VRIJE UNIVERSITEIT

# Monitoring the Engagement of Groups by Using Physiological Sensors

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam
op gezag van de rector magnificus
prof.dr. V. Subramaniam
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op woensdag 23 mei 2018 om 9.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Chen Wang

geboren te XuanZhou City, Anhui Province, China

| | |
|---|---|
| promotor: | prof.dr. D.C.A. Bulterman |
| copromotor: | dr. P.S. Cesar |

# Monitoring the Engagement of Groups by Using Physiological Sensors

Chen Wang

Promotiecommissie:

- Prof. dr. M. Ursu, York University, UK
- Prof. dr. T. Chambel, U. Lisboa, PT
- Dr. V. Kallen, TNO, NL
- Dr. N. Silvis-Cividjian, VU, NL
- Prof. dr. J. Heringa, VU, NL (Chair)

ISBN: 978-94-6259-028-1

Typeset with Microsoft Word

*Cover picture*:
Designed by Michael Williams, Rotterdam, The Netherlands

The work reported in this thesis has been carried out at the Centrum Wiskunde & Informatica (CWI), under the auspices of the group Distributed and Interactive Systems (DIS). The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. ICT-2011-287760, and the research funding from Xinhuanet (supporting period from 01/10/2014 – 30/09/2016).

# Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dick Bulterman for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank Dr. Pablo Cesar, for his insightful comments and encouragement, and also for the hard question which incented me to widen my research from various perspectives. His guidance helped me in all the time of research and writing of this thesis. He was an excellent mentor for my Ph.D study. This thesis would not have been possible without his help, support and patience. His good advice, support and friendship has been invaluable on both an academic and a personal level, for which I am extremely grateful.

My sincere thanks also go to Dr. Eric Pauwels and Prof. Marie-Colette van Lieshout, who provided me the opportunity to learn from them. Without their precious support, it would not be possible to achieve a breakthrough on this research.

I thank my fellow lab mates for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last 5 years. I also thank my friends at the following institutions: Falmouth University of UK, and Xinhuanet of China. In particular, I am grateful to Prof. Phil Stenton and Erik Geelhoed for enlightening me the start of this research.

I am also grateful to the Xinhuanet team: Prof. YangMing, Dr. Wangzhenzheng, Zhuxintong, Jacqueline, Thomas, Zhiyuan for their unfailing support for this research. It would not have been possible to write this doctoral thesis without the help and participation from them.

Last but not least, I would like to thank my family and my close friends for supporting me spiritually throughout writing this thesis and my life.

# TABLE OF CONTENTS

# List of Figures

# List of Tables

# ABSTRACT

This thesis reports on our research to design a GSR system that can support and enhance actors' awareness of remote audiences in a distributed theatrical environment. At the early stage of this work, we made concentrated efforts to design and develop appropriate sensor hardware. We did this to find a practical way to construct relevant hardware that could be worn in a theatrical environment. We thoroughly tested each hardware version in different user field studies. Then, we developed related algorithms to process the data delivered by our sensors. When we had enough knowledge about the relationship between GSR patterns and audience engagement, we developed the real-time algorithm and the feedback mechanism so that remote audience engagement could be visualized.

Our solution makes a technical contribution from both an engineering and a software design perspective in the creation of a system that allows theater stakeholders to explore the response data of an audience. This exploration has the potential to enhance the creative work of the theater stakeholders and to understand how audience members respond to their creative outputs. Although others have investigated audience response through the GSR systems, the deployment of a system of this scale in a live, large-scale theater setting is truly novel.

A first significant research result was the refinement of a set of heuristics for design and development of hardware. The heuristics were created during the first three years. We were able to use them for building and implementing the infrastructure in the developmental process of the sensor hardware. By following the heuristics, user cases conducted in the commercial theater performance and the dance performance proved that the sensor system was easily deployable in theaters. In addition, our studies also proved that our sensor system can be used for both real-time and offline purposes.

The second success was that we could use GSR data to (real time) infer audience engagement. We found that the responses from the engaged users concerning sensory pattern showed a strong correlation between lab and field studies. Interestingly, the responses from the non-engaged participants did not

correlate across user cases between the lab and the field trials. These results are consistent with a similar phenomenon mentioned in a previous research. A "boredom" state captured in a lab may have different patterns compared to one in a field study. In addition, our findings on sensor data may bring inspiration for researchers who intend to investigate user states crossing different scenarios. According to our learning experience from lab studies, we found that it is unlikely for users to generate a boredom state when watching short videos. This happens if every participant is seriously engaging in the task. Even though it is in an unknown language with a quite low resolution, they typically try to understand what it is happening in the video. Therefore, for certain user emotional states, it may be possible to obtain them in reality rather than just in lab conditions.

In our experimental studies, the development of a feedback mechanism has successfully enhanced the performance content for actors. During interviews, we have learned how to design and develop a feedback mechanism that suit a performance setting. However, it was not possible to develop a mechanism that will optimally fit all of our system requirements. Repetition will be required because the effect of a mechanism will be evaluated during the actual experiment. In general, one experiment alone cannot be used to make a concrete decision. If the intention is to design a mutual feedback mechanism, repetition and modification should be conducted.

The work reported in this thesis cover eight research questions and resulted in nearly a dozen scientific publications, of which three were awarded *best paper* distinctions at major scientific conferences.

*These subliminal aspects of everything that happens to us may seem to play very little part in our daily lives. But they are the almost invisible roots of our conscious thoughts.*

*- Carl Jung*

# 1

# Introduction

This dissertation reports on our research to design a feedback mechanism that would enhance the perception of artists' awareness with remote audiences. Of early interest was the application of this work to video communication platforms. A video communication platform provides many interesting possibilities to create novel applications. Numerous distributed performances using video platforms allow artists and remote audiences to collaborate from different locations. *Vconect*, a European project, has used distributed performances (Figure 1) as one of the use cases to analyze the functionality required when creating a video communication platform that promotes and aids the complex communication topologies that mediate conversations among associated group members. With this platform, both artists and audiences from different locations can interact with each other in real time using an online network as if they we located in the same room. The potential modes of interaction can include performances, rehearsals, and improvisation. The participants can be interconnected by "high fidelity multichannel audio and video links" and specialized collaborative software tools. In summary, distributed performances require novel technical support forms to enable artists and remote audiences to engage in a live stage performance.

Beyond distributed performances, even local measurements of affinity turned out to be useful for studying the core phenomena related to this work. For this reason, feedback capture and representation in both local and distributed performances were studied.

## 1.1 Motivation

Distributed performances challenge the artists' awareness of remote audiences. In a collocated performance setting, actors are aware of the audiences, of their appreciation, of their own performance, and of how immersed they are in the performance. Thus, ideally, a distributed performance system should facilitate a similar experience as this traditional setting. However, such links are lacking between the artists and audiences in a distributed environment. This drawback brings an interesting thought:

- *How can artists be aware of the responses of remote audiences in a distributed environment?*

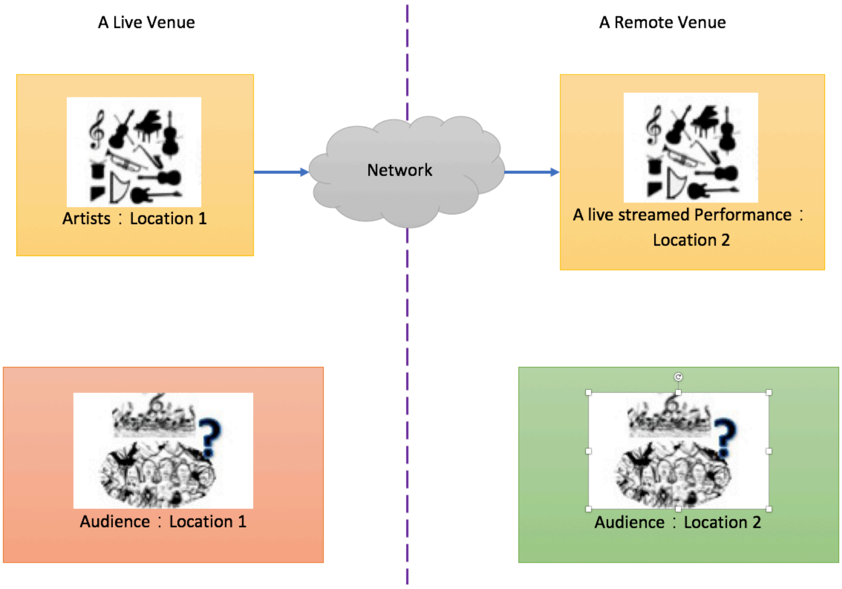


**Figure 1: The distributed performance**

The communications channel between artists and remote audiences is usually uni-directional, which isolates performers from the audience. If artists could get responses from the remote audiences, this can significantly enhance the awareness between the two parties. This means that without a facilitated feedback channel, the artists cannot identify what occurred in the remote locations, while the remote audiences feel they are watching a recorded video instead of a live performance. A feedback channel can simultaneously send responses, reactions, and experiences of the remote audiences to a live performance, connecting them to the venues.

The *Vconect* study on audience's responses showed limitations in a theatrical environment [11,21]. The traditional subjective evaluation methods (e.g., questionnaires, interviews) and behavioral observations (e.g., applauding) could only be illicited after a performance, and they could not be applied in a real-time situation. More immediate forms of user evaluation, such as eye tracking and facial expressions measurements, could not be deployed in a large (or at least, representative) scale for many practical reasons. A voting and labeling systems would definitely disturb the audience's watching experience. As a result, we suggest developing a non-intrusive mechanism to attain remote audience responses in a distributed environment. Such a mechanism allows the remote audience to view a performance without being interrupted. Also, their sensor data can be seamlessly obtained, processed, and manipulated based on application requirements. The collecting of feedback data would not interfere with the performance and tailored feedback mechanisms could be created for the needs of individual settings.

In this thesis, a physiological sensor system (PSN) was employed to monitor the responses from remote audiences. Specifically, a galvanic skin response (GSR) sensor system was used to measure audience responses at remote theaters; the collected sensor data was both analyzed offline or processed in real time. With this, researchers could explore the sensor data, and establish a necessary feedback mechanism to enhance the relationship between the performers and remote audiences in a distributed environment.

Two types of stakeholders are interested in the work. One group consists of the artists, directors, and producers who are particularly interested in continuous audience responses that are elicited during a play [13， 20], as they cannot get such data using the traditional methods. The parties are keen on exploring what kind of real-time mechanism can be developed during a play, so the new performing concept may be created, potentially supporting future artistic productions that use a sensor mechanism to create different story endings. The other stakeholder category is the audience themselves. Since the sensor devices measure their conscious and unconscious reactions, they may be curious about visualizing the experiences. Besides, they might be interested in comparing their own sensor signal patterns to the persons they already know or the strangers. In other words, whether other audiences will have similar experiences as they had, these may be reflected on the sensor readings. In summary, quantifying such abstract experiences through sensor readings provides meaningful information for both the artists and audiences.

## 1.2 Challenges

There are several challenges related to the GSR method in a distributed theatrical environment. We summarize them into six categories:

- **Understanding engagement from physiological sensor data.** Physiological sensor data is quite complex to understand as the inferring process requires a background from both psychology and computer science. Additionally, researchers working on physiological computing noted that in most cases, a significant correlation is not always found between physiological sensor data and subjective reports, which make it more difficult to validate the inferring process. Also, sensor signal patterns in a lab control environment cannot be easily replicated in a field study, and it remains unclear whether such patterns can be used across the different scenarios. Thus, a profound understanding of the different applications of the sensor data is required.
- **Developing appropriate live presentation mechanisms (visual and aural) from the inferred engagement data.** The study of this class of information is dependent on the end users' desire, interest, and performance style, with emphasis on the audience's profile. The different stakeholders (e.g., directors, producers, and artists) can have various definitions of an appropriate feedback mechanism. Thus, researchers need to initially identify who are the end users

and then communicate with them to know their insights and interests before designing a mechanism. In practice, such a learning process is a complex activity. A related challenge is interpreting scientific results into an artistic language that can be easily understood by the end users. This multi-level interaction is generally time-consuming and requires a close interaction and in-depth communication.

- **Environmental restrictions.** Conducting user studies in a theatrical environment is a complex task with many uncontrollable variables, such as varying rules from one theater to another. Also, generic theater environments are large open spaces, with many people sitting close together in a dimly lit and quiet environment. When implementing biosensor applications, all these issues have to be considered. All these factors affect the design and deployment of a sensor infrastructure, such as reducing the sensor packet loss because of the crowd presence.
- **Algorithm development for a real-time application.** Developing a real- time algorithm for the use of physiological signals requires a deep understanding of the sensor signal patterns. At the same time, the technical platform must sufficiently support both human performances (e.g., latency) and sensor networking performance (e.g., sampling rate). Before developing an evaluation algorithm, researchers should analyze the sensor features from the actual experiments, but the algorithm development is dependent on individual environments. This results in a learning process that starts from acquiring multidisciplinary background information. Without expertise from the different backgrounds, an effective real-time algorithm is unlikely to be developed.
- **Shortage of required hardware.** There are no available commercial products that can be used directly to provide sufficient functions (crowd measurement in real time) for theater audiences. Existing commercial physiological sensors generally use the Bluetooth protocol and they focus on individual measurements of limited scope. Also, their bulky size limits them from being deployed widely in the theaters. In addition, the majority of the commercial sensors are designed for the users to be worn in a lab environment, and they have standard designs without distinguishing user profiles. The uniform design will not bring

problems to the lab experiments, as participants are encouraged to finish a task, and they generally sit with limited body movements. But in theaters, audiences naturally have many bodily movements. In addition, different audience profiles require various designs for sensor hardware. For instance, children are unlikely to wear a technical appearance sensor. Therefore, sensors for a theatrical environment need a particular design to adapt to the environment and different audience profiles.

- **Scalability and reliability of the measurement system.** A measurement system developed for a theatrical environment should be reliable and scalable. The prototypes created in a lab can show unreliability in a theatrical environment. For instance, the users usually have many movements, and they may incidentally play with the sensors and accidentally destroy some mechanical connections. Also, each theater has its own characters, such as different sizes and structures. Thus, it is desirable that the sensor system portability can be adapted to the different theaters. In summary, a high scalable and reliable sensor system can guarantee an experiment efficiently that is set up and successfully executed at the theaters.

We return to these challenges in the various chapters of this document.

## 1.3 Research Path

This research covers hardware design, software development, user studies, and the development of a feedback mechanism design (Figure 2). It has four phases. It started on hardware prototyping with three versions, with each version tested in its respective experiment. Then, heuristics were built for the final robust hardware production. In parallel, the algorithms were developed for the respective application. With all the learning process, along with the end users, the feedback mechanism was developed, where the final experiment was carried to respond to the key research question.

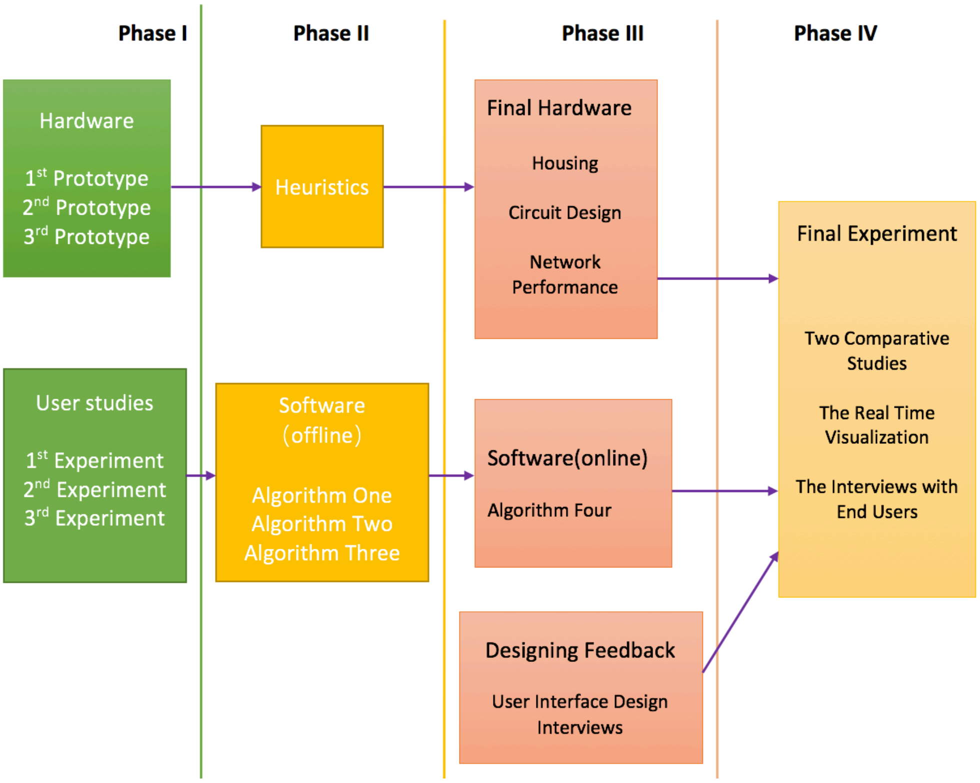


**Figure 2: The research path**

## 1.4 Research Area

The previous section cites the main challenges that can occur in a distributed theatrical environment. Our research approach seeks to solve the issues and the research methods can also be applied in some similar scenarios, e.g., distributed learning or museum studies. The main research question emerged from a distributed performance environment:

**Main Question:** *How can we support and enhance actors' awareness of remote audiences in a distributed theatrical environment?*

### 1.4.1 Hardware Design and Development

The prototyping of physiological computing systems have appeared in numerous fields, but few have made a leap from the lab studies to a widespread application in the theaters [32，64]. A theatrical environment requires a robust and scalable sensor system designed for the theatergoers at heterogeneous theatrical venues. We developed a prototype sensor system that reached the stage of the factory production and was tested in a real theater performance. These steps lead to the following research questions:

**Research Question 1:** *What are the heuristics behind building a GSR sensor system in a theatrical environment?*

Understanding the heuristics is vital for hardware design and development for a specific industrial application. The first research question deals with the process to generalize the guidelines for hardware development, particularly for a theatrical environment. It is unlikely to produce a mature hardware product for a specific artistic application without testing through an iterative process. The knowledge gained from responding this question can help the researchers to foresee some potential challenges in the theaters, even though they never had experiences to develop hardware for the theaters. They can learn from the past experiences and avoid making the same mistakes if they intend to create their own hardware for the similar field studies, e.g., distributed learning.

**Research Question 2:** *What are [the] key technical requirements of a GSR-based sensor for measuring audience engagement in a theatre?*

There are many methods to create a prototype GSR sensor circuit and design the wearing process of the sensors. Different methods can suit various applications. This research question closely looks at the technical requirement that GSR sensors should have in a theatrical environment, e.g., choice of developmental boards, wireless modules, sensor circuit design, and sensor housing design. This work has analyzed the different developmental methods to build different prototypes and explore the different wearing processes from the user's experience. The actual work was extensively tested in a lab and several field studies to identify their limitations, and then improved until it reaches the factory production. Finally, the best method to construct GSR sensors is chosen and validated in real theater performances.

**Research Question 3:** *What characteristics (e.g., features) such sensors should have?*

Theater audiences require wearable GSR sensors, particularly designed for their non-intrusive experience. This research question responds to the housing (appearance) design issues of the sensors to provide a comfortable wearing experience for various audience profiles. In particular, it involves designing different sensors for the two essential audience categories of theatergoers: adult and children. This work also showed the mature hardware product, the two different types of hardware, which is particularly designed for these two types of audiences.

### 1.4.2 Software Development

Software development mainly deals with the algorithm performance. Before developing the real-time algorithm, the extensive studies on understanding the features of sensor data were conducted, carrying out most of the data analysis offline. Also, the relationship is between the sensor data and questionnaires, and whether sensor signal patterns can be replicated in different scenarios were evaluated. Based on these fundamental frameworks, the real-time algorithm was developed, reporting its performance in an actual theatrical play.

**Research Question 4:** *Can audience engagement be inferred [in real time] from GSR data?*

This research question presents another challenge on the physiological sensors. The relationship between the sensors and subjective reports remains unclear, which makes inferring the user psychological states quite difficult. The fourth research question addresses the issue on the use of the user psychological states to interpret the sensor results. In this inferring process, the relationship between the sensor data and subjective methods is explored (e.g., questionnaires).

**Research Question 5:** *How is a live audience experience different from that with a remote audience?*

The fifth research question compares various audience experiences in a distributed environment. By using GSR sensors, the previous paradigm is extended and the different watching experience between a live audience and the remote audience is examined. The findings contribute to designing a feedback mechanism, and the results can also be useful to other similar studies, e.g., a distributed learning.

**Research Question 6:** *Are there patterns of audience engagement across productions and audiences?*

This is another unquantifiable research question on physiological computing. The methods are presented to compare the physiological patterns obtained from the different users' studies, and then get the quantified results to answer this question. With the methods, the researchers can further conduct an exploration of the physiological patterns.

**Research Question 7:** *Whether audience engagement can be inferred algorithmically from GSR data in a real-time application?*

This seventh research question deals with the development of a real-time algorithm. Different research questions or applications require various algorithms. In our case, we developed a real-time feedback mechanism, where the remote audience responses were visualized in a live venue. The developmental process heavily involved the iterations on the hardware, software development, end-user interviews, and field studies. The development of real-time algorithm was built on the knowledge gained from the different users' studies.

### 1.4.3 Feedback Mechanism

**Research Question 8:** *Can effective feedback mechanisms be developed for visualizing the remote audience responses in a distributed performance?*

The eighth research question addresses the feedback mechanism design. This work closely correlates with end user experiences (e.g., artists) to understand what kind of mechanism should be designed and developed during their performance. We show that the mechanism should be effective, i.e., they can immediately sense the remote audience responses, but not be so disruptive that it drains their attention. Extensive interviews were part of the actual work. The finalization of the feedback mechanism was tested in a large user study.

## 1.5 Contributions

This thesis presents the following contributions to the field of Human Computer Interaction (HCI):

- It presents a novel method that can be used to enhance the relationship between a remote audience and artists in a distributed performance. It uses a GSR sensor system to measure audience responses in real time, instead of traditional post-performance methods. The method can be used to infer audience engagement through rich and continuous sensor readings. At the same time, this method can be easily applied in real-time applications while other methods cannot. In particular, the real-time feedback mechanism can be explored through sensor readings. In such a manner, the missing link of the sensory feelings can be efficiently established between the artists and a remote location.
- Our work models the development of a real-time algorithm for an artistic application. The algorithm can be used to analyze and visualize audience sensory feedback. Typical sensors suffer from artefacts in theaters, which bring more problems to the data process. We overcome these difficulties and devote effort to sensor signal processing to minimize those effects. Without the knowledge of the physiological patterns, the sensor features cannot be easily extracted, and in turn, the development of a useable algorithm is impossible.

- Our work demonstrates how to design a wearable sensor for a theatrical environment. The theatergoers are a specific user group. Commercial sensors have generic designs that are unsuitable for broad audience profiles, i.e., child audiences who are unwilling to wear them while watching. Also, the majority of commercial sensors did not focus on research, e.g., the raw data were unreachable. This thesis work demonstrates an intensive iteration work on the hardware design and development, including lessons learned, and attaining a factory-ready production.
- Our work explores GSR sensors on a new and artistic application. Most physiological sensors have been extensively explored in a lab-controlled environment. However, they are unsuited for real artistic applications, especially in theaters. Each industrial sector has its own features and requirements, and thus, the related sensor product needs to be customized. Psychophysiological measurements in the wild have to overcome these challenges. In a theatrical environment, in particular, audience body movements and environmental factors may distort physiological activity. For example, individual movement can have an impact upon the GPS data, thus making the separation of engagement information from other facts difficult.

This thesis presents a detailed description of how we can develop our sensor prototype to prove its design concept, and how we test and validate it in a theatrical environment. The knowledge acquired in this research can guide others who are particularly interested in hardware design, development, and production for a specific industrial application. The study involves working closely with the people from arts groups such as *Holland Dance*, *ByBorre*, and *National Theater of China* to identify each community's needs. The research methods used are evaluated to explore the novel approaches that can provide a less biased and better understanding of the audience's engagement.

## 1.6 Thesis Outline

We have summarized the content and main contributions of each chapter below.

**Chapter 2** introduces the architecture of GSR sensor networks and the related work. This network architecture is specifically designed for a theatrical environment: robust and scalable for different theatrical venues. It gives the reasons we choose this architecture, including the specific wireless modules. Additionally, this chapter investigates the related work conducted in the past, and argues why our method is innovative, addressing the following key issues:

- Development path of the GSR sensor architecture;
- Related work.

**Chapter 3** describes our hardware design and development process. First, we detail the methodological approaches used in developing a set of heuristics, which works as the main structure for guiding the final hardware design. At the initial learning process, we prototyped the sensors by using the different techniques and tested them in respective user studies. We highlighted the limitations and advantages in each of the different prototypes, and how we learned from them and reach the stage of the final production. The contributions of this chapter, which directly addresses Research Questions 1, 2 and 3, can be summarized as:

- Introduction of motivation for building heuristics;
- Description of the developmental path;
- Sensor housing design;
- Sensor network evaluation.

This chapter extracts the information based on the following papers:

*a. C. Wang, J. Wong, X. Zhu, T. Roggla, J. Jansen, and Pablo Cesar, "Quantifying Audience Experience in the Wild: Heuristics for Developing and Deploying a Biosensor Infrastructure in Theaters," in Proceedings of the International Workshop on Quality of Multimedia Experience, (QoMEX2016), Lisbon, Portugal, June 6-8, 2016.*

*b. C. Wang, Pablo Cesar. Physiological Measurement on Students' Engagement In a Distributed Learning Environment. Proceedings of International Conference on Physiological Computing System 2015 (PhyCS 2015), Angers, France, 2015.*

c. *[Best Paper Award] C. Wang, Pablo Cesar, E. Geelhoed, I. Biscoe, P. Stenton. Sensing Audience Response - Beyond One Way Streaming of Live Performances. Proceedings of International Workshop on Interactive Content Consumption 2013, Como, Italy, 2013.*

**Chapter 4** describes the software development for this research. It focuses on how to infer user engagement by comparing sensor readings with user questionnaires and evaluates how the user experiences differed in distributed environments. This chapter also shows a quantified method by investigating whether physiological patterns can go across the different scenarios. Also, it shows how to define the significant intensity of the crowd emotions for both offline analysis and real-time applications. All the software developments are tested in different user studies and the results were reported in this chapter. This chapter addresses Research Questions 4, 5, and 6, which bring the following contributions:

- How to define audience engagement by using GSR sensors?
- How can we use GSR sensors to differentiate audience experience in a distributed environment?
- How to quantify whether GSR patterns can be replicated in different scenarios?
- How to define the significant intensity of the crowd emotions for both offline analysis and real-time applications?

This chapter is based on the following papers:

a. *C. Wang, E. Geelhoed, P. Stenton, Pablo Cesar. Sensing a live audience. Proceedings of ACM CHI Conference on Human Factors in Computing Systems 2014 (CHI 2014), Toronto, Canada, 1909–1912, 2014.*

b. *C. Wang, Pablo Cesar. Do we react in the same manner?: comparing GSR patterns across scenarios. Proceedings of Nordic Conference on Human Computer Interaction: Fun, Fast, Foundational 2014 (NordiCHI 8), 501–510, 2014.*

c. *C. Wang, and P. Cesar, "Measuring Audience Responses of Video Advertisements using Physiological Sensors," in Proceedings of the*

*International Workshop on Immersive Media Experiences (immersiveme2015), Brisbane, Australia, October 30, 2015, pp. 37-40.*

d. *[Best Paper Award] C. Wang, X. Zhu, E. Geelhoed, I. Biscoe, T. Roggla, and P. Cesar, "How Are We Connected? Measuring Audience GSR Response of Connected Performances," in Proceedings of the International Conference on Physiological Computing Systems, (PhyCS 2016), Lisbon, Portugal, July 27-28, 2016.*

**Chapter 5** shows the design method for developing a sensor feedback mechanism. Using extensive interviews, we investigated the artists' interests and perspectives when they performed in a distributed environment. The results addressed some important factors that should be considered when designing a mechanism. This chapter brings the following contributions:

- Challenges in a social co-presence;
- Performer attention to distributed liveness;
- Sensing engagement through subtle feedback;
- Representations of audiences;
- Design hybrid spaces for flexible social co-presence;
- Sense subtle feedback, convey abstractly;

This chapter is based on the following articles:

a. *C. Wang and P. Cesar, "The Play Is a Hit - But How Can You Tell?" in Proceedings of ACM Creativity and Cognition (ACM C&C 2017), Singapore, June 27-30.*

b. *[Best Paper Award] C. Wang, E. Geelhoed, and P. Cesar, "eTheatre: Connecting with the Remote Audiences," in Proceedings of the International Symposium of Chinese CHI (Chinese CHI), Guangzhou, China, June 8-9, 2017.*

c. *Andrew M. Webb, Chen Wang, AndruidKerne, and Pablo Cesar. 2016. Distributed Liveness: Understanding How New Technologies Transform Performance Experiences. In Proceedings of the 19th ACM Conference on*

*Computer-Supported Cooperative Work & Social Computing (CSCW '16). ACM, New York, NY, USA, 432-437.*

*d. T. Roggla, C. Wang, L. Perez Romero, J. Jansen, P. Cesar, "Tangible Air: An Interactive Installation for Visualising Audience Engagement," in Proceedings of ACM Creativity and Cognition (ACM C&C 2017), Singapore, June 27-30.*

**Chapter 6** demonstrates how to use our research method to answer the key research questions. We conducted the final experiment, where the interviews with artists were executed and these were used to design the feedback mechanism. The mature sensor hardware and the real-time algorithm were used to address the final issues studied in this research. We reported and reflected the pros and cons of the method, and the learned lessons can guide the following researchers to work on other mechanisms. This chapter presents the following contributions:

- Validating the research method;
- What did we learn?

**Chapter 7** provides open-ended discussions, questions and concluding remarks.

# 2

# Architecture

A scalable physiological sensor system particularly designed for theater audience provides a valuable mechanism for quantifying the engagement of audiences attending cultural events. In comparison to traditional methods, bio-sensors provide fine-grained timed data that can be used to infer the quality of the engagement of these audience members. However, there are no off-the-shelf GSR sensor system that could be deployed to the experiments in theatrical environment. As a result, this chapter demonstrates the architecture of our own GSR sensor system, particularly designed for measuring audience engagement in theaters.

Wearable systems for continuous monitoring of audience reactions could offer new avenues for theatrical experiences. They allow crowd experience to be measured in real time and provide feedback to producers, directors, and artists. If integrated, these sensor data can be manipulated locally or at distributed locations.

During the last few years, there has been a significant increase in the number and variety of wearable devices [76], such as smart watches with different sensors integrated, and they can be used to monitor user's physical activity, heartbeat, and even sleeping quality. However, there is no such off-the-shelf GSR sensor system available for theatrical environment. Besides, if such system is available, it would allow us to run several experiments to investigate whether we can us GSR sensors to infer audience engagement. profiles.

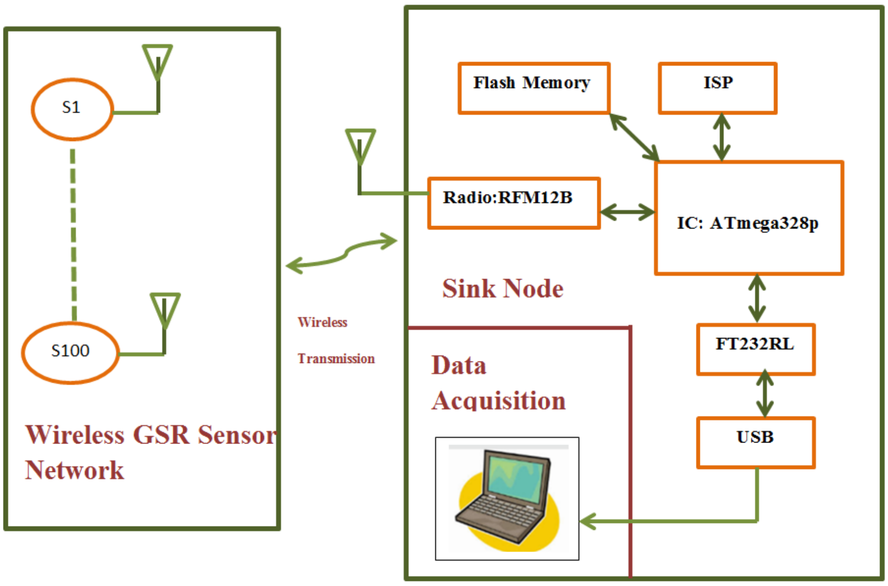


**Figure 3: Overall architecture of wireless sensor network consisting of 100 sensor nodes, sink node and laptop used as data acquisition.**

## 2.1 The Diagram of the Wireless Sensor Network

This thesis extends the experimental paradigm [36] by simultaneously measuring GSR responses of a group of participants during a live performance, aiming for envisioning a robust physiological measurement system in a theatrical environment (requiring anonymity and privacy, real-time gathering of data, support for large crowds of 30-100 people). Sensors simultaneously and independently deliver the data from multiple anonymous audience members directly to a central server that can process the data in real-time. The accompanying system offers not only offline processing of the data, but also a real-time analysis of that data, allowing for visualization or interactive installations where the (anonymous) data is aggregated, and analyzed based on the sensor id(s).

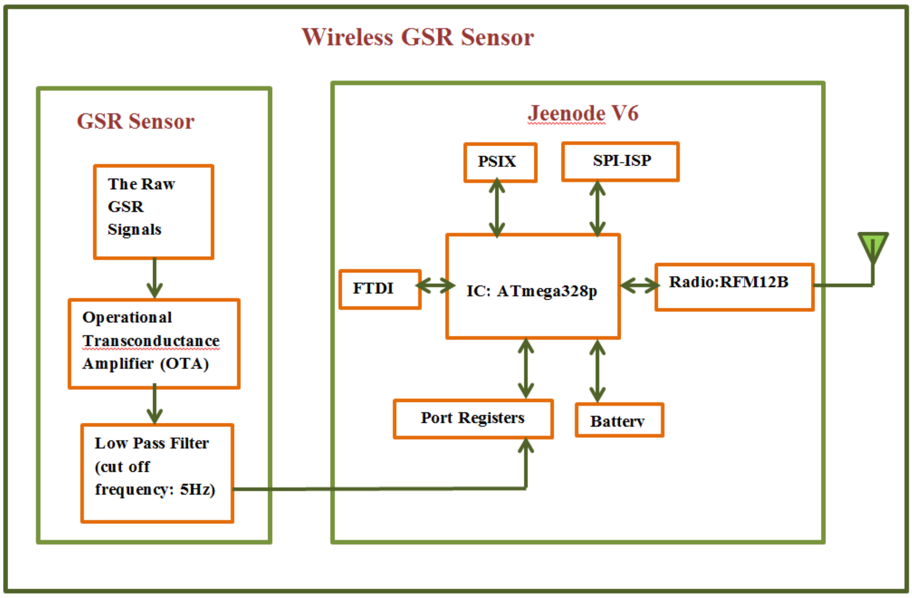


**Figure 4: The block diagram of the wireless GSR sensor**

The proposed wireless GSR sensor network for monitoring audience experience in theaters is illustrated in Figure 3 on the previous page. Each audience member wears a sensor that is strategically placed on his/her body. The primary function of the sensor node is to unobtrusively transfer the user GSR response to a sink node connected with a laptop through wireless transmission implemented using RFM12(B) protocol[1]. Once the sensor data is captured, it can be processed for both online and offline analysis.

