resultados han apoyado el desarrollo de perfiles de usuario altamente personalizados para servicios operativos del mundo real, contribuyendo a los avances futuros en servicios de redes de telecomunicaciones personalizados.

Executive Summary

This document presents the research activities carried out during the second phase of SORUS-RIS-A2.2.

During the first phase, the following actions were conducted: identifying the relevant psychological concepts related to latency perception and its impact on users; reviewing the methodologies available for its measurement; and mapping users' subjective sensitivity to latency using open datasets, along with a preliminary exploration of the association between psychophysiological signals and latency through a pilot study. The results of these activities suggested that a multimodal approach, combining self-reported metrics with psychophysiological metrics, is essential for more accurate user profiling

Building on this foundation, the second phase of SORUS-RIS-A2.2 focused on the systematic construction and analysis of psychophysiological datasets, which were central to complete task 3 addressed in this deliverable. These datasets captured user responses across three distinct scenarios reflecting real-world interactions: watching videos on YouTube, conducting online searches using a search engine, and interacting with a Large Language Model (LLM). By collecting and analysing these datasets, the cognitive, emotional, and behavioural variables of users were mapped to their multimodal psychophysiological signals. The results suggest that in more passive usage contexts, such as watching streaming videos or reading responses generated by language models, the impact of latency is primarily reflected in variables related to user motivation (EEG frontal alpha asymmetry signal), visual attention (occipital alpha), negative emotion (parietal beta), and emotional arousal (phasic EDA). However, when latency variations are minimal, the relationships between multimodal indicators and self-reported user responses are not strong. In contrast, interactive contexts, such as online searches, show a clearer mapping between latency levels and psychophysiological signals, particularly EDA, EMG, and HR.

To ensure the robustness of the collected data, the reliability of the multimodal psychophysiological signals was assessed through a combination of automated methods and expert analysis. The results indicate that EEG and EDA signals are sufficiently robust, while EMG signals often displayed noisy segments. For HR signals, the system based on ECG was found to be more reliable than the one based on PPG.

Finally, user profiles were defined by first determining individual sensitivity to distinct levels of latency in the analysed use cases, followed by clustering users based on their responses. The results indicate two distinct user groups: those who exhibit clear sensitivity to latency and those who do not express evident cognitive or emotional responses. This sensitivity manifests differently depending on the context: as phasic EDA responses (emotional arousal) in more passive use cases and as EDA, EMG, and HR responses in more interactive use cases.

In conclusion, through deliverable SORUS-RIS-A2.2-E2, the utility of using multimodal signals to evaluate individual responses to latency has been demonstrated. Central to this effort was the construction of datasets that enabled a detailed analysis of user responses across diverse contexts. These results have supported the development of highly personalized user profiles for real-world operational services, contributing to future advancements in personalized telecommunications network services.

1. Introduction

The central aim of the SORUS-RIS-A2.2 task "Modelo de comportamiento del UE" is the classification of user profiles according to their psychological and behavioural response to technical aspects of the telecommunications network, such as latency. The rationale of the activity is that identifying and modelling such response can be useful in future applications to propose customized solutions to different users, depending on their sensitivity to the technical performance of the system. In this way, for example, energy saving plans could be created that take advantage of the different capabilities and sensitivities of different users (e.g., prioritizing lower latency to those users more sensitive to latency, or offering cheaper plans, with higher latency, to those with less latency-sensitive users).

The previous deliverable (SORUS-RIS-A2.2-E1) described the first steps to achieve this objective, which included the definition and summarization of relevant user behaviours and responses, and the sensors available to monitor them. It also offered a first approximation to the modelling of user profiles based on these behaviours and psychophysiological signals. The current deliverable (SORUS-RIS-A2.2-E2) offers the completion of this work, addressing the content of Task 3 as described in activity A2.

Task 3 ("Mapping psychological constructs into low- and high-level indicators and generation of user profiles") has three main objectives. First, it aims to determine the optimal mapping of the behaviours identified in the previous tasks to the multimodal signals that can be obtained through the data channels considered in the project. Secondly, it seeks to analyze the reliability and robustness of each considered signal. Finally, based on the extracted signals and behaviours, it aims to identify user profiles according to their response to the technical performance of the system, particularly, to the system's latency.

According to this structure, the content of this Deliverable is organized as follows. Section 2 contains the research activities conducted to address the mapping of user' behaviours into multimodal data channels. Section 3 includes the assessment of the reliability and robustness of the psychophysiological multimodal signals employed in the project. Section 4 describes the proposed method and results obtained regarding user profiling based on psychophysiological signals clustering. Finally, section 5 summarizes the work conducted and the main extracted conclusions.

2. Mapping of user behaviours into multimodal data channels

Mapping user behaviors to the psychophysiological signals collected in this study allows for a range of analytical approaches, depending on the specific objectives and characteristics of the use case (e.g., Ajenaghughrure et al., 2020; Bach et al., 2018). Two dominant approaches in literature include classical statistical inference and machine learning. Classical statistical methods, prevalent in psychology, focus on explaining relationships between variables and quantifying these associations, often guided by theoretical frameworks. Conversely, machine learning methods, widely applied in engineering and affective computing, prioritize advanced modeling and predicting user states (cf. Yarkoni & Westfall, 2017). This deliverable integrates both approaches to provide a comprehensive understanding of the relationships between psychophysiological signals and relevant behavioral variables. Correlation analysis was used to assess redundancy among psychophysiological variables, multilevel mixed models provided insights into how these modalities explain subjective indicators of quality of experience and latency perception, while random forests evaluated the predictive power of the signals for behavioral outcomes.

Given the strong contextual influence on these relationships, this deliverable required the construction and analysis of datasets across three distinct scenarios to ensure robustness and relevance: (i) audiovisual content streaming, (ii) online information searching, and (iii) interaction with an Artificial Intelligence (AI) system, specifically a Large Language Model (LLM). While scenario (iii) represents a novel and less widespread form of interaction, its inclusion anticipates its potential future adoption. Developing these datasets was a central effort, forming the foundation for examining user responses in diverse contexts. Scenario (i), initially explored as a pilot study in deliverable SORUS-RIS-A2.2-E1, provided key insights that shaped the methodology. The addition of scenarios (ii) and (iii) introduced further complexity, enhancing our ability to understand latency effects across both active and passive tasks.

Subsequently, a dataset for each of the scenarios was built, combining psychophysiological signals collected in a controlled environment with corresponding behavioral variables of interest. Specifically, and building on the conclusions drawn from Tasks 1 and 2 from the SORUS-RIS-A2.2-E1 deliverable, EEG and EDA were collected using BitBrain-E32.A1 and BitBrain-BIO.A1 equipment, respectively. Particularly, an EEG cap with 32 electrodes was used, and electrodes placed on the participant's forehead. For the EDA signal recording, the wearable, small-sized device was attached to two fingers of the hand's participants, allowing mobility. The sampling rate used was 256 Hz. Parallelly, two different systems were used for HR and HRV recordings. For 2/3 of the scenarios (i.e., video streaming and LLM interaction), a photoplethysmography sensor was employed during the different tasks. For the remaining scenario (i.e., internet browsing), an electrocardiogram (ECG) system was employed, placing electrodes to the participant's wrists and forearms.

The presentation of the tasks, questionnaires when required, and the sending of marks for the synchronization of the signals was performed using Psychopy2. The recording and synchronization of the psychophysiological signals was performed using OpenVibe3. The contents were displayed using a ViewSonic VX3276-2K- MHD-2 (32-inch) screen.

The EEG signal preprocessing was performed using the MNE package in Python 4. The steps followed included: (1) a band-pass filter (0.3 - 40 Hz) applied to eliminate slow drifts; (2) the Fpz channel was taken as EOG, necessary for the use of the algorithm below; (3) each signal was divided into 2 s epochs with 1 s of overlap; (4) the FASTER method (Nolan et al., 2010), based on ICA was used to automatically detect, reject and interpolate those epochs with excessively noisy data; and (5) finally, different EEG-based metrics were derived. These included Frontal Alpha Asymmetry (FAA) as an indicator of motivational arousal; Occipital Alpha (alpha_occ) to measure visual attention; Parietal Beta (beta_par) as a marker of negative emotional response; and Engagement Index for both Frontal (eng_index_front) and Parietal (eng_index_par) regions, as indicators of engagement. The EDA signal was processed using the cvxEDA algorithm (Greco et al., 2015). The algorithm enabled decomposing the signal into its tonic (EDA_tonic) and phasic (EDA_phasic) components. Finally, for the processing of the HR and the EMG signals, the Neurokit25 package in Python 4 was employed. It is worth noting that signals were divided into epochs with a duration of 1s, obtaining several measures per participant and task. Finally, the HR and HRV values were derived using AcqKnowlege software (Biopac).

To complete each of the datasets, further behavioral subjective variables were asked to participants, including metrics for perception quality, attention, and enjoyment. All the collected data was fully anonymized to protect participants' privacy, ensuring that no personally identifiable information was stored or could be traced back to individual participants. This process adhered to relevant data protection regulations and ethical guidelines.

In the following subsections, specificities on the datasets and analyses conducted, and their results are presented for each of the datasets.

2.2 Video streaming dataset

First, as a representative dataset of the context of use of streaming audiovisual content consumption, the dataset collected in the pilot test reported in the deliverable SORUS-RIS-A2.2-E1 "4.2.1 Definition of use cases and experimental paradigms (Video streaming)" has been used for the analysis.

