Multi-layered platform demonstration of MuseIT co-creation services (I)

Document Information

Project Name:	Multisensory, User-centred, Shared cultural Experiences through Interactive Technologies
Project Acronym:	MuseIT
Grant Agreement No:	101061441
Deliverable No:	D5.2
Deliverable Name:	Multi-layered platform demonstration of MuseIT co-creation services (I)
Work Package:	WP5
Task:	T5.1, T5.2, T5.3, T5.4
Dissemination Level:	PU
Deliverable Type:	DEM
Lead Organization:	SHMU
Lead Member:	Nigel Osborn
Submission month:	M18
Date:	31 March 2024

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Document history

Revision	Date	Description / Reason of change	Submission by
v0.1	17-12-24	Structure proposal and initial draft	Moa Johansson
v0.2	07-02-24	First draft for internal review	Moa Johansson
v0.3.1	28-02-24	Second draft addressing review comments	Nigel Osborne
v0.3.2	15-03-24	Second version of v 0.3 including participatory session	Nigel Osborne
v0.4	25-03-24	Final draft addressing the PMB review comments	James Hanlon
v1.0	28-03-28	Final draft submitted to the EU	Renata Sadula

Authors

Partner	Name(s)
SHMU	Nigel Osborne, Moa Johansson
XSL	John Turner, Ruaraidh Osborne, James Hanlon, Jonathan Walton
CERTH	Maria Kyrou, Panagiotis Petrantonakis
CTL	Nikolas Petrou, Georgia Christodoulou

Contributors

Partner	Contribution type	Name
KCL	Review	Albert Merono Penuela
ACT	Review	Damien Faux
DANS	Review	Andrea Scharnhorst

Glossary

Acronym	Definition
ACCEL	Accelerometer
ACF	Affective Computing Framework
AI	Artificial Intelligence
AM	Amplitude Modulation
BVP	Blood Volume Pulse
CCRMA	Centre for Computer Research in Music Acoustics (Stanford)
CNN	Convolutional Neural Networks
ECG	Electrocardiogram

EEG	Electroencephalogram
FACS	Facial Action Coding System
FER	Facial Emotion Recognition
GSC	Galvanic Skin Conductance
HR	Heart rate
HRV	Heart rate variability
IBIs	Inter beat intervals
IRCAM	Institute for Research and Coordination in Acoustics/Music (Paris)
LOOCV	Leave-one subject-out cross-validation
LSL	Lab streaming layer
MAX/MSP	Max Signal Processing
MEA	Mood Estimation Algorithm
MIDI	Musical Instrument Digital Interface
ML	Machine Learning
MLP	Multi-layer Perception
MODA	Multiscale Oscillatory Dynamics Analysis
MTCNN	Multi-task Cascade Conventional Neural Network
QAM	Quadrate Amplitude Modulation
RNN	Recurrent Neural Networks
TEMP	Temperature
SVM	Support Vector Machine
VR	Virtual Reality
WESAD	Wearable Stress and Affect Detection
WP	Work Package

Table of contents

Executive Summary

The remote co-creation platform is the creative, pro-active component of a wider platform concerned with offering disabled users and others remote access to cultural assets. It is multi-layered because it offers simultaneously different levels of creative activity (ranging from entirely autonomous individual creativity to AI generated composition) different levels of sensing of states of mind and body, including behaviour of the heart, electrical brain activity and recognition of facial expression, and different ways of communicating this information, ranging from haptics to avatars.

Introduction

The Introduction is concerned with the purposes and function of the platform, the nature of various routings and the relationship of the project to disability

Background Architecture

In order to describe the multi-layered nature of the MuseIT Remote co-creation platform, and in order to explain the complex relationships between the many layers within it, we preface the Demonstrations with a description of the background architecture for the Dashboard, where all routings and connections are made clear.

This description includes -

- A diagram of the background architecture
- List of inputs
- Settings

Following the description of the Dashboard background architecture, there is a short report on progress with the choice of app (P.18) which will handle sensor data collection, processing, transmission through JackTrip, avatar display, and haptics drivers.

Demonstrations

Although significant progress has already been made with all layers of the system, the WP5 team agreed to present for Demonstration those that are developed to the point of being either ready, or close to being ready for integration.