Figure 4 displays the block diagrams of the wireless GSR sensor. The GSR sensor raw signals are captured by using electrodes, and then they are passed through an Operational Transitional Amplifier (OTA). The OTA is an amplifier capable of controlling output current based on input voltage, which can provide multiple output currents of identical potential and decrease required number of elements and power

[1] https://jeelabs.org/2011/06/09/rf12-packet-format-and-design/

consumption while applied to the analog filter and the sigma-delta analog-to-digital converter. In our case, we added a low pass filter with cut off frequency at 5Hz, because GSR signal frequency domain is 0.01 – 1 Hz. The output GSR signals are connected with port registers of Jeenode V6 (Arduino clone board). The analog readings of the GSR signals are transmitted to the sink node through RFM12(B) radio module.

## 2.2 Related Work

### 2.2.1 GSR Sensors

Electrodermal activity (EDA) is also known as skin conductance, Galvanic Skin Response (GSR), electrodermal response, skin conductance response (SCR), and skin conductance level (SCL)[2]. GSR refers to the changes in skin conductance at the surface, reflecting activity within the sympathetic axis of the autonomic nervous system (ANS). Autonomic responses in the skin, e.g., sweating, piloerection, and vasomotor changes, can thus be elicited by various emotional states via the Papez circuit in the limbic system [26]. Furthermore, it is widely recognized that increased GSR responses can be provoked by attention-related stimuli or tasks [36, 41].

GSR includes two variables. The first one is skin conductance level (SCL), indicating the slow and tonic changes measured across many discrete stimuli. The second one is electrodermal responses (EDR) related to specific stimuli, representing the quick and phasic changes imposed on shifts in tonic level in conductivity [17]. In general, there is a delay of 1-3 seconds between stimulus and SCR onset. Hands and feet can be used to measure GSR, as there the density of the sweat glands is the highest.

Commercial companies, such as *BioPac*, *Thought Technology* and *Q Sensors* offer this type of GSR sensor at a high price. Although such commercial sensors [1, 2] allow researchers to start experimenting immediately, they do not provide functions to measure groups of users simultaneously. The reason is that the communication

[2] https://en.wikipedia.org/wiki/Electrodermal_activity

protocol normally is Bluetooth, which has limitations on connecting cell nodes in wireless network, e.g., a master and up to 7 slave pico-net networks, which cannot support a simultaneous measurement with a large scale of audience members (e.g., 100 members).

A number of methods have been employed to gauge off-line audience feedback to performing arts, such as surveys collected after a performance.  GSR measures the excitation of the sympathetic nervous system, and it has been used to estimate audience engagement to video recordings of performances [10, 39]. Some studies apply interactive technologies to enhance the audience experience [49, 45, 23].

### 2.2.2 Physiological Sensor Networks

The current available personal sensor networks (PSNs) are mainly developed and designed for lab [1, 2] and home studies (e.g., monitoring elderly people) [18, 25]. In most cases, they communicate via Bluetooth, which is an excellent solution for measuring individuals [12, 42]. Besides, some studies [17, 49,14] developed a PSN by using open source alternatives (e.g., Arduino, connected with ZigBee/Xbee or WiFi), which have been used in the several use cases [15, 20]. Recently, with the advent of miniature electronics, radio technology provides alternative frequency bands to construct a wireless network [32, 31, 49]. Interestingly, most of these products were used for healthcare applications (i.e., monitoring the heart rate of people [20, 34, 36]).

### 2.2.3 Audience Responses

There are different definitions for audience response, depending on the author. Different research areas apply different synonyms such as audience engagement, audience feedback, audience interests, audience interaction and audience (or user) experience. Christopher Peters et al. [23] described the audience response as a combination of focus, interest, perception, cognition, experience and action. Accordingly, Heather L. O’Brien et al. [33] pointed out that audience engagement (response) could facilitate users with more enriching interactions in computer applications: engaged users tend to recommend the products (or service) to others.

Furthermore,  an affective computing users' emotional response [24] is obtained as an evaluation tool to define user engagement. In game application studies [26], audience engagement (response) refers to players' state of awareness and synchronization. Audience biofeedback, e.g., arousal, is also used as an indicator of the levels of players' engagement [22].

Audience response can be measured in two ways: explicitly and implicitly. Explicit methods normally require users' intentional inputs, for instance surveys or ratings, whereas implicit measurements generally obtain audience feedback by applying physiological sensors. During the process of data collection, sensor readings are acquired in real time without interrupting audience in their watching experience.

Visual analysis methods, e.g., eye movements, are also included into the scope of implicit methods of measuring audience response. In such studies, eye-gaze, eye movements, and head movement trackers are installed to define the users' interest in a video or other applications, e.g., game studies. Some other methods, e.g., physiological sensors, were also combined with visual analysis methods to obtain audience response data in order to characterize an audience state, e.g., boredom or fatigue. However, these studies could not avoid requiring intentional audience annotations constantly, in which they were used to label the attributes of the data [5].

Physiological sensors are the most common implicit methods to measure audience response. In 1995, Peter J. Lang combined GSR sensors, user ratings, and successful defined audience's emotions in the two-dimensional spaces: valence and arousal [32]. Such combination measurement tools (sensors and questionnaires) were also used in the studies to explore audience engagement in personal creative experiences [19], user experience in entertainment technologies [37, 54] and audience participation in movie viewing [47]. Some of the studies deployed more than one sensor in order to secure multi-physiological signals from the users in order to improve the accuracy of the algorithm prediction.

More generally, GSR data are good indicators for people's mental state and emotions in different application areas. Lin et al. successfully correlated galvanic skin response data with task performance in video game playing [47] and investigated fluctuations of GSR data during a 3D movie experience [46].

Nourbakhsh et al. [47] related GSR data to measures of cognitive load and emotions in both reading and arithmetic tasks.

### 2.2.4 Applications by Using Audience Responses

There are many quantitative studies executed in game research on the reliability and suitability of physiological sensors. Pejam et al. [60] have explored the possibilities of improving game design by providing user biofeedback; their results showed that combining user GSR feedback would help designers choose a proper design strategy resulting in high game play quality. Game technology companies, such as Sony, recently added a GSR sensor in their new game controller *DualShock 4*, where users' GSR-response is detected on interest level towards the game. Another example [22] is that users' GSR-response are used as the users' agitation in a game, and the more often a user feels agitated, the more enemies are produced in a game.

Previous studies have also shown that GSR sensor is one of the indicators on users' cognitive and emotional states. Lin et al. [52] successfully measured audience GSR-response in different movie sessions. In particular, the fluctuations of GSR data were linked to events during a 3D movie experience.

Matthew et al. [45] demonstrated a novel interaction by using GSR sensors in an audio stream bookmarking, where users' GSR-response was monitored as a response to external interruptions. GSR sensors were also applied in a wearable system in order to help users select the high arousal photos, which were most relevant to the users' ordinary daily life [25].

Web applications also benefit from audience response research. In the paper [14], users' biofeedback is used to distinguish the audience response within the different age scopes, in terms of a web 2.0 application. Audience affective states were also employed to investigate what kind of interaction technique on the web has a significant impact on users [15]. Transforming audience physiological signals (audience response) into a smile icon was implemented in an online chat [26]. Users' biofeedback on preference, e.g., like or dislike, was used as the input data in order to improve the accuracy of the online recommendation system [25, 79].

Audience response also plays an important role in some other applications. For instance, Olympic Games, audience clapping frequency was visualized on the display screen to encourage athletes' performance [29]; audience 'cheering meter" was measured to aid voting at rap competitions [52]. In sum, measuring audience response by using GSR sensors has been widely applied in many applications.

### 2.2.5 Audience Research in Performing Arts

Researchers have developed numerous ways to measure, interrogate and assess audience response to arts and culture, including biometric measurements, post-event surveying, qualitative post-event research, and longitudinal, or retrospective studies. Biometric measurements objectively manifest audience response to the aesthetic experience [14, 43, 48], even though interpreting such responses is difficult. Nonetheless, knowledge of audience bio-response will be of considerable value in advancing our conceptual understanding of impact on the individual [8, 15]. Post-event surveys have been widely applied to evaluate the short-term effects of specific cultural events [7, 9], and they can only capture the conscious experience of respondents. There is limited comparability across events, and they fail to capture effects that unfold over time. Qualitative methods allow informants to reflect on the areas that are most significant to them. Unlike surveys, qualitative studies can help researchers contextualize numerical data. Longitudinal impacts on audience studies add durable value to cultural experiences [40, 53]. While the retrospective identification of cultural events may not be helpful to assess the impacts of specific artistic works, it can inform us how cultural participation plays an important role within the large scope of people's lives [8, 65]. Therefore, GSR methods allow us to record audience both conscious and subconscious response towards performances, and the continuous sensor readings provide us an overall evaluation on audience engagement.

# 3

# Hardware Design and Development

As part of our research, we built heuristics to guide the design and the development of our hardware. Especially for a theatrical environment, the robust and the reliability of the hardware should be guaranteed, because audience naturally has many body movements and the environmental restrictions can have strong influence on the performance of the hardware. This chapter discusses our findings and development.

Designing a wearable sensor hardware requires user participation. With user involvement in the design phase, we can learn in which manner the hardware can give a non-intrusive wearing experience. In particular, in the early stage, designers, engineers, and users should work together and learn the limitations of different design concept and prototypes. Especially in a theatrical environment, the design and the development of a wearable could be different compared to other hardware design (e.g., lab experiments). In theatrical environment, the signal can be blocked by the presence of an audience, and the audience members may move a lot during a play. All these issues we need to take into account for the hardware design and development.

This chapter considers the first, the second and the third research questions:

**Research Question 1**: *What are the heuristics behind the building a GSR sensor system in a theatrical environment?*

**Research Question 2**: *What are [key] technical requirements of a GSR-based sensor for measuring audience engagement in a theater?*

**Research Question 3:** *What characteristics (e.g., features) such sensors should have?*

## 3.1 Heuristics for Hardware Design and Development

We describe the developmental path of building and deploying the biosensor infrastructure in the theaters by following heuristics. Heuristics are like mental shortcuts that enable people to make quick judgments and handle problems. The set of heuristics should provide a wide variety of perspectives on usability and be as complete as possible at explaining usability issues that occur in an actual environment [77]. There are many approaches in creating heuristics. Molich and Nielsen [44, 46] with their heuristics based on personal experiences and years of experience have gathered from teaching and consulting others on usability engineering.

A more structured approach was taken by Dykstra [27] in creating a set of heuristics in five steps: 1) listing all usability problems for each program and each participant using competitive analysis; 2) consolidating all the problems for each program; 3) categorizing the problems to be solved; 4) deleting duplicate problems and combining problems into fewer categories; and 5) developing the final heuristics. Similar to Dykstra's approach (but simplified), a heuristics version was developed by Madgunda, et al. [41]. They used a three-step approach: 1) identify problems from reviews of different users; 2) assign categories to groups of problems; and 3) develop the heuristics.

Firstly, we adopted a similar process as the one presented by Madgunda and colleagues, the *Convergent-Divergent model*, consisting of five steps. The first step was collecting requirements by discussing with various users including theater producers, performers, and audiences to understand the requirements to meet for deploying a biosensor infrastructure to measure audience response in a theater. The second step was the feasibility study in which we analyzed the requirements and took steps to develop a prototype. Based on the requirements, we developed prototypes for a biosensor infrastructure. Consequently, the prototype was tested in the wild and evaluations were made to improve the next prototype. Due to the vast differences in performances and environments, the second and third steps were repeated in an iterative process following an agile model. The fourth and fifth steps were validating the requirements and documenting them for later developmental stages.

An iterative methodology similar to an agile model was employed to accomplish steps two and three of the *Convergent-Divergent model*, resulting in four consecutive experiments. The experiments are described in the next section as the developmental path. Finally, putting together the lessons learned from the four experiments and our personal experiences as experimenters, we used the heuristics methodology to develop valuable knowledge to guide future work. Each of the experimenters identified problems encountered during the experiments individually.

Next, the experimenters discussed and eliminated problems that were already mentioned. The remaining problems were clustered and categories were assigned to each cluster.

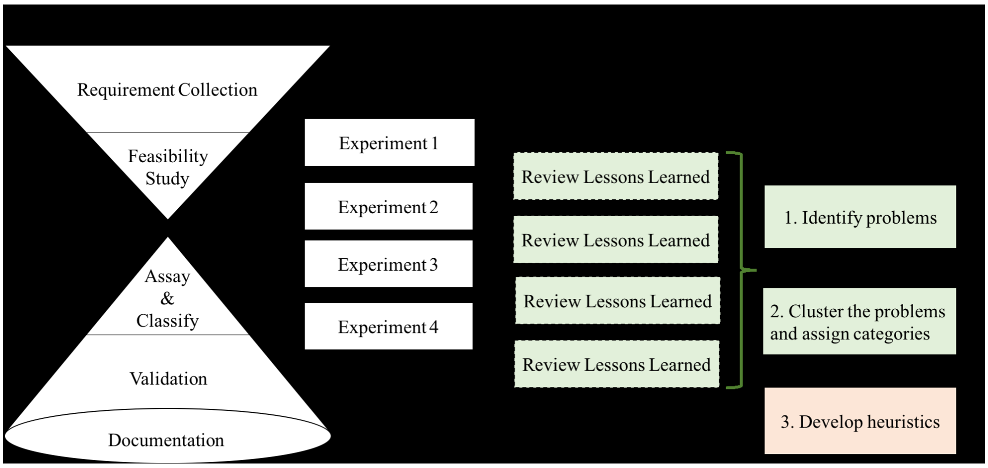


**Figure 5: Methodologies adopted in our research**

### 3.1.1 Developmental Path

In this section, we describe the developmental path of building and deploying the biosensor infrastructure in the theaters as shown in Figure 5. An iterative method was used where each experiment resulted in a set of lessons learned that were used to advance the design of the biosensor infrastructure in the following experiment. Concurrently, this methodology completes the two sequential steps two and three in the *Convergent-Divergent methodology*. The requirements and constraints that were initially gathered through discussions with end users (i.e., theater companies, producers, artists, audience) as well as issues from related work were analyzed and evaluated for feasibility (i.e., Step 2: Feasibility Study). The first prototype was developed to test whether the expected outcomes are achieved; problems encountered were later identified and categorized (i.e., Step 3: Assay and Classify).

Due to the many requirements of different nature and the complexity of conducting experiments in the wild, where constraints differ across theaters, four subsequent experiments were carried out to test whether the biosensor infrastructure deployed met the needs of the different end users. Through the four experiments, we gained valuable experiences that are translated into heuristics.

The initial requirements were to develop a biosensor infrastructure that can be used to measure audience response and to develop the mechanism to analyze the responses for understanding audience experience at different points in time during a performance. The first generation of our sensors consisted of one *Arduino UNO* board and one *Xbee* wireless module (for every five users), and a noise filter. Five sensors were connected to one main module. As such, we had to cluster audience members in groups of five. Biofeedback of 15 audience members was measured while they watched a play that was specially choreographed to elicit audience response. Based on the measurements that were synchronized with the video recordings of the performance [80, 81], it was concluded that GSR highly reflects audience response [15, 19].

From the first experiment, we learned that the data collected should preserve the unique characteristics of the signal we are collecting. There are much sources of artefacts that can influence the physiological signals (e.g., types of electrodes and placement of the electrodes). Being aware of these different sources of artefacts would influence the processing the data and their validity. Furthermore, the first generation of sensors was cumbersome to wear, and it may affect audience watching experience.

A second generation of sensors was developed to address some of these issues. We worked on scaling up the system with small form-factor developments in wireless technology and GSR measurements. The Arduino UNO board was changed to a *Jeenode* board so that the infrastructure could support up to 250 different groups. In other words, a larger group could be measured simultaneously. The Jeenode board also works at different frequencies and uses the *RFM12B* radio module that can operate at long distances with rather low power consumption. By putting together our own biosensor infrastructure, we can experiment with different solutions to create an adequate infrastructure.

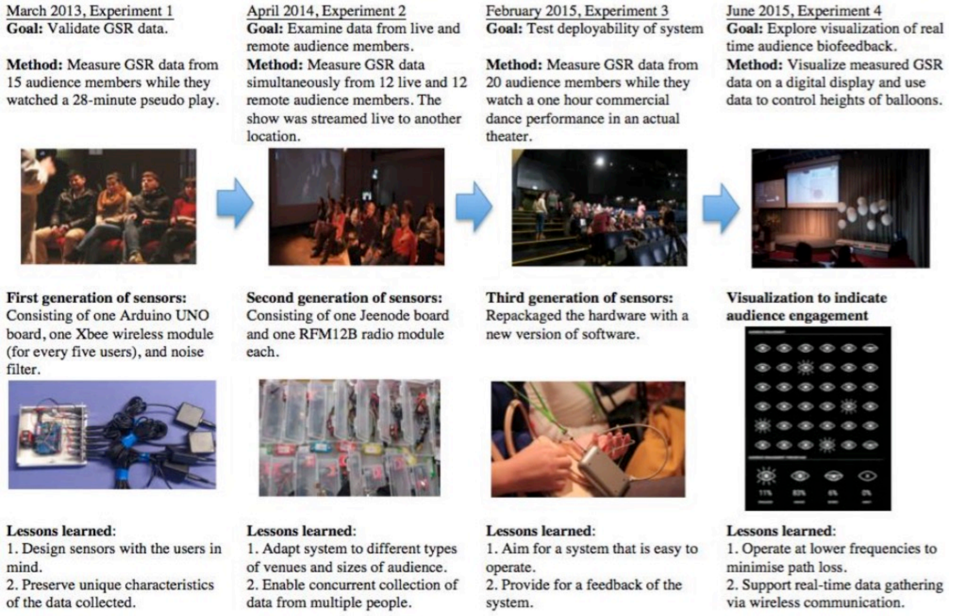
The improved sensors were tested in a second experiment intended to investigate the different responses of local and remote audience members attending a live theater play. At each location, each audience member wore an individual wireless sensor. The wireless sensors have the capability to simultaneously and independently send signals to a central server. The results showed that the new version was better than the first generation, because they are designed as a wearable prototype.

Although the hardware design of the second-generation of sensors made it more convenient for audience members to wear and we were able to simultaneously collect data from a larger group of audience and in different venues, we discovered that we had to improve the quality of the collected data.

Subsequently, a third generation of sensors was produced. The sensors were tested in a third experiment with 20 audience members watching a live one-hour commercial dance performance. A more advanced algorithm was adapted from Fleureau et al. [29] to process the data and to find significant SCR points in audience response. The network performance was also improved by setting up a transmission rate that maximized the capacity of the network.

The fourth generation of sensors were built ready for production. It has two versions: one type is for adults, and the other type is for children to wear. With this version sensors, we run the fourth experiment. We developed software that processed the captured data in real-time. During this experiment, 36 people wore the fourth generation of sensors that collected their physiological responses while they watched a artistic performance in a distributed environment. The physiological responses were manipulated to control the lighting condition. The information provided the presenter with an indication of audience engagement.

Steps from these previous studies were adopted and similar to [13, 78], we derived a set of heuristics in three steps: 1) identify problems through interviews with users, personal experiences, and lessons learnt from our experiments, 2) organize problems into categories by eliminating redundant problems and clustering similar problems into categories, and 3) develop heuristics. A total of seven heuristics had been derived and they are further categorized into three main areas: a) processes, b) data, and c) system.



March 2013, Experiment 1
**Goal:** Validate GSR data.

**Method:** Measure GSR data from 15 audience members while they watched a 28-minute pseudo play.

**First generation of sensors:** Consisting of one Arduino UNO board, one Xbee wireless module (for every five users), and noise filter.

**Lessons learned**:
1. Design sensors with the users in mind.
2. Preserve unique characteristics of the data collected.

April 2014, Experiment 2
**Goal:** Examine data from live and remote audience members.
**Method:** Measure GSR data simultaneously from 12 live and 12 remote audience members. The show was streamed live to another location.

**Second generation of sensors:** Consisting of one Jeenode board and one RFM12B radio module each.

**Lessons learned**:
1. Adapt system to different types of venues and sizes of audience.
2. Enable concurrent collection of data from multiple people.

February 2015, Experiment 3
**Goal:** Test deployability of system

**Method:** Measure GSR data from 20 audience members while they watch a one hour commercial dance performance in an actual theater.

**Third generation of sensors:** Repackaged the hardware with a new version of software.

**Lessons learned**:
1. Aim for a system that is easy to operate.
2. Provide for a feedback of the system.

June 2015, Experiment 4
**Goal:** Explore visualization of real time audience biofeedback.
**Method:** Visualize measured GSR data on a digital display and use data to control heights of balloons.

**Visualization to indicate audience engagement**

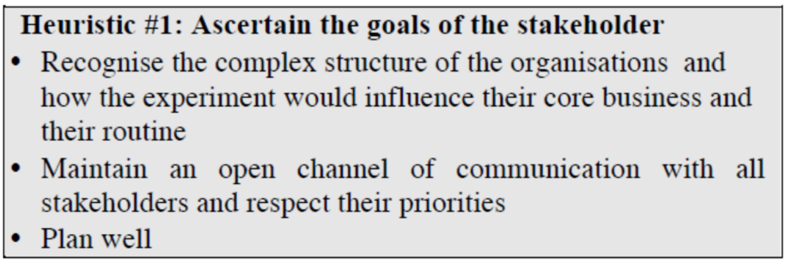
**Lessons learned**:
1. Operate at lower frequencies to minimise path loss.
2. Support real-time data gathering via wireless communication.

**Figure 6: Developmental path of building and deploying the biosensor infrastructure for audience research in the wild.**

### 3.1.2 Heuristics

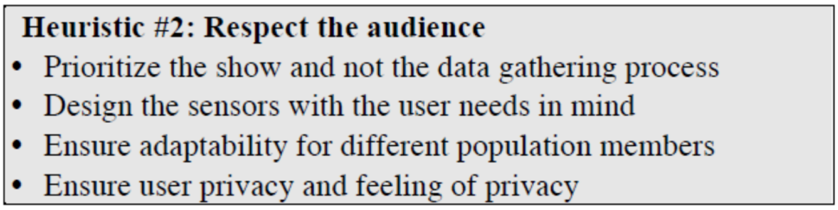
The aim of our heuristics is to guide future work and provide a foundation for an adequate infrastructure to measure audience experience in realistic conditions. We believe that an adequate infrastructure should cover at least three main areas (i.e., processes, data, and system) (Figure 6). Processes put the needs of the stakeholders and users before the needs of the audience research. It is pertinent to have a grasp of these needs so that the ultimate benefits of audience research can be realized. The underlying principle of Heuristic #1 is that the experiment does not cause inconvenience or becomes imposing because of a lack of understanding with the stakeholders. From our experiments, we learned that there are many professionals and each one in charge of a number of tasks (e.g., lighting, scripting…). While producers and directors may be interested in quantifying audience experience, they may not have the extra time and cycles for attending to researchers. To avoid unnecessary backlash, extra planning, improved communication, and good organograms would be advantageous apart from the basic guidelines for conducting experiments (e.g., consent forms, questionnaires, questions for the interviews). Other than stakeholders, audiences' needs should not be neglected which is the underlying principle of Heuristic #2. The implementation of a biosensor infrastructure should not affect theater-going experience in any way. We listed some benefits of enjoyable audience experience in our introduction and such benefits can only be obtained if audience needs are placed at the core of the design of the biosensor infrastructure.

*A) Processes:*



**Heuristic #1: Ascertain the goals of the stakeholder**

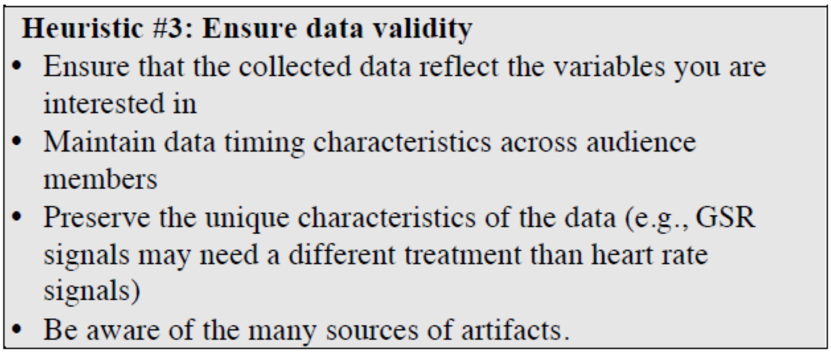
- Recognise the complex structure of the organisations and how the experiment would influence their core business and their routine
- Maintain an open channel of communication with all stakeholders and respect their priorities
- Plan well



**Heuristic #2: Respect the audience**

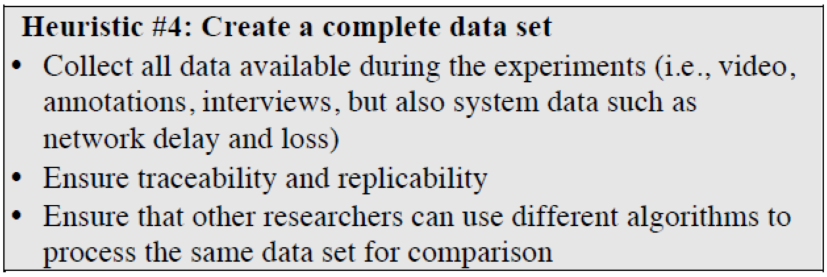
- Prioritize the show and not the data gathering process
- Design the sensors with the user needs in mind
- Ensure adaptability for different population members
- Ensure user privacy and feeling of privacy

*B) Data*



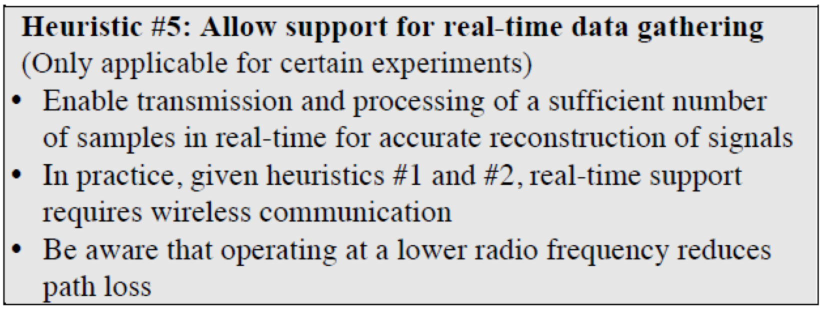
**Heuristic #3: Ensure data validity**

- Ensure that the collected data reflect the variables you are interested in
- Maintain data timing characteristics across audience members
- Preserve the unique characteristics of the data (e.g., GSR signals may need a different treatment than heart rate signals)
- Be aware of the many sources of artifacts.



**Heuristic #4: Create a complete data set**

- Collect all data available during the experiments (i.e., video, annotations, interviews, but also system data such as network delay and loss)
- Ensure traceability and replicability
- Ensure that other researchers can use different algorithms to process the same data set for comparison



**Heuristic #5: Allow support for real-time data gathering**

(Only applicable for certain experiments)

- Enable transmission and processing of a sufficient number of samples in real-time for accurate reconstruction of signals
- In practice, given heuristics #1 and #2, real-time support requires wireless communication
- Be aware that operating at a lower radio frequency reduces path loss

*C. Systems*

**Heuristic #6: Enable concurrency and scalability**
- Allow gathering of data from multiple people at the same time
- Allow adaptability to sizes of venues and audiences
- Do trial runs to ensure that the expected scale is met

**Heuristic #7: Aim at deployability and provide feedback**
- Be aware of the restrictions (e.g. different locations, audiences, performances…)
- Allow for easy installation and deployment of the system
- Allow for the system to provide feedback to the experimenters about its current operational state (e.g., faulty sensors), and the gathered data

Heuristics #3, #4 and #5 are related to data. A working knowledge of the unique characteristics of the data measured (e.g., GSR) is important so that these characteristics are preserved from the collection to processing and analysis. One should be prepared to cater to the many sources of bad or missing sensor readings due to poor sensor leads, sensor gels, packet loss during radio transmission or other sources. Without data validity, the experiment would be conducted in vain. Therefore, Heuristic #3 serves as a guideline for achieving valid data sets while Heuristic #4 addresses the importance of keeping data. Other than validity, having a complete set of data is helpful. A complete set of data refers to all forms of data collected during the experiment (i.e., video, interviews). These forms of data can help to verify findings and form a more complete picture of audience experience. Note that collecting system data is also important: future experiments with different data processing algorithms may well depend on having recorded detailed timing or packet loss information during the real experiment.

Although visualization may not be a requirement for all experiments, Heuristic #5 was developed to guide the real-time of data gathering. As mentioned in the introduction, visualization can enhance the feeling of shared experience and visualization requires real-time. As it would be difficult to see how running wires to individual audience members would fit heuristics #1 and #2, we need some form of wireless communication. Through our experience, wireless communication was effective and efficient in meeting the demands of real-time data gathering.

Design principles of a robust system are embodied in Heuristics #6 and #7. The emphasis of Heuristic #6 is concurrency and scalability. Individuals may have different experiences during the same performance. To effectively understand the effect of a performance on different audience members, concurrently measuring audience experience becomes a necessity. Having conducted the experiments in different venues with different sizes of audience, we appreciated a system that is adaptable to these differences. Trial-runs would help to ensure the scale of the actual experiment that could be achieved. Last but not least, Heuristic #7 accounts for the deployment of the biosensor infrastructure. There is only a short time window for experimenters to attach the sensors to audience members since they usually arrive at a performance venue close to the start of a performance. As a result, a system that provides feedback for experimenters to act upon and a system that can be easily deployed are crucial to the success of the experiment.

## 3.2 Three Versions of Hardware

This work developed the three versions hardware, which was guided by the heuristics. Each prototype was tested in the respective experiment, and the user experience and the sensor performance were analyzed. The limitations were improved in the next version.

### 3.2.1 The 1st version hardware

For a variety of practical reasons, GSR sensors with wireless communication modules are more applicable on group user studies. Many sensor nodes communicate with a sink node at the same time, which means that the experiment can measure audience members at a large scale. The first generation of the GSR sensors were constructed with Arduino boards, Xbee wireless transistors, and filters. The filters built with resistors and capacitors are used to reduce noise interference in a wireless channel. The communication range (10-20 meters) for this type of GSR sensor could be used for an experiment conducted in a free space (Figure 7). In addition, since the experiment was controlled to occur at one time, the random effects in terms of experimental settings cold be reduced to a minimum level, and in turn the gathered GSR data could report the most realistic and reliable results on audience response.

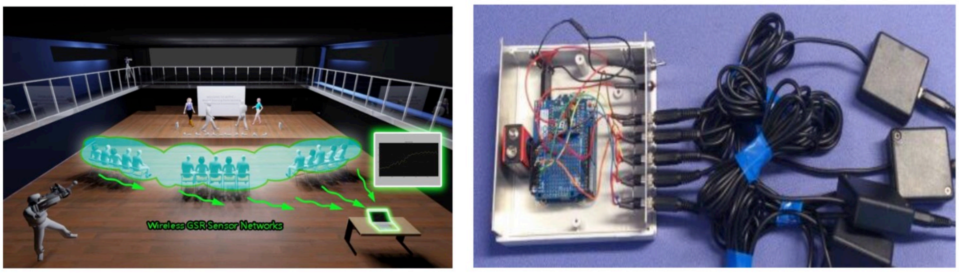


**Figure 7 : The first version hardware: Wireless GSR sensor network reporting to a centralized server (left). Individual GSR sensors (right)**

The measurement system consisted of 15 GSR sensors, 3 Arduino boards and 4 Xbee RF modules. Each Arduino board (sample rate is 1Hz) carrying 5 GSR sensors uses one Xbee module to send packets to the Xbee coordinator, which connects to a laptop server (Figure 7). The Xbee coordinator applies a polling scheme to receive packets from the different Arduino boards in order to minimize packet loss ratio in a wireless communication channel. Four cameras recorded the audience and the performance. By matching the video streams with the GSR data (time stamp method), we were able to use the video footage to link the audience GSR readings to the events during the performance. It is well known that GSR readings naturally have 3-6 second delay, and we also considered this when we linked GSR data to the performance events.

Before we conducted the real experiment, we established independent test and pilot studies (more than 50 users) to validate the usability of the new sensors in the different scenarios, i.e., video games and movie clips watching. First, we tested the function of noise proof of the sensors without users connected (no noise in our case after several hours testing). After that, we asked users wear the sensors, in order to observe the performance of the sensors in the different scenarios. In all scenarios, the distribution of the results demonstrated obvious linear patterns. The performance of phase skin conductance was also tested: when users faced sudden sounds, "spikes" appeared in the sensor readings.

The purpose of setting the sample rate 1Hz was due to the consideration in terms of the experimental design, by following the guiding from [30]: "the selection of the correct sample rate depends on the characteristics of both the raw signal and the translated signal". We first executed a pilot study in the Miracle Theatre to test whether 1Hz sample rate is sufficient with respect to data capturing and data analysis. After 30 minutes (1800 data points for each volunteer), we found that the sensors performed well after plotting all the sensor data. Secondly, in terms of the methodologies we applied in our studies, we consider 1680 data points for each user (28minutes play) as adequate to run our analysis algorithm. In addition, after receiving the advice from the published results, we realized that it would be better to increase the sample rate in terms of reliability and the process of the sensor data, but it should not influence the validation of the results reported here.

There were some limitations found in the first generation of the sensors. First, the sensors were not wearable. Especially, the electrodes were not comfortable for users to wear for a long time. Secondly, the communication range (10-20 meters) was rather short for a system to be used in a theatrical environment. Last, the battery consumption was fairly high (the battery was exhausted after roughly half an hour of use).

### 3.2.2 The 2nd version hardware

The second version hardware improved the performance and overcame the limitations found in the first generation of hardware. This version had reached the wearable stage, so that each user had his/her own independent, tiny sensor to wear while watching a performance. Furthermore, the sensor battery consumption was rather low, and the communication range (10-20 meters) could reach 100 meters in a free space and 40 meters in a crowd environment.

We choose Jeenode as our development board, as it is small, economic, and worked at 915, 868, and 433MHz. Jeenode was an open source Arduino clone with an Atmel 8- bit RISC microprocessor and an RFM 12B radio module (Figure 8). The major difference between Jeenode and Arduino is that the Jeeode had an integrated wireless module. Thus, it was easy to establish a wireless network by using Jeenode, while Arduino has to work together with separate wireless modules in order to establish a wireless network communication. As a result, the system integrated with Jeenode was much smaller and economical than the one using Arduino and other wireless modules (R6, R7). The standard Lithium Polymer battery (1100 mAh, 3.7 voltage) supported Jeenode attached with one GSR sensor (sampling rate: 10 Hz) to work for over 50 hours. Besides, Jeenode supported up to 250 different groups, each with up to 30 different node ID's. This meant that large crowds could be monitored simultaneously. Within each group, one sink node only communicated with the sensors from the same group, ignoring all the other incoming packets.