2.1.1 Dataset

This dataset includes data from 5 participants who were exposed to 5 videos of 4 minutes duration each, covering several themes, including sports, animation, musicas, and nature. The videos were presented with different latencies, synthetically introduced by means of a wrapper for Browsertime (https://www.sitespeed.io), which allowed to regulate latency and bandwidth. The 5 levels of latency, counterbalanced in the different videos for the different participants, were as follows: Q1(29.64 ms); Q2 (54.54 ms); Q3 (152.31 ms); Q4 (253.95 ms); Q5 (658.34 ms), and Q6 (1840.03 ms).

In the meantime, psychophysiological signals were collected, and self-reported measures of subjective experience by the participants were collected after each video, namely: perceived quality (subjective), using a single stimulus continuous procedure (Duanmu et al., 2016); attentional focus (using three items; Busselle & Bilandzic, 2009); and enjoyment (three items, Oliver & Bartsch, 2010). Complete details on the experimental paradigm can be found in deliverable SORUS-RIS-A2.2-E1.

2.1.2 Analyses and results

The first part of the analysis involved exploring the correlations between the different multimodal signals, to establish to what extent they may provide redundant or complementary information about the user's psychological processes. However, limited reliability was observed for the HR and EMG signals, primarily due to their sensitivity to participant movement during visualization tasks (see Section 3 for further details). As a result, these signals were excluded from subsequent analyses. As illustrated in Figure 1, the correlations among the remaining signals are generally low, suggesting a minimal level of collinearity.

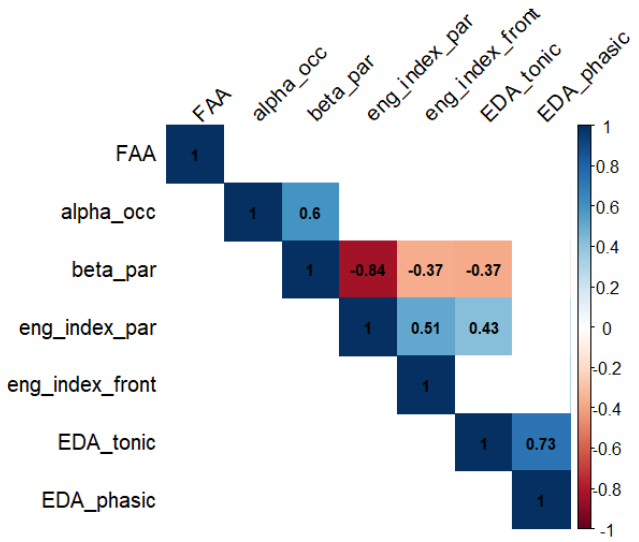

FIGURE 1. CORRELATION MATRIX FOR THE MULTIMODAL SIGNALS IN THE VIDEO STREAMING DATASET

Multilevel mixed models were then used to analyze the mapping between the set of psychophysiological signals and the subjective indicators of quality of experience in the different experimental conditions. An initial model was fitted for each subjective perception of experience variable, as outcomes of the model; based on the different psychophysiological signals as predictors. A second multilevel mixed model was then fitted in which additionally to the psychophysiological signals, the experimental latency was included as predictor. This approach allowed us to understand the extent to which the psychophysiological variables predict residual variations in user experience once the different latency levels are controlled. The results of these models are summarized in Tables 1 to 3.

Table 1. Summary of the models for subjective quality

	Model 1		**Model 2**		**Model 3**	
(Intercept)	3.342	***	3.163	***	3.849	***
FAA	0.020	***	0.020	***	0.009	***
alpha_occ	-0.010	**	-0.010	*	-0.007	***
beta_par	-0.016	***	-0.016	**	-0.004	
eng_index_par	0.052		0.057		0.024	
eng_index_front	0.000		-0.002		0.048	
EDA_tonic	-0.033		-0.052		0.100	***
EDA_phasic	0.033	*	0.030	*	-0.005	
order			0.049	*	0.025	*
latency					-0.001	***

* p <.05; ** p <.01; ***p<.001

The coefficients of the Models for perceived quality (Table 1) indicate that both FAA (related to motivational arousal) and measures of frontal alpha (visual attention) and parietal beta (negative emotionality) display a statistically significant association with perceived quality (in a context in which all other cues and variables are controlled for). However, when controlling for the effect of latency, parietal beta no longer is a significant indicator, suggesting that the variability it explains could be attributed to latency. Thus, the cues that show the strongest statistically significant relationship across the different models may be the most effective predictors of user-perceived quality. In the case of attentional focus (Table 2) the only signal showing a statistically significant relationship across the three models is the tonic EDA signal. Thus, user-reported attention may be more difficult to predict from other signals, such as EEG. However, it is also important to consider here that it is possible that users could not reliably report their level of attention. Finally, in the case of the reported enjoyment (Table 3), this is mainly related to the occipital alpha signal, as well as to the EDA signals. This is consistent with theoretical conceptualizations of enjoyment as a psychological construct, in which it is related to levels of attention (indicated by the occipital alpha signal) and emotional arousal (indicated by the EDA).

Table 2. Summary of the models for attentional focus

	Model 1		Model 2		Model 3	
(Intercept)	4.159	***	4.298	***	4.763	***
FAA	0.006		0.006		-0.001	
alpha_occ	0.000		0.001		0.003	
beta_par	-0.010		-0.009		-0.001	
eng_index_par	-0.098		-0.101		-0.124	
eng_index_front	-0.028		-0.027		0.007	
EDA_tonic	0.119	***	0.133	***	0.237	***
EDA_phasic	-0.011		-0.009		-0.033	*
order			-0.038		-0.055	*
latency					-0.001	***

* p <.05; ** p <.01; ***p<.001

Table 3. Summary of the models for enjoyment

	Model 1		Model 2		Model 3	
(Intercept)	4.182	***	4.936	***	5.287	***
FAA	0.005		0.004		-0.002	
alpha_occ	-0.013	**	-0.012	**	-0.010	**
beta_par	-0.003		-0.001		0.005	
eng_index_par	0.035		0.018		0.001	
eng_index_front	-0.043		-0.034		-0.009	
EDA_tonic	0.079	*	0.155	***	0.233	***
EDA_phasic	-0.055	**	-0.041	*	-0.059	***
order			-0.208	***	-0.220	***
latency					-0.001	***

* p <.05; ** p <.01; ***p<.001

The next step, to go beyond the statistical inference approach, was to explore the predictive power of multimodal signals to predict the subjective values of quality of experience (in terms of perceived quality, attentional focus, and reported enjoyment). For this purpose, the Python package scikit-learn was used to generate a random forest model (random forest regressor) for each of the subjective variables (perceived quality, attentional focus, and reported enjoyment), and the psychophysiological signals were taken as predictors. The dataset was divided into training dataset and testing dataset (20% of the original), and 5-fold cross-validation was used. The metric used to assess the result was R^2. The R^2 values suggest that the combination of psychophysiological signals has considerable predictive power (Quality R^2 = 0.64; Attention R^2 = 0.66; Enjoyment R^2 = 0.63), although the remaining unexplained variability suggests that there may still be considerable room for improving these predictions with the incorporation of other data sources or signals.

2.2 Language Model Responses dataset

An analytical approach similar to that employed in the streaming video viewing dataset was carried out with a dataset collected in a laboratory interaction context between users and Large Language Models (LLM).

2.2.1 Dataset

To create this dataset, eight participants were asked to read and rate responses given by an LLM to a certain prompt. One of them performed the test in two sessions. To control the quality of the responses, prompts and user-rated responses available in the Open Assistant Conversations dataset (https://huggingface.co/datasets/OpenAssistant/oasst1) were used. For each participant, between 20 and 30 prompts were presented, and for each prompt, two responses previously annotated as best and worst quality, were offered consecutively. Each participant read the responses and rated them in terms of their quality and was also asked to rank them in order of quality. The psychophyisiological signals described above were jointly collected during the study. In order to analyze the impact of latency on the participant's subjective experience (i.e. their assessment of the quality of the responses), as well as on their cognitive and emotional response (i.e., assessed through psychophysiological signals), the different levels of latency introduced and recorded by the software in the presentation of the language model responses were analyzed. These latencies had values between 1 ms (values lower than this were adjusted to 1 ms) and 103 ms, with a mean of 19 ms and standard deviation of 16 ms.

2.2.2 Analyses and results

The analytical approach employed was similar to that described for the video steaming dataset. First, correlations between psychophysiological signals were explored (Figure 2), and then multilevel models were fitted for each subjective variable (Tables 4 and 5).

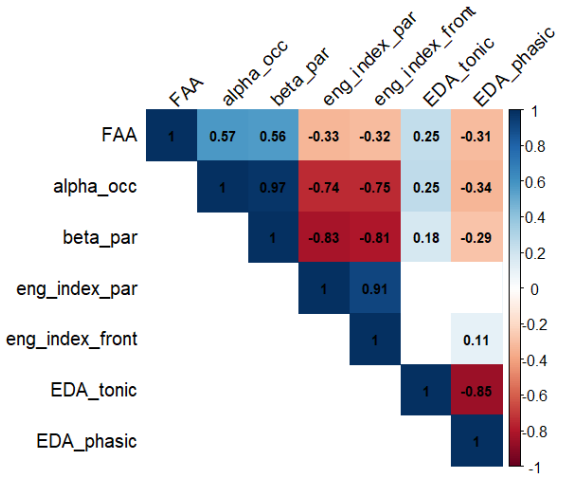

FIGURE 2. CORRELATION MATRIX FOR THE MULTIMODAL SIGNALS IN THE LLM RESPONSES DATASET

As shown in Tables 4 and 5, none of the signals appears to be statistically significantly related to either the quality of the response reported by the participant or the ranking assigned to it. This occurs in all three models for each subjective variable (even when controlling for the effect of presentation order and latency). The reason for this may be that in this case we considered the signal averaged per response, rather than a time series of the signal per second as in the case of the streaming video dataset (so that many fewer samples per stimulus are available). It is also possible that, in this case, since the latency values were much lower than some of the conditions used in the previous dataset, they had no observable impact on the cognitive and emotional aspects of the user, as measured by the psychophysiological signals.