- The first demonstration (XSL) is the signal channel, intended to add sensor data to spare capacity in the JackTrip channel, which allows data to be communicated together with JackTrip audio data and at the same speed.
- The second demonstration (CERTH) is concerned with sensor diagnostics, and in particular using sensor data for emotion recognition and mood induction.
- The Third Demonstration (CTL) involves mood estimation through the use of Facial Emotion Recognition, including partially occluded faces.
- The fourth Demonstration (CTL) is concerned with stress estimation, through calculation of Heart Rate Variability
- The fifth Demonstration (XSL) is concerned with a computational model of the musical brain capable of predicting the neurophysiological effects of individual tracks of music.
- The sixth Demonstration (XSL) shows how the platform “audifies” users’ EEG, and then uses its computational model of the musical brain to search for existing music in the world repertoire that is closest to the users’ EEG.

Participatory Workshop

The participatory workshop was in effect a “seventh demonstration”. It explored the potential of Heart Rate signals, both audio and haptic to communicate emotions and states of mind and body

between co-creators both in proximity and remotely. It also marked the beginning of the process of the design of avatars.

Appendices

Most of the Demonstrations involve detailed description and referencing. The advice of our reviewers was to include this more detailed work in the Appendices.

1. Introduction

Work Package 5 is concerned with the design and proof of concept of a remote co-creation platform, focused on the needs of those with disability, but also intended for universal use. The main concerns are zero latency - that is to say, no delay in the sound signal between users - and enhanced expressive and emotional communication. When musicians play in close proximity, a whole series of vitality affects, intuitions and intentional and emotional cueing signals are shared. At a distance this important information is lost; the intention of Muse-IT is to replace it by means of relevant new available sensor and communication technologies. The same technologies may be used to support users who cannot speak or move. Muse-IT is capable of helping users generate music from their minds and bodies without verbal or gestural communication. It also uses AI tools to support a wide range of creative compositional processes.

The principal areas of technology implemented on the platform may be described as:

1. Effective low latency, allowing co-creators to work in "real time" without the delays on standard communication platforms.
2. Sensor and communication technologies, allowing users to cue one another and "share" states of body and mind as they would in physical proximity in "real life".
3. Sensor and communication technologies to enhance users' creative self-expression, particularly in the case of users with challenges in verbal or physical communication.
4. AI tools to support users in creative processes.

The platform is multi-layered, in the sense described above, of different layers of creativity, ranging from autonomous to AI-supported, and different levels of sensing and communication. In order to present complex and frequently overlapping layers in a comprehensible way we have prefaced the demonstration with a diagram and summary of the background architecture for the Dashboard, somewhat in advance of schedule. In this way we can demonstrate routing of layers throughout the system from input to output, where they interlock and where they diverge, their relationship to dashboard controls and of course their function. In order to make this clear, we summarise the whole system, including layers such as composition algorithms and AI tools, not yet ready for integration.
There follows a short discussion about app technology choices, including options for UI (game engines, Electron etc.) and avatar display (Godot, Unreal etc.).

The Demonstrations describe layers that are currently functioning and ready for integration. In the case 3.2, the layers, once integrated, will bifurcate, with data routed on the one hand to a), the communication of states of mind and body between co-creators and on the other hand to b), support of processes of self-expression as well as input for AI tools to assist in the process of composition.

Under User-engagement/prototype testing we describe the latest participatory workshops, examining the use of HR and HRV sensors, haptics and avatars, and presenting the results. It has clear relations to demonstration case 3.1.

2. Background architecture for the Dashboard

2.1 Background architecture diagram

2.1.1 Alternative Systems

Video calling and conferencing systems have become increasingly popular in recent years due to the pandemic and the move to remote/hybrid working. Systems such as Zoom, Skype and Teams allow voice and video communication, but fall short when people try to use them for music due to a mix of high latency and audio quality that has been optimised for speech. Additionally these systems do not have ways to send additional data synchronised with the audio stream to augment sessions with sensor or other data. A big advantage of JackTrip's use of JACK is that we can control and mix data into the audio being transmitted.

2.1.2 Dashboard Architecture

The dashboard diagram is based on decisions concerning the WP5 architecture taken during valuable partner meetings when the consortium met in Cyprus, October 2023. The diagram tracks the routing of layers throughout the system, from initial input to final output. An important feature is the local sensor hub where data is gathered and then directed either to further processing or to outputs. The diagram distinguishes between a) communication of states of mind and body between co-creators and b) co-composition and performance (points 2. and 3. of the introduction above). But there is a single sensor hub; and there are emotional detection algorithms and other layers that serve both a) and b).