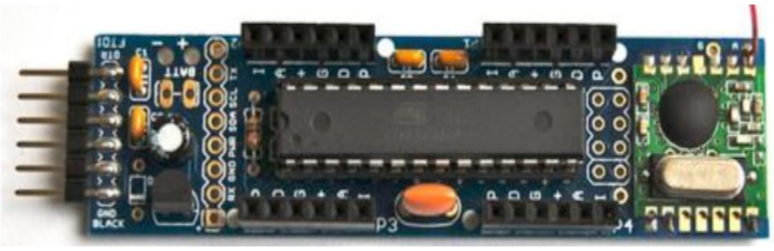


**Figure 8: The Jeenode board**

The integrated wireless module RFM12(B) on Jeenode worked on the ISM-band at the 433, 868, or 915MHz. Different countries have different regulations to use the three different frequencies. For instance, 915MHz can be used in North America, Australia and Asia; 433MHz is available in many countries; while 868MHz is for Europe. Since our experiments were located in Europe, we choose the 868MHz Jeenode version.

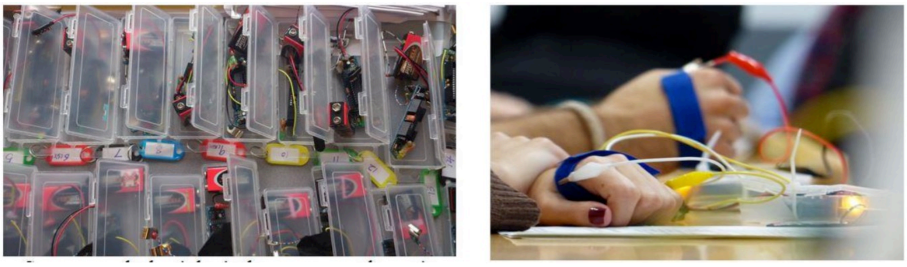


Figure 9: The second version GSR sensors

The second version still had some disadvantages in terms of audience usage in a theatrical environment (Figure 9). First, users did not feel comfortable in terms of sensor housing and the electrodes. Second, the soldering work in the lab did not yield reliability when sensors used in reality. In particular, some mechanical wires in the system design created some connection problems in the experiments.

Last, sensor housing did not show a professional image to users. If we want to use such sensors in a real theatrical performing environment, the sensors needed to have a professional look.

### 3.2.3 The 3rd version hardware

We improved the circuit design and produced the sensors through factory production in the third version. We designed the sensor circuit and then asked the factory to produce the PCB (printed circuit board) board of the sensors. In addition, we could control the size and the weight of the integrated product and made it more convenient and comfortable for users to wear. Surprisingly, the factory production turned out to be cheaper than producing these outselves. Each GSR pcb board cost $2.50, including an appropriate interface which can easily be connected to the developmental board Jeenode (Figure 13). Instead of buying a commercial sensor, the factory produced GSR sensors ensured the quality is robustness and reliable, which could be further integrated into a professional housing.

Inspired by the construction of the GSR sensor from *Bitalino*, we built our own GSR sensors by using an operational trans-conductance amplifier (OTA) and a low-pass filter (LPF). One of the functions of the OTA is to increase the amplitude of the weak potential differences generated from the biological electric signals. In addition, the OTA is ideal for measuring signals from low level output transducers in noisy environments, which amplifies the difference between two input signal voltages while rejecting any signals that are common to both input terminals. This is known as Common-Mode Rejection (CMR). The CMR function cancels common-mode signals from both alternating current (AC) and direct current (DC), reducing the undesired source errors that are difficult to be removed afterwards. As we know, the frequency for the EDA sensors is between 0.01 – 1Hz. Thus, after the OTA, we applied a second order low pass Butterworth filter (Gain = 2; fc (cut-off frequency) = 5Hz) to filter out the frequencies that lie outside the desired range.

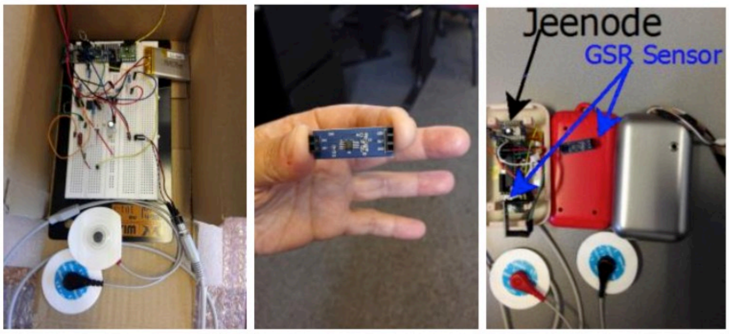


**Figure 10: The third version GSR sensors (adult version)**

We first prototyped the circuit on a breadboard and connected it to the Jeenode (Figure 10: left). After that, we produced a GSR board to obtain a compact GSR sensor with the interfaces needed by Jeenode (Figure 10: middle). The Jeenode board and GSR sensor are connected and mounted in a customized 3d printed box (Figure 10: right). Two electronodes are connected via a cable. We used only one analog port of the Jeenode, so three remaining ports could be used to connect other types of sensors (e.g., Electrocardiogram (ECG)).

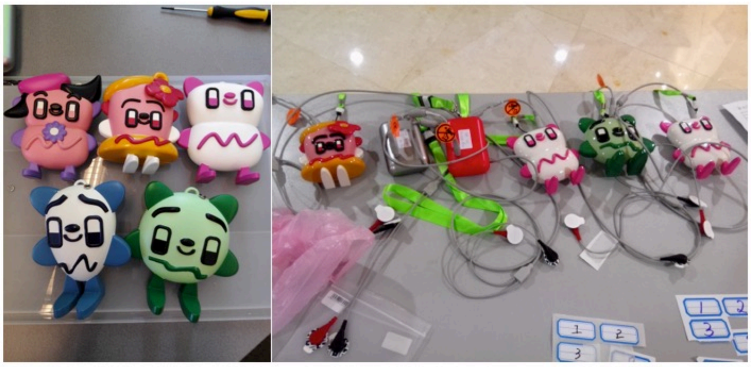
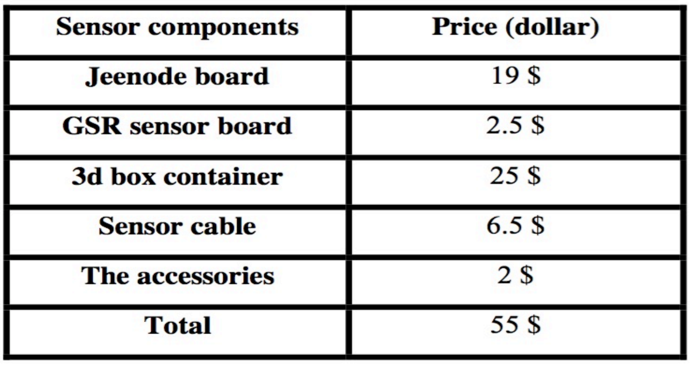


**Figure 11: The third version GSR sensors (child version)**

In this version, we also took into account of the audience profile. We produced two versions of the sensors: one is for adult, and another one is for children (Figure 11). Our sensors were rather small, and each sensor fit in a box of 10cm*5cm*2cm. At 50g each, they were lightweight. The size and the weight of our sensors make them well suited for audience to wear, and feedback from the commercial dance performance indicates they are non- intrusive. Besides, the prices were based on producing a small number of them: 300 pieces. Establishing a GSR sensor network with 20 nodes would cost around $1100, which is below the commercial price (Figure 12). Thus, theatres could easily incorporate them into their basic equipment.

In order to validate our sensors, we compared them with the commercial GSR sensor from *Bitalino* in a cognitive task. We used the two sensors at the same time, in the same experiment, for the same person. After that, the Spearman correlation coefficient method was used to calculate the correlations between the two GSR sensors. The correlation method validated not only the data distribution, but also the phasic changes between the two sensors. All the tests had a sampling rate of 10Hz and 100Hz.



| Sensor components | Price (dollar) |
|---|---|
| Jeenode board | 19 $ |
| GSR sensor board | 2.5 $ |
| 3d box container | 25 $ |
| Sensor cable | 6.5 $ |
| The accessories | 2 $ |
| Total | 55 $ |

**Figure 12: The sensor cost**

We chose the video game “Spade A” as the cognitive task to validate our GSR sensor (Figure 13: left). The experiment was divided into three phases. At the beginning, the participants listened to 5 minutes mediation music. When this meditation phase was finished, the participants played the video game, called “Spade A”. In the Spade A game, users are asked to indicate the correct position of the Spade A. In each round the playing cards changed positions, and the users have to pay attention to recognizing in which position the Spade A ends. When the user ticks out one of the cards, the screen displays the correct answer. If the result is positive, the user obtains a point, otherwise the user does not gain a point. There was a counter displayed at the top left side of the screen, and the game duration was 96 seconds. When the game finished, the participants went again for a phase of 4 minutes of meditation music. In this way, we could observe the sensor performance during the meditation, the cognitive task, and the recovering phase.

During the validation procedure, the participants wore our GSR sensors in the index finger and the middle finger of the left hand, and another commercial GSR sensor was attached at the bottom of the index finger and the ring finger from the same hand. All the users wore the headphone during the test, and all of them used the right hand to interact with the video game.

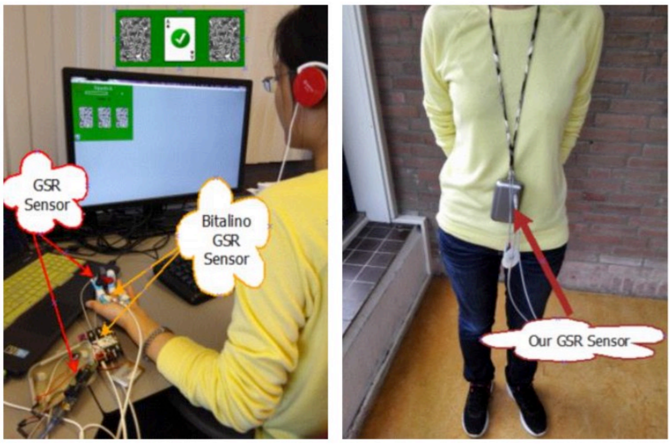
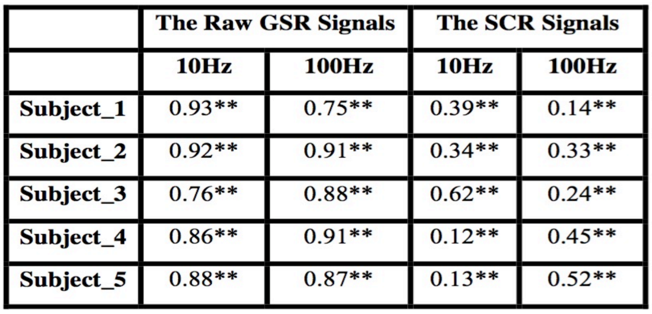


**Figure 13: The *Spade A* video game was used to validate our GSR sensors (left); One of the users wearing the sensor around her neck (right).**

The results from the spade A game showed a strong correlation in terms of performance between our GSR sensors and the commercial ones (Table 1). During the meditation procedure, the arousals of all the users gradually decreased. When the video game started, we could see a substantial increase in the GSR data distribution.

**Table 1: The Experimental Correlation Results (r**: r is correlation result, and ** indicates that p value is significant at the 0.01 level)**



| | The Raw GSR Signals | | The SCR Signals | |
|---|---|---|---|---|
| | 10Hz | 100Hz | 10Hz | 100Hz |
| Subject_1 | 0.93** | 0.75** | 0.39** | 0.14** |
| Subject_2 | 0.92** | 0.91** | 0.34** | 0.33** |
| Subject_3 | 0.76** | 0.88** | 0.62** | 0.24** |
| Subject_4 | 0.86** | 0.91** | 0.12** | 0.45** |
| Subject_5 | 0.88** | 0.87** | 0.13** | 0.52** |

Finally, the GSR readings displayed a steady decline, when the user finished the game and went into the recovering stage.

In the experiments, we tested the GSR sensors with two different sampling rates (10Hz and 100 Hz), and the results were both strongly correlated (Table 1) (Figure 14). Besides, we also checked the correlation between the extracted SCR signals, and they were both significantly correlated in the conditions of the two sampling rates.

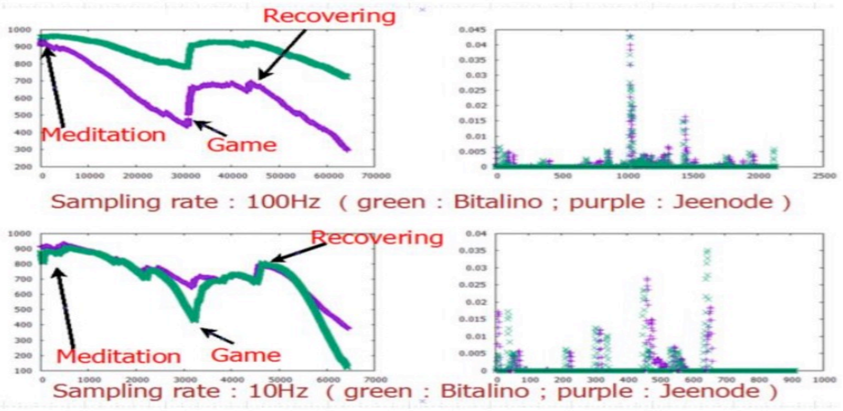


**Figure 14: An example of the data distribution between our GSR sensor and the one from the Bitalino. The left graphs (top and down) were plotted from the raw GSR data, and the right graphs (top and down) were plotted from the extracted SCR signals (after truncation procedure).**

### 3.3 The Network Performance of Hardware

This chapter reports our work on the design and development of a robust physiological measurement system that can be easily deployed in theatre-like environments. This system is adaptable to different types of sensors, audience sizes and environmental conditions such as theatre size and layout. It ensures the anonymity and privacy of the audience members, allows for real- time gathering of data, and supports large crowds of over 100 people. The contributions of this work are twofold. First, we provide an architecture for the system-based infrastructure that allows easy deployment of biosensor audience measurements. Second, we provide a predictive model for this architecture that is able to compute its performance during the experiment. The model accounts for two relevant issues: scalability with the number of audience members and delay and frequency of individual measurement values. This work represents one important first step towards the deployment of robust physiological measurement systems, which will radically influence the cultural sector on audience research.

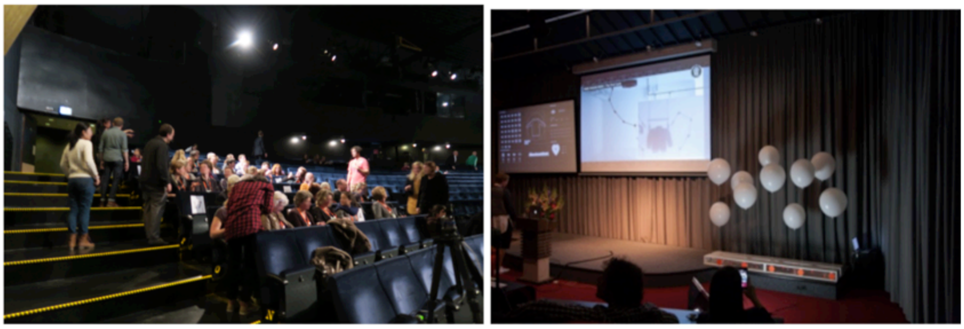


**Figure 15: （left）Audience members wearing wireless GSR sensors before the start of a performance. (right) Providing direct audience feedback by using movable helium balloons (Artist: Lilia Perez)**

The majority of available physiological sensors are designed and developed for lab studies [6] or for home environments [12,21]. A theatre-style building is rather different (see Figure 15): a large open space, full of people sitting close together, dimmed lights, quiet environment. There are a number of issues to be addressed for this use case of biosensors:

- A theatre is an environment that is less under control than a laboratory, with sources of radio interference and absorption, and the system should not interfere with the primary theatergoers of watching a performance.
- To obtain meaningful data on audience reaction to the performance using physiological sensors. The sensors should not adversely impact the experience, so they need to be small and lightweight.

- We must be able to correlate the individual measurements, and match them to what happens on the stage.
- If direct feedback is required the sensor readings should be collected in real-time.

One may argue that wearable technology paired with the mobile phones already exists (e.g., *Google Smartwatch* or *Empatica*). The phones can then seamlessly broadcast the physiological signals to a web server installed in the theatre. We argue that such solution explicitly excludes a large part of the population, who do not have such equipment, may bias the results. Moreover, the anonymity of the bio signals may not be preserved (since the mobile phone will have to sign on to use the web server). Instead, we foresee that the trend will be theatres providing the audience with their own lightweight sensors and deploy a specific network for the purpose of better understanding and quantifying the audience experience. Each of the sensors will deliver the data from an anonymous audience member to a central server that will process them in real-time.

We have in fact implemented a number of such systems, among others in collaboration with *Holland Dance* in the Netherlands and the *National Theatre of China* in Shanghai, and we have seen that these systems are similar, but different in details. The differences are such that a one-size-fits-all system seems unattainable. We have also experienced that various performance aspects of these systems have an impact that cannot be easily judged before the whole system is built and deployed in-situ.

In this work we present a scalable architecture that can be used to implement systems to do live audience research. We also present an accompanying model for determining the performance of this system before building and deploying it. Through this model, the system can be dimensioned to meet specific requirements from the stakeholders, such as number of individual sensor streams captured, or sensor sample rate.

### 3.3.1 The Workflow of the Sensor Architecture

The workflow in which our architecture should operate consists of 5 steps:

- **Requirements gathering**. This step consists of general design of the experiment, and happens in close cooperation with the stakeholders. The stakeholders can be managerial, creative or most probably both. The mix depends on whether the experiment is primarily aimed at measuring audience satisfaction or at using audience physiological feedback as part of the show. Obviously, selecting the type of biosensor to use (GSR, heartrate, etc) is part of this step, which gives rise to specific requirements on sample size*,* sample rate *and* measurement latency. Another important outcome of this step is determining the minimum required number of valid sensor streams. When quantifying audience reaction, the number must be selected in a way that the final outcome can be considered statistically valid.
- **Determining deployment environment**. Our system is based on radio communication, and a base station deployment strategy has to be selected. Radio signals are susceptible to interference and attenuation from many sources. Physical distance between a sensor and its base station is an obvious one, as is interference from furniture and people. There may also be other radio sources (such as radio microphones or in- house automation systems) that cause interference. The physical space will also determine where base stations can be placed: if overhead rigging is available and accessible, this can ameliorate the interference. If audience members can move around freely, this can cause multipath fading, which is less of a problem if people are seated. Furthermore, if seating is fixed we can pre-assign sensors to seat numbers, and we can position base stations physically close to the sensors they connect with.
- **System design**. Based on the information gathered in the previous two steps, and possibly after building a prototype sensor to determine exact timing characteristics, we can determine how many sensors we need and what the base station deployment options are. Now we can select the most cost-effective option that satisfies the requirements.
- **Build and test**. The required hardware needs to be bought, assembled and programmed. Practically speaking, there is a good chance that most, if not all, can be reused from previous experiments. The whole system is tested in the lab.

- **Deploy and run**. This involves installing the hardware before the performance, selecting the audience members to receive a sensor and attaching sensor electrodes and such, running the experiment and analyzing the results after the show. This step falls outside the scope of this study.

Based on previous experiments, and together with the workflow outlined above, we have determined a set of heuristics to follow, the full set of heuristics is shown in Chapter 3. For the remainder of this study, which concentrates on the system architecture for collecting the data, we are mainly interested in the *data* and systems groups.

### 3.3.2 Problem Statement

From the workflow and heuristics in the previous section we can formulate our problem statement:

- We need an architecture for gathering real-time data on how audience members react to an event such as a theatre performance, a concert or a lecture, and the resulting system must have a predictable performance.

The highlighted phrases in that statement require a bit of explanation, we will refer to heuristics with parenthesized numbers.

We want an architecture because we think this research area is one that will continue to grow for some time to come, and we do not want to re-invent the wheel for every experiment. We want real-time data because in some cases we may want to provide live visualization of the data to the performers, the director or even the audience themselves. This real-time requirement means that we cannot use sensors that record data during the performance for later offline analysis. We want to gather biosensor data from members of the audience, and potentially from large numbers of people. This means we need to be able to support a large number of sensors, and that those sensors need to be relatively cheap. Our focus on events, finally, means that we cannot rely on fixed infrastructure in the performance space, such as wiring in seats or near them.

The architecture should be scalable in various dimensions. Some experiments will require a larger number of sensors, some will require higher sample frequencies, some will require larger physical distances to be covered. Applying scalability measures may come at a cost, such as increasing the number of participants at the expense of sample rate, or even at a real cost (as in: requiring more hardware to be bought), but this choice should be left to the experimenter.

We want to be able to model performance of the system, so that key parameters such as number of participants, sensor reading frequency and radio range can be used as inputs to the design process. Again, the goal is to put a priori decision capability in the hands of the experimenter (and ultimately the stakeholders), as opposed to having to modify the experiment in an improvised way (such as by lowering the number of participants, or making do with a lower sample rate) at the time the experiment is ready to run.

A final design goal is deployability, and minimizing the workload on the experimenter during the experiment. Experiments with theatre audience members, and potentially a large number of them, require all sorts of logistics like explaining what the experiment is for, making sure release forms are signed and attaching sensor probes. Our architecture cannot help with this, but it can make sure that other tasks such as replacing batteries does not have to be done in that same hectic hour before the show. Our system also keeps a complete record of all data, including metadata such as radio signal strengths and packet loss and such but this falls outside the scope of this study.

Note that we are interested in aggregate results over groups of people, not in specific individuals, so we can afford to lose individual sensor streams or drop them in case of excessive data loss. Therefore, being able to record the data of a given individual sensor is an explicit non-goal.

### 3.3.3 System Architecture

Our system architecture (Figure 3, page 18) has been designed with real-time requirements in mind, so that we can know in advance that we can satisfy the requirements for a number of available sensor readings. In general, we have selected predictable technologies and structure and adhered to the KISS principle. Because of this principle and because we can afford to lose (or disregard) individual sensor streams we over-dimension the number of deployed sensors. In this way, we have a single solution for problems ranging from failing hardware to bad radio reception and packet loss.

The physical sensor nodes worn by the audience members consist of the actual sensor hardware, which has predictable timing for all bio signals we envision measuring, a radio module, and a small processor controlling it. These sensors are grouped with a single master node (radio module used as a base station, small processor, network interface) to collect the data from all group members. The protocol used is round-robin polling over all nodes in the group, with a strict timeout. The radio modules have been selected to have deterministic MAC and physical layers. The software in the sensor nodes and master nodes are simple busy-wait loops, with no multithreading and limited interrupt handling. All this leads to deterministic performance of a sensor node group.

For supporting multiple groups, we have chosen to use frequency division multiplexing (FDM), because this ensures independence between groups, and therefore maintains the deterministic performance. The effectiveness of FDM with the limited range radios we envision depends not only on the distance in frequency between the groups, but also on the physical separation, the separation in space. In other words: if two groups are spatially non-overlapping they can use frequencies that are closer together without interfering. Because the total available frequency range is fixed (by hardware capabilities as well as spectrum regulations) this means we can support more groups if we are able to cluster the group members in space. Assignment of nodes to groups is static (during an experiment) so there is no requirement for an arbitration or handover protocol that could interfere with the determinism.

The master nodes communicate with the central data collection system using an out-of-band network technology with known performance bounds.

### 3.3.4 Performance Model

An essential part of the third step in our workflow, system design, is determining how many sensors we should deploy, and how much we should over-dimension so that we still meet our requirements in case of problems. We also need to know how many we can support in a single group, how many groups we therefore need and how we can deploy these groups.

For this we first need to determine the duration of a poll cycle of a single sensor node, Tpoll. The details on measuring this duration are given later, in section 7.3, for now it suffices to know that this number is measurable and has a fixed upper bound.

From Tpoll and the requirements on sample rate and measurement latency we can determine how many nodes we can support in a single group: $N_{1group}$.

Depending on the seating arrangement and physical size of the theatre determined in workflow step 2, we select either the clustered group communication configuration or the multi-group configuration. Based on this, and on the legal requirements for radio spectrum usage in the deployment country, we can determine $G_{max}$, the maximum number of groups we can support. The details of such a computation are given below.

To determine the number of sensor nodes needed we obviously need to start with the number of valid sensor streams needed, from the requirements, $N_{wanted}$. The actual number of sensors we will have to deploy, N, is going to be larger than this to cater for our over-dimensioning. How much we should over-dimension $N_{wanted}$ to get N is the product of a number of factors:

- fraction $F_r$ of sensor nodes expected to be out of range of their master node, given static placement of nodes (measured during workflow step 2) fraction $F_l$ of sensor nodes expected to have packet loss rates that are too high (statistically determined during workflow step 2)
- fraction $F_h$ of sensor nodes expected to suffer hardware failure, for example battery exhaustion (upper bound based on previous experience)
- fraction $F_e$ of sensor nodes expected to suffer from electrode connection problems (upper bound based on previous experience).

Of these, the inclusion of $F_l$ depends on an assumption that packet loss is more correlated with individual sensors than with events that influence packet loss for all sensors. Below we show a measurement that supports this assumption.

Given these factors we can compute how many sensors we need to deploy:

$$N = N_{wanted} / ((1-F_r)(1-F_l)(1-F_h)(1-F_e))$$

Given N and $N_{1group}$ we can determine G, the number of groups we need to deploy. We already know $G_{max}$, the maximum number of groups we can support, based on radio characteristics and seating requirements. If $G \leq G_{max}$ we can start building the system. If $G > G_{max}$ we know that we cannot build a system that will satisfy all the requirements and we will have to go back to step 1 of the workflow and re-negotiate the requirements.

### 3.3.5 Implementation of Wireless Communication

*Communication Technology*

A common solution when wirelessly gathering sensor data is to use a smartphone to collect the data from the sensor, e.g. via Bluetooth, and then use a smartphone app to analyze it, visualize it or send it to a cloud service for storage or further processing. While this is a good choice for personal use, such as fitness tracking, it has disadvantages when used for groups. Smartphones are nowadays personal devices that univocally identify us. Thus, the anonymity required or desired is broken by using it as proxy. Even if precautions are taken in the software architecture to ensure anonymity, users may still feel uncomfortable about potential privacy issues. For this reason, we have chosen to design a system that does not rely on users' smartphones. Instead, we propose the use of standalone devices that incorporate sensing and communication capabilities. Thus, these devices can be handed out to people without linking them to their identity. Apart from anonymity, this model eases the real-world deployment of the experiment. For instance, audience members can get a sensing device when they come in and leave it when leave, just like the way museum visitors do with audio guides nowadays. Whereas if using their smartphone, they would have to install an app and link the sensor to it, which may stop people from using it.

Since we are not using smartphones, we need to select a wireless networking infrastructure for our communication. Choosing the right frequency band and radio technology are key to designing a wireless network. Not all bands are legally available, and different bands imply different propagation distances, energy consumption and bitrates. The radio bands used by WiFi or Bluetooth (e.g. 2.4 GHz) require a relatively high amount of power to obtain an acceptable range, which is not convenient for small portable sensors running on batteries. Furthermore, the large quantity of devices already operating in these frequencies and the number of sensors that we expect in our type of deployments would increase the chances of interference. Thus, we have selected to use the radio bands from the Short Range Devices (SRD) recommendation. Specifically, as we are based in Europe, we have selected the 868 MHz band, switching to 433 MHz for our China deployments (where 868 MHz is not available). Some existing technologies, like home automation systems, ZigBee and LoRaWAN, operate in these bands. Nonetheless, they are still not as saturated as the other possibilities. These frequency bands give us enough range to cover the potential application scenarios, e.g. theaters. In addition, they require much less power than bands in the GHz range.

The regulation for the SRD bands is established on a country by country basis, following the recommendations of international entities. We have mainly taken regulations in EU and China into account, as it is where our experiments took place. In general, when devices transmit with a duty cycle of less than 0.1% or use Listen Before Talk (LBT) to ensure minimal interference with other users of the sub-band, things are okay. The transmission power allowed ranges from 10 mW e.r.p., for 433 MHz, to 25 mW e.r.p in the other frequencies. Here e.r.p stands for effective radiated power, which is difficult to measure but an upper bound can be determined from the data sheet of the radio chip used. The most restrictive bandwidth limitations ask for channels smaller than 100 KHz. Local regulations have been taken into account to choose compliant hardware and design parameters (bandwidth, transmitted power, etc.).

We also have to decide a media access mechanism and a network topology. For the MAC layer there are two main options: either nodes compete for the airwaves in an Aloha-like fashion or communication is handled in a coordinated way. Given our real-time requirements and the availability of a master node in each group a coordinated approach was the obvious choice: the master node turns on its transmitter and sends a poll to a single sensor node, and only then does that sensor node turns on its transmitter to send a reply. All other sensor nodes leave their transmitters off (until they are addressed themselves). Other coordinated approaches such as distributed control based on timed slots were considered but did not result in enough performance gain and failed the KISS principle (Figure 16).

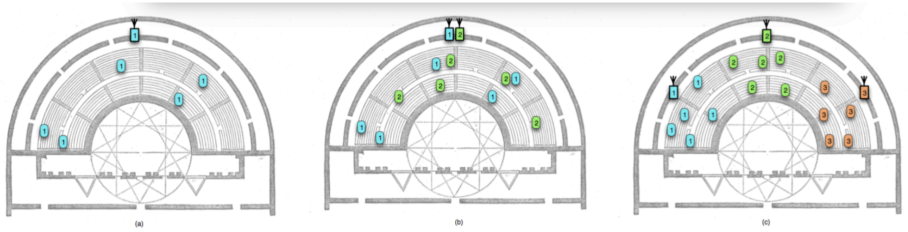


**Figure 16: Communication configurations: single group (a), multi-group (b) and clustered group (c).**

*Sensor Nodes*

For the sensor hardware platform, we have chosen to use JeeNodes designed by JeeLabs[3]. The JeeNode is an Arduino-compatible open hardware platform with a 16Mhz ATmega328 and an RFM12B radio board. It comes in various models with different I/O options (with or without USB, number of general purpose I/O pins, etc).

This platform was selected based on the fulfillment of a number of requirements:

- It is available for 433 MHz, 868 MHz or 915MHz, depending on local radio band regulations.

[3] http://jeelabs.org

- Power consumption is low: we measured 19 mA running our software during a typical experiment, down to 18 mA between experiments. We have implemented a standby mode, where between experiments the whole node is powered down for 2 minutes, after which it automatically comes up for 10 seconds (during which it can be told a new experiment is starting). In standby, power consumption is 5 μA, resulting in an average power consumption of under 1.5 mA between experiments.
- Radio range is sufficient: more than 100m unobscured outside, and no discernible packet loss and interference in a fairly empty 150-seat theatre (with no special considerations for antenna placement and alignment, and with radios both statically placed and moving around).
- Compute power is sufficient for our needs and price point of about €25 is acceptable.
- The open source JeeLib software package handles all the details of communicating with the RFM12B, configuring the devices, storing configuration data in flash memory.

For the sensor nodes we use a JeeNode v3 or a JeeNode SMD, with a 1000 mAh LiPo battery. These should allow us to run experiments of up to 24 hours, and a sensor shelf life of more than a week. The latter is important because it alleviates the logistics of charging and replacing a large number of batteries just before show-time. For the GSR sensor experiment we have used as a test case we added a small board with an amplifier and a low-pass filter to preprocess the electrical signal from the sensor hardware before passing it to an analog input.

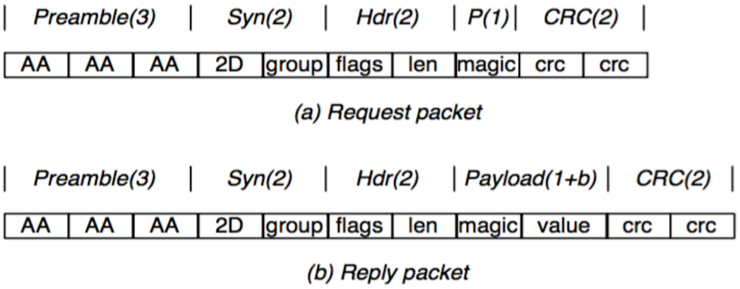
For the master nodes we use a JeeLink v3, which have a USB interface, connected to a Raspberry Pi 2. The Raspberry Pi has a builtin ethernet interface which is the preferred way to connect the master node to the central data collection system. However, if this is physically or logistically impossible (for example because we cannot run ethernet cables through the theatre) we can fall back to WiFi by adding a USB WiFi dongle, at the expense of a bit more timing uncertainty. The Raspberry Pi runs the standard Raspbian distribution, with only an RFC2217 telnet server added to allow the data collection system to communicate with the JeeLink. Our master node software runs completely on the JeeLink, ensuring our realtime requirements. Ethernet and USB throughput are at least three orders of magnitude better than our radio protocol so we can safely ignore their effect on our timing.

The master node communicates with the data processing application using simple CSV (comma separated value).

*Protocol*

Our packet format has 1 byte dedicated to a magic number than we vary between experiments, this allows us to detect sensor nodes that are inadvertently still running software for another experiment. The skeleton software also handles entering and leaving standby mode, and it reacts to an "anybody out there?" message that we use to find forgotten sensor nodes lying in a cupboard somewhere that could otherwise disrupt an experiment. This additional functionality is triggered by different magic numbers.

The resulting packet formats are shown in Figure 17. Syn, Hdr and CRC are defined by the RFM12B hardware and the JeeLib protocol and handle framing, addressing and error detection. Payload is determined by us. The request contains only the magic number, giving a total packet size of 10 bytes. The reply additionally contains a single sensor value, with size b equal to 2 bytes in case of a GSR sensor, giving a packet size of 12 bytes.



*(a) Request packet*

*(b) Reply packet*

**Figure 17: Packet formats**

The JeeLib protocol has a limit of 30 nodes per group. But because we use a strict request-reply message structure, we can sidestep that limit by using multiple JeeLib-groups to implement one of our node groups, and have the master node switch groups to the sensor node it currently needs to communicate with.

*Performance Model Parameters*

The performance model from chapter 6 requires that we determine Tpoll, the time for a single poll cycle. For the hardware given in the previous subsections we can now measure this number with the exception of a single contribution: the time needed by the biosensor hardware to take a single reading (H, in the formula and diagram below). However, this number is often known (for example: a GSR measurement takes a single analog value conversion, which takes around 100 µS on an ATmega328) and otherwise can be measured after constructing a single sensor.