Table 4. Summary of the models for response quality

	Model 1		Model 2		Model 3	
(Intercept)	3.863	***	4.030	***	3.899	***
FAA	-0.315		-0.314		-0.357	
alpha_occ	0.024		0.042		0.057	
beta_par	-0.023		-0.044		-0.056	
eng_index_par	-2.085		-2.881		-2.914	
eng_index_front	1.319		1.592		1.523	
EDA_tonic	-0.057		-0.068		-0.064	
EDA_phasic	-0.089		-0.101		-0.102	
order			0.007		0.010	
latency					6.886	

* p <.05; ** p <.01; ***p<.001

Table 5. Summary of the models for response ranking

	Model 1	Model 2	Model 3
(Intercept)	0.415	0.406	0.287
FAA	-0.019	-0.023	-0.036
alpha_occ	-0.006	-0.007	-0.004
beta_par	0.007	0.009	0.008
eng_index_par	0.013	0.075	0.163
eng_index_front	0.149	0.121	0.167
EDA_tonic	-0.007	-0.007	-0.005
EDA_phasic	-0.006	-0.004	-0.001
order		-0.001	0.000
latency			1.936

* p <.05; ** p <.01; ***p<.001

Finally, a random forest model was implemented for each of the subjective variables, including reported quality and ranking. Twenty percent of the data was used as test dataset, and a 5-fold cross-validation was conducted. As expected, in the view of the results of the previous multilevel models, in neither case was a relevant predictive capacity achieved (Quality R^2 = 0.10; Ranking R^2 = 0.15), which suggests that the capacity of these signals to predict these subjective variables is quite limited when simply considering the average signal per response, instead of considering the complete time series, as demonstrated in the previous case.

2.3 Online search dataset

A third context of use in which the mapping between multimodal psychophysiological signals and users' subjective perception was analyzed is online search.

2.3.1 Dataset

In this case, a total of 19 participants were asked to search for different aspects in an online search engine. A client-side script was used to artificially manipulate the latency in the presentation of search results. Such introduced latency presented four distinct levels: 0 ms, 500 ms, 750 ms and 1,000 ms. Then, each participant performed four tasks, with each of the latency levels. These tasks were presented in a random order, and varied among participants. The search tasks required performing as many queries as possible, within a list of web domains, and retrieve the URL associated with the results, within eight minutes per task. While participants performed the tasks, psychophysiological variables were recorded. In this case, we excluded EEG measures, focusing solely on peripheral nervous system signals to evaluate their standalone potential. This decision was driven by the goal of prioritizing signals that are more feasible to collect outside laboratory settings (cf. Tronstad et al., 2022) and in real-world environments. For instance, such signals can be captured using wearable devices on the wrist (e.g., the Empatica E4 wristband, https://www.empatica.com/en-eu/research/e4/), while facial activity measures can be derived from camera recordings as a proxy for EMG measures (Inzelberg et al., 2018; Perusquía-Hernández et al., 2019).

2.3.2 Analyses and results

While some basic aspects of the analysis in this case are similar to those reported with the other two previous datasets (i.e., the sequence of exploring the correlations between the signals, performing multilevel models with a statistical inference approach, and ending with a prediction task using random forest), in this case there is some notable difference. In the absence of data on the subjective perception of the participants, we have chosen to focus on the mapping between the users' latency levels and the psychophysiological responses obtained. First, as in previous analyses, the correlations observed between the different psychophysiological signals are reported (Figure 3). As it is shown in the figure, there are certain correlations between the different signals, but these are limited in scope, which suggests that they may contribute to explaining diverse sources of variability in user experience.

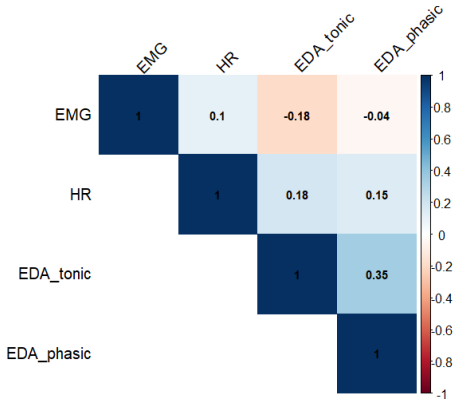

FIGURE 3. CORRELATION MATRIX FOR THE MULTIMODAL SIGNALS IN THE ONLINE SEARCH DATASET

In a succession of three multilevel models, we analyzed whether multimodal signals (tonic and phasic EDA, EMG facial, and HR) act as significant predictors of the observed latency in the search engine response. To this end, the values of these signals (sampled at 1 Hz) were analyzed within seven seconds after each of the searches. Three models were calculated, controlling successively for temporal order effects (order of presentation of the searches, and second -epoch- after performing the search). The results of the models, summarized in Table 6, show that both the tonic EDA activity, the facial EMG signal and the heart rate are significant predictors of the observed latency, indicating that they are affected by it. Moreover, with all three signals maintaining statistical significance in these models where all other signals are taken into account, this suggests that they capture different aspects of the user's experience (e.g., emotional arousal in the case of EDA, attentional responses in the case of HR, or negative valence of emotion in the case of EMG, cf. Rajendra Acharya et a., 2006; Bolls et al., 2019; Boucsein, 2012).

Table 6. Summary of the models for the multimodal signals in the online search task

	Model 1		Model 2		Model 3	
(Intercept)	0.543	***	0.546	***	0.545	***
EDA_tonic	-0.064	***	-0.063	***	-0.064	***
EDA_phasic	-0.002		-0.002		-0.002	
EMG	-0.175	***	-0.175	***	-0.175	***
HR	0.000	**	0.000	**	0.000	**
order			-0.001		-0.001	
epoch					0.000	

Finally, a random forest model, similar to those described above (random forest regressor, 20% test set, five-fold cross-validation) was used to analyse the potential of these signals for predicting latency (in this case, given the absence of records of subjective ratings, it was decided to try to directly predict the latency observed in the tests). The results (R^2 = 0.76) point to a considerable potential of these signals for predicting observed latency levels, and, consequently, to their usefulness for profiling users based on their response to latency.

3. Reliability and robustness of the multimodal signals

After the mapping of psychophysiological signals and the values of subjective perception of latency, the next task described in Task 3 of the proposal focuses on the analysis of the reliability and robustness of the psychophysiological signals. Analysing the reliability and robustness of a signal involves assessing its consistency, accuracy, and resilience to noise or interference. There are several approaches that can be considered to evaluate the reliability and robustness of multimodal, psychophysiological signals like the ones considered in our tests. First, the most obvious option is to focus on the quality of the signal itself. Through visual inspection and spectral analysis - usually employed in psychophysiological analysis - artifacts, noise or distortions affecting the quality of the measurement can be identified. Other approaches in this regard include examining test-retest reliability (i.e., to collect multiple recordings of the same signal under similar conditions and compare them to assess the level of agreement between measurements, employing metrics such as intraclass correlation coefficient). One can also focus on the internal consistency of the signal (i.e., if the signal comprises multiple components or features, assess the internal consistency of these components), either consider the sensitivity of the signal, compare it to certain gold standards, or perform replication studies.

Each of these approaches has certain advantages and disadvantages and presents certain requirements. When establishing our approach, we have considered the existing limitations in our project (e.g., the absence of gold standards in this specific field of latency response, the existence of limited studies and multiple conditions of use -which limits the performance of replication studies- as well as the high individual variability in these signals. Based on this consideration, we opted for a different approach depending on the signal analyzed to adjust to its characteristics.

In the case of EEG signals, we have chosen to use an algorithm (FASTER; Nolan et al., 2010) for the automated detection of artifacts and noisy channels and epochs in the EEG signal, capable of automatically estimating the quality of the signal, based on an Independent Component Analysis. In the case of peripheral psychophysiological signals (i.e., EDA, facial EMG, HR), expert inspection has been chosen because, although some automatic techniques exist to detect poor quality psychophysiological signal segments, there is still no satisfactory automatic solution for all signals used, in various contexts. Therefore, visual inspection by experts is still considered one of the most effective systems (e.g., Kleckner et al., 2017). In our case, this involved graphing each segment of the signal and visual exploration of these by an expert with over 10 years of experience in

psychophysiological methods in an experimental context. Details on the reliability and robustness analyses for each signal and scenario are provided in the following sections.

3.1 EEG signals

As mentioned above, the analysis of the reliability and robustness of EEG signals was performed using the FASTER algorithm (Nolan et al., 2010). The algorithm produces an estimate of the signal quality of channels and epochs (i.e., segments) of the EEG signal. Then, based on statistical thresholds, it points out which channels and epochs present poor quality (e.g., excessive noise in the signal). This algorithm was applied to our two datasets containing EEG signals (i.e., the dataset collected while participants viewed streaming content with different latencies, and the dataset collected while participants interacted with large language models).