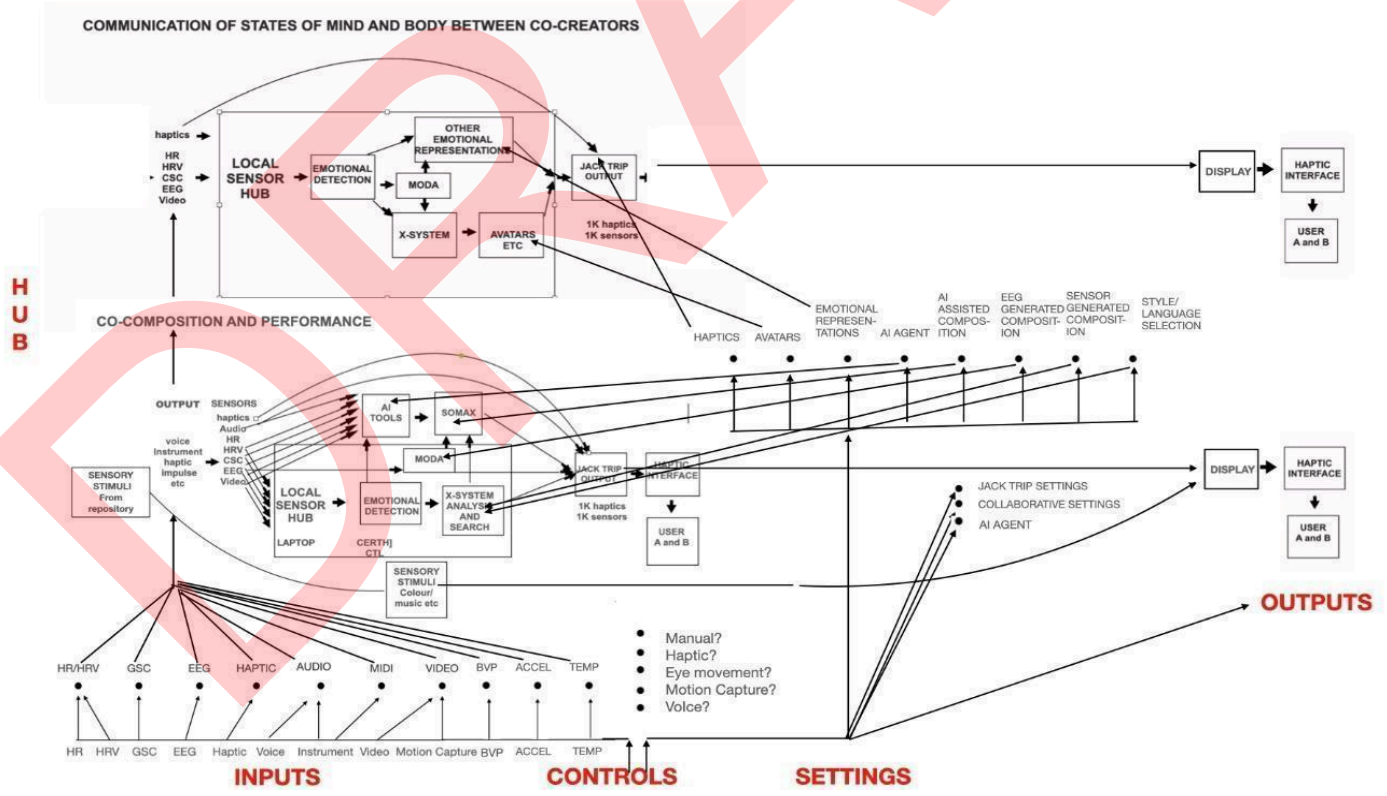


Figure 1: Background architecture for dashboard

2.2 Background architecture – layers

In the subsequent sections, we will describe the multi-layer nature of the platform, through the optics of controls, input and settings.

2.2.1 Controls

The nature of the controls will be determined by the participatory session with the co-designers and potential users. The March workshops are concerned with heart rate and haptics and reported in this paper. The April workshops will test all sensors to be used in the system, including facial emotion recognition. The process will not be completed until next year, when all of the possibilities ranging from conventional manual controls to haptics, eye movement, movement capture or vocal cues (voice) will have been fully explored and tested.

These controls will need to be able to select and activate inputs, settings, and outputs, and to control levels. There may be more than one form of control.