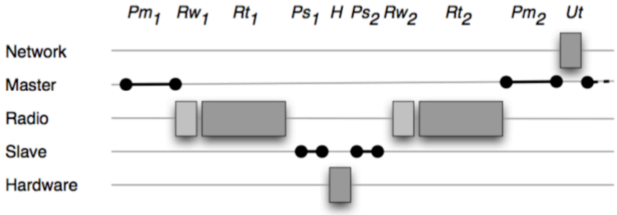


**Figure 18: Timing of single sensor poll cycle**

Figure 18 gives the complete timing of a single poll cycle. $Pm_1$ is the time the master node uses to prepare the request for transmission. $Rw_1$ is the time the master needs to wait until the radio channel is available, using LBT, and is non-deterministic but we give it a known upper bound through a timeout mechanism. $Rt_1$ is the time spent transmitting the poll request. $Ps_1$ is the time the sensor needs to process the request, H the time to do a sensor reading and $Ps_2$ the time to prepare the reply. $Rw_2$ is the time the sensor needs to wait for the radio channel to become available, and is again non-deterministic but bounded. $Rt_2$ is the time needed to send the reply packet.

$Pm_2$ is the time the master needs to process the reply, and finally Ut is the time needed to send the reply to the data processing application.

We implemented a sensor node that returned 2 16-bit numbers, the real GSR sensor reading and the sensor-perceived poll interval in milliseconds, and we proceeded to calculate and measure the timing parameters, as outlined in section 6. Request packet size is 10 bytes, transfer rate is 50Kbps, therefore $Rt_1$ is 1.6 ms. Similarly, reply packet size is 14 bytes given $Rt_2$ is 2.2 ms. Tpoll was measured to be 6.30 ms (with a standard deviation of 0.53). To validate, we repeated the experiment with the GSR sensor code disabled and got 6.22 ms for $Tpoll_0$ (standard deviation 0.45), which seems reasonable: it is close enough to the 100 µS for an analog analog reading from the ATMega328 data sheet. Finally we ran the master in send-only mode, without waiting for sensor node replies. This took 2.91 ms (standard deviation 0.45). We can now compute $Pm_1+Pm_2+Ut$ to be 1.3 ms and $Ps_1+Ps_2$ to be 1.2 ms.

All of these experiments were done in lab conditions to keep $Rw_1$ and $Rw_2$ as small as possible, and the standard deviations seem to indicate that we succeeded in this. So this leaves the experimenter free to adjust the timeouts, the upper bounds for $Rw_1$ and $Rw_2$, where the tradeoff is that a lower timeout results in a higher polling frequency at the expense of higher packet loss.

*FDM Parameters*

To test that our FDM strategy works for isolating groups and determine its limits we ran an experiment. We built the hardware and software and ran an experiment with two groups consisting of a master and three sensor nodes each, using 868 MHz radio boards, running software as described in the previous subsections. These radios use LBT, listen before talk, to ensure the channel is free before transmitting. We then tarted by measuring the poll intervals of each individual group member, which was a very consistent 18.86 ms (Figure 19). To determine the worst-case for interference between two groups we put two groups in the same frequency (Figure 19) and the poll intervals have gone up by a factor of 6, with very high standard deviation.

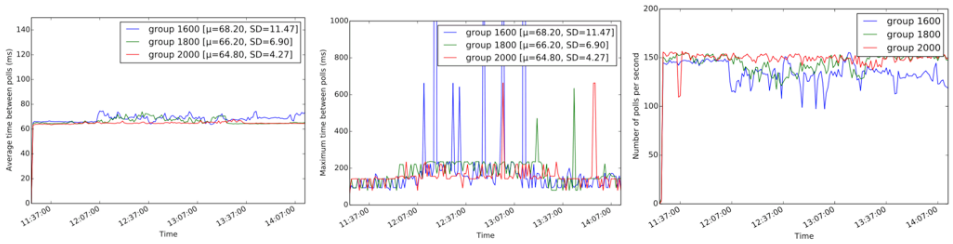
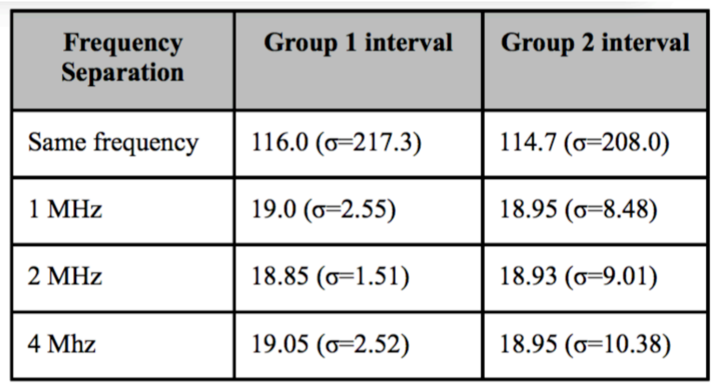


**Figure 19: Polling rate performance in a crowded space:**
**3 clustered groups of 10 nodes each**

Next we put each group into its own frequency band again, and varied the distance between the base frequencies. We tried this with various distances, the interesting results are summarized in Table 2 (with the results for running two groups in the same frequency added for reference).

**Table 2: Effect of spatial separation on average poll interval**

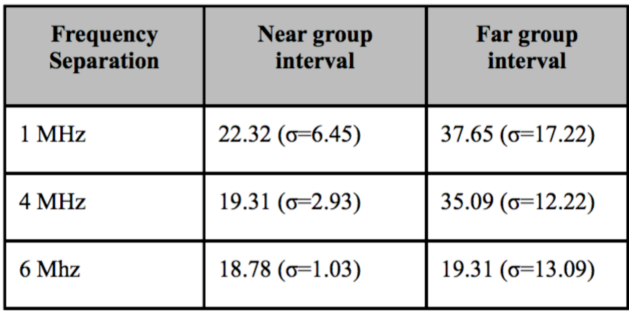


| Frequency Separation | Group 1 interval | Group 2 interval |
|---|---|---|
| Same frequency | 116.0 (σ=217.3) | 114.7 (σ=208.0) |
| 1 MHz | 19.0 (σ=2.55) | 18.95 (σ=8.48) |
| 2 MHz | 18.85 (σ=1.51) | 18.93 (σ=9.01) |
| 4 Mhz | 19.05 (σ=2.52) | 18.95 (σ=10.38) |

With 1 MHz separation we have room for 10 groups, which seems enough for our purposes. In addition, 1 MHz is well above the preferred channel spacing of 100 KHz. Finally, we started spacing out the groups physically, i.e. moving some sensor nodes further away from their master node. Something interesting happened: for

larger frequency separation things worked more or less as before, but if the groups were closer in frequency the results suffered. Specifically, when we kept the two master nodes and one group of sensor nodes fairly close together and moved the other group of sensor nodes to a distance, it was the group that was far away that suffered much more than the sensors that were near their master node.

**Table 3 : Effect of frequency separation on average poll interval**



| Frequency Separation | Near group interval | Far group interval |
|---|---|---|
| 1 MHz | 22.32 (σ=6.45) | 37.65 (σ=17.22) |
| 4 MHz | 19.31 (σ=2.93) | 35.09 (σ=12.22) |
| 6 Mhz | 18.78 (σ=1.03) | 19.31 (σ=13.09) |

Table 3 shows the results: with a frequency separation of less than 6 MHz we have to be wary of interference if group members are far away. Note that these particular results (3 groups spaced at 6 MHz for non-clustered grouping, 10 groups spaced at 1 MHz for clustered grouping) depend in part on the ISM band used (433 or 868 MHz) are specific legal requirements in the deployment country, so at least part of this calculation will have to be repeated for a new experiment.

*Packet Loss Measurements*

Our strategy to use over-dimensioning in deployed sensors to overcome radio interference and packet loss, depends on an important factor: we presume that radio problems will occur at specific sensors, more than at all sensors at specific times. We have conducted an experiment to see how random obstruction of the signal path influenced packet loss, at different power levels. For this experiment, we created a group with one master and 5 sensors. We placed the master on one side of a busy corridor with a coffee machine and meeting area and the 5 sensors closely together at the other side, carefully placed where we were seeing only a little packet loss when conditions were optimal. We also had a camera near the master node looking towards the coffee area and the sensor nodes, and we took a picture every minute. We inspected these pictures to determine how busy the corridor and coffee area were at each time. The sensor nodes were close together, less than λ/2 apart.

Table 3 shows the cumulative packet loss per power level over a 5-hour period, and (non-rigorous) inspection of the pictures showed that there was some correlation between packet loss and the number of people in the radio signal line of sight. As expected, the packet loss increases as the power level decreases. But the interesting finding is the packet loss figures for the individual nodes. Figure 19 (right) shows the loss data at the 0 dBm level (the dark blue line in Figure 19): nodes 3 and 6 are the only ones having problems, the other three nodes experience below 2% loss. A similar pattern showed up at lower power levels: as power drops the nodes start experiencing packet loss one after the other.

This corroborates our assumption that packet loss is more auto-correlated for each node than correlated with obstructions and other temporal effects.

*Grouping measurements*

To validate our parallelism measures we ran an experiment with three clustered groups of 10 nodes each. We ran this over the lunch break in the staff restaurant at our workplace. Master nodes were placed at three locations in the restaurant, and sensor nodes (with no sensor hardware) were affixed to adjacent tables so that the closest master for each sensor node was its own master. Generally, people start coming in for lunch around noon, and the peak time is between 12:15 and 13:45 with 100-200 people coming and going in the area where we had set up the experiment, so we ran the experiment from around 11:45 until 14:15. The master nodes communicated with the data collection computer over WiFi, because running Ethernet cables was not convenient.

Figure 19 on page 58 shows the results of this experiment (note that the group numbers are for identifying purposes only: "group 1600" is a group that had its sub frequency set to 1600, sub frequencies are measured in 5KHz intervals) graphed as average time between polls of each individual sensor, measured per minute, maximum time between polls of each individual sensor, again measured per minute and number of sensors polled per second. We measured the intervals between successive polls as seen by the sensor node, because this is a measure for both packet loss and other delay factors.

If we compare the average time between polls with the 6.22 ms we measured in the lab the results look reasonable: before people arrive that is what we measure, and as more people come in we see an increase in response time, due to packet loss and interference. We have no good explanation for the anomalous behavior of group 1600 after the rush hour is over: the maximum poll time goes back to the pre-lunch level, but average poll interval and number of polls do not. We presume some furniture may have been moved around.

It is also clear that there is some correlation between the groups as people come and go, but much more correlation within each group (as tables to which nodes for that group are attached get occupied).

To validate that clustering groups is worth the extra effort we ran a second experiment with two groups of 25 nodes each, this time without clustering our sensors. We ran this over another lunch break in the staff restaurant. Sensor nodes (with no sensor hardware) were affixed to adjacent tables, and the 2 master nodes, together with the data processing computer, were placed at the far end of the groups. The results are shown in Figure 20.

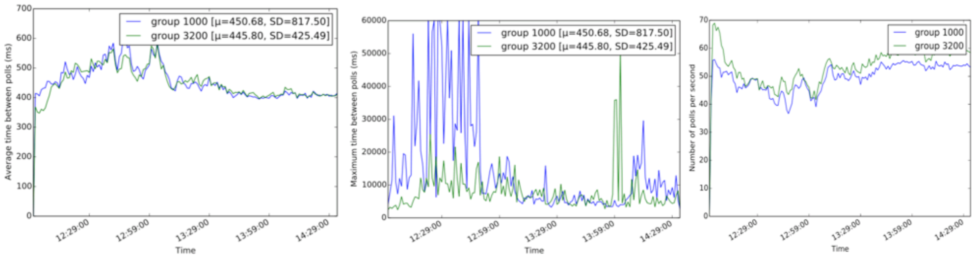


**Figure 20: Polling rate performance in a crowded space: 2 groups of 25 nodes each**

(Please note that the absolute values of the poll times are not comparable, because the second experiment was run with a slightly different instrumentation code). The effect of people entering the restaurant is more pronounced than in the clustered group experiment, because the average distance between sensor and master node is bigger. Moreover, there is much more correlation between the two groups (because the group members are interspersed in the same area).

But the data for the maximum poll time is even more telling: when it is busy there are often nodes that have not been polled for 10 seconds, or even up to a minute for group 1000 members (whose members were, on average, further away from the master nodes).

If we compare the trends in Figure 20 (right) we see that the clustered group configuration reacts more gracefully to increasing number of people, especially when we look at the maximum poll intervals. This shows that the added complexity pays off, and can indeed help us increasing the number of participants in an experiment.

## 3.4 Conclusions and Future Work

We have shown the design of an architecture for wirelessly obtaining real-time biosensors readings from audience members, and we have implemented one instance of that design, to take GSR readings.

We have shown that this system has a predictable performance, that this performance can be measured before the whole system is constructed, and that the final system can therefore be built at the correct scale to match predetermined requirements such as the number of valid sensor streams.

While we have not used the architecture for a real audience experience experiment yet, we have shown that we can take the parameters from such an experiment and use our performance model to show the feasibility of the experiment.

We have shown that our grouping measures indeed allow to increase the number of sensors that can be polled at a given frequency, and that our clustered grouping model performs better than the simple non-clustered grouping parallelism.

We have found that, for our fairly static deployment of sensor nodes in a crowded setting, packet loss is tied more to individual sensors than to events that influence all sensors at a given period in time. Therefore, increasing the deployed number of sensors is a good strategy to cope with packet loss.

Some of our specific findings may be relevant outside our immediate area of interest, specifically the results on packet loss, frequency spacing and physical spacing. Specifically, the result that interference decreases with increased physical distance as well as with increased frequency distance is probably more widely applicable (with the caveat that we have only tested with stationary nodes).

There are a number of areas in which we will continue to work on this architecture. We would like to be able to implement dynamic assignment of a sensor node to a group, if we can do it in a way that does not seriously interfere with the timing predictability. This may also open the way for allowing sensors to move around in space, which enlarges the application area for our system. The ease of deployment is another area of interest, and we are experimenting with sensors nodes that can completely shut down, effectively extending their shelf life to months, if not years.

# 4

# Software Development

In this chapter, we present our software development for different user studies. We spent considerable effort on understanding the sensor data, i.e., how to understand the sensor features, how to interpret the sensor data and whether the sensor signal patterns can across different scenarios. Generally, the raw data are smoothed first in order to remove artefacts occurred during an experiment. After that, the required sensor features will be extracted. In order to link the physiological data into a meaningful user's psychological state, an algorithm needs to be developed or applied. However, inferring the subject's psychological state by using physiological signals is a challenge, which requires knowledge of both psychology and computer science. First, one psychological elements affect many physiological responses, which means the connections between physiology and psychology are more likely to be one-to-many. Second, there are no standardized methods for the interpretation of psychophysiological responses, and each study needs to find a suitable algorithm to process the sensor data. Third, inference validation is problematic. Once we developed an algorithm capable of inferring a person's psychological state, we need to test it and see if it works correctly. In general, we use questionnaires, observable behavior, or standardized stimuli that are expected to induce the same psychological state. But, questionnaires measure conscious process, while physiological responses are based on both conscious and unconscious processes. Therefore, subjects may not be aware of their psychological states. Observable behavior can be an alternative, but physiological changes can occur at any time although there is no corresponding behavior. Last, induction of psychological states in lab conditions is difficult to be generalized in other scenarios. For instance, we use mental arithmetic task to induce frustration, but it may not evoke the same physiological patterns as frustration occurred in a traffic jam. Thus, finding appropriate validation methods is likely to be an application dependent affair. Some researchers take an alternative approach by calculating the accuracy of the psychophysiological inference instead of validation part.

In this chapter, we will introduce the algorithm development and present the analysis results in their respective studies. This chapter particularly consider the following research questions:

**Research Question 4:** *Can audience engagement be inferred [in real time] from GSR data?*

**Research Question 5:** *How live audience experience is differentiated from remote audience?*

**Research Question 7:** *Are there patterns of audience engagement across productions and audiences?*

## 4.1 Psychophysiological Measurements

We applied a *Multidimensional Scaling Method* (MDS) and *Correspondence Analysis* (CA) to explore the relationship between physiological sensor data and subjective ratings, and investigate whether GSR sensor signal patterns can go through different scenarios. Multidimensional scaling can be used to explore data structure (visual representation) in a set of distance measures, e.g., dis/similarities, between objects/ cases. This method includes a series of techniques that help analysts to identify key dimensions underlying respondents' evaluations of objects. Marketing research often uses it to identify key dimensions underlying customer evaluations of products, services or companies. When we have sensor data on hand, we can use this method to determine how the audience's sensor signal patterns are related perceptually. The resulting perceptual maps show the relative positioning of audience members' bio responses.

MDS is a means of visualizing the level of similarity of individual cases of a dataset, in particular to display the information contained in a distance matrix [44, 52]. Furthermore, MDS technique aims to place each object in an N-dimensional space such that the between-object distances are preserved as well as possible. In our analysis, we used a two–dimensional space to display the similarities between the objects.

MDS has been widely applied in psychological research [21, 52, 83], but it is a new research technique when applied to physiological computing. Unlike other statistical techniques that test hypotheses that have been proposed a priori, MDS is an exploratory data method that explores data for which no specific hypotheses have been formed. The difference is that MDS is a means of visualizing the level of similarity of individual cases of a data set, in particular to display the information contained in a distance matrix, i.e., Euclidean distance.

In order to examine how the results are generated by using MDS, we reported the overall fit statistics: Kruskal's stress and R Square. For a better interpretation of results, we choose a two-dimensional based perceptual map to interpret the results. Before conducting MDS algorithm, Pearson product-moment correlation coefficient was used to analyze the similarities or dissimilarities among sensor data. After that, MDS was applied to display the data clusters on a perceptual map.

The CA originated approximately 50 years ago， and has been referred to by a variety of names, such as dual scaling, method of reciprocal averages, and categorical discriminant analysis [34]. CA has become most of popular in fields as ecology, where data is collected on the abundance of various animal species in specific sampling units/areas. However, this statistical method merits further attention within the field of psychological research.

Unlike the many statistical methods that test hypotheses, CA is an exploratory data technique that explores categorical data for which no specific hypotheses have been formed. More specifically, CA analyses two-way or multi-way tables with each row and column becoming a point on a multidimensional graphical map, also called a biplot. In terms of continuous physiological senor readings, the data could be categorized and subsequently analysed as discrete data.

CA uses the chi-square statistic – a weighted Euclidean distance to measure the distance between points on the biplot. In other words, CA was operated and first obtained a chi-square measure of similarity for the values in a contingency table, and these chi-square values were standardized and converted to a distance value that would be presented in the perceptual map. Once the dimensional graph has been established, we can identify a category's association with other categories by their proximity on the perceptual map. In other words, on the perceptual map the categories can be compared to see if two can be combined (they are in close proximity on the map), or if they do provide discrimination (they are located separately in the perceptual space).

CA is also a good way to examine data validity and facilitates the treatment of outliers. In our case, the result of research question 1 would recognize non-engaged audience member, and the outcome of the research question 2 was related to the different performance strategies. In addition, we categorized the GSR sensor data with response to the different types of performance, so that we obtained the discrete sensor readings. In the next, we applied CA to generate a contingency table where the row and column representing audience id and the different performance strategies respectively. Therefore, in the generated results, we could clearly define audience outliers was linked to which performance strategy.

We applied these two methods in three user studies to analyze the GSR sensor data. In one user study, we applied it on sensor data to explore the proximity of audience members' bio responses and reveal the effects of different performance strategies. While in the other two user studies, we investigated whether GSR sensor data can across different scenarios. This work addresses research questions 4 and 6.

### 4.1.1 User Study: Sensing a Live Audience

Psychophysiological measurement has the potential to play an important role in audience research. Currently, such research is still in its infancy and it usually involves collecting data in the laboratory, where during each experimental session one individual watches a video recording of a performance. We extend the experimental paradigm by simultaneously measuring Galvanic Skin Response

(GSR) of a group of participants during a live performance (The 1st version hardware). GSR data were synchronized with video footage of performers and audience. In conjunction with questionnaire data, this enabled us to identify a strongly correlated main group of participants, describe the nature of their theatre experience and map out a minute-by-minute unfolding of the performance in terms of psycho-physiological engagement. The benefits of our approach are twofold. It may provide a robust and accurate mechanism for assessing a performance. Moreover, our infrastructure can enable artists, in the future, to obtain the real-time feedback from remote audiences for online performances.

The primary motivation for the current study is to explore the viability of using Galvanic Skin Response (GSR) to monitor audience feedback during a live performance. We took GSR measurements of 15 people watching a live theatre performance simultaneously. The readings were synchronized with video recordings of the performance and the audience. The audience filled out questionnaires, and we used them to evaluate the emotions evoked by the performance. This resulted in a high volume of useful data of around 1680 data points for each participant.

Results indicate that our approach – gathering GSR data during the play - is valid, as such data accurately reflects the engagement of the audience members. Moreover, it proves to be a useful tool for temporally unfolding the experience of the public, as the reactions of the public can be mapped to specific events during the play. In principle, we can conclude that our solution of using GSR data for monitoring audience feedback is novel and very valuable.

### *Experimental Design*

Seven females (mean age 28.29) and eight males (mean age 23.13) formed the audience for a 28-minute theatre performance. Their GSR was measured every second throughout, resulting in 1680 data points for each participant. Actors devised and performed a comedy that was aimed at audience participation and produced occasional “shocks” (e.g. a popping balloon) to elicit the occurrence of GSR spikes during the performance (Figure 21).

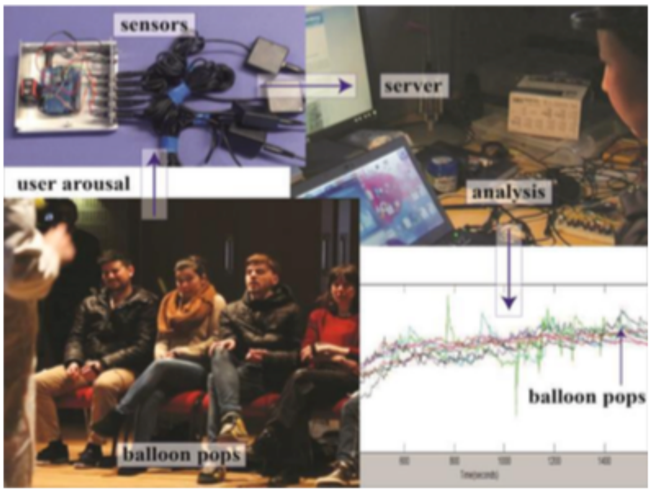


**Figure 21: GSR system (the 1st version hardware)**

Groups of five sensors were each connected to one of three Arduino UNO boards (sample rate 1Hz) (Figure 21). Xbee RF modules were used to create a wireless network such that the GSR data were sent directly to a laptop. This ensured the synchronization of all GSR readings. Cameras recorded the audience and the performance. Video streams were synchronized (post production) with GSR data.

Before the performance, participants filled out a short questionnaire asking about the type and intensity of the emotions they had experienced during the day. Afterwards participants filled out a similar questionnaire asking about emotions experienced during the play. The questionnaires were in the form of graphic rating scale and measured 100mm. Participants were asked to make a mark between two extremes, i.e. between "not at all" and "very much".

Participants were seated in one row with three sections of five seats each, arranged in a semi-circle around the stage. GSR modules were attached to the palm of the left hand. Before the performance started, participants took part in a meditation exercise to establish a baseline GSR level.

Questionnaires were analyzed using Analysis of Variance (ANOVA) and correlations. The synchronized GSR and video streams enabled us to relate events during the performance to corresponding GSR readings. GSR readings were analyzed using the MDS method. Correlations and ANOVA had some limitations to do a complete interpretation of the readings. They are fairly suitable if the audience is being treated as a whole, but they cannot properly explain relationships – similarity and dissimilarity - between objects in a multi-dimensional space. In our case, we were interested in understanding the relationships between 15 objects (each audience member) GSR responses'. We calculated the dissimilarities between the objects using Pearson Correlation Coefficients, and two-dimensional scaling was chosen for scaling. After 30 iterations, the final configuration graphs were achieved and Kruskal's stress reported in the results.

*Audience Clustering*

Questionnaires were analyzed through Analysis of Variance (ANOVA) and correlations. The synchronized GSR and video streams enabled us to relate events during the performance to corresponding GSR readings. GSR readings were analyzed by MDS method. Correlations and ANOVA had some limitations to do a complete interpretation of the readings. They are fairly suitable if the audience is being treated as a whole, but they cannot properly explain relationships – similarity and dissimilarity - between objects in a multi-dimensional space. In our case, we were interested in understanding the relationships between 15 objects (each audience member) GSR responses'. We calculated the dissimilarities between the objects using Pearson Correlation Coefficients, and two- dimensional scaling was chosen for scaling. After 30 iterations, the final configuration graphs were achieved and Kruskal's stress reported in the results (Figure 22).

Five participants displayed different patterns. Two showed an initial rise in GSR followed by a decrease, i.e. after an initial engagement with the performance their attention waned; for one this was related to receiving sad news during the day. For the other it showed an initial lack of rise in GSR followed by an increase; they reported to be confused initially by the purpose of the play and as such it took them a while to get into the performance. One participant displayed a consistent drop in

GSR and reported not liking the performance. These characteristics enabled us to label the extremes of the X and Y-axes.

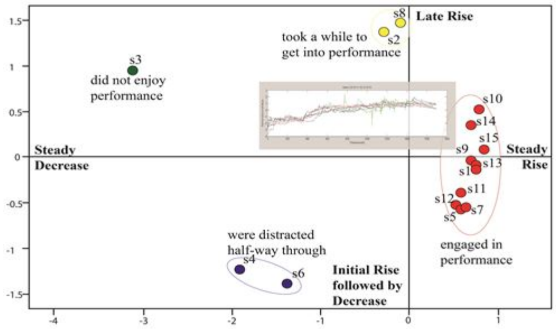


**Figure 22: Audience Clustering**

*Unfolding of the Performance*

For each minute, the GSR readings were averaged for each participant (Figure 22). Here MDS (Kruskal’s stress: 0.05) yielded an almost chronological minute by minute unfolding of the play (anti-clockwise in Figure 23) up to minute 19. Using the video footage we were able to identify the clusters based on the content of the performance. Thus, initially the GSR readings are low (minute 1) followed by a steady rise (minute 2 – 19) after which the intensity of the GSR flattens (minute 20 – 28). The first part of the performance (minute 2 – 16, in red in Figure 23) built up to an active and physical participation during which the participants were asked to raise either their left or their right leg in response to (silly) questions by the actors. In minute 17 – 19 (in green), the results of a competition were revealed, where the relatively higher GSR readings might indicate anticipation. After that the audience was not required to interact as actively as they listened to a trumpet player (dark blue) and watched a juggling act (yellow). The Y-axis reflected levels of GSR intensity and the X-axis ran between low and high audience participation. Spikes were identified that corresponded to the intended “shocks”, e.g. balloon popping, the sudden sound of a (badly played) trumpet.

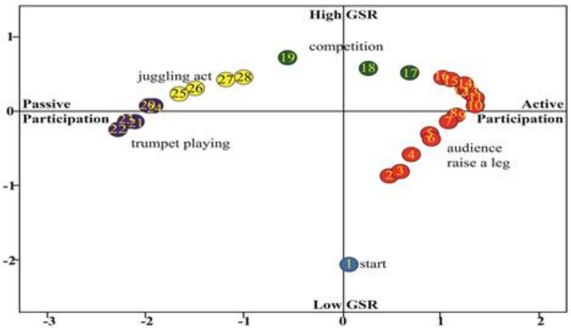
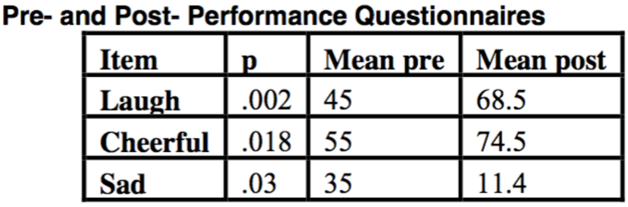


**Figure 23: Unfolding of the Performance**

The minute average GSR readings during this comic play correlated positively with participants being (very) cheerful (on average r = .62) and correlated negatively with participants being sad (on average r = -.60) at different stages of the performance, in particular from minute 16 onwards the average GSR readings showed strong correlations with audience's "cheerful" ratings.

Table 4 summarizes the significant differences between pre- (asking about their experiences during the day) and post-performance questionnaires. The ratings were given on a scale between "not at all" (= 0) and "very much" (=100). Thus participants rated that during the day, on occasion, they had a laugh with a mean intensity of 45 (Mean pre in Table 4) and they reported that the intensity of laughter generated by the performance was rated on average as 68.5 (Mean post), resulting in a significant difference, F （1， 14） = 14.68, p = .002. Similarly, for cheerfulness, the difference between pre-and post- ratings was significant, F (1， 14) = 7.12, p = .018. On average, participants had a reasonably cheerful day (Mean = 55) but these ratings increased to an average 74.5 after the performance. Lastly, although participants did not have a particularly sad day (with the exception of one participant) yielding a mean of 35, this was significantly reduced to a mean of 11.4 after the performance F (1， 14) = 5.82, p = .03. There were also significant effects for gender and whether a participant knew another participant sitting in their row or not. However, due to the low numbers in each “cell”, we refrain from reporting these in this thesis.

**Table 4: Significant differences between pre- and post- questionnaires**



Pre- and Post- Performance Questionnaires

| Item | p | Mean pre | Mean post |
|---|---|---|---|
| Laugh | .002 | 45 | 68.5 |
| Cheerful | .018 | 55 | 74.5 |
| Sad | .03 | 35 | 11.4 |

*Linking audience members and performance strategies*

We performed the CA between the performance strategies and the audience clusters ($\chi^2$ (56, N = 25200) = 6498.29, p < .01). Figure 24 shows the detailed mapping between the audience members and the performance strategies, which can help performers assess the relationship between the devised performance strategies and the different audience members. In other words, we also could identify the non-engaged audience being linked to which performance strategy. Obviously, the GSR feedback showed that a3 is closest to the beginning of the performance (s1). In other words, the GSR response of participant 3 remained synchronized with the beginning of the performance, since he did not enjoy the performance. Furthermore, both participants a2 and a8 are adjacent to s4 indicating that they may be fond of the trumpet act than of the other strategies, which it is also consistent with their self-reports: they clarified that it took them a while to understand the performance.

Both a4 and a6 were nearby s2 which makes much sense as they were active in the interaction with actors (as seen in the video), but they got distracted during the half of the performance. In contrast, the engaged users (in the red cluster) stayed closest to the main body of the performance - an approximate equilateral triangle constructed by s2, s3 and s4. Although some of them, i.e., a10 & a15, where the nearest neighbors of s3, which might be interpreted as that the contest strategy affected them to the most since their group won the prize. Accordingly, both a14 and a9 were around the corner of s5, which may indicate that the trumpet playing and juggling act aroused them in the end of the play, and they reported these events as special moments in the questionnaires.

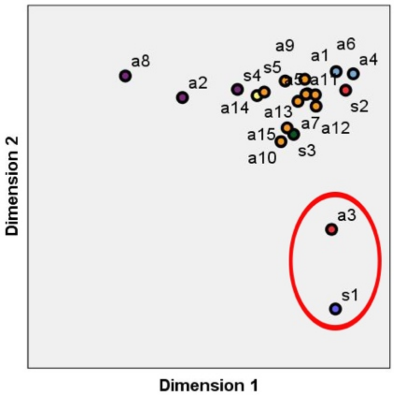


**Figure 24: The correspondence between the audience members and the performance strategies.**

### 4.1.2 User Study: Do We React in the Same Manner?

Is the physiological response from participants different between a lab experiment and a field study? This question represents another challenge in physiological computing. In this study, we exhaustively compare the GSR patterns between two different scenarios. The first one was conducted in a theatre during a performance, while the second one in a laboratory during a video watching session. Questionnaires, interviews, and video recordings helped us to interpret sensor signal patterns, and to map them to user engagement. When comparing the GSR responses, we found a strong positive correlation between all engaged users of the two scenarios. Interestingly, such correlation was not present between the responses of non-engaged users. These results show the homogeneity of positive responses across scenarios, when compared to the variability of negative ones. The results corroborate as well that sensor data results obtained in lab studies cannot be easily generalized to real-world situations.

Physiological sensors provide valuable and reliable data about the responses of users to products and experiences. Lately, it has attracted the interest of the HCI community, becoming one more tool to help evaluations. Unlike subjective approaches like surveys, sensors provide objective data, do not interfere with the activity, and can be instrumented for different purposes (e.g., to visualize sensor data in real-time).

Even though the many benefits of sensor technology for evaluating user experience, comparative studies across scenarios are missing [58]. Most of the previous studies targeted one situation, one context, and one activity. There have been articles reporting the use of physiological sensors as objective evaluation mechanisms [16, 5]. Others have explored the relationship between biosensor data and subjective methods [4, 18]; and most of the work has been targeted to label the users' affective states [10, 57] (e.g., fatigue). Moreover, in most of the cases the user studies were conducted in lab settings and the results were based on the averaged performance of the group (and not on the performance of the individuals).

This lack of comparative studies raises one important question then. Are user responses in controlled lab studies comparable to those obtained in the field? For example, can we use GSR data patterns reflecting non-engagement in a lab as a baseline for identifying non-engagement in the field? In general, authors discourage such generalization [43, 48, 49], since the response of users might be different in different contextual situations.

This work compares two scenarios, and the particular case of physiological sensors, exhaustively comparing the GSR data obtained in two different studies. In order to perform the comparisons, we need to classify user response into clusters, representing different types of feedback. We must avoid previous pitfalls of averaging data readings [46, 29], which do not provide the required level of detail and concreteness. Machine learning can play an important role for classifying data [27]. Nevertheless, we decided not to use such approach because of the annotations required in the training data. These annotations may alter the sensor readings and cannot be detected in the data set. In our experiments we prefer not to explicitly assign tasks during the experiment, which might help machine learning, but surely will disturb the experience of the users.

We classify engagement following the method initially proposed by Peter Lang: GSR sensor and audience subjective reports are used together in order to identify the feedback from the users. Similarly, Celine Latulipe et al., used the same model to describe audience engagement for recorded videos of performing arts [4]. They extended the model linking the GSR sensor data to two self-reported scales. Their results indicate that GSR readings are a valid approach for measuring audience engagement. Furthermore, researches in affective computing and HCI have shown interesting results between GSR and engagement [18].

In particular, we conducted an exhaustive comparison between the GSR sensor signal patterns across two user studies, aiming at a better understanding of whether engagement follows similar patterns. One experiment was run in the field, during a theatre play, and the other one was run in the lab, with users watching videos. The same experts, using exactly the same sensors and software, ran the two experiments. In both cases, apart from the sensor data, several other materials were recorded (interviews, questionnaires, videos) in order to identify user engagement.

These materials helped us to interpret the sensory patterns, and to subjectively map them into types of audience engagement. For example, terms defined in the questionnaires were used to check audience emotional states, such as cheerful or enjoyable. Furthermore, group interviews provided us detailed information about the experience (e.g., a bad day may distract audience attention). Finally, video recordings were used to more accurately analyze the data, for example to recall what happened during the experiment and to examine, in synchronicity, particular events during the experiment and the behavior of the users. We believe the research reported in this thesis can help bridging a gap between lab and field studies, helping to better understand what to expect from physiological sensors.