3.1.1 Video streaming dataset

From this dataset, EEG data were analyzed for four participants, who viewed six videos each (i.e., a total of 24 viewings) with a duration of 4 minutes each (i.e., a total of 144 minutes of EEG signal). The signal was collected using a wet/gel electrodes system, including 32 electrodes in positions according to the standard system. Each of the video signals was divided into 240 epochs of two seconds duration, with a one-second overlap. These epochs were analyzed using the FASTER algorithm.

Table 7. Number of noisy channels and epochs per participant in the video streaming dataset

Channels							
	video 1	video 2	video 3	video 4	video 5	video 6	M(SD)
p1002	1	1	2	2	2	1	1.50(0.55)
p1003	3	4	4	4	3	3	3.50(0.55)
p1004	2	3	3	2	2	2	2.33(0.52)
p1005	3	2	2	2	1	2	2.00(0.63)

Epochs							
	video 1	video 2	video 3	video 4	video 5	video 6	M(SD)
p1002	7	14	9	10	9	4	8.83(3.31)
p1003	9	9	7	8	12	8	8.83(1.72)
p1004	8	10	12	9	12	11	10.33(1.63)
p1005	4	5	7	10	9	7	7.00(2.28)

As Table 7 shows, the number of channels and epochs the algorithm considers low quality or noisy is small. It ranges between 1 and 4 channels (out of 32; average 2.33, standard deviation 0.92) and between 4 and 14 epochs (out of 240; average 2.33, standard deviation 0.92). This suggests that the signal is robust enough to provide reliable results on most occasions.

3.1.2 Large Language Model Responses dataset

A similar analysis to that described above was carried out on the dataset concerning the interaction with LLM responses. In this case, the eight participants read and rated between 40 and 120 responses provided by an LLM, with varying durations, from just a few seconds to more than a minute. As can be seen in Table 8, the results are broadly similar to those obtained in the previous test. The number of noisy channels is on average low, and something similar occurs with the number of noisy epochs: only between 1 and 5% of these are identified as having low quality. Therefore, it can be concluded that the EEG signal is sufficiently robust in this domain.

Table 8. Summary of the automatic detection of bad channels and epochs in the LLM responses dataset

Channels				
Participant	**Min. N bad channels**	**Max N bad channels**	**Mean (across responses)**	**SD**
1	0	5	2.44	1.26
2	0	2	0.68	0.62
3	1	5	3.60	0.90
4	1	6	3.20	1.22
5	2	4	3.08	0.76
6	2	6	4.00	0.78
7	1	4	2.10	0.93
8	2	5	2.93	0.76
			Average across participants	
			2.69	**1.33**
Epochs				
Participant	**Total epochs**	**Bad epochs**	**Bad/Total**	
1	2283	18	0.01	
2	678	7	0.01	
3	782	9	0.01	
4	1009	51	0.05	
5	633	16	0.03	
6	803	14	0.02	
7	500	6	0.01	
8	498	17	0.03	

3.2 Peripheral physiology signals

For the analysis of the peripheral psychophysiological signals a different approach was followed. Given the absence of a robust and reliable tool for an automated analysis, this approach relied primarily on expert inspection and quality assessment. Expert opinion, while subjective, remains the most reliable tool for this aim. In this case, an expert with over 10 years of experience in psychophysiological methods in an experimental context reviewed the collected signals. This analysis was applied to each of the peripheral signals recorded in the three different scenarios.

3.2.1 EDA

The first of these peripheral measures was EDA. In the case of the video streaming dataset, the signal was processed employing the cxvEDA algorithm (Greco et al., 2015). The algorithm reduced noise present in the original signal. Subsequently, minimal evidence of noise was reported in the resulting signals by the expert. The same applies for the results on EDA signals for the Large Language Models Response dataset. See Figure 4 for further details.

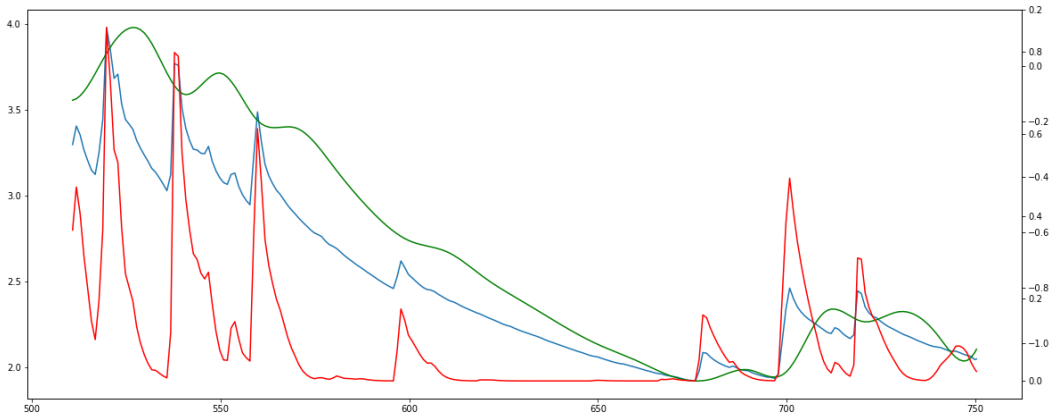

FIGURE 4. SAMPLE EDA SIGNAL, SHOWING THE ORIGINAL EDA SIGNAL (BLUE), THE TONIC COMPONENT (GREEN) AND THE PHASIC COMPONENT (RED).

In addition, the resulting EDA signal was also online search dataset was used, in which a different system (Biopac MP-150) was employed. In this dataset, the recording of the EDA signal was associated with times when users perform searches. The average of the phasic component of the EDA signal was calculated in the seconds after the searches were performed. In these seconds, it is expected to see a Skin Conductance Response, a momentary increase in EDA, which typically appears a few seconds after stimulus presentation (Dawson et al., 2007). This type of SCR response was common in the dataset, confirming that the signal responds reliably to the stimuli, as would be expected.

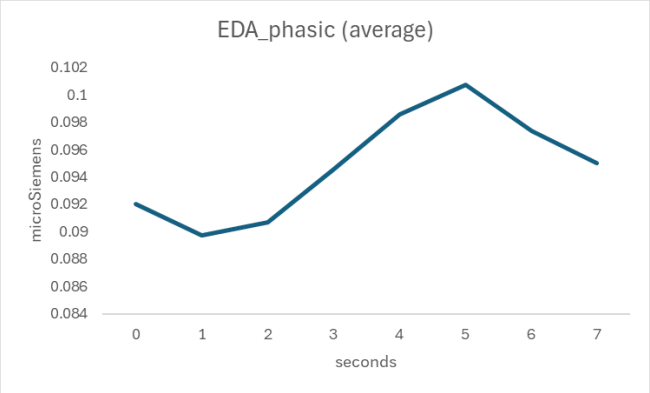

FIGURE 5. AVERAGE PHASIC COMPONENT OF THE EDA SIGNAL IN THE ONLINE SEARCH DATASET, FOR SEVEN SECONDS AFTER THE SEARCH

3.2.1 EMG

In the case of the EMG signal, our results suggest that the signals are more affected by noise than EEG in all datasets. Some of this noise is attributable to poor fixation of the electrodes on the participants' faces. See Figure 6 for further details. Filtering techniques were effective in removing some of these artifacts, although they were not successful in all instances. Therefore, the results suggest that the use of this signal requires expert supervision to ensure its quality, which could pose challenges in terms of cost and feasibility for analyzing user responses.

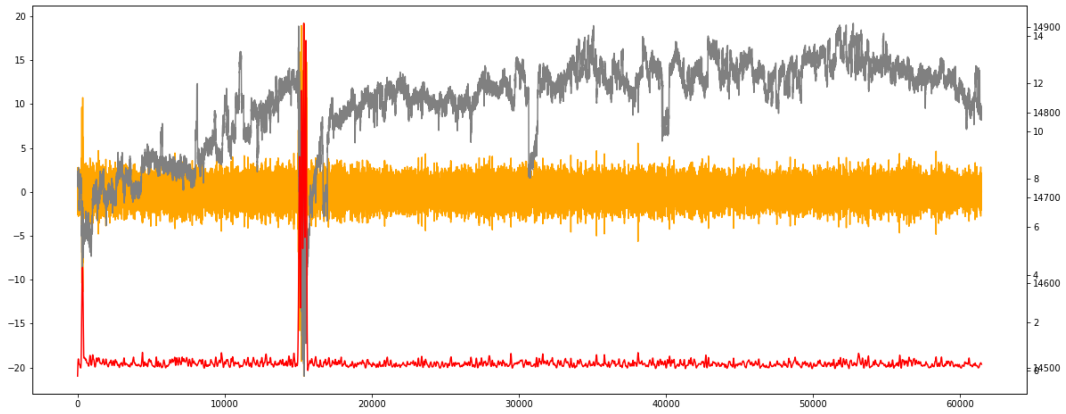

FIGURE 6. SAMPLE OF EMG SIGNAL (RED), SHOWING LARGE PEAKS DUE TO NOISE

3.2.2 HR

In the case of the **HR signal**, this was obtained from a photoplethysmography sensor. The results suggest that the sensor is overly sensitive to the noise caused by the participant's hand movements (Figure 7). Visual inspection by the expert showed that in many of the signals collected, noise and distortions were present and could not be corrected with a classic filtering approach.