2.2.1.1 Inputs

INPUTS	HR/HRV		GSC	EEG	HAPTIC	AUDIO		MIDI	VIDEO/IMAGE		BVP	ACCEL	TEMP
	•		•	•	•	•		•	•		•	•	•
SENSORS	↗	↖	↑	↑	↑	↗	↖	↑	↗	↖	↑	↑	↑
	HR	HRV	GSC	EEG	Haptic	Voice		Instrument	Camera	Motion capture	BVP	ACCEL	TEMP

2.2.1.2 HR and HRV

HR is Heart Rate and describes the speed at which the heart is beating, usually in beats per minute. HRV is a measure of Heart Rate Variability. When we have negative emotions our hearts tend to beat in a rigid manner with low variability. When we have positive emotions the heart tends to beat with higher variability. HRV may therefore act as a measure of valence or vagal power. An ECG is an Electrocardiogram that measures the heart's rhythm and electrical activity.

The Heart Rate and Heart Rate Variability input will route ECG sensor data to the sensor hub, where it will be analysed by algorithms, developed by partners CERTH and CTL, to extract HR and HRV values both to contribute to CERTH AI identification of emotions for communication to co-creators and for use in music generation. HR data may also be routed via the sensor hub or directly to Jack Trip outputs to the co-creator's haptic interface. Furthermore, the HR and HRV data will be routed to CTL's stress estimation algorithm.

ECG → Sensor hub → HR and HRV extraction (CERTH/CTL)

→ Emotion identification AI (CERTH)

→ HR signal directly to-creator's haptic interface

→ Music generation

2.2.1.3 BVP

BVP, or Blood Volume Pulse, is a method of detecting heart beats by measuring the volume of blood passing the sensor in either red or infrared light.
Blood Volume Pulse input will route BVP data to the sensor hub where it will contribute to CERTH extraction of HRV values and to CERTH AI identification of emotions, and possibly to music generation.

BVP → Sensor hub → HR and HRV extraction (CERTH) → Emotion identification AI (CERTH)

2.2.1.4 GSC

GSC, or Galvanic Skin Conductance is a method to measure the electrical conductivity of the skin in response to stimuli. When we experience something particularly emotional in some way, we trigger our sweat glands in very small ways that we are not aware of, whereby our skin becomes more conductive to electricity. In general high conductivity is related to wet skin and high arousal autonomic activity and low conductance to dry skin and low autonomic arousal.
The Galvanic Skin Conductance input will route GSC sensor data to the sensor hub where it will contribute to CERTH AI identification of emotions and possibly contribute to music generation.

EEG → Sensor hub → Emotion identification AI (CERTH)
→ Music generation

2.2.1.5 EEG

EEG , or Electroencephalography is the recording of electrical brain activity, usually related to different levels of consciousness and wakefulness.
The Electroencephalography input will route multi-channel EEG data to the sensor hub where it will contribute to CERTH AI identification of emotions and will be further routed to MODA/XSL and/or to SU brain stethoscope technology for audification and use in music generation.

EEG → Sensor hub → Emotion identification AI (CERTH)
→ MODA (XSL) and/or Brain Stethoscope for audification and music generation

2.2.1.6 Haptics

The Haptic input will route data from haptic sensors to the sensor hub, and then directly to JackTrip outputs, and on to the receiver haptic interfaces of co-creators.

Haptics → Sensor hub → JackTrip outputs

2.2.1.7 Audio

The Audio input will receive signals from a pair of stereo microphones and might also receive signals from contact mikes. This may involve voice or live instruments. The signal will be routed to the sensor hub, then directly via the JackTrip output to co-creators. The signals may also be used to accompany or modulate music generation.

Audio → Sensor hub → JackTrip outputs
→ Accompaniment and modulation of music generation

2.2.1.8 MIDI

MIDI, or Musical Instrument Digital Interface is a standard to transmit and store music, originally designed for digital music synthesisers. MIDI does not transmit recorded sounds. Instead, it includes musical notes, timings and pitch information, which the receiving device uses to play music from its own sound library.

The Musical Instrument Digital Interface input will route MIDI data to the sensor hub, then directly via the Jack Trip output to co-creators. The signals may also be used to accompany or modulate music generation, in particular SOMAX-based AI.