*Method*

In order to select the videos participants watched, we conducted two rounds of interviews with experts: a professor at the media faculty of one of the top Chinese universities and six of his students. Based on the collected opinions, we chose two advertisements: one is a fitness product, and the other one promotes one coffee brand. Unlike previous studies about video watching, we used videos as the stimulus to capture the GSR response from users, and not to identify their emotional state

[37]. This lab experiment was done in the computer labs from the same Chinese university (Figure 25).

This experiment was organized in the UK, during an interactive theatre play that lasted around 30 minutes. Four actors devised a comedy with different types of performance: juggling, asking the audience questions, and trumpet playing. Fifteen audience members participated in the play at a local theatre in the UK (Figure 25).

GSR sensors with wireless communication modules are better suited for running studies with larger groups of users. Such infrastructure should be capable of handling several sensors at the same time, since certain times it is not possible to repeat the experiment for each user (e.g., theatre play). Moreover, several rounds of repeated experiments might bring some undesirable random effects to the data collection.

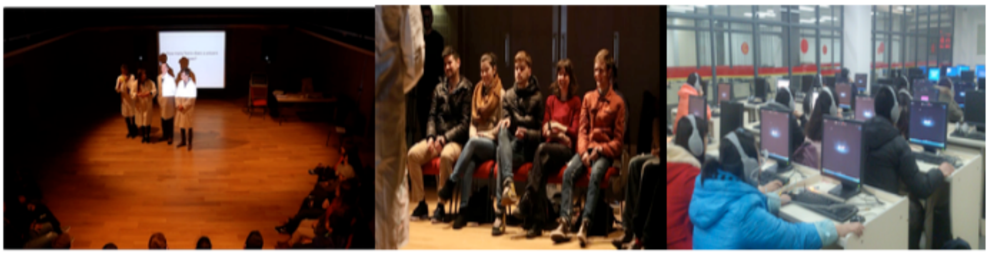


**Figure 25 : Two scenarios are considered: performance and video consumption. The first one (left and middle images) was studied during a theatre play, while the second one (right image) was studied in a large lab session. In both cases, sensor data was collected using exactly the same sensors and software.**

In the studies we used the same home-built, with Arduino, GSR sensors. The wireless module was different in each study: RF12 for the video watching experiment and Xbee for the theatre play (Figure 26). This implied different sample rates due to the different protocols executed at the MAC layer. The data was sampled at 1Hz using polling scheme (theatre play) and 50Hz using ALOHA (video consumption), respectively. Since we knew that ALOHA would bring more collisions, we increased the sample rate in this case. Both settings were extensively tested in the pilot studies.

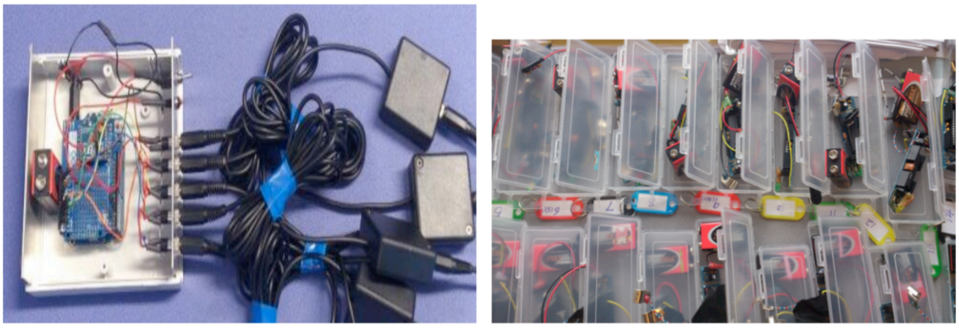


**Figure 26：Data gathering system. Our own built hardware to gather GSR data in the two scenarios (the 2nd version hardware)**

The Xbee wireless module works on 2.4 GHz with relatively short communication range (roughly 10 meters). While RF12 works on 868 MHzwith a maximum range of 30 meters. In both cases, we had a sink node connected with the laptop to receive all the data packets.

All the GSR sensors were tested independently in terms of reliability and robustness before the real experiment. For example, we invited more than 50 users to watch video clips and to play video games, how our GSR sensors performed during these events. Furthermore, we also plotted the data distribution of each GSR sensor, since we know that the data patterns from GSR sensors should be a linear function.

*Participants*

In live theatre play, we recruited 15 users in total: seven females (Mean age = 28.29, SD = 4.85) and eight males (Mean age = 23.13, SD = 8.21). None of them had performance experience before. During the performance, the audience was required to attach the GSR sensor in their left palm. In the lab environment, the two videos were displayed one after the other to two different groups of participants. The first group, Video A, consisted of 15 users: seven females (Mean age = 22.67, SD = 3.01) and eight males (Mean age = 20.3, SD = 1.25). In the second group, Video B, 14 users took part in the experiment: seven females (Mean age = 21, SD = 2.08) and seven males (Mean age = 21.17, SD = 1.47). All the participants had GSR sensor attached in their non-dominant hand during video playback.

The culture and background of the participants in the two scenarios were rather different. In the live theatre play, all the participants were from the UK with different backgrounds. On the contrary, the participants recruited for the lab experiments were undergraduate and master students from one of the top universities in China. We agree that this might be a limitation of the comparison, since culture and background might play a role in the GSR patterns.

*Questionnaires*

In both cases, a pre-questionnaire and a post-questionnaire were provided before and after the experience. The majority of the questions in the post-questionnaires were related to emotions derived from either the theatre play or the videos. In the theatre experiment, questions in the pre-questionnaire were mainly about the type and intensity of the emotions they had experienced during the (working) day. In the video watching experiment, we also examined whether the participants had watched the videos before, and their previous knowledge and experience on video design. The questions were in the form of “Graphic Rating Scales” in which users were asked to make a mark on a line between two extremes, e.g.,

How often did you laugh during the performance?

|________________________________________|

Not at all                                                  Very often

The line measured 100 mm and responses were accurately measured to 1 mm.

*Experimental Procedure*

In the play, within group design was applied. When the audience arrived, they were required to fill the questionnaires. Before the play, we explained to the audience what we were measuring during the performance, and some notes that they should pay attention to, such as not taking off the sensors during the performance. After the performance, they had the post-questionnaires.

In the video consumption, a between group design was conducted, and the experiments were run in the two rounds with the two different group participants. All of them were required to fill the pre-questionnaires before the experiments. In addition to the pre-questionnaires, we explained to the users about the purpose of the experiments, and some actions should be avoided during the experiments, for instance questions. When the first group finished the video, they filled the post questionnaires before they left. After that, we had the second group participants watching the second video.

*Data Analysis*

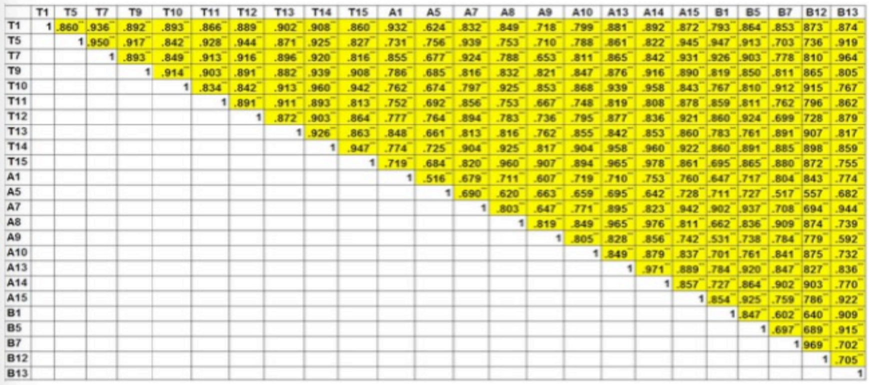
MDS was applied to analyze the sensor data. All the data analysis was done using SPSS. Pearson product-moment correlation coefficient was used to analyze the similarities or dissimilarities between the GSR readings. After that, MDS was applied to visualize the clusters of the responses on a perceptual map. In order to interpret each audience cluster, we took into consideration the subjective data for identifying the actual different types of response.

Regarding to the audience arousal level, we used the first reading coming into the receiver as the baseline for our calculation. We also investigated the arousal level in each cluster, and we displayed this result in Figure 27.

The exhaustive comparison was done by performing several times the Pearson product-moment correlation coefficient. In this way, we could examine whether the same type of responses, i.e., engaged users, was correlated with the sensory patterns across the two scenarios, and thus providing an answer to our research question. Taking into consideration that the two cases (theatre play and video playback) had a different duration, we averaged the time before we performed the algorithm. This averaging procedure did not change the data distribution of sensor readings.
sensor

In the results of Pearson product-moment correlation coefficient (Table 5), we used one star "*" representing 95% confidence level and two stars "**" indicating 99% confidence level. In addition, the overall fit statistics (Kruskal's stress and R Square) in MDS were provided to reveal how the algorithm fitted the input data.

***Table 5:* The correlation of the responses across the red clusters: "liked the performance very much" (*: p<0.05; **: p< 0.01)**



| | T1 | T5 | T7 | T9 | T10 | T11 | T12 | T13 | T14 | T15 | A1 | A5 | A7 | A8 | A9 | A10 | A13 | A14 | A15 | B1 | B5 | B7 | B12 | B13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T1 | 1 | .860** | .936** | .892** | .893** | .866** | .889** | .902** | .908** | .860** | .932** | .624** | .832** | .849** | .718** | .799** | .881** | .892** | .872** | .793** | .864** | .853** | 873** | .874** |
| T5 | | 1 | .950** | .917** | .842** | .928** | .944** | .871** | .925** | .827** | .731** | .756** | .939** | .753** | .710** | .788** | .861** | .822** | .945** | .947** | .913** | .703** | 736** | .919** |
| T7 | | | 1 | .893** | .849** | .913** | .916** | .896** | .920** | .816** | .855** | .677** | .924** | .788** | .653** | .811** | .865** | .842** | .931** | .926** | .903** | .778** | 810** | .964** |
| T9 | | | | 1 | .914** | .903** | .891** | .882** | .939** | .908** | .786** | .685** | .816** | .832** | .821** | .847** | .876** | .916** | .890** | .819** | .850** | .811** | 865** | .805** |
| T10 | | | | | 1 | .834** | .842** | .913** | .960** | .942** | .762** | .674** | .797** | .925** | .853** | .868** | .939** | .958** | .843** | .767** | .810** | .912** | 915** | .767** |
| T11 | | | | | | 1 | .891** | .911** | .893** | .813** | .752** | .692** | .856** | .753** | .667** | .748** | .819** | .808** | .878** | .859** | .811** | .762** | 796** | .862** |
| T12 | | | | | | | 1 | .872** | .903** | .864** | .777** | .764** | .894** | .783** | .736** | .795** | .877** | .836** | .921** | .860** | .924** | .699** | 728** | .879** |
| T13 | | | | | | | | 1 | .926** | .863** | .848** | .661** | .813** | .816** | .762** | .855** | .842** | .853** | .860** | .783** | .761** | .891** | 907** | .817** |
| T14 | | | | | | | | | 1 | .947** | .774** | .725** | .904** | .925** | .817** | .904** | .958** | .960** | .922** | .860** | .891** | .885** | 898** | .859** |
| T15 | | | | | | | | | | 1 | .719** | .684** | .820** | .960** | .907** | .894** | .965** | .978** | .861** | .695** | .865** | .880** | 872** | .755** |
| A1 | | | | | | | | | | | 1 | .516** | .679** | .711** | .607** | .719** | .710** | .753** | .760** | .647** | .717** | .804** | 843** | .774** |
| A5 | | | | | | | | | | | | 1 | .690** | .620** | .663** | .659** | .695** | .642** | .728** | .711** | .727** | .517** | 557** | .682** |
| A7 | | | | | | | | | | | | | 1 | .803** | .647** | .771** | .895** | .823** | .942** | .902** | .937** | .708** | 694** | .944** |
| A8 | | | | | | | | | | | | | | 1 | .819** | .849** | .965** | .976** | .811** | .662** | .836** | .909** | 874** | .739** |
| A9 | | | | | | | | | | | | | | | 1 | .805** | .828** | .856** | .742** | .531** | .738** | .784** | 779** | .592** |
| A10 | | | | | | | | | | | | | | | | 1 | .849** | .879** | .837** | .701** | .761** | .841** | 875** | .732** |
| A13 | | | | | | | | | | | | | | | | | 1 | .971** | .889** | .784** | .920** | .847** | 827** | .836** |
| A14 | | | | | | | | | | | | | | | | | | 1 | .857** | .727** | .864** | .902** | 903** | .770** |
| A15 | | | | | | | | | | | | | | | | | | | 1 | .854** | .925** | .759** | 786** | .922** |
| B1 | | | | | | | | | | | | | | | | | | | | 1 | .847** | .602** | 640** | .909** |
| B5 | | | | | | | | | | | | | | | | | | | | | 1 | .697** | 689** | .915** |
| B7 | | | | | | | | | | | | | | | | | | | | | | 1 | 969** | .702** |
| B12 | | | | | | | | | | | | | | | | | | | | | | | 1 | .705** |
| B13 | | | | | | | | | | | | | | | | | | | | | | | | 1 |

All the GSR sensor data were post-processed through the smoothing and filtering Matlab function, in order to minimize the impact of hand movements. However, we found that thanks to the well-prepared design of the experiments, the data were of high quality. Overall, there was no significant difference in the data before and the data after the smooth procedure.

*Audience Clustering*

In Figures 26, we display the MDS results from the three user studies. We used T plus a numerical number (sensor id), e.g., T3, to represent the participants from the theatre play, and either A (the video A: fitness product) and B (the video B: coffee brand) accompanied with the numerical numbers (sensor id) to represent to a user who joined the lab experiments. We used four different colors (red, yellow, green and orange) to distinguish the different clusters, where the same color in all the figures represents the same category of responses.

Figure 27 (left) displays the four audience clusters in the theatre play, showing the different experiences. Users in the red clusters reported the highest scores in the post questionnaires in terms of cheerful, enjoy and like. On the other hand, participants in the green clusters scored the lowest in the questionnaires.

In addition to the questionnaires, subject T3 in the interview told us that he did not like the play at all, and this can be seen in the recorded video. Users in the yellow and orange clusters had a different experience: T2 and T8 took a while to get into the performance, which might imply that they did not understand the beginning of the performance. In contrast, T4 and T6's attention waned in half through the performance, one of the users in the interview stated that she started to recognize one of the actors at certain moment of the play and that is why her attention shifted.

Figure 27 (right) displays the clusters from the responses of users watching video A. In this particular case, we only observed three types of feedback. Similarly, the red cluster represents users that rated the highest scores in terms of immersion, attention level, and concept design (encouraging them to purchase the product). By contrast, participant A11 (the green dot) was not so interested in buying the product, and he labeled his attention level as the lowest possible. Interestingly, the participants in the yellow cluster reported that they had no previous experience, and their knowledge was limited on the video design.

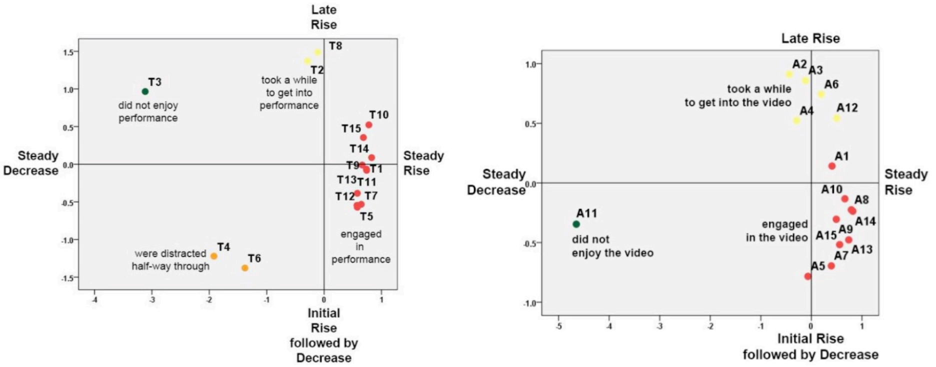


**Figure 27: (left) Feedback clusters for the theater play (Stress: 0.03, RSQ: 0.99); (right) Feedback clusters for Video A (Stress: 0.03, RSQ: 0.99)**

In Figure 27 (left), we find four clusters. Similarly, the participants in the red cluster rated the highest score in terms of immersion, and attention level; the participant B10 (the yellow point) reported that he had no previous experience and knowledge on video design. During the group interview, the students from the orange clusters were rather active, and show interest in this video, but they all reported a busy day and this might explain why their attention declined after a while after the video started playing.

Figure 27 (right) shows the arousal level in each cluster from the two user studies. Obviously, the theatre play evoked a higher arousal compared to consuming video. On the other hand, the participants from both the red cluster and the orange one all had positive arousal levels, which were much higher than the ones from the rest of the clusters: the arousal levels of the green clusters were all negative values, and the yellow clusters had relevant low arousal scales, displaying a negative value for video B.

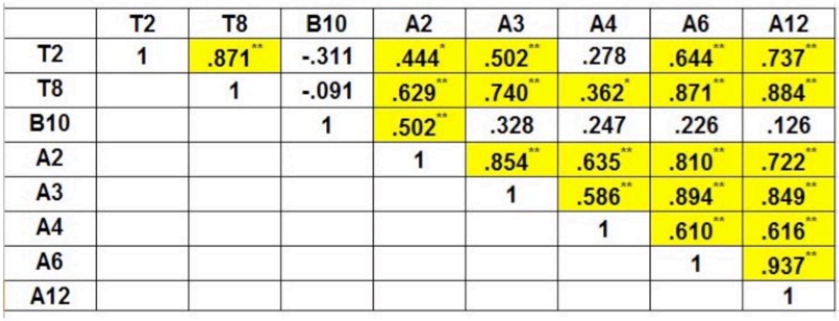
By observing the distribution of data in the four clusters in both studies, we found that there were roughly four different patterns: a steady increase in the red clusters; a steady decrease in the green clusters; a late rise in the yellow clusters; and initial rise followed by a decrease in the orange clusters. These descriptions were used to label the figures generated by MDS.

Through the combination of the GSR feedback and the subjective data, we could clearly identify two main types of responses: the engaged audience versus the non-engaged audience, which were in the red clusters and the green clusters respectively. However, with respect to the users in the yellow and the orange clusters, we concluded that personal reasons interfered their watching experience.

We can assume that participants in the same color cluster had a similar watching experience. Our interest was on examining whether their GSR sensory patterns showed correlations across the two scenarios.

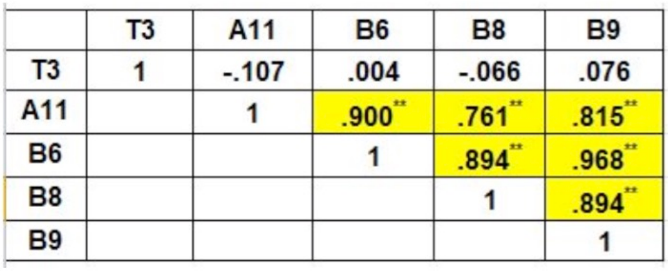
Regarding the red cluster, we were surprised to see that all the users' GSR responses (24 users) were all significantly correlated with each other (Table 5): averaging 0.831. We can safely conclude that the sensory patterns were strongly synchronized across the scenarios.

**Table 6: The correlation of the responses across the green clusters : "did not like the performance" (*: p <0.05; **: p<0.01)**



| | T2 | T8 | B10 | A2 | A3 | A4 | A6 | A12 |
|---|---|---|---|---|---|---|---|---|
| T2 | 1 | .871** | -.311 | .444* | .502** | .278 | .644** | .737** |
| T8 | | 1 | -.091 | .629** | .740** | .362* | .871** | .884** |
| B10 | | | 1 | .502** | .328 | .247 | .226 | .126 |
| A2 | | | | 1 | .854** | .635** | .810** | .722** |
| A3 | | | | | 1 | .586** | .894** | .849** |
| A4 | | | | | | 1 | .610** | .616** |
| A6 | | | | | | | 1 | .937** |
| A12 | | | | | | | | 1 |

**Table 7: The correlation of the responses across the yellow clusters : "took a while to understand the performance " (*: p <0.05; **: p<0.01)**



| | T3 | A11 | B6 | B8 | B9 |
|---|---|---|---|---|---|
| T3 | 1 | -.107 | .004 | -.066 | .076 |
| A11 | | 1 | .900** | .761** | .815** |
| B6 | | | 1 | .894** | .968** |
| B8 | | | | 1 | .894** |
| B9 | | | | | 1 |

For the orange and yellow clusters, the GSR responses were partially correlated (Table 6 &Table 7). For instance, the GSR response of participant T4 was strongly correlated to all the users that attended the video experiment, user T6 was correlated to most of the users in the orange clusters. Nevertheless, both T2 and T8 were both correlated to most of the users located in the yellow clusters of the video consumption.

For the green clusters on Table 4, we found that the correlation was not strong. Correlations only existed between the two watching video experiences: averaging 0.872, p< .01, and there was no significant correlation between video consumption and theatre performance. In addition to the correlation checking, we also found that the GSR response of T3 experienced fluctuations during the performance, although all the users displayed a steady decrease on their sensory pattern.

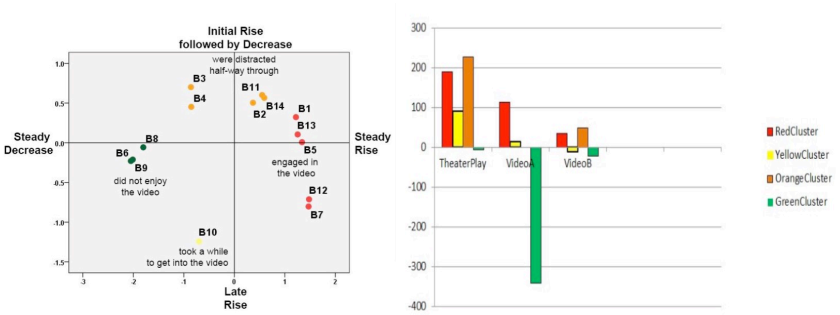


**Figure 28: (left) Feedback clusters for Video B (Stress: 0.04, RSQ: 0.99); (right) Arousal levels from the clusters in each experiment**

We performed MDS on the GSR responses of the non-engaged participants (Figure 28 (left)) and the GSR response of all them (Figure 28 (right)), in order to investigate the distance, in a perceptual map, between the non-engaged participants and the proximity between the non-engaged ones and the rest of the them.

In Figure 28 (left), we found that the distance was large between the responses of T3 and the responses of the other users. This result is consistent with the results displayed in Table 6.

In Figure 28 (right), we could clearly see a massive cluster formed in the left part of the map, mainly coming from the red, orange and yellow clusters. In contrast, the responses from the non-engaged video consumers formed a cluster on the right side of the map, where the green points were closed to each other. Regarding the responses from subject T3, his geometrical location was more adjacent to the orange clusters, and positively correlated to the responses from B2: 0.549, $p < 0.01$; negatively correlated to the ones from B10, A2 and A3: averaging -0.501, $p < 0.01$.

*Discussion*

In this study, we have compared the GSR responses from participants in two different use cases. In particular, we used MDS and CA to successfully classify audience responses in clusters and subjective data (interviews, questionnaires) to interpret what each cluster represents. Based on these techniques, we could differentiate between engaged and non-engaged participants, and then perform exhaustive comparison across the two use cases.

We found that the responses from the engaged users showed a strong correlation on their sensory patterns between lab and field studies. Interestingly, the responses from the non-engaged participants did not correlate across use cases (between lab and field trial), but correlated between the two lab experiments. This result is consistent with a similar phenomenon mentioned in previous research: a “boredom” state captured in a lab may have the different patterns compared to the one in a field study. Still, these previous studies did not quantify such results and did not report any comparative data for the more engaged users.

Even though we could use “boredom” to define the state of the participants in the green clusters, we decided to apply the more general term, non-engaged, to define this type of responses. We followed the learning from our previous studies: it is unlikely to generate a boredom state when people are watching short videos in a lab situation, since every participant might take the task seriously. They typically try to understand what it is happening in the video, ignoring its quality even if the video is in an unknown language. On the other hand, “boredom” is a state that can instead happen in longer field trials. Therefore, we preferred to refer the responses in the green clusters as non- engaged, which better describes the cases for the lab situation and the field trial.

In our studies, we only had 5 out of 44 users with non- engaged responses. We certainly may require larger amount of this type of responses to better understand how similar are sensory patterns across a lab study and a field trial. Nevertheless, it is almost impossible to estimate the number of non-engaged users that would result from an experiment, even though a large number of participants (44 in our case) are involved.

It would be a strong backup to our findings if we can provide some results of the subjective evaluations, for example whether the questionnaires of the engaged users were also strongly correlated. However, in our case we had some difficulties to run the correlations across the two scenarios. First, the questions designed in the two use cases were not exactly the same ones, so that it is not possible to correlate the answers from the one participant from the lab experiment with one from the theater play. Second, the correlation method requires the sufficient and equal length of data sets to run the algorithm, but in questionnaires each participant gave one score for each question, which is unlike the sensor data: each participant had a sequence of sensor readings (from the beginning of the theater play/video consumption to the end). Therefore, the subjective data crossing the two scenarios cannot satisfy these requirements, but we can compare the scores of engaged users to the non-engaged ones. Besides that, the video recordings and the interviews also helped us interpret the clustering results of the MDS and the CA.

We believe that comparing a video - recorded performance and a live performance is another interesting research topic. In our case, we had the different experimental settings: the theater play versus the video consumption. However, we think that the results are valid and innovative. First, the sensory patterns of engaged users are strongly correlated crossing the different scenarios, even though the experimental settings are different, and this phenomenon has never been mentioned in the previous studies. Second, more interestingly, the sensory patterns of the non-engaged users are different, which motivates us to take a further step to investigate the reasons that may cause this result. For instance, whether the experimental settings have effects on the non-engaged users, or the matters of the performance itself, as we have already seen the sensory patterns of the non-engaged users were strongly correlated in the lab experiment. At the current stage, we will leave this box open for the future exploration.

The methodology we have applied for reporting the results does not require intentional inputs from the participants, in order to guarantee high quality of the sensor data. However, it is still a challenge to interpret sensor data, which requires well-designed experiments, questionnaires, proper-organized interviews, and high quality of video recordings.

## 4.2 Benchmarking of the SCR and SCL Signals

GSR, is also known as galvanic skin response, electro dermal response (EDR), psych galvanic reflex (PGR), skin conductance response (SCR), or skin conductance level (SCL). GSR sensors measure the users' electrical conductance of the skin, where users' sweat glands are varied and controlled by the sympathetic nervous system. Therefore, GSR sensors are normally considered an indicator of psychological or physiological arousal or stress. When users are highly aroused, users' skin conductance is increased in turn. In general, there is a delay of 1-3 seconds between stimulus and SCR onset. Hands and feet can be used to measure GSR, as there the density of the sweat glands is the highest.

According to the physiological attributes of GSR signals, we can use SCL and SCR to process the data. When we conduct SCL analysis, the first sensor readings of each participant were used as the baseline, which was then subtracted from the raw data, to remove individual differences. Then, the Pearson product-moment correlation coefficient was used to check whether there was a significant correlation between the responses from the audiences at different locations. In addition, a t-test was used to compare the SCL data of live audience members and remote audience members.

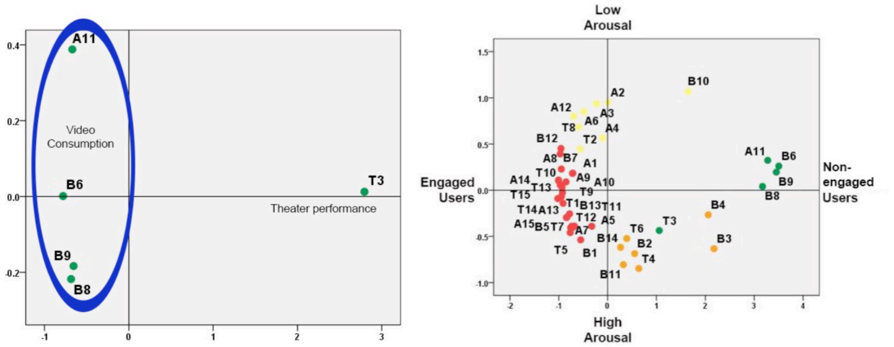


**Figure 29: (left) MDS result when applied to all the responses in the green clusters (Stress: 0.05, RSQ: 0.99); (right) MDS result when applied to all the responses across the two scenarios (Stress: 0.03, RSQ: 0.99)**

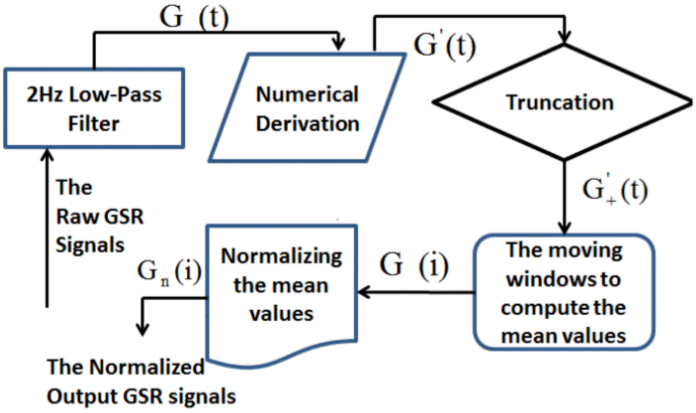

G (t)
G'(t)
2Hz Low-Pass Filter
Numerical Derivation
Truncation
G'+(t)
The Raw GSR Signals
Gn(i)
Normalizing the mean values
G (i)
The moving windows to compute the mean values
The Normalized Output GSR signals

**Figure 30 : The description of the different steps of the algorithm for processing the raw GSR signals**

We adapted the method developed by Fleureau [29] by observing SCR signals invoked by stimulus (Figure 30). The method first extracts the SCR signals based on the timeline, and then defines which SCR responses were statistically significant when compared to the background noise. In such a manner, we can provide links between the significant audience bio-responses and the related performing events (e.g., the appearance of Joey), so our stakeholders can have a better understanding in terms of which events substantially affected the audience and which ones did not.

**Equation 1: The smoothing procedure on the GSR signals**

$$G_n(i) = \frac{G\ (i)}{\sum_{j=1}^{k} G\ (j)} = \frac{\int_{W_i} G'_+(t)dt}{\sum_{j=1}^{k} \int_{W_j} G'_+(t)dt}$$

Extracting the SCR signals from the sensors requires signal processing, filtering, smoothing, and derivative procedures, as shown in Heuristics: 3.3. First the raw signals are scaled and filtered by a 2Hz low-pass filter (G(t)). After that, the numerical derivation performed, but only the positive phasic changes are kept while the negative ones are ignored. The reason is that the negative phasic changes are not of our interest, as they only reflect the physical recovering from a stimulus. In the next step, we apply a moving window (Wi,Wj :i, j = 1......k, where k is the number of windows) with the window size of 120 samples (30 seconds), and the overlapping size of 60 (15 seconds) to smooth and compute the mean values of the derivative signal G (i) . Finally, (since each individual may have a different amplitude for the derivative GSR signals), we normalize the computed signals by using the sum of subsampled skin response values as the denominator to calculate the individual mean arousal value (Gn(i): n=1......N, where N is the number of users) (2).

The bilateral Mann-Whitney-Wilcoxon (MWW) test is performed to define the significant SCR responses (Equation 1). The mean p values are computed by averaging the p-values of the bilateral MWW test performed between the latent unknown distribution of Gn (i) and the background noise. We define 10% of the computation results with the lowest mean as background noise. Only an associated p value below 5% is considered as significantly different from the background noise.

The process on both SCL and SCR can be done both offline and online. For a real-time application development, we calculate the SCL value based on the defined time duration, and the baseline is default as the first sensor data reading. While for the significant SCR moments, we count how many significant SCR events appeared in a pre-defined time domain. Both parameters can be visualized through a graphical design and be vividly presented to end users. This work directly addresses the research question 5 and 7.

### 4.2.1 User Study: Shanghai War Horse Experiment

Research has shown that physiological sensors provide a valuable mechanism for quantifying the experience of audiences attending cultural events. In comparison to alternatives (the strength of applause or questionnaires), bio-sensors provide fine-grained timed data that can be used to infer the quality of the experience of the audience members. Unfortunately, available commercial sensors are designed for lab or home usage and studies, focusing on the individual and not on the reactions of a crowd of people. Audience research calls for a robust physiological measurement system that overcomes the challenges of using and deploying sensors in a theatrical environment (anonymity and privacy, real-time gathering of data, support for large crowds of 30-100 people). In this work, we report that our GSR sensors (and network) can be used to simultaneously measure audience response and successfully capture the sufficient data to understand the reactions from the audience.

Arts add value to the lives of individuals and to society as a whole. Attendance and participation in arts can help individuals and society develop internal cognitive and emotional process. As a result of a cultural experience, such processes can have impact on external outcomes, e.g., increased educational attainment, reduced crime rates, health and overall well-being [26]. Besides, arts engage audience at the emotional and intellectual as well as the aesthetic level. Understanding how the communication from artists to audience works can help artists appreciate the value of arts.

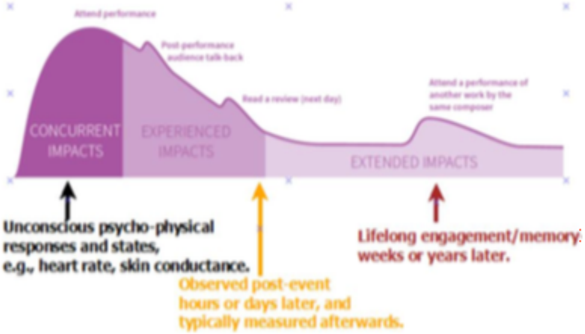

Attend performance
Post performance audience talk back
Read a review (next day)
Attend a performance of another work by the same composer
CONCURRENT IMPACTS
EXPERIENCED IMPACTS
EXTENDED IMPACTS
Unconscious psycho-physical responses and states, e.g., heart rate, skin conductance.
Observed post-event hours or days later, and typically measured afterwards.
Lifelong engagement/memory: weeks or years later.

**Figure 31: Three discernable types of impacts of arts defined by their temporal proximity to a cultural event.)**

The literature defines the impacts of arts on an audience as a progression of three stages based on their temporal proximity to the cultural event: concurrent impacts, experienced impacts, and extended impacts [12]. As shown in Figure 31, concurrent impacts occur during the event, and can be measured through biometric research. Experienced impacts can happen before, during and after the event, and are typically measured through post-event surveys and interviews. With respect to extended impacts, they are measured through retrospective interviewing and longitudinal tracking studies.