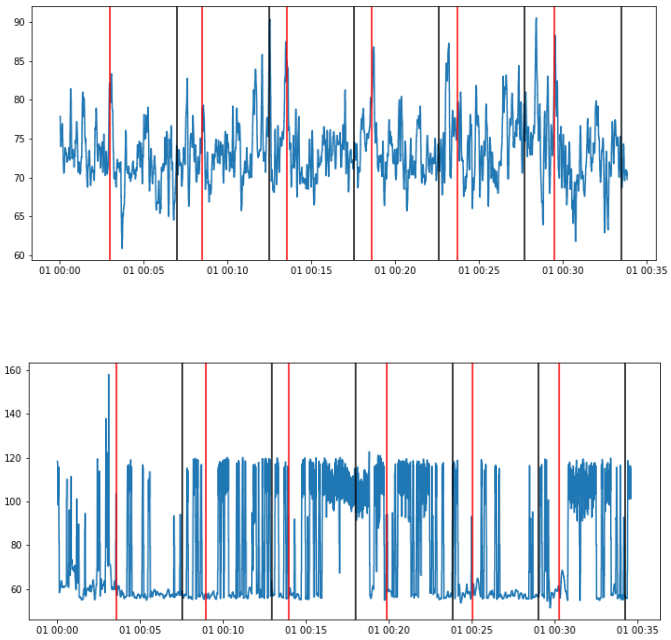

FIGURE 7. HR SIGNALS FROM TWO PARTICIPANTS, SHOWING A GOOD QUALITY SIGNAL (UP), AND A VERY NOISY SIGNAL (DOWN))

The poor quality of the HR signal affected up to one third of the samples collected in the video streaming dataset. These artifacts were not equally distributed across conditions and participants, unbalancing the experimental design, and making the use of this signal unfeasible in practice. Something similar happened to the HR signal collected in the LLM response dataset, limiting the utility of the data derived from this signal in this context. However, in the case of the online search dataset, a different system was used. Instead of using a photoplethysmography (PPG) sensor, an electrocardiogram (ECG) was performed by placing electrodes on the participant's wrists and forearm. The HR signal on this was then automatically computed using AcqKnowledge software (Biopac). The signal was analyzed by an expert, and those segments (downsampled to 1 Hz) that showed unrealistic values were eliminated. The result showed that about 98% of the samples provided signals of acceptable quality. This clearly indicates that the ECG-based system is much more reliable and robust than the PPG-based system, which provided a very low quality signal.

4. User profiles based on psychophysiological signals

In section 4 of SORUS-RIS-A2.2-E1, a preliminary methodological approach was established for the profiling of users based on their response to latency in different network usage contexts. In essence, the proposed methodology was divided into two steps:

1. First, **individual sensitivity** to system latency was assessed. Based on the literature review and the experimental results reported in the previous deliverable, sensitivity to latency has been shown to vary depending on the context and the latency ranges experienced within that context. Various approaches exist for reporting this sensitivity. For example, latency sensitivity can be defined in terms of subjective individual responses (e.g., perceptual quality reported at different latency levels). However, it can also be understood in a more comprehensive manner, as a function of the cognitive and emotional responses it induces, consciously or unconsciously, in the user. In the first deliverable, we followed the first paradigm, relying on self-reported responses. In this stage of the project, however, we focused on the second paradigm, utilizing psychophysiological signals to measure users' sensitivity to latency across the different contexts analyzed.
2. Once a sensitivity value has been established for each user, the next step is to use these values to create **user profiles** and define distinct typologies based on their responses to latency. Following the experimental design explored in the first deliverable, techniques such as clustering or Latent Profile Analysis (LPA) methods (Spurk et al., 2020) were considered to be used for comparative purposes. However, although LPA may provide some advantages, such as the ability to accommodate partial membership to a given cluster, the method also requires a large number of participants. Given the complexity associated to collecting psychophysiological data, the number of participants in our experimental setting is low compared to the open, self-reported datasets used in the first deliverable. This makes the use of LPA unfeasible in our experimental context. By contrast, when using non-latent methods, e.g. k-means, power to detect clustering is primarily dependent on cluster separation, and much less on sample size (Dalmaijer et al., 2022). Therefore, in the analyses conducted in this section we only opted for non-latent approaches.

Regarding point 2, numerous non-latent clustering approaches have been proposed in the literature. As a result, it is neither straightforward to determine the best method for a specific dataset in advance, nor is it trivial to identify the optimal number of clusters in the profiling. Therefore, in this study we have taken advantage of the approach provided by the clValid package (Brock et al., 2008) in R. This package allows, in a straightforward way, to compare multiple clustering algorithms and identify which performs best. It also enables exploring the optimal number of clusters.

In our case, three different methods were tested, including **hierarchical clustering**, **k-means method**, and **Partitioning Around Medoids (PAM)**. In short, hierarchical clustering is a method that builds a hierarchy of clusters, allowing the researcher to visualize the data in a tree-like structure (i.e., dendrogram). It is particularly useful for discovering the structure of the data without needing to predefine the number of clusters. The k-means method, on the other hand, is a popular technique that partitions the data into a predefined number of clusters, aiming to minimize the variance within each cluster. It is efficient and well-suited for large datasets but requires the user to specify the number of clusters beforehand. Finally, Partitioning Around Medoids (PAM) is similar to k-means but uses actual data points (medoids) to represent the clusters, making it theoretically more robust to noise and outliers. Testing these three methods provides a comprehensive approach, allowing us to assess diverse ways of structuring the data and compare their effectiveness in identifying meaningful user profiles in our context.

The proposed package also provides **internal validation** measures, which evaluate the quality of the estimated clusters based on intrinsic properties of the data. Other validation methods are also provided by the package, but not discussed here. The internal measures that we specifically explored are, on the one hand, a **connectivity** measure that indicates the degree of connectedness between clusters, based on k-nearest neighbors. The method ranges such a connectivity between 0 and infinity, and the objective is to minimize this value. On the other hand, average **Silhouette width** and **Dunn index**, were also tested. These metrics were designed to evaluate the quality of clusters. In particular, the Silhouette width ranges from -1 to 1, and it measures how well each point is clustered. It considers both cohesion (closeness within the same cluster) and separation (distance from other clusters). A higher score indicates better clustering. The Dunn index, ranges from 0 to infinity, and measures the separation between clusters and the compactness within clusters. A higher value means better separated and more distinct clusters.

The profiling results for each dataset are presented in the following subsections.

4.1 User profiling with the video streaming dataset and the LLM interaction dataset

Given the costs associated with collecting psychophysiological measures, studies that include such data typically involve a small number of participants. To analyze existing profiles, we aimed to identify clusters of participants based on their psychophysiological responses to latency. We combined two datasets: one from video streaming, which includes data from five participants, and one from LLM interaction, which includes data from eight participants. Considering these two datasets together is appropriate, as they both involve passive tasks and utilize the same set of EEG and peripheral psychophysiological signals. This comparison is meaningful, especially when contrasted with interactive tasks (such as the online browsing task), where we hypothesize that the impact of latency may differ significantly (Doherty & Sorenson, 2015).

We used the datasets to analyze whether different clusters can be identified according to their sensitivity to latency, expressed in terms of the impact of latency on multimodal signals. Based on the results collected in section 2.1; the most suitable multimodal signals for this purpose were the **FAA**, **Alpha occipital** and **beta parietal** signals, as well as the **phasic EDA** signal.

Following the methodological approach explained above to calculate the individual sensitivity to latency, the average value of each signal during the viewing of each video was calculated for each participant. Then, the correlation between the latency values and those of each signal per participant was calculated, and the probability values (p-value) associated with this correlation. This data was used to perform step two, the exploration of the emerging clusters. Table 9 and Figure 8 below show the results of this analysis.

Table 9. Results of the cluster validation analysis (video streaming and LLM interaction datasets)

		n-clusters							
		2	3	4	5	6	7	8	9
hierarchical	Connectivity	3.28	6.38	14.74	18.20	19.77	24.22	27.21	30.08
	Dunn	0.66	0.74	0.65	0.65	0.65	0.65	0.74	0.91
	Silhouette	0.26	0.25	0.12	0.09	0.06	0.10	0.09	0.08
kmeans	Connectivity	9.58	6.38	17.68	22.66	22.79	26.74	28.83	31.08
	Dunn	0.45	0.74	0.54	0.58	0.60	0.77	0.83	0.92
	Silhouette	0.22	0.25	0.11	0.11	0.12	0.14	0.12	0.11
pam	Connectivity	10.25	13.32	16.15	19.44	22.65	24.22	27.21	29.21
	Dunn	0.42	0.51	0.57	0.57	0.65	0.65	0.74	0.74
	Silhouette	0.08	0.07	0.08	0.07	0.11	0.10	0.09	0.06

Optimal Scores:			
	Score	Method	Clusters
Connectivity	3.28	hierarchical	2
Dunn	0.92	kmeans	9
Silhouette	0.26	hierarchical	2

As the results indicate, both the validation based on connectivity and the Silhouette method converge in identifying the hierarchical clustering with two clusters as providing the best results.