MIDI	→ Sensor hub	→ JackTrip outputs
		→ Accompaniment and modulation of music generation

2.2.1.9 Video

The Video input will be linked to a video camera and will route visual facial data to the CTL mood estimation algorithm. The results will be either represented by haptics or avatars or combined with CERTH emotional identification AI. The video may also be used for motion capture.

Video	→ Sensor hub	→ Motion Capture	
		→ Mood Estimation Algorithm (CTL)	→ Direct to haptic/avatar representation
			→ Combined emotional identification AI (CERTH)

2.2.1.10 ACCEL

The accelerometer input will route data to the CERTH emotional identification AI.

ACCEL	→ Sensor hub	→ Emotional identification AI (CERTH)

2.2.1.11 TEMP

The temperature input will route peripheral skin temperature data to the CERTH emotional identification AI.

TEMP	→ Sensor hub	→ Emotional identification AI (CERTH)

2.2.1.12 Eye Tracker

If the Eye tracker is also used as a control, then non-control data may be sent to the sensor hub and CERTH emotional identification AI. If not, then there will be a dedicated eye tracker input on the dashboard.

Eye tracker	→ Sensor hub	→ Emotional identification AI (CERTH)

2.2.2 Settings

Haptics	Avatars	Emotional representations	AI agent	AI assisted composition	EEG generated composition	Sensor generated composition	Style/ Language selection
•	•	•	•	•	•	•	•
↑	↑	↑	↑	↑	↑	↑	↑

2.2.2.1 Haptics

Haptics is defined as a technology that transmits tactile information using sensations such as vibration, touch, and force feedback.

Haptic and vibrational data (e.g. musical cueing, HR, selected low frequency sound etc.) will be routed from haptic, HR and HRV sensors by way of the relevant INPUTS to the sensor hub, and then on to JackTrip outputs. The haptic signals will be received by co-creators by way of the receiver haptic interface.

Haptic output	→ Haptic input	↘		
Audio	→ Audio input	→ Sensor hub	→ JackTrip output	→ Co-created haptic interface
HR sensor	→ HR/HRV input	↗		

2.2.2.2 Avatars

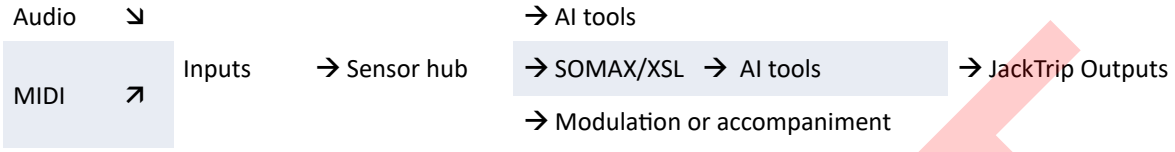
An avatar is an electronic image that may represent a person or an emotion and is manipulated by a computer user (as in a computer game.)
Avatars will be generated from the full range of sensor inputs as well as musical/audio inputs, from CERTH emotional identification AI and/or CTL mood estimation and/or XSL autonomic arousal/vagal power colour circles. Avatar data will be directed to JackTrip outputs, and then on to the co-creators' receiver display.

		↗ Emotional identification AI (CERTH) ↘				
Inputs	→ Sensor hub	→ Mood estimation algorithm (CTL) →	Avatar generator	→	JackTrip outputs	→ Co-creator display
		→ Audio Input (XSL) →				
		↘ MODA EEG Input (XSL) ↗				

2.2.2.4 AI Agent

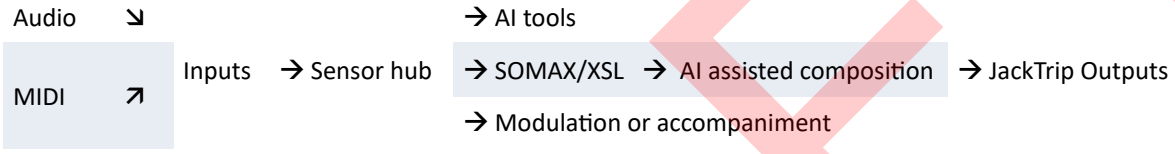
AI agents usually control or optimise devices, or enable robots to perform tasks. In music they may be regarded as artificially intelligent creative agents, capable of entering creative dialogues with human beings. XSL, or X-System is a computational model of the musical brain capable of predicting the neurophysiological effects of music and identifying music close to the electrical brain activity of individuals. SOMAX (Somax2) is an application for musical improvisation and composition. It is implemented in Max and is based on a generative model using a process similar to concatenative synthesis to provide stylistically coherent improvisation, while in real-time listening to and adapting to a musician (or any other type of audio or MIDI source).