Interestingly, the majority of the previous studies used subjective methods, (e.g., questionnaires [7, 11], interviews [50, 32], labeling systems [11, 9]) for measuring both concurrent and experience impacts. This makes the boundaries between the two different types of impact blurry. Individual's concurrent impact may not be always tracked by using subjective measurements, as individual may respond to cultural events without being consciously aware of it, and any conscious reflection (e.g., using a labeling system during a performance) on the individual's state may interfere with the experience (i.e., interrupt their sense of flow or absorption).

Physiological and pre-cognitive psychological responses can be used to measure the aesthetic experience, and the biological functions provide objective measures of concurrent impacts [20]. Besides, physiological responses can be measured non-intrusively at the very moments at which they occur, because intellectual and emotional reactions from the audience can take place at any time during the events. A number of studies have sought to capture biometrical data as a means of gauging audience responses to arts, by tracking the eye movements [43], heart rate [47], skin conductance [48], emotional response [22] and aesthetic judgments [48] during performances. However, the complexity of setting up the biosensor infrastructure, the bulky form factors, and high prices of the commercial products limit their widespread usage in theaters.

Biometrical data can also bring liveness to the performance. Some artistic productions explore appropriate manners to enhance the audience experience by visualizing their responses or by offering interaction. During the event, audience may hold their breath, their heart rate and skin conductance may increase, they may also lose track of time, or experience chills. Such non-tangible experiences from the audience are difficult to be conceptualized. It is possible that through the manipulation of biometrical data, the audience experience can be vividly represented in the real world.

Commercial GSR sensors are not suitable for theaters. First, they are generally designed for lab studies, using technologies like Bluetooth which are intended for individual rather than for group users. Second, several commercial sensors use radio technology [1, 2], and some use a SD card to locally store the sensor data [14, 19], but none of them is particularly designed for audience members to wear in theaters, and there are no algorithms that can be used for real-time visualizations of data. Third, some commercial smart watches have sensors integrated. In this case, the anonymity of the sensor signals is not preserved (since the mobile phone may have to sign on to use the web server). Instead, we foresee that theaters companies will provide the audience their own lightweight sensors and deploy a specific network for the purpose of better understanding and quantifying the audience experience.

Some studies in the wild used customized sensors to fit the user needs (i.e., customized sensors for monitoring elderly people at home [15, 39]) or for long term ambulatory field studies [40]. However, such customized hardware is not suited for simultaneously measuring audiences, as the majority of them still use Bluetooth technology intended for individual measurements.

**Figure 32: Our physiological sensors were used during the War Horse performance Chinese version (the 3$^{rd}$ version hardware).**

Motivated by the intrinsic and instrumental benefits from the arts, we collaborated with the China National Theater Company in particular for the Chinese version of War Horse (Figure 32). The main contribution of this study is to report on the developments needed for setting up a biosensor network for studies in the wild in theaters. By using such a sensor network, audience bio-response can be used for both real-time applications and offline analysis. In this study, we tested the third generation of GSR sensors. The knowledge gained from this work will be beneficial for researchers who have similar interests in this creative field.

*Participants*

We recruited 90 participants (Male: Avg.: 33, STD: 12.4; Female: Avg.: 28.9, STD: 8.04) who joined 3 rounds of experiments (the same performance but at the different date). Before the performance started, we explained the purpose of the experiment, and provided pre-questionnaires and consent forms. After this, the audience wore the sensors at their non-dominating hand, and they went inside the theater 15 minutes before the performance. All the seats for the participants were reserved. There was a 15 minutes break between the first part and the second part of the performance. When the performance finished, we helped the audience members take off the sensors. After that, there were post-performance questionnaires and interviews.

*Interviews*

This research was conducted in collaboration with the National Theater of China, aiming at better understanding what are their needs for the next-generation audience monitoring systems.

We interviewed (semi-structured) the producer, the director, and the stage manager from China National Theater Company. The interviews lasted 30 minutes, and video and audio were recorded. Five researchers took observational notes during interviews. After the interviews, researchers discussed the observational notes and identified the potential interesting topics. More specific research questions were concluded after the discussion.

The Chinese version of War Horse is the first co-production between the National Theatre of Great Britain and the National Theatre of China. The original British version was first premiered in London in 2007, and the show has been seen by over

6 million people worldwide and been performed over four thousand times. In the year of 2015, War Horse took its biggest jump into China.

**Figure 33：The interview with the producer Li Dong**

Li Dong, producer of the Chinese version of Warhorse, together with the stage manager and the director, were interviewed in Shanghai, China (Figure 33). Li Dong stated that there is a need for quantified audience feedback to resolve ambiguity. For instance, he was unsure of the impact of the Warhorse performance in China. In England, the production was popular for all ages. Even though the script was adapted from a children's novel, it dealt with history and complex emotions. Therefore, he was not sure how audiences in China would receive the show. Furthermore, he felt that there is a need for concrete understanding of audiences' reception rather than their interests in how many performance scenes related to Joey (a horse, the main character of the play) would arouse the Chinese audience. With this understanding, they could better plan marketing strategies and promotional-related materials. Besides, both producer and director felt that a quantified and effective audience feedback could help them make informed choices. In particular, Li Dong mentioned that almost 70% of the theatregoers in China are females and they bring along their male friends and children to watch a show, as such, it will be useful to know what elements of the show are engaging for females and the difference in engagement level between female and male audience members. In addition, by understanding the elements of a play that are engaging the majority of the audience (e.g., the thrilling climax scene about the competition between Joey and Topthorn,), the director can make more astute choices in the later production and create more successful performances. Last, all of them stated that a quantified audience experience should not limit artistic expression but can help to form the basis of certain choices. For ambiguous perception of how popular the show is. For instance, they all mentioned instance, selecting elements or narrating the story from which character's point of view, it may help generate audience interest. Besides, they foresee the possibilities to use audience response in real-time to explore new performing areas (e.g., an interactive play that is particularly designed to have the different endings based on the feedback of audience).

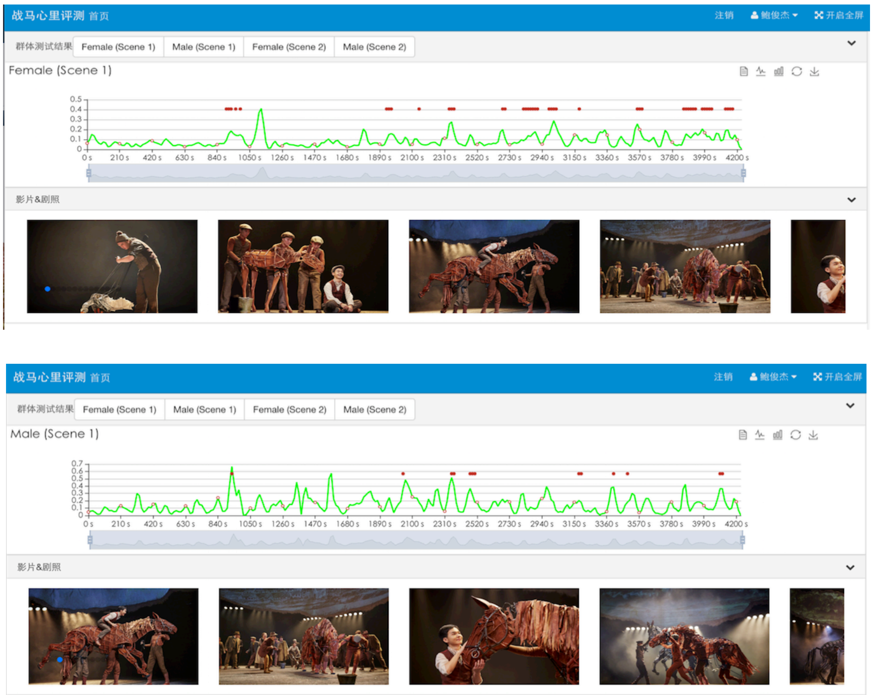


**Figure 34：The significant arousal events compared between the female audience and the male audience in the scene 1.**

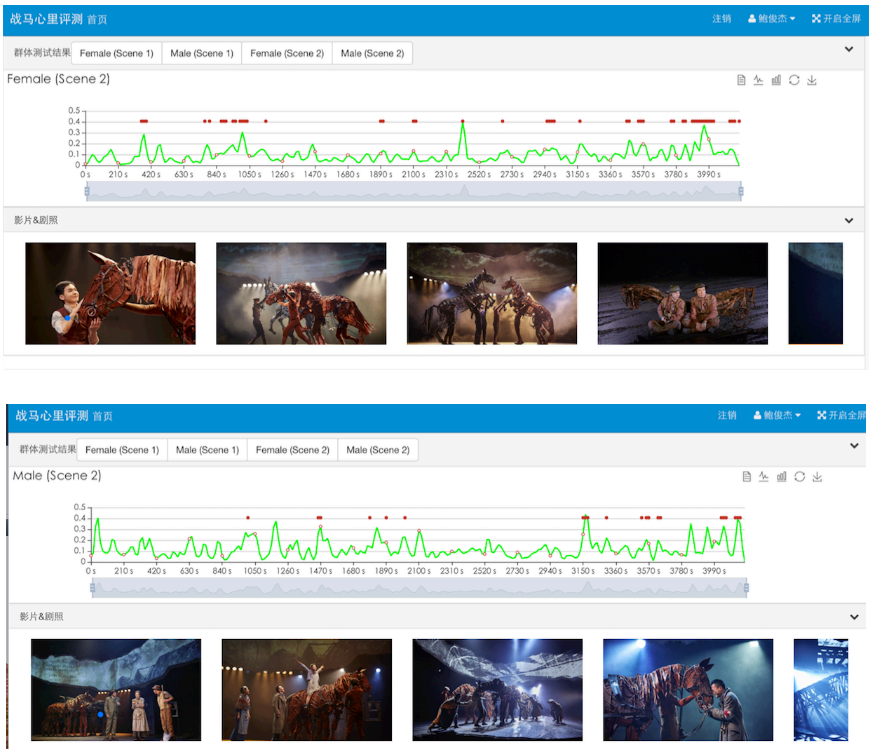


**Figure 35: (top) The significant arousal events compared between the female audience and the male audience in the scene 2.**

*Results*

As shown in Figure 34 & 35, we found that both male and female participants were significantly aroused during the thrilling scene where the competition between the two main protagonists Joey and Topthorn happened, but the duration of this significance lasted longer for females than for males, as indicated by the length of red bars. In addition, there were 226 scenes related to Joey. The audiences were significantly aroused in about 32 of those. Interestingly, by observing the p values, we found that in some of the scenes both male and female audiences were stimulated, but in some of the scenes females were more responsive than males. By comparing the results from the first and the second part (Figure 34, 35), we found that significant responses from females were more intense compared to the ones from males. Besides, the duration of the responses of the female audiences was generally longer than those of the males.

The results on the family audience were surprisingly interesting. Based on the marketing investigation results from China theater company, the majority of the parents thought that the performance would be difficult to understand for children. However, based on our results, we found that the emotional intensities elicited from the children audiences were significantly stronger than their parents (Figure 36 & 37). Besides, from the subjective reports and the interviews with the children, they enjoyed the performance even though they might not fully understand. In comparison, the parental experience revealed that they had less emotional intensities compared to the children. According to their report, they were extremely tired during the day time, because they had to take care of their kids before they went to the theater. During the show, they were distracted by their children instead of enjoying a performance; some of them even slept during the play.

The sensor results also showed that during the whole performance parental audiences had less arousal moments compared to their children during the whole performance. For instance, there are in total 226 scenes where the Joey appeared: the children were significantly evoked by 73 ones, while their parents were only stimulated by 35 ones. Furthermore, compared to their parents, the children also had more substantial responses towards the emotional scenes when compared to their parents. For instance, there are in total 35 emotional scenes, where the children were statistical elicited by 13 ones, which was as twice as their parents.

The results also guide the theater company to adjust their marketing strategy. Before the experiment, the theater company arranged a visiting event to the back-stage exhibition, where Joey and Thompton, and the family audience can have a direct contact with the horse puppet. One of the group participated in the visiting events before they watched the show, and their emotional intensities from the sensor readings were significantly reduced compared to the other. We assume that the familiarity has decreased the emotional intensities during the watching experience after visiting the back-stage exhibition. Based on our reports, the theatre company has adjusted their marketing strategy: inviting visit at the back stage only after the show.

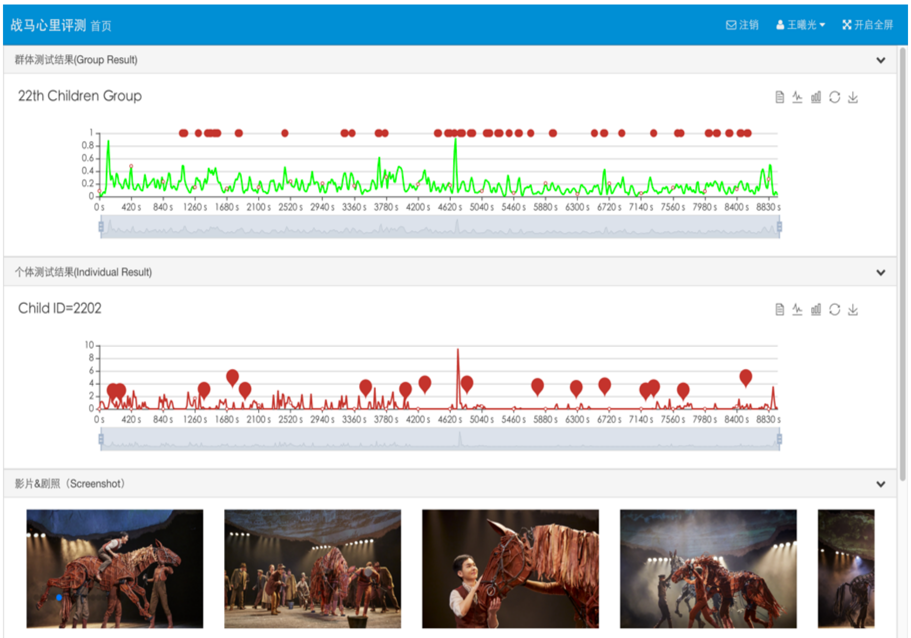


**Figure 36: the visualization results based on the performance of 22th, November. The first row displays the significant arousal moments from the group children. The second row shows the significant performing events evoked from the child id (2202, his parental id: 2206 in Figure 10). By clicking the red dots in the second row, the related performing scenes (the screenshot) will be displayed in the below.**

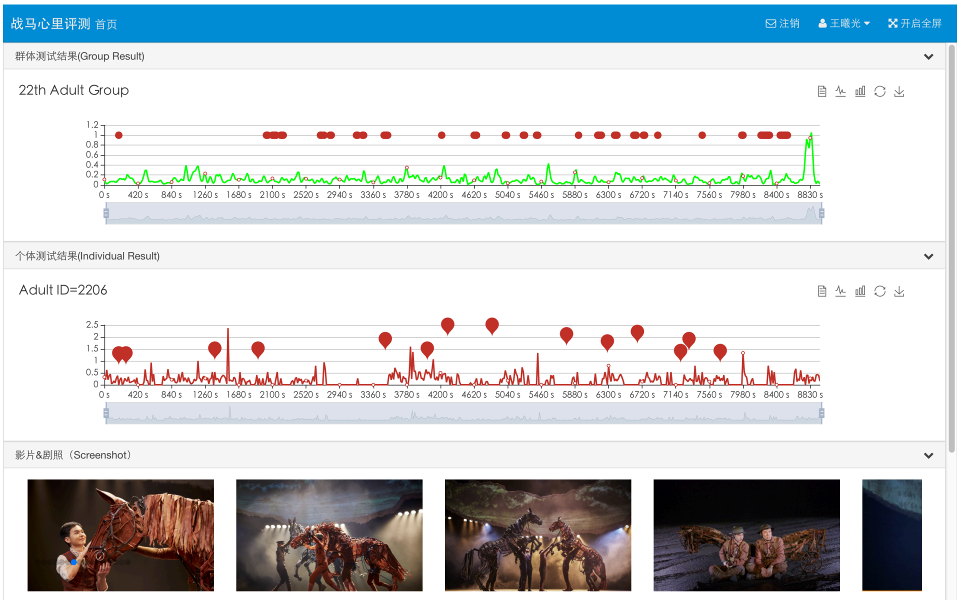


**Figure 37: the visualization results based on the performance of 22th, November. The first row displays the significant arousal moments from the adult group. The second row shows the significant performing events evoked from the parental id (2206). By clicking the red dots in the second row, the related performing scenes (the screenshot) will be displayed in the below.**

We invited theater experts to interpret the results. In general, they write journals on a performance based on their own feedback, or they interview several theatergoers in order to obtain the audience opinions. Even though sometime experts combine the internet information and generate a report, but such manner has a big delay. The traditional way of collecting data is based on small samples. If big samples are required, they have to sort out other methods, e.g., spreading out questionnaires. Still, for news reporting, the timing is extremely important. The experts think that our method can provide rich and timing data for them to write reports.

The results from the gender difference helped the experts prove their assumptions on audience watching experiences. In general, the female audience is easily touched by the emotional stories, while male is more linked to action scenes. The experts pointed out that the results revealed that the emotional scenes triggered the female audience the most, while the male audience's interest was more aroused in the action scenes in the first part of the show. While in the second part of the show, both genders were more connected to the ending in which the soldier Albert met Joey again, but the important scene describing the girl Emilie and Joey significantly elicited the female audiences.

The producer Li Dong received our interview again. He was extremely excited about the results. He told us that this method well solved the contradictions between him and the directors, as sometimes they may have different thoughts about how a performance should be constructed. The results also stimulated them to be more creative in the future artistic production, i.e., creating an interactive performance, where sensors can be used for visualization or interactions. The most important inspiring thing is that this method could help him define a marketing strategy. For instance, if he knew the children were intensively attached by the show, he would have invested the budget to promote on the children market.

### 4.2.2 User study: Distributed Performance at Falmouth

Accurately measuring the audience response during a performance is a difficult task. This is particularly the case for connected performances. In this study, we staged a connected performance in which a remote audience enjoyed the performance in real-time. Both objective (galvanic skin response and behaviors) and subjective (interviews) responses from the live and remote audience members were recorded. To capture galvanic skin response, a group of self-built sensors was used to record the electrical conductance of the skin. The results of the measurements showed that both the live and the remote audience members had a similar response to the connected performance even though more vivid artistic artefacts had a stronger effect on the live audience. Some technical issues also influenced the experience of the remote audience. In conclusion, we found that the remoteness had little influence on the connected performance if there is a channel by which they can interact with the artists.

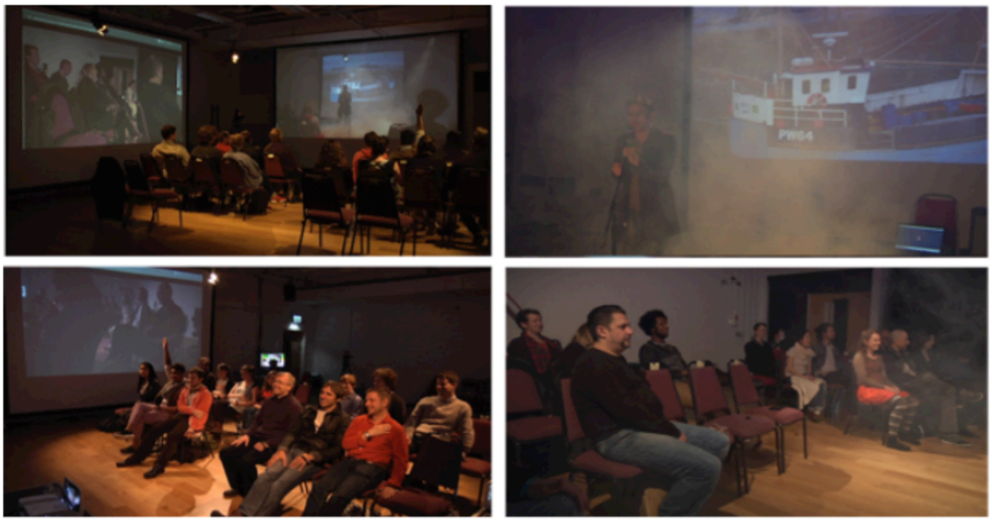


**Figure 38: The performance: the two photos on the left show the remote location (the performance and the live audience were both displayed on the big screens (top) and the remote audience actively interacted with the actor (bottom)); two photos on the right were taken at the live location (the artist with special effect smoke (top) and the audience watching the play (bottom))**

One-Way delivery of live theatre performances to cinemas or other theatres is a relatively recent phenomenon, as well as still relatively small-scale. However, it has already been a commercial success for well-funded companies using expensive and not readily available infrastructure (e.g. satellite communication). For example, the National Theatre in UK often applied NT Live technology to broadcast live performances to digital cinemas. The long-term vision is that over the next few years, smaller companies will follow suit to reach wider audiences beyond their local community. In addition, we foresee the development of the technology to enable remote audiences to play a much bigger role during live performances. Remote audiences may interact with performers across space and provide feedback, promoting a sense of audience community on a larger scale.

A number of previous studies focused on how to enable connected performances to engage the audience [30, 13]. However, audience response to connected performances has only been investigated in a few papers. Our study was conducted to better understand the effect of remoteness on audience experience during a connected theatre performance. This is a first step in evaluating audience response to connected performances. This work aims to address the research question 4.

To answer this question, an experiment in highly realistic conditions was conducted. Together with a small theatre company, exploratory work was done on synchronous watching (live streaming) of one theatre play, which was called "Styx Boat on the River". It was staged at the University of Falmouth in Falmouth, United Kingdom. The performance was live streamed to another studio located at the same building, which meant that the audiences at the two locations watched the same performance at the same time (Figure 38). The experience of the audience was captured by our second version of GSR sensors, video recordings, and interviews.

*Participants*

All the participants were recruited at the university, and they all were university staff without any visual or acoustic problems. There were 12 audience members in each location (24 participants in total).

*Stimuli and Apparatus*

The performance for this experiment was carried out by a single actor. The play, called "Styx Boat on the River", was interactive including a number of pieces like singing, effects using theatrical smoke and a vacuum cleaner sound effect. The whole performance lasted 25 minutes. For this experiment, we used the second generation GSR sensors, which were worn at the user's palm, holding the electrodes.

Both, actors and audience members were interviewed after the performance. The interview of audience members mainly focused on three parts: the overall evaluation of the performance and the reasons behind their opinions, the awareness they felt to the actors, and the awareness they felt to the audience at the other location. The interview with the actors discussed the overall evaluation of the performance and the reasons behind their opinions, and how they felt with respect to the audience.

*Other Apparatus and Software*

The performance was live streamed to another performance studio located in the same building, which meant that the audiences at the two locations watched the same performance at the same time. The technical research team developed the live streaming system. There were three cameras deployed in total, so that the remote audience could see the actor and the live audience through three projector screens. At the live venue, there were only two projector screens installed, so that the actors could see the reaction of the remote audience during the performance (Figure 39).

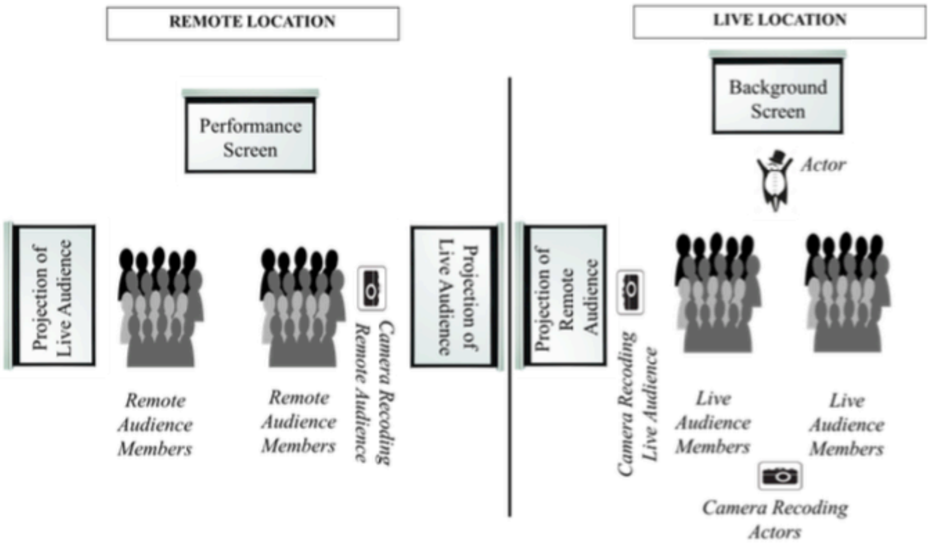


**Figure 39: Conceptual sketch of the experimental facilities at each location. Left: at the remote location, there was one screen showing the performance from the live location in front of the audience members. Another two screens, which showed the live audience members, were both on their left and right. One camera at the right of the audience was used to record them. Right: at the live location, the actor was performing in front of the live audience. There was a camera in the back, which recorded the performance. The projection of the remote audience was placed in a screen on the left of the live audience. The camera recording the live audience was on the left side. This set up allowed that audience at both locations felt as if they were in the same space.**

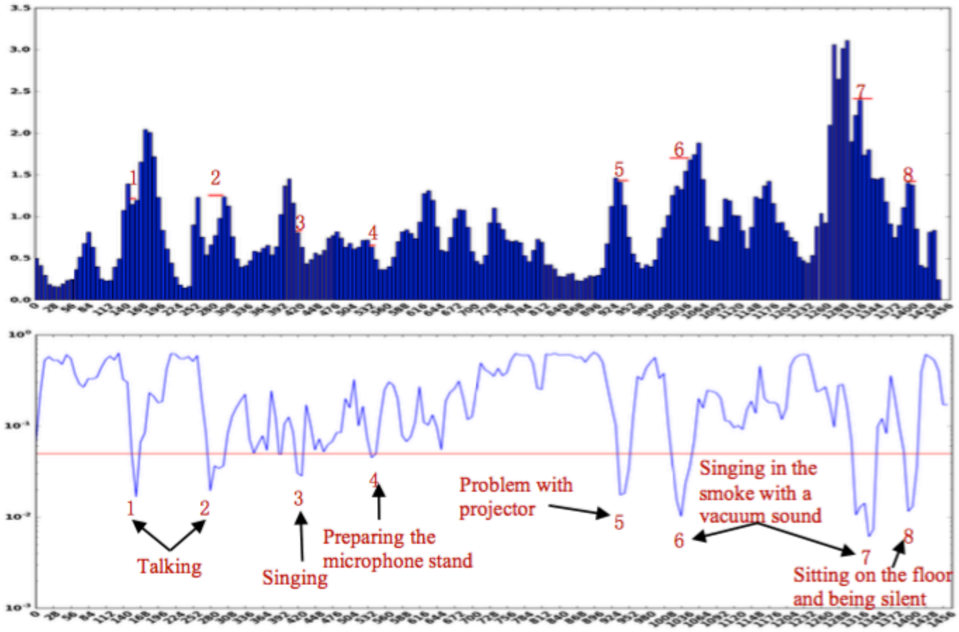


**Figure 40 : The extracted SCR signals of the live audience members during the performance, where points 1, 2, 3, and 4 are the significantly different SCR responses identified by the algorithm. In the top graph, the y-axis is the mean derivative value. In the bottom graph, the y-axis is the mean p value of the bilateral MWW test. The x-axis of both two graphs is the time in seconds. 1, 2, 3, and 4 are events performed were the live audience SCR response is significantly different from the background noise. 1: the smoke event; 2 and 3: the interaction between the actor and the audience; 4: the actor is sitting in the audience and talking.**

During the rehearsal, the latency of the live streaming system was measured, to be around 150 milliseconds, so that the audiences at the two locations could hardly feel the influence of delay.

The software for controlling the cameras, recording the data and networking was written in C and Python. All the data analysis was done through SPSS and Python.

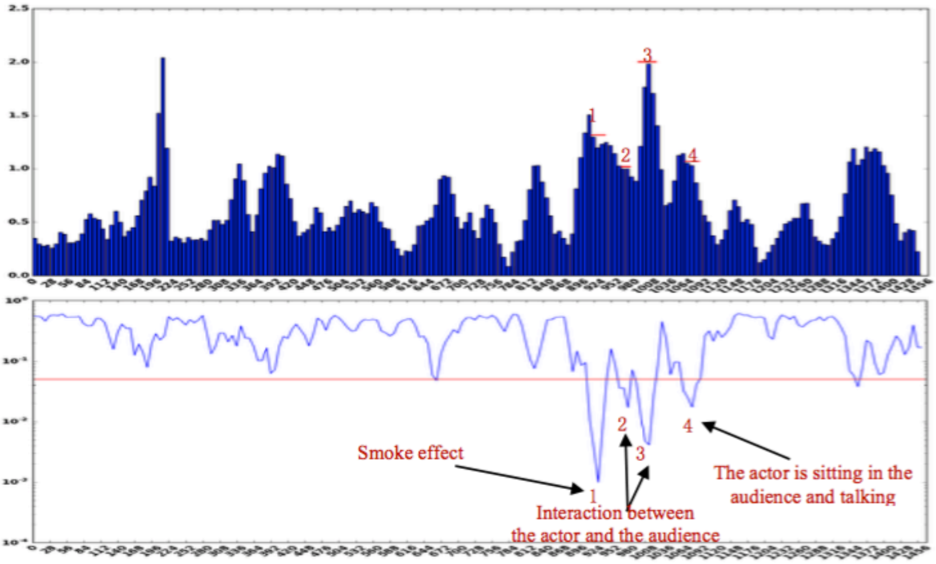


**Figure 41：The extracted SCR signals of the remote audience members during the performance, where 1, 2, 3, 4, 5, 6, 7, and 8 are the significantly different SCR response defined by the algorithm. The meaning of x-axis and y-axis is same as Figure 4. 1, 2, 3, 4, 5, 6, 7, and 8 are the events performed while the remote audience SCR response is significantly different from the background noise. 1: the actor is talking with his arms hurling; 2: the actor is talking to the remote audience; 3: the actor is singing; 4: the actor is preparing the microphone holder for singing; 5: there is some problems of the projector and the audience members raised their hand; 6 and 7: the actor was singing in the smog effect with a vacuum sound; 8: the actor was sitting on the floor and being silent.**

*Experimental Procedures*

Before the experiment started, the participants filled an informed consent form. Then oral instructions were provided. After that, the audience members from both locations attached the sensors to their non-dominant palm. At the end of the play, there was a small group interview at each location. Both the audience behavioral response during the whole performance and the performance were video recorded in order to better recall the experiments when we analyze the sensor readings.

*SCR Results*

The event-related SCR results, extracted from the 24 participants (two groups: 12 live audience members and 12 remote audience members) during the whole performance, are shown in Figure 40 and Figure 41. In the top graph of each figure, the blue columns represent the average value.

The red bar means that the *p* value in this moment is less than 0.05, i.e. significantly different from background noise. The concept is mirrored in the bottom graph where the p-value (blue line) goes below the critical value (red line).

The algorithm detected a number of moments where the event-related SCR signals were significantly different from the background noise, which means that the audience members were more engaged. For example, significantly different audience SCR response can be seen during the theatrical smoke effect in the graph of the live audience. During the smoke effect, also the remote audience was significantly engaged. The remote audience members were more absorbed when the actor was singing, while the live audience members were more engaged during the interaction event. In addition to that, it is interesting to see that the number of engaging moments of the remote audience members is higher than for the live audience members.

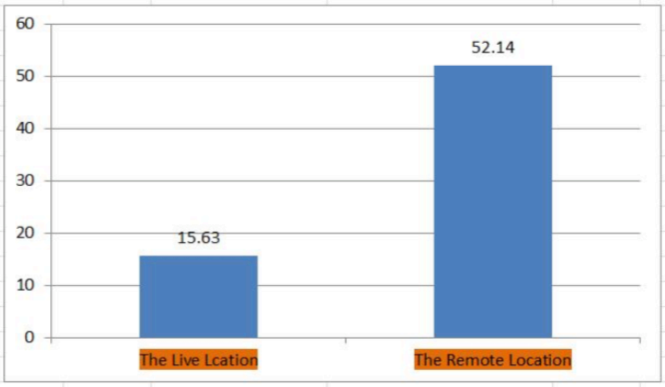


**Figure 42：The SCL difference during the singing event**

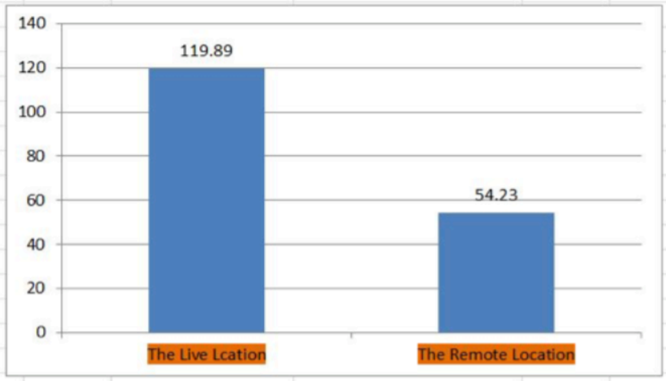


**Figure 43: The SCL difference during the theatrical smoke**

*SCL Results*

First, we compare the SCL data of the audience at different locations during the whole performance. There is a strong positive correlation between the data from the live audience and the remote audience ($r = 0.535$, $n = 12$, $p < 0.01$), which indicates that the skin conductance response pattern at both locations was synchronized. Additionally, the result of the t-test showed that there was no significant difference between the response from the live audience and the remote audience ($t = 1.18$, $p >$ .05).

Although the SCL data at the two locations was similar, we found that the two audiences responded significantly different to different events. These findings may help performers to better understand what kind of effects could arouse a remote audience. When the actor was singing, we found that the remote audience was more absorbed ($t = -4.04$, $p < 0.01$) (Figure 42). Additionally, both the theatrical smoke and the interaction were more engaging for the live audience (smoke effect: $t = 3.35$, $p < 0.01$; interaction: $t = 4.37$, $p < 0.01$) (Figure 43 and Figure 44).

*Interview and Video Recordings*

The data from interviews and video recordings is summarized and presented in Table 8. In the video recording, we found that eye contact between the actor and the audience at both locations was constant during the performance. Besides, most of the time, the audience at both locations were smiling.