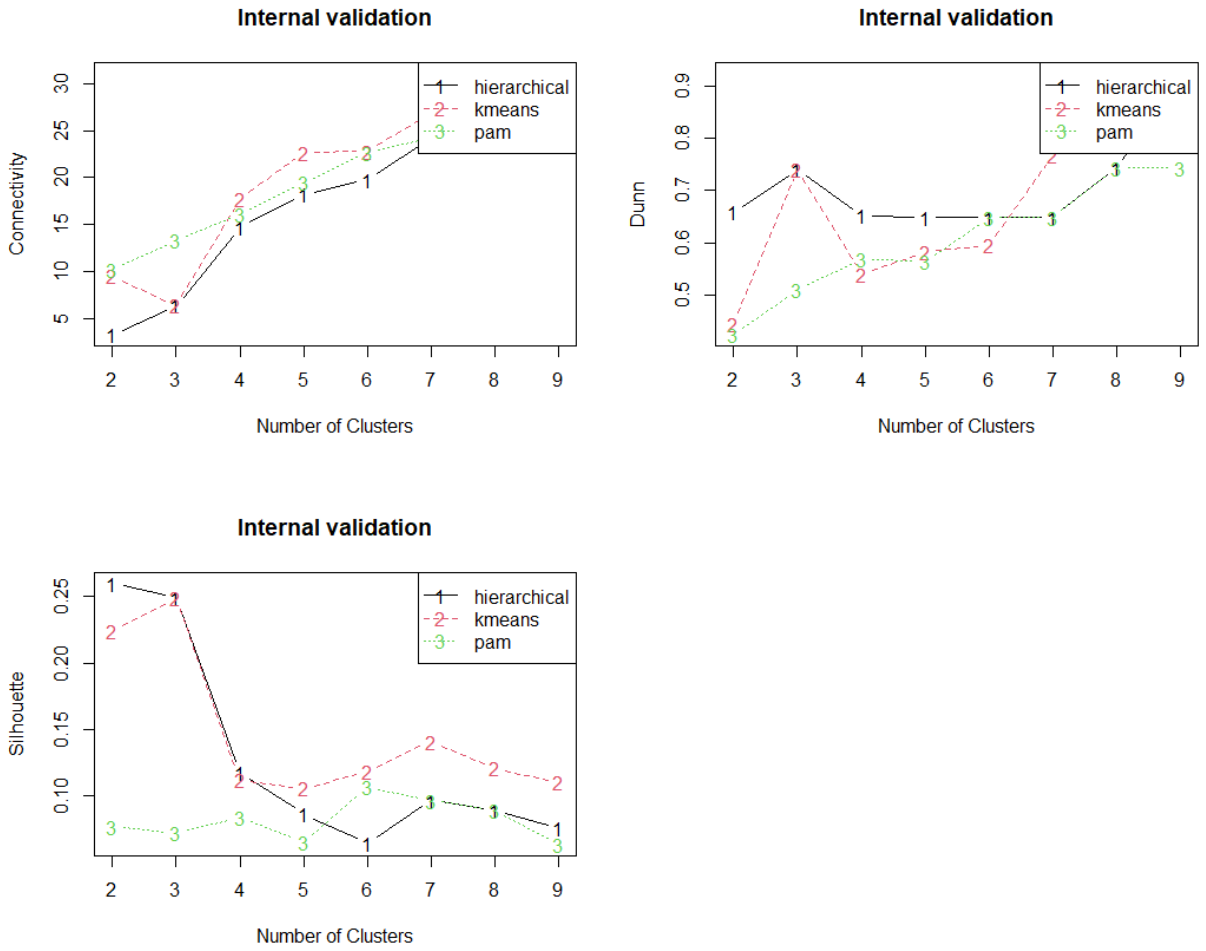

FIGURE 8. GRAPHICAL REPRESENTATION OF THE INTERNAL VALIDATION METHODS FOR DIFFERENT NUMBERS OF CLUSTERS IN THE VIDEO STREAMING AND LLM INTERACTION DATASET

Thus, hierarchical clustering, with two clusters was employed in this study, using a dissimilarity matrix based on Euclidean distance. Subsequently, the mean values of each of the measures per defined groups were evaluated. The results are shown in Figure 9, with the descriptive statistics for these two clusters presented in Table 10 and Figure 10.

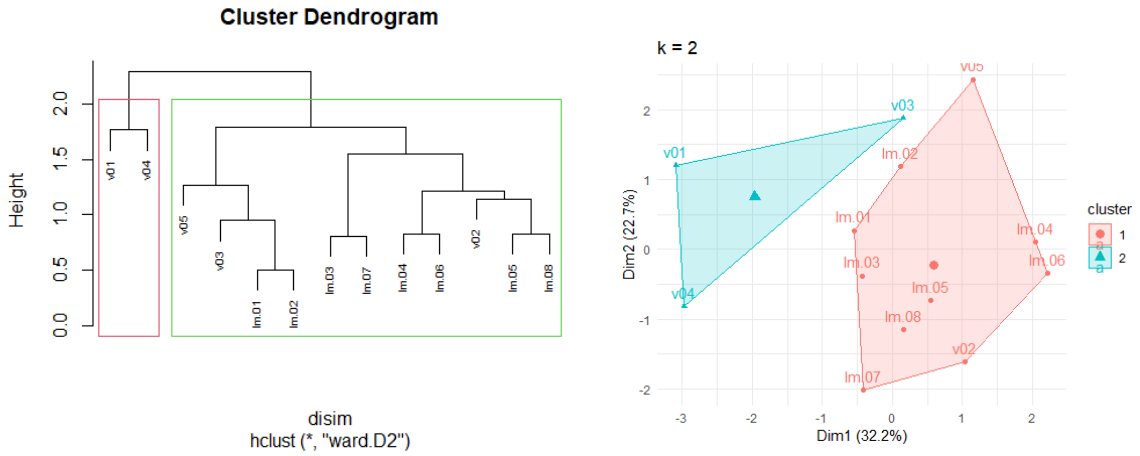

FIGURE 9. GRAPHICAL REPRESENTATION OF THE CLUSTERING SOLUTION: DENDROGRAM (LEFT) AND VISUAL REPRESENTATION IN TWO DIMENSIONS (VIDEO STREAMING AND LLM INTERACTION DATASETS)

Table 10. Cluster statistics (video streaming and LLM interaction datasets)

	FAA		Alpha occ.		Beta par.		EDA phasic	
Cluster	**Pearson r**	**p-value**	**Pearson r**	**p-value**	**Pearson r**	**p-value**	**Pearson r**	**p-value**
1	-0.08	0.55	-0.12	0.46	-0.08	0.46	-0.05	0.49
2	-0.44	0.43	0.06	0.30	-0.28	0.36	0.80	0.06

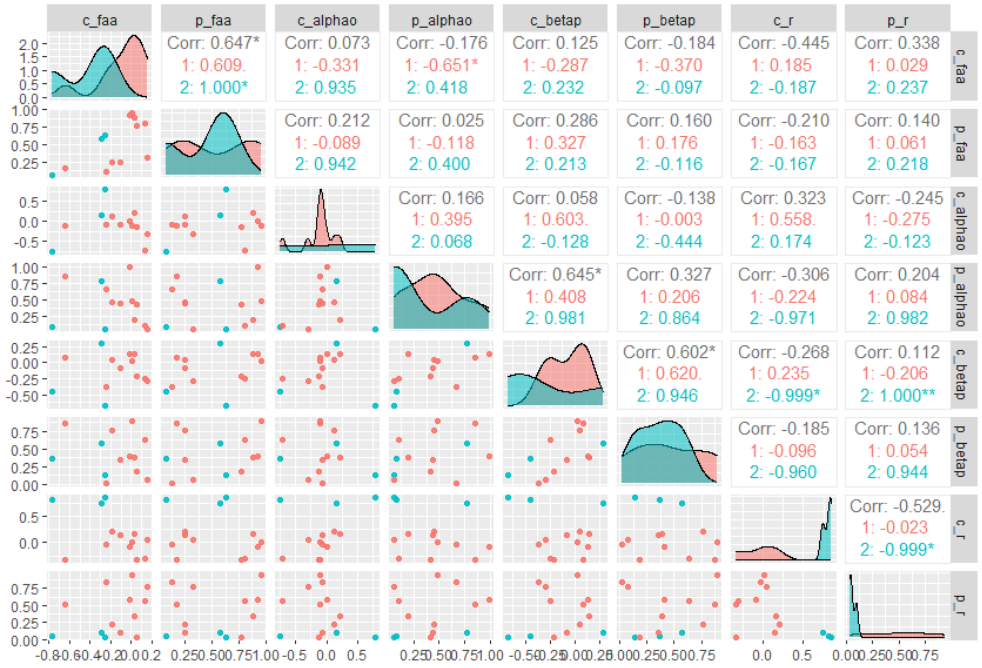

FIGURE 10. CORRELATIONS BETWEEN CLUSTERS' COMPONENTS (VIDEO STREAMING AND LLM INTERACTION DATASETS)

These results suggest that the main difference between the two clusters is in the phasic responses of electrodermal activity: the larger group of participants shows more consistent responses, as suggested by the high correlation coefficient and low p-value, whereas the minority group lacks such responses. Thus, the main difference between participants in their response to latency seems to be in terms of the emotional arousal they produce: one could distinguish between a group sensitive to latency and one that is not, and one could, in principle, discriminate between the two on the basis of the emotional arousal that latency produces. However, it is important to note that the group with lower sensitivity to latency consists of only a small number of users. Therefore, it would be beneficial to replicate these results with larger sample sizes in future studies.