User or users may select an AI agent as co-creator. The AI agent will "react" to the user's musical output and/or "co-improvise" and "co-create". Audio and MIDI signals from the user will be routed through INPUTS to the sensor hub, and then either directly to AI tools, or by way of XSL analysis and search, then on to SOMAX, or to other AI composition resources.



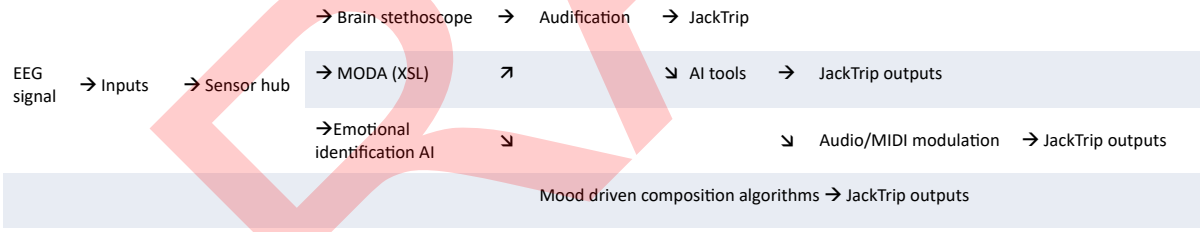
2.2.2.5 AI assisted composition

Here audio and MIDI signals from the user will be routed through INPUTS to the sensor hub and then directly to AI composition resources; once again, this procedure may include X-System analysis and searches. Audio inputs may also modulate or accompany AI assisted composition.



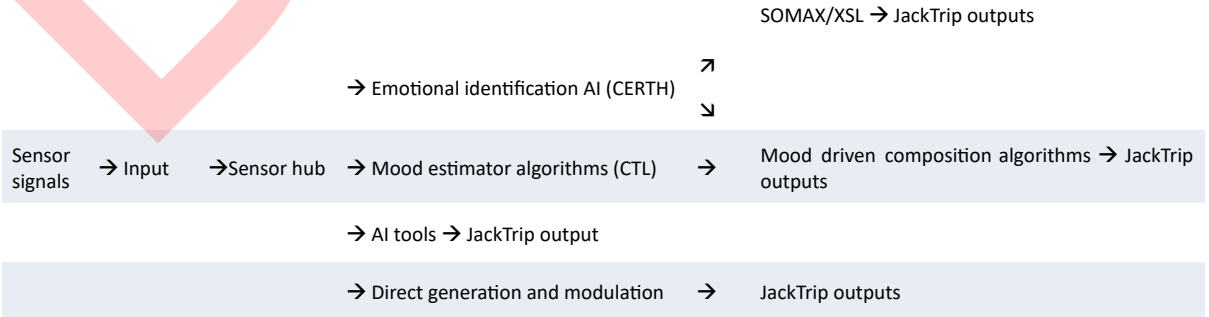
2.2.2.6 EEG generated composition

EEG signals are routed through INPUTS to the sensor hub, then to MODA/XSL or to Brain Stethoscope for audification; these signals may be further routed to AI composition resources. EEG signals may also be directed to CERTH emotional identification AI, and then on to mood-driven compositional algorithms.



2.2.2.7 Sensor generated composition

Sensor signals may be routed through INPUTS to the sensor hub and then subsequently either directly on to AI tools and/or direct music generation or modulation, or by way of emotional identification AI and/or the mood estimator to mood-driven composition algorithms such as CTLS's appliance of the MusicGen system, and/or XSL and SOMAX and possibly more.



2.2.2.8 Style/language personal selection

There are several interfaces in the unfolding of co-creation procedures where musical style and language choices may be made - in relation to AI tools, AI composition resources, Brain stethoscope and X-System searches. This setting will connect directly to these interfaces.

2.3 App Technology

2.3.1 Choices

The WP5 team have explored various technology choices for an app which will handle sensor data collection, processing, transmission through JackTrip, avatar display, and haptics drivers. Below we outline the advantages and disadvantages to consider for different choices, as well as some technology options.