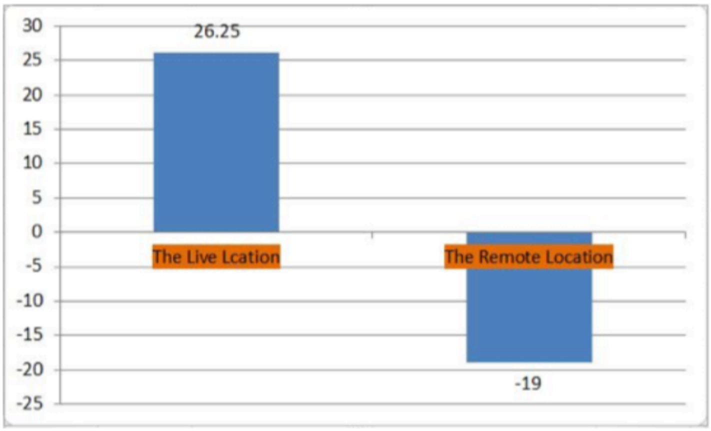
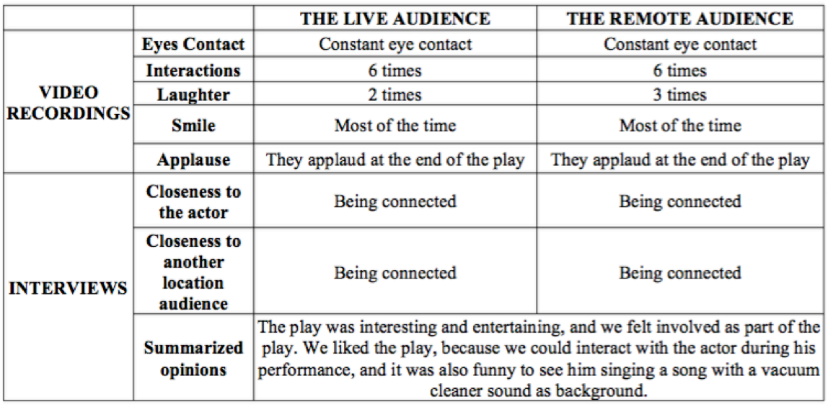


**Figure 44: The SCL difference during the interaction**

According to the results of the interview, all of the audience members felt connected to both the actor and the audience at the other location. Thus, we can conclude that both the live and the remote audiences were similarly immersed during the performance.

**Table 8: Summaries of interviews and video recordings**



| | | THE LIVE AUDIENCE | THE REMOTE AUDIENCE |
|---|---|---|---|
| VIDEO RECORDINGS | Eyes Contact | Constant eye contact | Constant eye contact |
| | Interactions | 6 times | 6 times |
| | Laughter | 2 times | 3 times |
| | Smile | Most of the time | Most of the time |
| | Applause | They applaud at the end of the play | They applaud at the end of the play |
| INTERVIEWS | Closeness to the actor | Being connected | Being connected |
| | Closeness to another location audience | Being connected | Being connected |
| | Summarized opinions | The play was interesting and entertaining, and we felt involved as part of the play. We liked the play, because we could interact with the actor during his performance, and it was also funny to see him singing a song with a vacuum cleaner sound as background. | |

# 5

# Feedback Mechanism

We identify the emerging phenomena of distributed liveness, involving new relationships among performers, audiences, and technology. Liveness is a recent, technology-based construct, which refers to experiencing an event in real-time with the possibility for shared social realities. Distributed liveness entails multiple forms of physical, spatial, and social co-presence between performers and audiences across physical and virtual spaces. We interviewed expert performers about how they experience liveness in physically co-present and distributed settings. Findings show that distributed performances and technology need to support flexible social co-presence and new methods for sensing subtle audience responses and conveying engagement abstractly.

We investigate how performers' experiences of liveness are transformed by technology in distributed performance, where performers and audiences are not all present in the same physical space. By *liveness*, we mean experiencing an event in real-time with the potential for shared social realities among participants [8, 82]. The concept of 'live' performance emerged in the 1930s with the introduction of radio as a way to distinguish from broadcasts of recorded performances [3, 4]. Initially, live performances involved only physically co-present performers and audiences. The Internet and social media advanced new forms of online liveness in which performers and audiences are socially co-present, but not physically [63].

We observe that as technologies for representing performance evolve, notions of live and recorded evolve with them. It becomes important to investigate the impact of this co-evolution on both performers and audiences. Prior HCI researchers have focused on audience experiences [2, 3, 6, 4]. We instead focus on performers and how they experience liveness.

This investigation examines traditional forms of performance: theater, dance, and music. To broaden exposure, live art performances have begun using technologies for distributing across the world [13, 18]. A resident of a small town in the UK can listen to the Metropolitan Opera at home or attend an expensive production by the National Theatre in London at the local theater. One might think that this was enabled by television, but television is one-way. The goal of distributed performance is to join performers and audiences in a shared sensory experience through bi-directional connections.

Emerging from our investigation, we identify the phenomena of *distributed liveness*, involving new relationships among performers, audience, and technologies that have the potential to transform live experiences. Distributed liveness encompasses various forms of physical, spatial, and social co-presence. For example, in *Can You See Me Now?* physically and socially co-present performers run through city streets, chasing online players who are spatially and socially co-present in a virtual game space. Distributed liveness is supported by hybrid spaces [4], which connect the physical and virtual to create shared experiences.

We conducted a qualitative investigation of performance artist experiences. We interviewed artists experienced in both physically co-present and distributed settings. Through data analysis, we discovered four themes: challenges in social co-presence, performer attention to distributed liveness, sensing engagement through subtle feedback, and representations of audiences. Findings motivate implications for design of hybrid spaces that promote distributed liveness for performers.

This chapter presents our qualitative methodology, followed by a discussion of findings, and implications for design.

This chapter particularly considers the eighth research questions:

**Research Question 8**: *Whether audience engagement can be inferred algorithmically from GSR data in a real-time application?*

## 5.1 Methodology: Interviews

To construct an understanding of how performers sense audience engagement and the differences between physically co-present and distributed performances, we performed a qualitative investigation together with performance artists.

We conducted semi-structured interviews with eleven artists including five musicians, three actors, two dancers, and one director (Table 9). All of them are successful professionals, with training and careers in performance. Their levels of experience with distributed performance varied. Participants were recruited from institutions already engaged in exploring creative intersections between art and technology.

**Table 9: List of participants, ID, performer type, gender, and expertise in performance and with distributed liveness. Performance expertise describes participants' experiences in the roles of performer and designer. All performance designers were also expert performers. Distributed liveness expertise is based on number of distributed performances where participant was involved, and the role that participant took in designing performances.**

| ID | Type | Sex | P. Expertise | Distributed Liveness Expertise |
|---|---|---|---|---|
| A1 | Actor | F | Expert | *Novice.* Interactive theater piece with mobile video; some performers were not physically co-present with other performers and the audience. |
| A2 | Actor | F | Expert | *Expert.* Combined theater and music pieces across two physical locations. Performers and audience in both locations. |
| A3 | Actor | F | Expert | *Novice.* Instructional performance for planning a distributed theater performance; she was not physically co-present with audience. |
| D1 | Dancer | F | Designer | *Expert.* Distributed performance designer. Art performances in which she was not physically co-present with other performers or the audience. |
| D2 | Dancer | F | Expert | *None.* |
| M1 | Musician | M | Expert | *Expert.* Many performances; various levels of physical, social, and spatial co-presence. |
| M2 | Musician | M | Expert | *Novice.* Improvisational music performances; he was not physically co-present with other performers or audiences. |
| M3 | Musician | M | Expert | *Novice.* Improvisational music performance; he was physically co-present with audience and several performers. Other performers were not physically co-present. |
| M4 | Musician | M | Designer | *Expert.* Distributed performance designer; physically co-present with audience and several musicians. Other musicians were not physically co-present. |
| M5 | Musician | M | Expert | *Novice.* Improv. music performances; not physically co-present with others. |
| R1 | Director | M | Designer | *Novice.* Interactive theater piece with mobile video; some performers were not physically co-present with other performers and the audience. |

Interviews were conducted via video chat, with the exception of one participant who was interviewed via e-mail. Interviews lasted from 30 to 60 minutes. Participants were asked about their experiences in sensing a live audience both in physically co-present and in distributed performance. Video and audio were recorded. Researchers took observational notes during interviews.

After each interview, researchers discussed observational notes. Potential interesting phenomena were identified. Interview questions were revised to help ask clearer, more specific questions directed at emerging phenomena.

Interviews were transcribed. Transcripts were broken down into units of meaning. Over 600 units were derived from the data. We performed open coding [16] on units to iteratively derive emergent themes. We initially developed over 20 codes. Codes were categorized into four themes which are presented in the following section.

## 5.2 Findings and Discussions

Our analysis of interview data discovered several themes. First, how spaces are connected can prevent social interaction despite the intention of supporting social co-presence. Second, distributed liveness requires active attention from per- formers, in order to sense audience engagement. Third, performer sense engagement of physically co-present audiences through subtle physical cues that are lost in distributed settings. Fourth, abstract representations of audiences can be effective at conveying engagement to performers.

### 5.2.1 Challenges in Social Co-Presence

Performers experience problems connecting with distributed audiences when using technologies intended to support social interaction among performers and audiences, such as video and audio communication channels. Musicians, in particular, described how in physically co-present settings, they are able to socially interact with audience members during setup, sound check, between songs, and after the performance. However, they felt socially disconnected in distributed settings, despite video and audio channels supporting two-way communication with the audience.

*M2: It makes a difference that I'm aware that the audience is there, and I try to make sure that what I perform is something that I would like to hear, and I'm not just experimenting the whole time ... It's not the same as going to a venue. Talking to the people there. Making sure that the wires are well connected. Having people come in slowly and realizing," Oh! This is actually a real thing."*

Being physically co-present supports performers developing social awareness about a space, even without paying close attention to what is going on in that space.

*M1: When you are in the room, you're together in a space with the sound, the physical movement of the sound. You are in the same space as the musicians and the audience, so even if you are not necessarily paying attention to it, I think we must be aware of what's happening in that space.*

Performers want to socially interact with audience members, but the technology does not always facilitate such interactions with distributed audiences. Providing two-way communication does not insure that performers and audiences will experience remote participants as live. In one example, actors in Korea talked to an audience in New York via live video, but audience members failed to recognize the actors' liveness.

*A2: I thought it would have been really great to react and talk to the audience, because I wasn't even sure if they were aware of us being in Korea. I got feedback from New York afterwards, and they said some of the audience members were actually confused. When they were informed later that we were actually in Korea performing at 6 in the morning, they were like, "Oh my god! That was Korea." Some of the people weren't aware of us being in Korea and doing telepresence. They thought that was just a video clip of something. Although, we talked in it with them.*

While performers described the importance of social interactions in the moments before and after a performance, the amount of interaction with the audience during the performance varied by performer and context. Several of the musicians and actors expressed concern about certain audience reactions that they believed could harmfully affect the performance, such as audience members ignoring the performance and talking to each other. The participants wanted flexible control in distributed settings over the kinds and amount of audience response transmitted to them.

> M3: *It's nice to have that constant connection with the venue. In terms of visual feedback, I would always like to have it there in some form. The form could change in little ways that the data was being displayed. ... I would always like something there, so that I might not have to be looking at it all the time, but if you were just to glance across at the screen you could quickly see how things were with the audience.*

Audience feedback is delayed due to technological limitations. This disrupts continuity of feedback loops, creating cyclic periods of performing and watching. A participant performs an action that requires a response, and then must wait to observe reactions from the audience or another performer.

> D1: *Because there is always a slight delay, there's always an element of I do it and I watch. So, there's a little bit of that tiny time segment of what is performance. Do and watch. You know in that sense that we are witnessing each other as well as engaging with each other.*

### 5.2.2 Performer Attention to Distributed Liveness

The lack of physical co-presence in distributed performances challenges performers' abilities to be attentive to remote audience engagement. A performer must adapt how she directs her attention towards audiences in other spaces. The common approach for connecting spaces is to provide video or audio streams of the other spaces. These connections filter what a performer is able to perceive about an audience, providing a limited, focused representations of the space. Performers must actively direct visual attention at a display with a video stream. Each connected space is often represented by an individual display, requiring performers to divide their attention among several different displays. As a result, participants expressed favoring auditory feedback more than visual in past distributed performances.

*A2: More audio than video. Because of the nature of the telepresence, it is really hard to focus on different screens at the same time. So, I can only focus on the other performer or on the other location.*

When physically co-present with others, performers will build connections with those around them (e.g., technicians in a studio or performers in the same space). The shared physical space and the directed attention of those observing creates a localized performance within the larger hybrid space.

*D1: The audience changes, because even when you are doing a remote performance, the reality is that you do have an audience. You have people in the lab, and you can't help but connect with the people that you are with …*

Thus, the model of distributed liveness consists of smaller, localized performances that are combined to constitute a whole, connecting spaces, performers, and audiences.

### 5.2.3 Sensing Engagement through Subtle Feedback

Participants reported using sensory feedback involving vision, hearing, and kinesthesia to sense audience engagement in physically co-present settings. Much of the feedback described involve subtle physical responses from the audiences reflecting changes in emotional state or engagement, such as facial expressions, tightening of muscles, or a shared energy. In particular, dancers and actors described this physical feedback as kinesthesia or proprioception, sensing through the positioning of body parts and movement. When an audience is engaged, the performers and audience are physically synchronized (e.g. sharing similar respiratory)

*D1: If I do something that has constant inhales, [inhales deeply several times] and I keep doing this; the way human beings are designed is that we mimic, so the audience will start doing that. When you are doing extreme things on the stage that involve breath or risk taking, you can feel the audience's kinesthetic engagement with you; and that is a powerful thing. That really makes you feel connected.*

Performers want to feel the audience's presence and engagement. The audience's physical presence gives energy to the performers, creating a unique live experience for all.

> D2: *When there are a lot of people gathered around you, of course, there is an energy. That's a natural energy of human beings being together as a collective, which is an experience that we don't have very often.*

Participants primarily experienced audience presence in distributed performance through cameras and microphones, used to capture views and sounds of the audience and to stream those to the performer. Participants reported problems with this approach for conveying the subtle feedback of human expression and engagement of distributed audiences.

> M4: *The artist-audience relation becomes even more pronounced. Which is unfortunate, because you want to think of the internet as very egalitarian, democratic. But when you put it in that context, only the local audience is so specialized. You really feel that difference even more so. Yes, it is harder to tell whether or not they are actually engaged. It's difficult to know if people are smiling. A smile is a difficult thing to capture, even with a good camera. Particularly, when people are moving, and they don't want to be on camera themselves. Those little differences in peoples' faces, such as the look of excitement, are very hard to communicate through a screen. Those are things that are really missing for a performer. They just can't get those subtleties of human expression.*

In traditional stage performances, such as theater, opera, and ballet, the absence of sound can be an indicator of audience engagement. During intense moments, a performer expects the audience to be on the edge of their seats and silently engaged. If the performer hears the audience rustling, it could be an indication that audience members are disinterested.

> D1: *Silence can also be an incredible indicator. [The audience] retracts in a way. That's also really powerful. You feel they give you the space to go deeper into your moment, which might sound like a contradiction. That*

> *they kind of recede and you feel even more alone or quiet and silent, but in a way it connects you even more*.

## 5.3 Implications for Design

We present implications for the design of new performance environments to support distributed liveness. Designing hybrid spaces that give performers flexible, directed control over social interactions among audiences will improve experiences of liveness. The physical separation of spaces in distributed liveness requires new methods for sensing subtle visual, auditory, and kinesthetic reactions from distributed audiences, and conveying that feedback abstractly to performers.

Distributed performances bring together performers and audiences across different spaces, both physical (e.g. theaters) and virtual (e.g. YouTube streams). Design for distributed liveness creates hybrid spaces [4] that mix the physical and the virtual. Physical spaces situate performers, audience members, and objects. Virtual spaces are comprised of representations of people and their engagement.

We need to design hybrid spaces to effectively support social co-presence in distributed performances. The technology used to connect spaces impacts how performers and audiences form connections with each other. Despite the provision of visual and auditory channels, meant to convey social cues, performers reported difficulty socially interacting with distributed audiences. Performers may feel isolated or be at a loss if there is not much interaction with remote audiences for a live performance.?

Several participants wanted ways to engage in direct social interaction with distributed audiences. One way that is presently supported is text chat. Examples of hybrid spaces that effectively support social co-presence through text chat can be found in the virtual game spaces of *Can You See Me Now?* [4] and live streaming environment of Twitch [10]. In performance art contexts, reading text chat during the performance would be difficult for many of our participants. Yet in these contexts, technology should still support performers interacting with audiences before, during intermissions, and after the performance. For example, performers should be able to view audience text chat, toggle on video and audio feeds, and talk to audiences directly, addressing questions, chatting about the performance, as well as see and hear how other performers attract specific audiences.

### 5.3.1 Sense Subtle Feedback, Convey Abstractly

Performers identified the importance of sensing audience engagement through subtle visual, auditory, and kinesthetic feedback. The physical separations and clear boundaries effected by video screens and speakers make prior teleconferencing systems inadequate for conveying this feedback.

We need to develop new techniques for sensing and conveying audience engagement. Physiological sensors, such as those for GSR, respiration rate, electromyography, provide means for measuring audiences' and performers' bodily responses. Participants described forming connections with the audience and other performers through similar bodily experience, such as shared breathing patterns, heart rates, or muscle tension. Physiological sensors have been previously shown as effective measures of audience engagement [66,75].

While invasive sensing technologies are suitable for experimentation, deployment in the wild requires non-invasive techniques that address privacy concerns. New commercial devices for health and fitness, such as Apple Watch and Fitbit, provide personal sensing of physiological data, such as heart rate and body movement. These sensing technologies can be combined with mobile applications that operate only in local areas. This would enable audience members to login in to collection of anonymous sensor data in commercial performance venues, such as theaters, to participate in distributed liveness. This approach enables performance venues, such as the theater broadcasting the Met Opera, to serve as distributed liveness venues. Standards and compliance certification for how-to collect such data and guarantee that privacy would ameliorate, but not eliminate, privacy issues.

We advocate representations that convey subtle sensory feedback. Some musicians we interviewed had experience performing in response to abstract representations of distributed audiences. They found these representations helpful for perceiving audience engagement. Abstract representations avoid overwhelming attention. They are ambiguous and allow performers to form their own interpretations. Gaver et al. point out that “ambiguity can be frustrating, to be sure. But it can also be intriguing, mysterious, and delightful. By impelling people to interpret situations for themselves, it encourages them to start grappling conceptually with systems and their contexts, and thus to establish deeper and more personal relations with the meanings offered by those systems” [67, 69].

Representing kinesthetic feedback requires using physical devices beyond screens and speakers. For example, heart rate data could be represented with a pulsating arm band that expands and contracts to mimic beats of the heart. This representation is a form of wearable *kinetic garment*, containing mechanical components, such as actuators, that move in response to physiological data [68, 70]. Such garments will pronounce kinesthetic engagement in new ways to performers, as compared to traditional physically co-present settings. Devices, such as TVs and headphones, which convey visual and auditory feedback, reproduce light and sound waves for our eyes and ears, as if physically co-present where original stimulus was produced. Devices for kinesthetic feedback produce a new sensation that seeks to mimic the original stimulus, but it is not the same. Performers will have to train their kinesthetic senses to interpret feedback from these devices. Conversely, remote audience members can wear kinetic garments, creating bi-directional kinesthetic feedback loops.

## 5.4 Summary

We coined the term *distributed liveness*, to refer to an emerging aspect of computer-supported collaborative performance with broad impact on creative human experiences. The Internet provides means to connect performers with audiences from around the world. Yet, this technological connection of- ten fails to provide a shared sensory experience. Performers and audiences are physically and often temporally separated. Performers become unaware of remote audience experiences. We contextualized this historically, noting that liveness has been a socio-technical construct for almost a century.

We conducted interviews to understand performer experiences of liveness in different settings. Analysis of our findings contribute implications for design of distributed liveness environments. Hybrid spaces for distributed performances need to support flexible social co-presence, enabling performers to switch among levels of social interaction at different moments in a performance. We need to develop new ways to sense subtle physical cues of audience engagement, and communicate audience response without overloading attention.

As an emergent arena of phenomena, distributed liveness provides avenues for exploration of diverse, new forms of computer supported collaborative work and play. In addition to the performing arts and games, we envision designing new classroom environments for distributed liveness, in which virtual spaces connect classrooms of students at multiple institutions, as well as students at home, in shared learning experiences. Abstract representations of aggregated sensory data from remote participants, in concert with live streaming, will enable teachers to sense remote student engagement and ad-dress gaps in student attention.

A result of the seams inherent in experiences of distributed liveness is that we must co-design performance and technology, because the tandem fundamentally prescribes participant experiences. Seamless design strategies, in which performance is composed while taking into account limitations, such as delay, is one part of this. Another is to build presentations that holistically combine sensory feeds, using a strategy such as information composition [72, 73, 74, 61].

Co-designed approaches to distributed liveness have the potential to transform the nature of performance events, connecting participants physically, spatially, temporally, and socially in new ways. As technology changes performance, so performance must change, creating new hybrid forms [71]. Performers will need to broaden their skills, learning to interact with new technologies and engage in new types of performance. Likewise, writers and directors will need to account for the characteristics of seams and the situated technologies that produce them. We look forward to new hybrid spaces and works that engage collaboration through distributed liveness to materialize compelling, participatory forms of performance.

# 6

# Final Experiment

This chapter provides a discussion of an experiment particularly designed for validating the GSR method, which is used as an effective feedback mechanism. The experiment was carried out in a distributed environment, where the real-time response from the remote audience was visualized by the artists in a live venue. The feedback mechanism was created based on the interview results with the artists. After the experiment, the artists helped us to evaluate how the mechanism was employed during the performance.

## 6.1 Objectives

This chapter is the concluding part of this research. Prior to this experiment, extensive work was done regarding the user studies: analyzing the experimental environment, sensor features, hardware performance, where the data analysis was mainly conducted offline. In this experiment, the real-time algorithm was developed to visualize the remote audience's response to the live venue. The interview results with the artists could help in understanding the pros and cons of the mechanism. For researchers who may be interested in studying further this subject matter, they can follow the proposed method and establish their own real-time algorithm for any specific applications. This chapter addresses directly the key research question.

**Main Question:** *How can we support and enhance actors' awareness of remote audiences in a distributed theatrical environment?*

## 6.2 Method

The feedback mechanism was developed based on the results from the extensive interviews with the artists. Two artists with previous performing experience in a distributed environment were interviewed initially. Then an interview with the artist who provided the actual performance during the experiment. From the interview results, the mechanism was designed, which was used to visualize in real-time the remote audience response.

**Figure 45: (Left) The technical team conducted a rehearsal with the artist Ma Xue to test the lighting conditions; (right) the interview was conducted with the artist.**

The evaluation procedure was carried out after the experiments. The procedures were repeated thrice. In each round, the participants were asked to sit in the two locations (Figure 45 and Figure 46). After each play, a post-interview with the artist and the two location audiences was conducted. The objective of the post-interview was to understand the effects and the limitations of the feedback mechanism.

**Figure 46: Ma Xue, the artist**

### 6.2.1 Interviews

*Pre-interviews*

The two rounds of semi-conducted interviews were conducted with the artists. The first was executed with the two artists who had previously distributed performance experience while the second was the artist who actually participated in the experiment (Figure 44).

The interview consisted of four questions. The purpose was to analyze the appropriate feedback mechanism they would like to have during a performance.

- **Q1**: What did you feel when you performed in a distributed environment?
- **Q2**: What do you think are the different feelings in a live audience and a remote audience?
- **Q3**: What would you care about on audience responses (from both venues)?
- **Q4**: What kind of mechanism would you like to have in a remote audience during a distributed performance? (For instance, lighting changing, a digital display or a vibration method)

From the results of the first round of interviews, the two actors noted there was a huge difference between a live audience and remote audience. In a live venue, the actors can sense the audience's interest through eye contact, but sensory feelings are lost in a remote audience. They had to pay more attention to their performing speed to let them more involved in a remote audience. Since the actors cannot see a remote audience, the performance content is rather vital and should be interesting. Otherwise, a remote audience will not focus on the performance.

Both actors believe that an appropriate feedback mechanism is essential. Otherwise, it may interfere with their focus. The performance requires the actors to fully engage in it, and they are less likely to check at a digital display. A vibration method can be distractive during a performance. Thus, a direct and physical visualization method is preferred. For the three mechanism options, they chose the lighting changing as the feedback channel where they can directly sense the arousal response from a remote audience.

In the second round of interviews, the artist had similar opinions with the other two artists. But, based on her performing design and requirements, she suggested having two lighting conditions during her play. Each represents the response from a live venue and a remote location, respectively.

*Post-interviews*

The post-interviews were conducted after each experiment, involving the artist and the two location audience members (Figure 49). The results were used to evaluate the lighting mechanism effects.

### 6.2.2 Participants

A total of 30 participants attended the experiment. Among them, there were 20 females and 10 males. The 18 participants came from the company, while the others were recruited from two universities. The actual valid sensors data were from 25 participants and there were five-person sensor data that were not successfully received.

### 6.2.3 Performance

It was a solo performance. The artist Ma xue (Figure 46) studied broadcasting and hosting, and theatrical performance was one of the courses she studied in the university. The play was interactive. First, the artist narrated a story. From the story, the two location audiences had to think about who were the play's characters. Then, the artists gave a note to each person during the live venue. On the note, each audience member has to identify whether his/her role was the character in the play or not. The host then divided the audience at the live venue into two groups, and they have to continue playing the puzzle to guess the characters. During the guessing, the live audience had to interact and communicate with each other and with the artist at the same time. The artist gave the hint through singing, dancing, and using verbal language.  Although the remote audience could not participate in the play, they had to watch carefully and guess who the characters were. Also, the remote audience could receive the hint from the artist through the display. After the play, the artist would announce whether the covered persons won or not, and the remote audience would elect the best performer from the live audience.

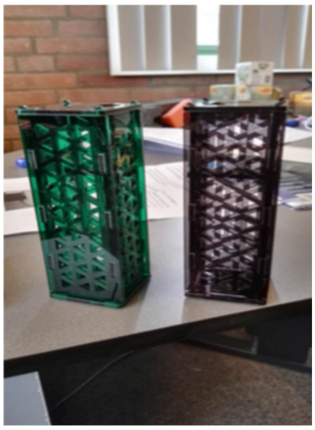


**Figure 47: The GSR sensor system**

### 6.2.4 Apparatus

The newest version of sensors was used in the experiment (Figure 47). The sensors were built based on Jeenode developmental platform and used the RFM (12) for the wireless module to construct the group communications. The GSR sensors were factory-printed PCB board, which can be easily connected with the Jeenode interface. The sensor housing was laser cut to fit all the components.

Also, the monitoring system was developed for watching the working state of the sensor system (Figure 48), through which, for instance, it can be seen which node works or not so that we can take an action to check or replace it. Before the actual experiment, the time between the monitoring and the sensor system needs to be synchronized.

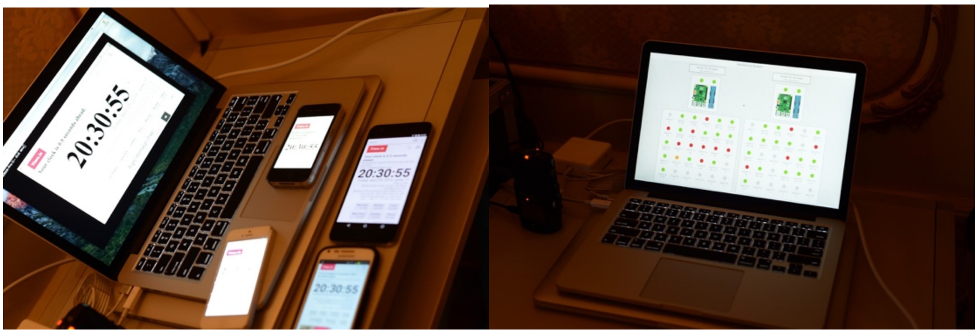


**Figure 48: The monitoring system.**

### 6.2.5 Software

The real-time algorithm was developed to use the sensor readings to control the lighting condition. The two levels of the lighting condition are red and blue. When the mean sensor value was higher than the baseline (the readings obtained from the 10 minutes' meditation) by 20%, the lighting condition was blue. Otherwise, the lighting condition was set to red. The calculation procedure was automatically performed every five minutes. Every five minutes, the baseline was recalculated.

### 6.2.6 Network Platform

The distributed performance platform was established through the real-time recording camera. There was HDMI cable used to transmit the video live stream from the live venue to the remote location (Figure 49 & Figure 50). This was the best solution to guarantee the video quality and reduce the delay issue, while other platforms, i.e., Skype or WeChat, did not give the qualified technical performance. During the pilot studies, there were heavy delays found and low resolution appeared on the live video stream, although a cable internet connection was used.

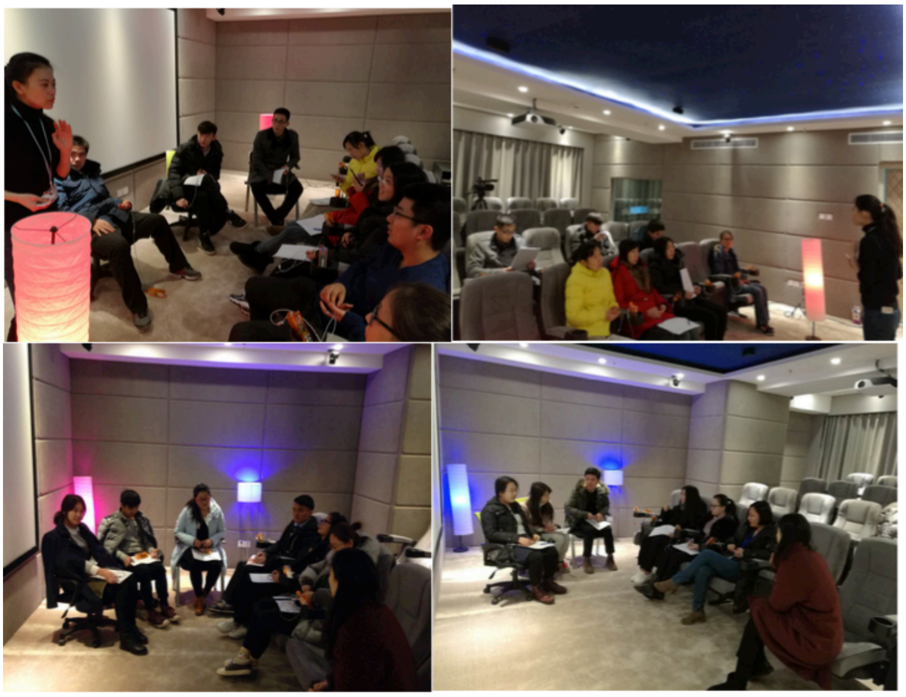


**Figure 49: The live performance location.**

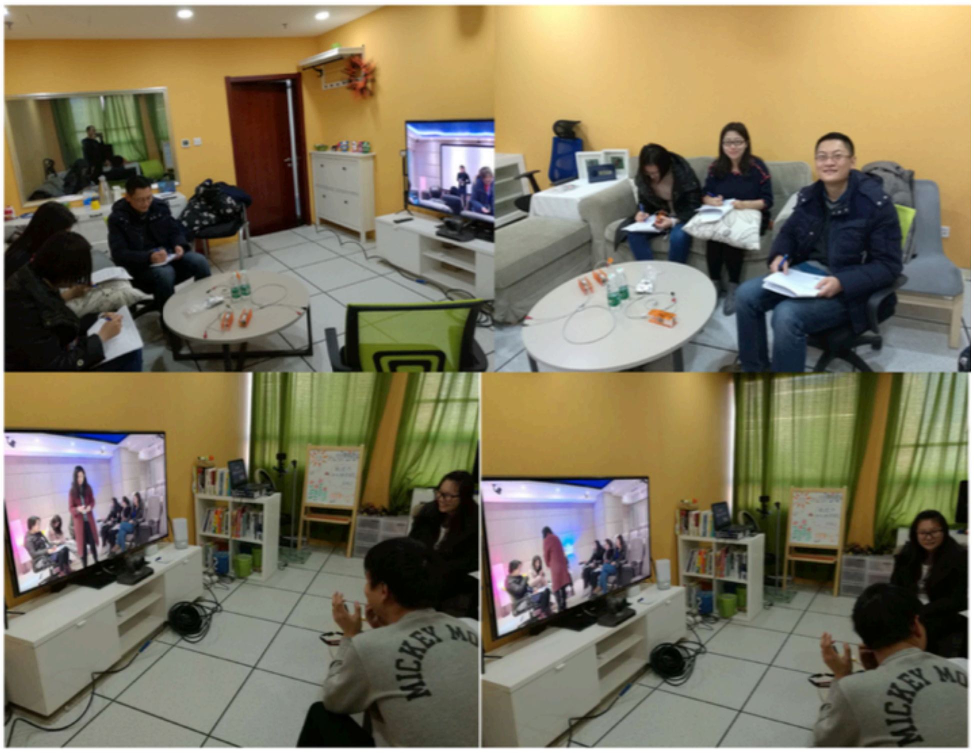


**Figure 50: The remote performance location.**

### 6.2.7 Interviews

The pre- and post-questionnaires were conducted during the experiments. All questions were designed for a seven-point scale. The pre-questionnaires were used to analyze the physical and emotional states of the participants, while the post-questionnaires were adapted from the paper [22] and used to form the reports that reflected the audience's subjective viewing experience.
All questions were listed below:

*Pre-questionnaires:*

- **Q1**: I am in a good mood today;

- **Q2**: I am energetic;
- **Q3**: I am feeling cheerful today;
- **Q4**: I am interested in watching the performance;
- **Q5**: How often do you watch the theater performance?

*Post-questionnaires :*

- **Q1**: The play was interesting;
- **Q2**: The play was exciting;
- **Q3**: The play was pleasant;
- **Q4**: I forget the world around me during the performance;
- **Q5**: I give my attention to my surroundings during the show;
- **Q6**: I am completely captivated;
- **Q7**: I will definitely turn to another performance;
- **Q8**: I will recommend it to my friends;
- **Q9**: The performance is worth watching;
- **Q10**: The performance cheered me;
- **Q11**: This performance made me energetic;
- **Q12**: This performance made me happy;
- **Q13**: The sensor for the show was natural;

The interview questions conducted with the artist Ma Xue are in the following:

- **Q1**: How did the lighting condition help you during your performance?
- **Q2**: Could you feel the connection with the remote audience during your performance?
- **Q3**: How did you apply the lighting feedback during your performance?
- **Q4**: Were there any improvements that could have done during your performance?

In the following section, we consider the results of the interviews based on the questions above.

## 6.3 Results

*Interviews*

Although the experiment was particularly designed for getting the artist's feedback, the interview was not only conducted with the artist, but live and remote audiences were also interviewed (Figure 51). The results showed that the GSR mechanism was extremely helpful for the artist to understand the arousal level of the remote audience, and the artist could establish the connection with the remote audience via the lighting indication. The artist intentionally intervened in the remote audience when she noticed that the lighting color from the remote audience became red. Also, the artist felt that the distance with the remote audience became narrower. With the help of the lighting facility, it looked as if they were in the same space.

The remote audience felt more engaged and pleased with the use of the lighting facility. During the play, they immediately noticed the changing lighting, and they started to laugh and guess the reasons why the lighting color somehow related to the live audience became red. When they saw the color turned from blue to red to their location, they blamed themselves for not focusing so much on the play. Although the different lighting conditions interfered with their attention, they were tolerant of such disturbance during their viewing experience. They also commented that the lighting condition made them feel close to the performance as if they watched the show at the same location.