4.2 User profiling with the online search dataset

In the case of the online search dataset, a similar strategy was followed. First, the correlations and p-values of these between the latency values of each search and the tonic EDA, EMG, and HR values were calculated. Again, these signals were selected according to their statistically significant relationship with latency, as shown in section 2.3. Next, the clValid package in R was used to explore the three clustering methods proposed (i.e., hierarchical, k-means, and PAM), to define the optimal number of clusters (values examined between 2 and 8), and using internal validation methods. The results of this exploration process are reflected in Table 11 and Figure 11.

Table 11. Results of the cluster validation analysis (online search dataset)

		n-clusters						
		2	3	4	5	6	7	8
hierarchical	Connectivity	3.33	8.18	15.65	16.58	22.18	24.88	31.23
	Dunn	0.58	0.35	0.46	0.46	0.55	0.55	0.51
	Silhouette	0.28	0.23	0.29	0.26	0.27	0.25	0.25
kmeans	Connectivity	11.55	13.21	15.65	16.58	27.47	30.17	31.23
	Dunn	0.32	0.42	0.46	0.46	0.48	0.48	0.51
	Silhouette	0.28	0.31	0.29	0.26	0.27	0.26	0.25
pam	Connectivity	7.46	11.57	21.58	23.89	26.34	28.99	31.69
	Dunn	0.31	0.41	0.33	0.33	0.42	0.42	0.42
	Silhouette	0.29	0.31	0.24	0.24	0.23	0.26	0.23

Optimal Scores:			
	Score	Method	Clusters
Connectivity	3.3317	hierarchical	2
Dunn	0.576	hierarchical	2
Silhouette	0.3113	pam	3

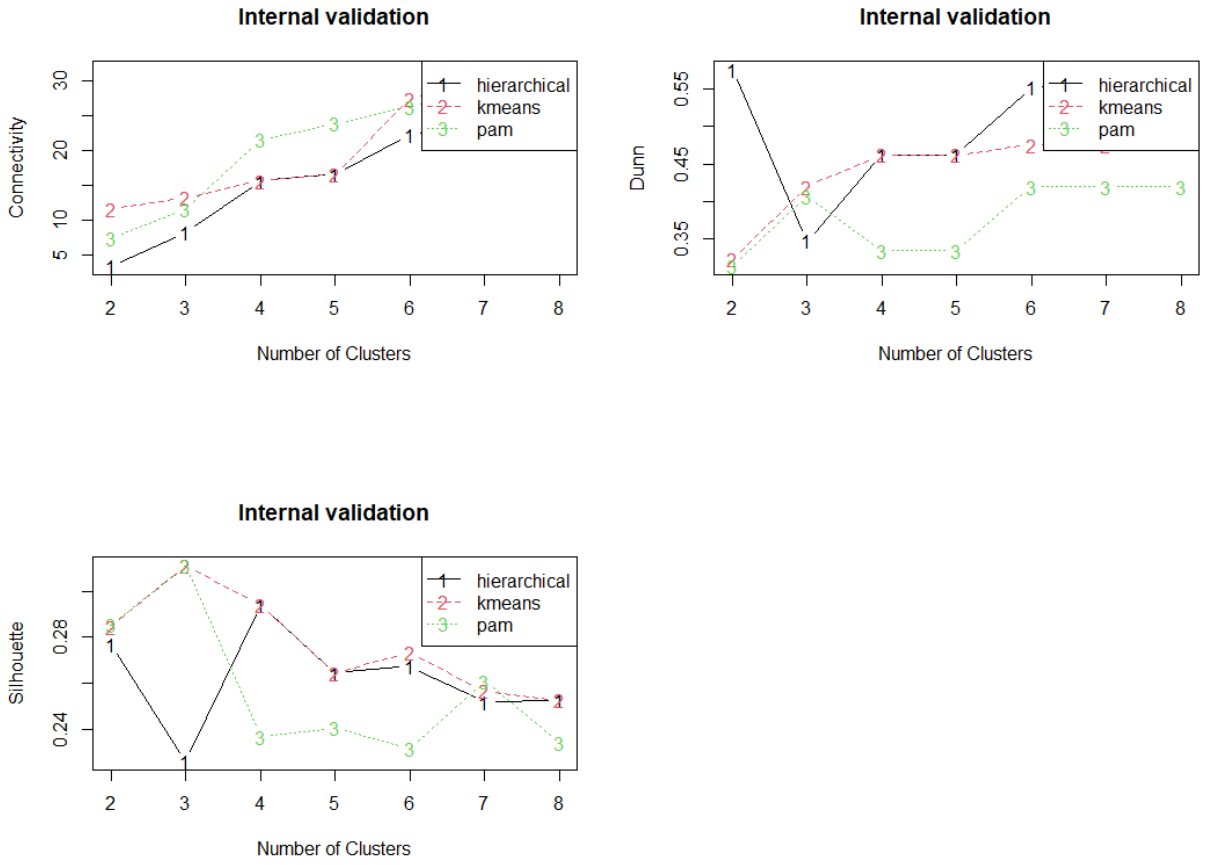

FIGURE 11. GRAPHICAL REPRESENTATION OF THE INTERNAL VALIDATION METHODS FOR DIFFERENT NUMBERS OF CLUSTERS IN THE ONLINE SEARCH DATASET

The connectivity and Dunn validation methods agree that the best solution in this case - as in the previous case - is to apply a hierarchical clustering method, defining two clusters. Thus, this method was applied to the observations of the 19 participants in this study, using a dissimilarity matrix based on Euclidean distance. As can be seen in the graphical representation of the results (Figure 12), a larger cluster is described (approximately 2/3 of the cases) and a second smaller cluster, which also contains an apparently larger dispersion of the different values.

In the statistics of the different components of both clusters, we see that one of the groups seems to have a pronounced and clear response (higher values in the correlations) to the latency present in the searcher responses. This is evidenced in positive correlation values in tonic EDA, negative correlation in HR (which is compatible with an attentional orienting response) and increased facial EMG (which was measured in the corrugator supercilii, thus indicating negative valence of emotions). All this contrasts with negative or closer to zero values presented by the other group, which indicates a lower sensitivity (lower arousal and negative responses) to latency in the searcher.

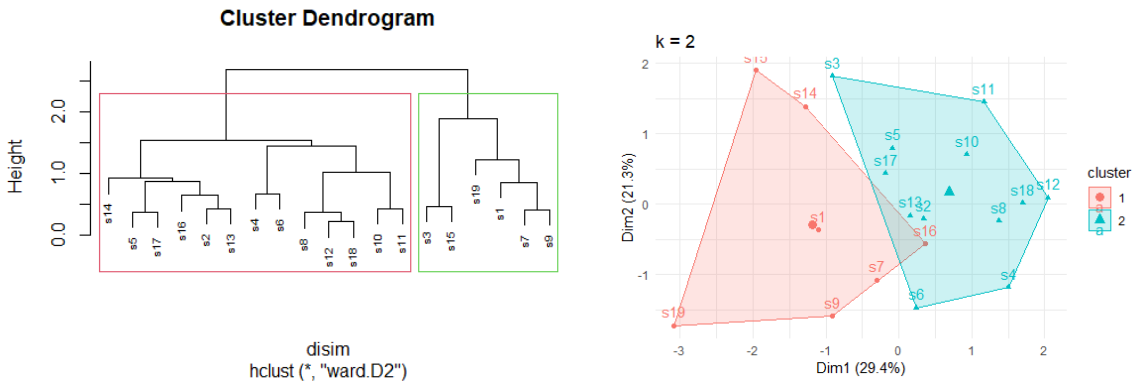

FIGURE 12. GRAPHICAL RESPRESENTATION OF THE CLUSTERING SOLUTION: DENDOGRAM (LEFT) AND VISUAL REPRESENTATION IN TWO DIMENSIONS (ONLINE SEARCH DATASET)

Table 12. Cluster statistics (online search dataset)

	EDA tonic		HR		EMG	
Cluster	**Pearson r**	**p-value**	**Pearson r**	**p-value**	**Pearson r**	**p-value**
1	0.202	0.112	-0.097	0.165	0.421	0.126
2	-0.225	0.148	0.028	0.078	-0.390	0.040