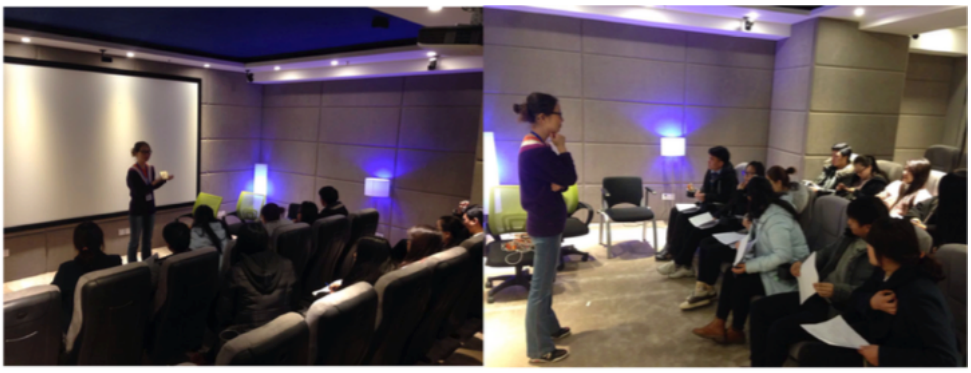


**Figure 51: A post-interview was conducted with the two location audiences and the artist.**

The live audience did not care as much about the lighting condition as the remote audience. They paid much attention to the play and the interaction with the artist. They engaged themselves more into the play instead of checking the lighting condition. Even though they knew that there was a remote audience watching the play at the same time, they did not have such a time to think about it. When the artist started to interact with the remote audience, they regarded it as a part of the performance.

However, such mechanism raised several issues, which can be further analyzed in future work. First, the setting of the lighting condition was not so natural for a theater play. The interviewees thought that the physical setting of the lighting condition should be merged into the environment, i.e., the lighting could be set up in the ceiling, giving a natural lighting atmosphere. Second, they believed that there should be a transition color between the two colors, as the transition process would have less interference with their attention instead of disturbing the audience with subtle changes. Third, they pointed out that the representation of each color was still vague. For instance, red color indicated a low arousal, but there would be a difference between the two locations when the two lights displayed the same red color. Finally, they suggested that a physical bar with the level indication would be a better option to replace the lighting because a physical bar with the specific arousal level could give a direct impression of the audience's response.

*Questionnaires:*

The reliability of the questionnaires is tested by alpha reliability statistic, and the Cronbach's alpha is 0.86 ($a > 0.7$). The descriptive statistics of the pre-questionnaires and the post-questionnaires are shown in Figure 52 and 53, and Table 3. According to the pre-questionnaire, the group of participants (from the two locations) have slightly positive mood before the experiment (good mood: mean = 4; cheerful: mean = 3.4; and interested: mean = 4), and the energy level is roughly neutral (mean: 2.9).

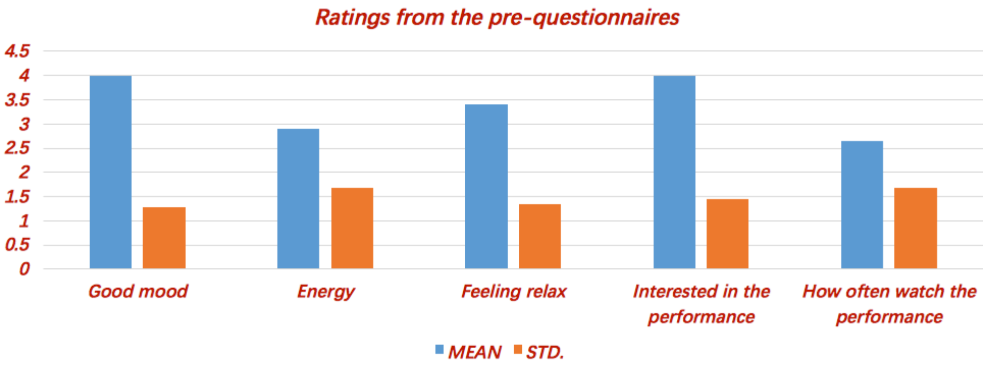


Figure 52: The descriptive statistics from the pre-questionnaires.

The test of homogeneity of variances indicated that the four items (1: I had attention to my surroundings; 2: it is worth to see; 3: the play has cheered me; 4: there is no interference from the sensor) could not be tested using the analysis of variance (ANOVA), and they were tested using the non-parametric method. But the results showed that there was no significant difference between the two location audiences.

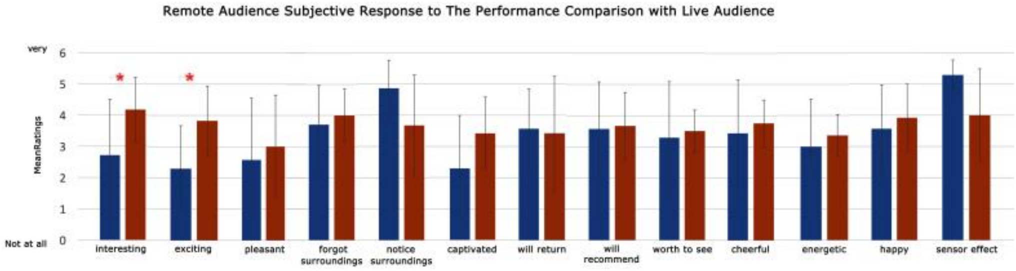


Figure 53: The descriptive statistics from the post-questionnaires (blue: the remote audience; red: the live audience); red star indicates the significant level at 0.95 confidence level.

For the rest of the items, the one-way ANOVA revealed that the live audiences felt more interested (F (1,18) = 5.1, p =0.04) and more excited (F (1,18) = 7.17, p =0.02) compared to the remote audiences. This was where the live audience could directly interact with the artist, while the remote audience only watched the performance. Even though the responses from the remote audience could be visualized at the live location, watching the performance on the display was unlikely to bring the same excitement and interest as the live audience experienced.

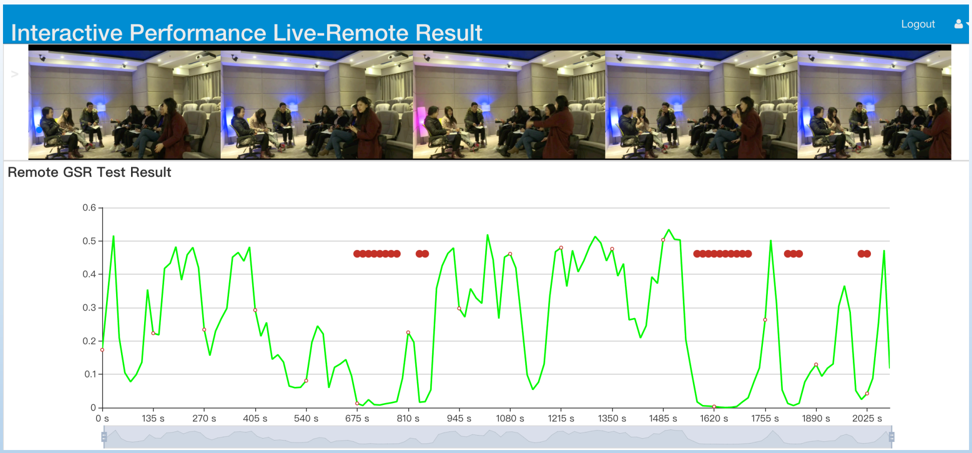


**Figure 54: The visualization results display the significant arousal moments from the remote audience. The red dots in the second row indicate the significant performing events evoked from them. By clicking the red dots in the second row, the related performing scenes (the screenshot) will be displayed on the above.**

*Sensor Results*

The visualization platform was developed to present the sensor results (Figure 54). Also, the results could be mapped to the specific performance activities through the platform functions. This means that the two location audiences' emotional intensities were linked to the respective performing event. Through this, it can be investigated how the same or different performing events affected the two location audiences (Figure 55 and 56).

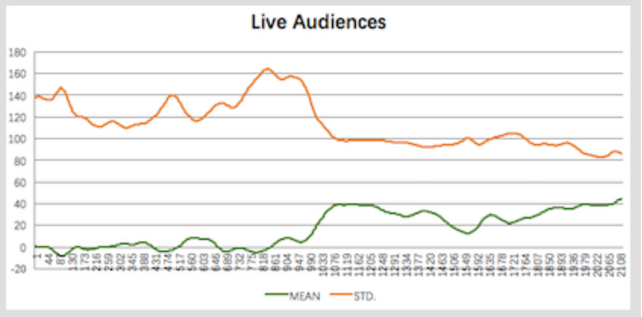


**Figure 55: The distribution of the mean and stand deviation of sensor readings from the live audiences.**

The platform is a web-based application, which was developed with NodeJS and MongoDB. It could dynamically load the raw data and generate the interactive charts in real time. It can be divided into several major models, including raw data.

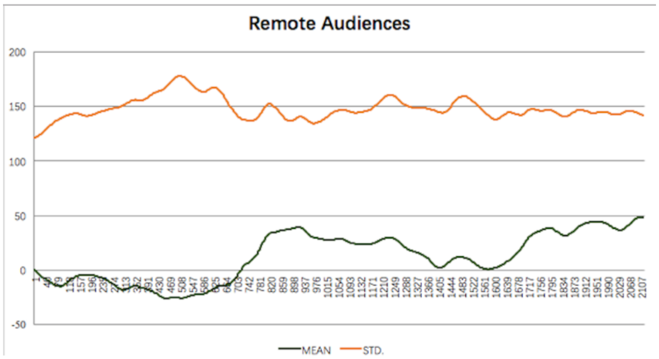


**Figure 56: The distribution of the mean and standard deviation of sensor readings from the remote audiences.**

The results showed that the interest from both location audiences was quite equally aroused. There were 26 significant emotional intensities (as red dots shown) at the live location against 25 at the remote location. In particular, the interest of the two location audiences was both significantly linked to the performance ending, where the artist Ma Xue tried to guide the audiences to investigate the mocked person during the play. However, when the lighting changed the color, the remote audience was substantially affected, but it did not show significant effects on the live audience. The results were also consistent with the findings from the interviews, where the remote audience reported that they paid attention to the lighting condition but the live audience did not notice that as they were immersed in the play.

However, the sensor results also revealed that the start of the performance (up to 675 seconds) did not significantly affect the remote audience at all, although the artist Ma Xue tried to interact with the remote audience. For the live audience, there were few moments where their interest was substantially aroused, but still, there was a long interval which appeared in the duration between 220 seconds and 800 seconds.

Based on our experience, this may indicate that at the start of the performance, this may be a bit boring, especially for the remote experience. The reasons behind were that the remote audience needed more time to adapt to a live play, as the hearing condition, video quality, and understanding of the performing content were rather different to the conditions at the live location.

The results brought the different perspectives to the artist. When the results were presented to the artist Ma Xue, she felt that the sensor results were extremely helpful for her to reflect the design and how to perform in the future. She believes that the design and performing style should be properly considered in a distributed environment. Some traditional performing styles may be suited to one location, but they may not be fit for a distributed environment. She also pointed out that the GSR mechanism could become the part of the performance and used it as a hint to get the attention of the audiences. But she noted that the online GSR mechanism could not fundamentally change the audience experience, and it may bring a distraction to an audience. She suggested that the attention should be paid to the performance design in a distributed environment, and the GSR method could work as an effective evaluation tool to provide the objective feedback to the artists.

The statistical analysis shows the consistent results on the emotional intensities from the two location audiences (Figure 54 & Figure 55). The bio-responses from the two location audiences were significantly correlated. The correlation between the two location means was 0.682** (**: at 0.01 level), and the correlation between the two location STDs was 0.195** (**: at 0.01 level). The mean findings indicated that the two location audiences had roughly similar response towards the performance, and the STD result showed that the emotional fluctuations from the two locations demonstrated the significant similarities during the performance.

# 7

# Conclusion

## 7.1 Summary of Contributions

This thesis reports on our research to design a GSR system that can support and enhance actors' awareness of remote audiences in a distributed theatrical environment. At the early stage of this work, we made concentrated efforts to design and develop appropriate sensor hardware. We did this to find a practical way to construct relevant hardware that could be worn in a theatrical environment. We thoroughly tested each hardware version in different field user studies. Then, we developed related algorithms to process the data delivered by our sensors. When we had enough knowledge about the relationship between GSR patterns and audience engagement, we developed the real-time algorithm and the feedback mechanism so that remote audience engagement could be visualized.

Our solution makes a technical contribution from both an engineering and a software design perspective in the creation of a system that allows theater stakeholders to explore the response data of an audience. This exploration has the potential to enhance the creative work of the theater stakeholders and to understand how audience members respond to their creative outputs. Although others have investigated audience response through the GSR systems, the deployment of a system of this scale in a live, large-scale theater setting is truly novel.

A first significant research result was the refinement of a set of heuristics for design and development of hardware. The heuristics were created during the first three years. We were able to use them for building and implementing the infrastructure in the developmental process of the sensor hardware. By following the heuristics, user cases conducted in the commercial theater performance and the dance performance proved that the sensor system was easily deployable in theaters. In addition, our studies also proved that our sensor system can be used for both real-time and offline purposes.

The second success was that we could use GSR data to (real time) infer audience engagement. For instance, we compared the GSR responses from participants in two different user cases [16]. In terms of the sensor psychophysiological states, we used a MDS method to differentiate sensor signal patterns between engaged and non-engaged participants. We found that the responses from the engaged users concerning sensory pattern showed a strong correlation between lab and field studies. Interestingly, the responses from the non-engaged participants did not correlate across user cases between the lab and the field trials. These results are consistent with a similar phenomenon mentioned in a previous research [30]. A "boredom" state captured in a lab may have different patterns compared to one in a field study. In addition, our findings on sensor data may bring inspiration for researchers who intend to investigate user states crossing different scenarios. According to our learning experience from lab studies, we found that it is unlikely for users to generate a boredom state when watching short videos. This happens if every participant is seriously engaging in the task. Even though it is in an unknown language with a quite low resolution, they typically try to understand what it is happening in the video. Therefore, for certain user emotional states, it may be possible to obtain them in reality rather than just in lab conditions.

Moreover, the development of the feedback mechanism has successfully stick to the performance content. During the interviews, we have learned how to design and develop a feedback mechanism that suits a performance. However, it is unlikely to develop a mechanism that will optimally fit the final requirements. Repetition will be required because the effect of a mechanism will be evaluated during the actual experiment. In general, one experiment alone cannot be used to make a concrete decision. If the intention is to design a mutual feedback mechanism, repetition and modification should be conducted.

## 7.2 Limitations and Future Research

In spite of its overall comprehensiveness, our study has shown some limitations in terms of covering all the aspects of the main research question.

**Main Question:** *How can we support and enhance actors' awareness of remote audiences in a distributed theatrical environment?*

The solutions for answering this main research question are dependent on multiple factors. These factors are performing style, the desire of artists, and the environmental restrictions. We have done one scenario experiment where the GSR signals were manipulated to control the lighting conditions. We found out that multiple repetitions were required in this setting. The pre-design strategy appeared not to have the optimal solution during the real experiment. Thus, both artists and audiences have come up with new thoughts during their participation. Furthermore, when the experimental scenario was altered, the requirements from stakeholders varied as well. In order to adapt to new requests, a new design needed to be made. Finally, the choice of the performance in a distributed environment is rather different compared to a location play where artists do not care about the communication with multiple locations. This phenomenon will not be simply solved by applying technology. It actually opens a new research area for artistic production. The GSR solution can work as an objective evaluation tool to help artists select the appropriate script. The best approach would be to involve such technology from the outset during the preparation of the show until its conclusion at the end of the public performance. Running several experiments to test whether a solution works or not cannot solve the fundamental problem. For example, it cannot be determined whether a show is actually not suitable for a distributed environment. If a show itself has less engaging factors to attract audiences, it cannot expect technology to magically change the impact of a performance. To sum it up, how to design and present a performance in a distributed environment should be considered at the initial stage of the research. This is an important factor in terms of audience engagement. GSR sensors can play a role as one of the factors and work as an objective evaluation tool to test what does and does not work. However, it definitely is not a saver to the fundamental issue that exists in a distributed environment.

We have presented a path where a sensor hardware prototype is roughly transformed into a relevant mature commercial product (research questions 1,2, and 3). Moreover, we ultimately obtained results by considering a user experience providing the two versions of sensors. The sensor system shows scalability and robustness in a theatrical environment. However, the system turns out a better performance in the engineering aspect rather than on user experience. The design restricts the usability for operators who have no technical background. For instance, using a command line to control the sleep status of the sensors is not a friendly interface for operators. The design could have been done by simply applying a switch on the sensors. If it was anticipated that a system was to be used by non-technical operators, the user experience should be considered as a priority. In addition, the engineering work should stick to the requirements of the user experience. If the design concept is done in reverse, the system will eventually be replaced by a new user-friendly interface thus the time and energy expended on it will be wasted.

We have developed several software solutions for different user studies (research questions 4, 5, 6, 7, and 8). The results brought us different perspectives of the sensor data. We could use them to infer the user psychological states. However, our work only touched the tip of the iceberg in the sensory data area. First, there is no universal software solution for all scenarios. A specific algorithm needs to be developed for each application. Second, we explored both conscious and unconscious responses from the audiences by using GSR sensors. However, our study still did not answer the relationship between the user psychological states and their subliminal reactions. This is probably due to subliminal responses not always being consistent with subjective reports. Third, the algorithm developed did not deeply investigate the noise issues that appeared in the sensor data. In particular, in field studies, the resources of noise can show different patterns. Some of them can destroy the quality of the raw signals. Smart noise-proof processing methods need to be developed in the future to combat this. Finally, although it satisfied the experimental requirements, the real-time algorithm applied in the experimental scenario seemed to be trivial.

Intelligent solutions need to be developed for sophisticated applications (e.g., defining user bored states in real time).

Instead of just inferring specific user emotions such as happiness, our study addressed both the user emotional arousal and the emotional intensities. This is because we only used one type of sensor. However, the trial experiments provided a productive and valuable new approach that other teams might follow when doing audience research. In the future, we will continue to investigate how to make a sensor product based on the user engagement, while, at the same time, satisfying the technical details. Since the development platform still has the capability to carry more sensors, we will investigate how other types of sensors (e.g., ECG sensor and acceleration sensors), can be integrated into our sensor network in order to provide a complete representation of a group user engagement.

Security is an emerging issue for wearable technology. In many applications, the user sensor raw data were transferred to a cloud service. In doing this, they are vulnerable to access by a third party. In order to protect a user's privacy, it was suggested that the raw data not be stored in the cloud. The required features could be sent to a mobile app or a cloud service for algorithm development. In extreme cases where the user raw data has been sent to an authorized party, the data can be cyphered and the reverse engineering work cannot be done. In this way, we can protect the user's privacy, and, at the same time, wearable products could be developed for daily life.

Our solution can be deployed in any crowded event where it is valuable to understand the audience experience. Since we have already successfully tested the system during some public lectures we are planning to expand the reach of the hardware to other venues. These include listening to music at a concert, visiting a museum, and watching movies at the cinema. Our toolbox allows for not only a better understanding of the audience engagement but also how to use the quantified information in real-time for visualization or interactivity purposes.

# REFERENCES

1. Shimmer sensor: http://www.shimmersensing.com/
2. Empatica: https://www.empatica.com/e4-wristband
3. National Theatre of London. 2015. NT Live. http://timeandspace.org/ntlive.
4. The Metropolitan Opera. 2015. The Met: Live in HD. http://www.metopera.org/metopera/liveinhd/ LiveinHD.aspx.
5. Ashish Kapoor, Winslow Burleson and Rosaline W. Picard. Automatic prediction of frustration. Int. J. Human-Computer Studies (2007).
6. Anmol Madan, Ron Caneel, and Alex "Sandy" Pentland. 2004. GroupMedia: distributed multi- modal interfaces. In Proceedings of the 6th international conference on Multimodal interfaces (ICMI '04). ACM, New York, NY, USA, 309-316. DOI=10.1145/1027933.1027983
7. Andruid Kerne, Andrew M. Webb, Steven M. Smith, Rhema Linder, Nic Lupfer, Yin Qu, Jon Moeller, and Sashikanth Damaraju. 2014. Using Metrics of Curation to Evaluate Information-Based Ideation. ACM Trans. Comput.-Hum. Interact. 21, 3, Article 14 (June 2014).
8. Anselm Strauss and Juliet M. Corbin. 1998. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. SAGE Publications.
9. Bakhshi, H., Mateos-Garcia, J. and Throsby, D. 2010. Beyond Live: digital innvoation in the performing arts. [Accessed 10 February 2016]. Available from: http://eprints.brighton.ac.uk/7234/.
10. Bennett, S. 2013. *Theatre Audiences*. Routledge.
11. Bunting, Catherine, and John Knell. "Measuring quality in the cultural sector. The Manchester metrics pilot: Findings and lessons learned." Arts Council England (2014).
12. Bradley, M.M., Greenwald, M.K., Petry, M.C. and Lang, P.J. 1992. Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology. Learning, memory, and cognition*. 18(2), pp.379–390.
13. C. Wang and P. Cesar, "The Play Is a Hit - But How Can You Tell?" in Proceedings of ACM Creativity and Cognition (ACM C&C 2017), Singapore, June 27-30.
14. C. J. Stevens, E. Schubert, R. H. Morris, M. Frear,J. Chen, S. Healey, C. Schoknecht, and S. Hansen. 2009. Cognition and the temporal arts: Investigating audience response to dance using PDAs that record continuous data during live performance. International Journal of Human- Computer Studies, 67(9):800 – 813.

15. Chen Wang, Erik N. Geelhoed, Phil P. Stenton, and Pablo Cesar. 2014. Sensing a Live Audience. In *Proc. CHI.* 1909–1912.

16. Two cases

17. Chanel, G., Rebetez, C., Bétrancourt, M. and Pun, T. 2008. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era – MindTrek '08., p.13.*

18. Celine Latulipe, Erin A. Carroll, and Danielle Lottridge. 2011. Evaluating longitudinal projects combining technology with temporal arts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11). ACM, New York, NY, USA, 1835- 1844. DOI=http://dx.doi.org/10.1145/1978942.1979209

19. Celine Latulipe, Erin A. Carroll, and Danielle Lottridge. 2011. Love, Hate, Arousal and Engagement: Exploring Audience Responses to Performing Arts. In Proc. CHI.

20. Carnwath, John D., and Alan S. Brown. "Understanding the value and impacts of cultural experiences." Manchester, United Kingdom: Arts Council England (2014).

21. C. J. Stevens, E. Schubert, R. H. Morris, M. Frear,J. Chen, S. Healey, C. Schoknecht, and S. Hansen. 2009. Cognition and the temporal arts: Investigating audience response to dance using PDAs that record continuous data during live performance. International Journal of Human- Computer Studies, 67(9):800 – 813.

22. C. Martella*, E. Gedik*, L. Cabrera-Quiros*, G. Englebienne, and H. Hung, "How Was It? Exploiting Smartphone Sensing to Measure Implicit Audience Responses to Live Performances", ACM MM 2015 (oral) *These authors contributed equally to this article

23. Christopher Peters, Ginevra Castellano, Sara de Freitas. An exploration of user engagement in HCI. AFFINE '09, November 6,2009. Boston,MA, USA.

24. Charalampos Liolios, Charalampos Doukas, George Fourlas, and Ilias Maglogiannis. 2010. An overview of body sensor networks in enabling pervasive healthcare and assistive environments. In Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '10), Fillia Makedon, Ilias Maglogiannis, and Sarantos Kapidakis (Eds.). ACM, New York, NY, USA, , Article 43 , 10 pages. DOI=10.1145/1839294.1839346

25. Dance, T. 2010. Exploring the Design Space in. Evolution.,

26. Dmochowski, J.P., Bezdek, M.A., Abelson, B.P., Johnson, J.S., Schumacher, E.H. and Parra, L.C. 2014. Audience preferences are predicted by temporal reliability of neural processing. Nature Communications. 5, pp.1–9.

27. Dykstra, Dean Julian. A Comparison of Heuristic Evaluation and Usability Testing: The Efficacy of a Domain-Specific Heuristic Checklist. A & M Univ., Texas, 1993.

28. Eva Oliveira, Mitchel Benovoy, Nuno Ribeiro, Teresa Chambe. Towards Emotional Interaction: Using movies to automatically learn users' emotional states. INTER ACT 2011, part I, lncs 6946. Pp. 152-161,2011.

29. Fleureau, J., Guillotel, P. and Orlac, I. 2013. Affective benchmarking of movies based on the physiological responses of a real audience. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*., pp.73–77.

30. Fairclough, Stephen H. "Fundamentals of physiological computing." *Interacting with computers* 21.1-2 (2008): 133-145.

31. Gonzalez, B., Carroll, E. and Latulipe, C. 2012. Dance- inspired technology, technology-inspired dance. *Proceedings of the 7th Nordic Conference on Human-Computer Interaction Making Sense Through Design - NordiCHI '12*., p.398.

32. GREENFIELD, Adam. Everyware: The dawning age of ubiquitous computing. New Riders, 2010.

33. Heather L. O'Brien and Karon E. Maclen Measuring the User Engagement Process. Engagement by Design.

34. Hoffman, D. L., Franke, G. R. (1986). Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research,* 23, 213-227.

35. Ingwer Borg and Patrick J.F. Groenen. Modern Multidimensional Scaling: Theory and Applications. 2005 Springer Science+Buiness Media, Inc. ISBN-10: 0-387-25150-2.

36. Latulipe, C., Charlotte, C., Carroll, E. a and Lottridge, D. 2011. Love , Hate , Arousal and Engagement : Exploring Audience Responses to Performing Arts. *Performing arts*., pp.1845–1854.

37. Louise Barkhuus, Arvid Engstro ̈m, and Goranka Zoric. 2014. Watching the Footwork: Second Screen Interaction at a Dance and Music Performance. In *Proc. CHI*. 1305–1314.

38. Lunn, D. and Harper, S. 2010. Using galvanic skin response measures to identify areas of frustration for older web 2.0 users. *W4A 10 Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility W4A. (January), pp.1–10.*

39. Lindgaard, Gitte, and TW Allan Whitfield. "Integrating aesthetics within an evolutionary and psychological framework." Theoretical Issues in Ergonomics Science 5.1 (2004): 73-90.

40. *Mandryk, R.L. 2004. Objectively evaluating entertainment technology. Extended abstracts of the 2004 conference on Human factors and computing systems., p.1057.*

41. Madgunda, Sneha, Upasna Suman, and Sai Praneeth G. Raunak Kasera."Steps in Requirement Stage of Waterfall Model." International journal of computer & mathematical sciences, pp. 86-87, 2015.

42. Matarasso, F. "Creative Progression: Reflections on quality in participatory arts." UNESCO Observatory Multi-Disciplinary Journal in the Arts 3.3 (2013): 1-15.

43. Mohammad Adibuzzaman, Niharika Jain, Nicholas Steinhafel, Munir Haque, Ferdaus Ahmed, Sheikh Ahamed, and Richard Love. 2013. In situ affect detection in mobile devices: a multimodal approach for advertisement using social network. SIGAPP Appl. Comput. Rev. 13, 4 (December 2013), 67-77. DOI=10.1145/2577554.2577562

44. Molich, Rolf, and Jakob Nielsen. "Improving a human-computer dialogue." Communications of the ACM 33, no. 3, pp. 338-348, 1990.

45. Matthew K.X.J. Pan, Gordon Jih-Shiang Chang, Gokhan H. Himmetoglu, AJung Moon, Thomas W. Hazelton, Karon E. MacLean, and Elizabeth A. Croft. 2011. Galvanic skin response-derived bookmarking of an audio stream. In CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11). ACM, New York, NY, USA, 1135-1140.

46. Nielsen, Jakob. "Enhancing the explanatory power of usability heuristics." In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 152-158. ACM, 1994.

47. Nargess Nourbakhsh, Yang Wang, Fang Chen, Rafael A. Calvo. 2012. *Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks*. In Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI '12), Vivienne Farrell, Graham Farrell, Caslon Chua, Weidong Huang, Raj Vasa, and Clinton Woodward (Eds.). ACM, New York, NY, USA, 420-423. DOI=10.1145/2414536.2414602

48. O'Brian, H. and MacLean, K.E. 2009. Measuring the User Engagement Process. Advances., pp.1–6.
Picard, R.W. 1995. Affective Computing.

49. Picard, Rosalind W. 2009. "Future affective technology for autism and emotion communication." Philosophical Transactions of the Royal Society B: Biological Sciences 364.1535: 3575-3584.

50. Posner, J., Russell, J.A. and Peterson, B.S. circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology. 17, pp.715–734.

51. Radbourne, J., Johanson, K., Glow, H. and White, T. 2009. The Audience Experience: Measuring Quality in the Performing Arts. International Journal of Arts Management., 11(3), pp.16–29.

52. Pinelle, David, Nelson Wong, and Tadeusz Stach. "Heuristic evaluation for games: usability principles for video game design." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1453-1462. ACM, 2008.

53. Reason, M. 2010. The Young Audience: Exploring and Enhancing Children's Experiences of Theatre. *Trentham Books Ltd*.

54. Roto, V., Obrist, M. and Väänänen-Vainio-Mattila, K. 2009. User experience evaluation methods in academic and industrial contexts. In: Interact 2009 conference, User Experience Evaluation Methods in Product Development (UXEM'09).

55. Philip Auslander. 2008. *Liveness: Performance in a mediatized culture* (2nd ed.). Routledge.

56. Ruan, S., Chen, L., Sun, J. and Chen, G. 2009. Study on the Change of Physiological Signals During Playing Body- controlled Games *In*: *Proceedings of the International Conference on Advances in Computer Enterntainment Technology.*, pp. 349–352.

57. Russell, J.A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*. 39(6), pp.1161– 1178.

58. R.Mandryk. Objectively evaluating entertainment technology. In CHI'04, pages 1057–1058. ACM Press, 2003.

59. Rahman, Md Mahbubur, et al. 2014. "Are we there yet?: Feasibility of continuous stress assessment via wireless physiological sensors." Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM.

60. Pejman Mirza-Babaei, Lennart E. Nacke, John Gregory, Nick Collins, and Geraldine Fitzpatrick. 2013. How does it play better?: exploring user testing and biometric storyboards in games user research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13). ACM, New York, NY, USA, 1499-1508.

61. Sidharth Nabar, Ayan Banerjee, Sandeep K. S. Gupta, and Radha Poovendran. 2010. Evaluation of body sensor network platforms: a design space and benchmarking analysis. In Wireless Health 2010 (WH '10). ACM, New York, NY, USA, 118-127. DOI=10.1145/1921081.1921096

62. Sauro, J. and Lewis, J.R. 2012. Chapter 2 - Quantifying User Research *In*: *Quantifying the User Experience.*, pp. 9–18.

63. Sawchuk, A.A., Chew, E., Zimmermann, R., Papadopoulos, C. and Kyriakakis, C. 2003. From remote media immersion to distributed immersive performance *In*: *2003 ACM SIGMM Workshop on Experiential Telepresence, ETP '03.*, pp. 110–120.

64. Shneiderman, Ben. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India, 2010.

65. Sheppard, R.M., Kamali, M., Rivas, R., Tamai, M., Yang, Z., Wu, W. and Nahrstedt, K. 2008. Advancing interactive collaborative mediums through tele- immersive dance (TED): a symbiotic creativity and design environment for art and computer science. *Digital Media*., pp.579–588.

66. Stevens, C.J., Schubert, E., Morris, R.H., Frear, M., Chen, J., Healey, S., Schoknecht, C. and Hansen, S. 2009. Cognition and the temporal arts: Investigating audience response to dance using PDAs that record continuous data during live performance. *International Journal of Human-Computer Studies*. 67(9), pp.800–813.

67. Steve Benford, Andy Crabtree, Martin Flintham, Adam Drozd, Rob Anastasi, Mark Paxton, Nick Tandavanitj, Matt Adams, and Ju Row-Farr. 2006. Can You See Me Now? *ACM Trans. Comput.-Hum. Interact.* 13, 1 (March 2006), 100–133.

68. Steve Benford, Gabriella Giannachi. 2011. *Performing Mixed Reality*. MIT Press.

69. Susan Broadhurst. 2006. Intelligence, Interaction, Reaction, and Performance. In *Performance and Technology: Practices of Virtual Embodiment and Interactivity*, Susan Broadhurst and Josephine Machon (Eds.). Palgrave Macmillan.

70. Schiffman, Susan S., M. Lance Reynolds, and Forrest W. Young (1981), Introduction to Multidimensional Scaling: Theory, Methods, and Applications, NY: Academic Press.

71. Stuart Reeves, Steve Benford, Claire O'Malley, and Mike Fraser. 2005. Designing the Spectator Experience. In *Proc. CHI*. 741–750.

72. Trevor F.Cox and M.A.A. Cox. Multidimensional Scaling, Second Edition. ISBN 1-58488-094-5.

73. Tao Lin, Akinobu Maejima, and Shigeo Morishima. 2008. Using subjective and physiological measures to evaluate audience- participating movie experience. In Proceedings of the working conference on Advanced visual interfaces (AVI '08). ACM, New York, NY, USA, 49-56. DOI=10.1145/1385569.1385580

74. Tschacher, Wolfgang, Steven Greenwood, Volker Kirchberg, Stéphanie Wintzerith, Karen van den Berg, and Martin Tröndle. 2012. 'Physiological correlates of aesthetic perception of artworks in a museum'. Psychology of Aesthetics, Creativity and the Arts 6:1. 96- 103.

75. Teresa Cerratto-Pargman, Chiara Rossitto, and Louise Barkhuus. 2014. Understanding Audience Participation in an Interactive Theater Performance. In *Proc. NordiCHI*. 608–617.

76. Thecla Schiphorst, Wynnie (Wing Yi) Chung, and Emily Ip. 2013. Wo.Defy:

Wearable Interaction Design Inspired by a Chinese 19th Century Suffragette Movement. In *Proc. TEI*. 319–322.

77. Wang, C., Geelhoed, E.N., Stenton, P.P. and Cesar, P. 2014. Sensing a live audience *In*: *CHI '14*., pp. 1909–1912.

78. Wang, C., Wong, J., Zhu, X., Roggla, T., Jansen, J. and Cesar, P. 2016. Quantifying Audience Experience in the Wild: Heuristics for Developing and Deploying a Biosensor Infrastructure in Theaters. *In: Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEX2016).*

79. William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity As a Resource for Design. In *Proc. CHI*. 233–240.

80. William A. Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on Twitch: Fostering Participatory Communities of Play Within Live Mixed Media. In *Proc. CHI*. 1315–1324.

81. Walmsley, Ben. 2013. '"A big part of my life": a qualitative study of the impact of theatre.' Arts Marketing: An International Journal 3:1. 73 – 87.

82. *Yang, Z., Yu, B., Wu, W., Nahrstedt, K., Diankov, R. and* Bajscy, R. 2006. A Study of Collaborative Dancing in Tele-immersive Environments *In*: *Eighth IEEE International Symposium on Multimedia, 2006. (ISM'06)*., pp. 177–184.

83. Young, Forrest W., and Robert M. Hamer (ed.) (1987), Multidimensional Scaling: History, Theory, and Applications, Hillsdale, NJ: Erlbaum.