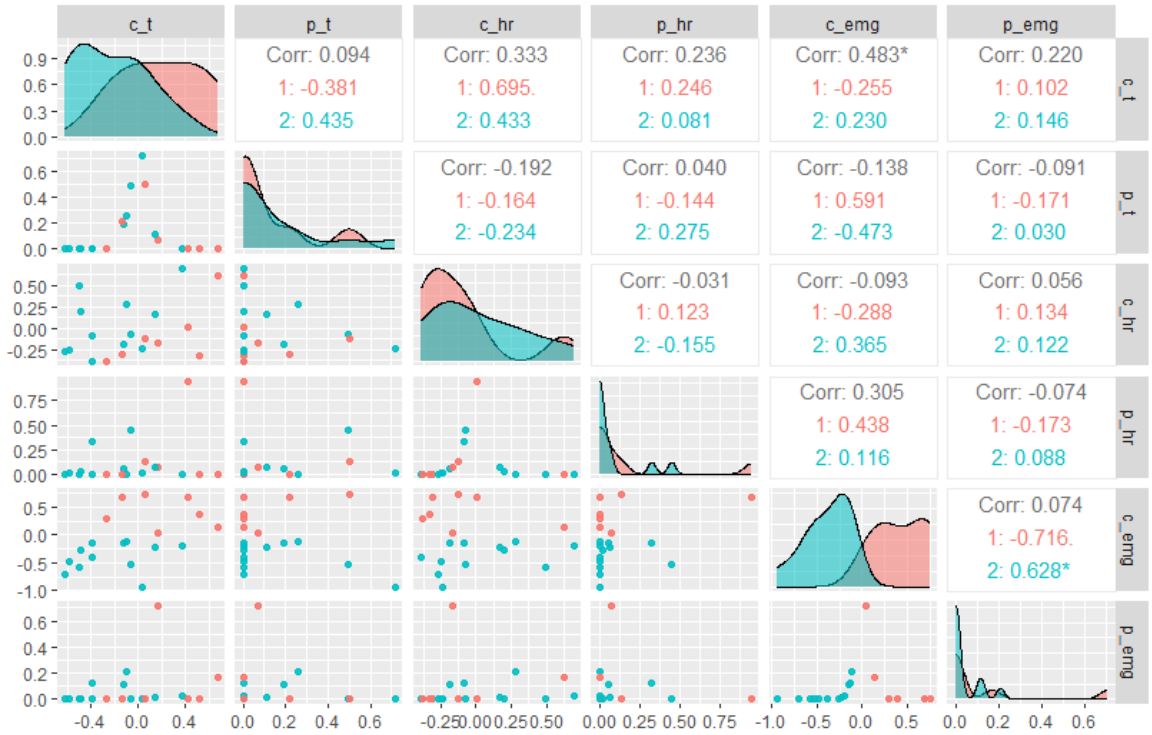

FIGURE 13. CORRELATIONS BETWEEN CLUSTERS' COMPONENTS (ONLINE SEARCH DATASET)

5. Conclusions

Building upon the results presented in the previous deliverable, Task 3 of SORUS-RIS-A2.2-E2 ('Mapping Psychological Constructs to Low- and High-Level Indicators and Generating User Profiles') was addressed in this phase. In prior tasks, key psychological concepts related to latency perception and its impact on users were identified, along with available methodological approaches for their measurement. Initially, in SORUS-RIS-A2.2-E1, user profiling was conducted using self-reported metrics. However, a pilot study indicated that combining self-reported subjective measures with psychophysiological data was particularly valuable for more accurately identifying distinct user profiles. As a result, this phase focused on conducting a more comprehensive analysis of psychophysiological measures for detecting latency, in conjunction with self-reported metrics. For this aim the construction of appropriate datasets in different scenarios was of s. Additionally, the task aimed to assess the reliability and robustness of each signal considered. Finally, based on the extracted signals and behaviours, the task sought to generate more complete user profiles based on their responses to the system's technical performance, with particular emphasis on latency, across various contexts. The general conclusions drawn from each of these subtasks during this second period are presented below.

The first objective of this activity was to analyse the relationship between multimodal psychophysiological cues and the perception and subjective impact of latency on users across different contexts of real-world services. The results suggest that in less active usage scenarios, such as watching online videos or reading responses generated by language models, the effects of latency are primarily reflected in changes related to user motivation (as seen in frontal alpha asymmetry - FAA - in the EEG), visual attention (occipital alpha), negative emotion (parietal beta), and emotional arousal (phasic EDA). However, when variations in latency are minimal, correlations between multimodal indicators and users' self-reported responses become less consistent, indicating that the impact of latency is not as easily perceived in such conditions. In more interactive environments, such as online browsing, the connection between latency levels and psychophysiological signals is more evident. Here, latency has a more noticeable impact, with clear relationships observed between the latency and signals such as EDA, EMG, and HR, indicating a more robust and reliable mapping between the two.

In the second part of the analysis, the reliability and robustness of the psychophysiological signals used in the study were thoroughly evaluated by an expert to assess their potential as reliable indicators for the research methodology. The findings suggest that both EEG and EDA signals demonstrate sufficient robustness, making them effective tools for addressing the research objectives. Specifically, the EDA signal stands out due to its ease of use and its potential for integration into wearable devices, highlighting its promising applicability in future research and real-world contexts.

However, the analysis of the EMG signal revealed frequent segments with noise, which were likely caused by electrode movement during the experimental sessions. This finding underscores the importance of reviewing EMG data carefully, with expert oversight, to ensure its reliability before being used in further analysis. In light of these challenges, alternative approaches could be considered to complement or replace EMG measurements. For instance, detecting facial expressions through camera-based systems could offer a non-intrusive and reliable way to assess emotional and cognitive responses. Techniques such as facial action coding systems (FACS) or computer vision algorithms could be used to analyse facial muscle movements, providing valuable insights into user reactions without the need for physical sensors. These methods could serve as a potential solution for reducing the impact of noise and improving the overall reliability of psychophysiological measurements, especially in cases where traditional sensors may be prone to interference.

Regarding the HR signal, the electrocardiogram (ECG)-based system provided reliable results, in contrast to the photoplethysmography (PPG)-based system, which exhibited less consistent performance. As a result, the ECG-based system is recommended for future use due to its superior reliability in capturing heart rate data.

Finally, the third objective was to establish user profiles according to their response to latency, understood as the impact of latency on different psychophysiological indicators, which are associated with cognitive and emotional processes. The results obtained from this task indicate that the sets of users examined can be categorized into two distinct groups. One of them comprises those users who show a greater sensitivity to latency, while the other includes those who do not seem to react in a clear cognitive or emotional way to this factor. This sensitivity manifests itself in diverse ways depending on the context: in the form of physiological EDA responses (emotional arousal) in the more passive use cases (e.g., watching videos or reading LLM responses), and in EDA, EMG, and HR responses in more interactive use cases (performing online searches).

To identify these user profiles, several clustering methodologies were tested, including hierarchical clustering, k-means, and Partitioning Around Medoids (PAM). Among these, hierarchical clustering with two clusters provided the best results, clearly differentiating between the two groups based on their latency sensitivity. This approach, combined with an evaluation of various psychophysiological indicators, proved to be an effective method for profiling users according to their responses to latency, allowing for the identification of distinct user typologies that can inform more personalized service offerings in the future

In summary, this phase of the project successfully expanded upon the initial profiling work by integrating psychophysiological data with self-reported subjective measures to improve the accuracy of user profiles. The analysis not only clarified the impact of latency on user experience across various contexts but also validated the reliability of the signals used, with promising results for the use of EDA and EEG signals in future research. Furthermore, the generation of user profiles based on psychophysiological responses lays the foundation for developing more personalized and context-sensitive services that can adapt to individual user needs and preferences in real-world scenarios. This approach offers valuable insights for enhancing user experience, particularly in

latency-sensitive environments. Additionally, the methodological advancements made in this phase underscore the potential for continued exploration of multimodal psychophysiological data, promising further improvements in understanding user behaviour and refining interactive technologies for diverse user groups.

References

Ajenaghughrure, I. B., Sousa, S. D. C., & Lamas, D. (2020). Measuring trust with psychophysiological signals: a systematic mapping study of approaches used. Multimodal Technologies and Interaction, 4(3), 63.

Bach, D. R., Castegnetti, G., Korn, C. W., Gerster, S., Melinscak, F., & Moser, T. (2018). Psychophysiological modeling: Current state and future directions. *Psychophysiology, 55*(11), e13214.

Bolls, P. D., Weber, R., Lang, A., & Potter, R. F. (2019). Media psychophysiology and neuroscience: Bringing brain science into media processes and effects research. Media effects: Advances in theory and research, 195-210.

Busselle, R., & Bilandzic, H. (2009). Measuring narrative engagement. *Media Psychology, 12*(4), 321-347.

Boucsein, W. (2012). Electrodermal activity. Springer Science & Business Media.

Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2022). Statistical power for cluster analysis. BMC bioinformatics, 23(1), 205.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. Handbook of psychophysiology, 2, 200-223.

Doherty, R. A., & Sorenson, P. (2015). Keeping users in the flow: Mapping system responsiveness with user experience. *Procedia Manufacturing, 3*, 4384-4391.

Duanmu, Z., Zeng, K., Ma, K., Rehman, A., & Wang, Z. (2016). A quality-of-experience index for streaming video. *IEEE Journal of Selected Topics in Signal Processing, 11*(1), 154-166.

Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., & Citi, L. (2015). cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering, 63*(4), 797-804.

Inzelberg, L., Rand, D., Steinberg, S., David-Pur, M., & Hanein, Y. (2018). A wearable high-resolution facial electromyography for long term recordings in freely behaving humans. Scientific reports, 8(1), 2058.

Kleckner, I. R., Jones, R. M., Wilder-Smith, O., Wormwood, J. B., Akcakaya, M., Quigley, K. S., ... & Goodwin, M. S. (2017). Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. IEEE Transactions on Biomedical Engineering, 65(7), 1460-1467.

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: fully automated statistical thresholding for EEG artifact rejection. Journal of Neuroscience Methods, 192(1), 152-162.

Oliver, M. B., & Bartsch, A. (2010). Appreciation as audience response: Exploring entertainment gratifications beyond hedonism. *Human Communication Research, 36*(1), 53-81.

Perusquía-Hernández, M., Ayabe-Kanamura, S., Suzuki, K., & Kumano, S. (2019, May). The invisible potential of facial electromyography: A comparison of EMG and computer vision when distinguishing posed from spontaneous smiles. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-9).

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior, 120*, 103445.

Tronstad, C., Amini, M., Bach, D. R., & Martinsen, Ø. G. (2022). Current trends and opportunities in the methodology of electrodermal activity measurement. *Physiological Measurement, 43*(2), 02TR01.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12(6), 1100-1122.