Add to JackTrip's UI:

Pros	Cons
Would have just a single UI for everything	Stuck using their technology stack (Qt in particular was being a pain to set up)
May get good support form JackTrip themselves.	Unknown accessibility options
	Need to be careful not to break the app for connecting to servers.
	Will have to keep our code working with updating versions of JackTrip

Electron UI:

Pros	Cons
Good, mature accessibility options	Need to use native plugins for sensor drivers and audio or run another background process to do those.
Familiarity with the technology stack	Need to be careful about performance issues.
	Some overhead if using WebGL vs Vulkan/etc.

Technology Options for Overall UI includes Vue JS and React Native and for avatar display https://www.babylonjs.com/, Godot HTML export and Godot OpenGL captured in a window.

Game engine UI:

Pros	Cons
Easy to develop 3D avatar options quickly.	Less mature accessibility options
Very easy to deploy cross-platform.	Somewhat less suited for UI development

Technology Options for Game engine UI includes Godot, Monogame/FNA and Unreal.

Mono/.NET UI:

Pros	Cons
Relatively good performance	Likely worse experience on Mac
Easy access to native code for audio and sensor drivers	Less familiarity with it in the technical team

2.3.2 Recommendation

While we are happy to receive further feedback on this and have further work package discussion, our current recommendation is to proceed with an Electron app. The main drivers for this are the known and well tested accessibility features which should allow better support for the many ways people use computers, as well as our developers' familiarity with the web technologies used in it.

3. Demonstrations

3.1 JackTrip Channels (XSL)

JackTrip[1] was originally developed at CCRMA Stanford University, but as part of MuseIT, is now adapted for the purposes of the co-creation platform (as reported in D5.10). The WP5 team are currently in the process of adding the signal channel layer to JackTrip which will allow sensor and haptic data to be communicated together with, and at the same speed as, JackTrip low-latency, co-creation data.

At the core of the co-creation platform is the ability for users to send and receive biological sensor data in real time, and at low latency. The current plan is to do this by encoding the data into the music signal that is already sent and received at low latency over JackTrip.

Various different encoding options were explored, including Amplitude Modulation (AM) and Quadrature Amplitude Modulation (QAM) of a carrier tone, as well as simply up-sampling and adding low frequency sensor data (such as ECG) directly into the music signal.

All three of these methods were demonstrated to work locally, meaning we were able to encode additional data into a music signal then read out the data and the music without the latter being significantly altered on a single machine.

The next step was to get this working over JackTrip. JackTrip is built on top of JACK[2] (Jack Audio Connection Kit), which is a sound server API. Using an open-source JACK client for Python,[3] we were able to pass our carrier tones (carrying encoded sensor data) into JackTrip and read them back out into Python on another machine. Due to our remote work situation, we were able to test this process between Zagreb (Croatia) and Edinburgh (Scotland).

Various encoding/decoding methods, as mentioned above (AM, QAM, direct signal addition) over JackTrip, have been tested. During the participatory session, 11-12th of March 2024, in Gothenburg, we were able to send an ECG signal in real-time into JackTrip, via JACK, and have that data decoded and transmitted to a haptic receiver connecting to a different machine. Images and further outlining of the session are included in chapter 4, User engagement and prototype testing. This served as a starting point for our exploration into using sensor data in conjunction with music in real time.

3.2 Affective Computing Framework service for Music (ACF-Music) (CERTH)

The Affective Computing Framework service for Music (ACF-Music) currently under development by CERTH comprises a grouping of AI emotional recognition algorithms. Results are plotted on Russell's two-dimensional valence-arousal space model, serendipitously the same approach as XSL. An important sensor input is Galvanic Skin Response (GSR) measuring galvanic conductance across the surface of the skin, dependent on sweat glands - arousal leads to increased gland activity, more moisture and higher conductance; counter-arousal leads to reduced activity, drier skin and less conductance.

The team performed tests on 15 subjects using an Empatica E4 wristwatch-like wearable. A median filter was used to eliminate artefacts, and minmax normalisation to account for individual differences. The GSR time series were grouped in 30 second windows, and statistical metrics extracted from 15

[1] https://www.jacktrip.com/

[2] https://jackaudio.org/

[3] https://jackclient-python.readthedocs.io/en/0.5.4/

features. A machine learning algorithm - a Support Vector Machine with an RBF kernel - was employed to classify features extracted from GSR in the form of arousal estimation.

To facilitate real-time monitoring, the Lab Streaming Layer (LSL) system was used. LSL is a comprehensive system designed for gathering measured time series data in research settings, encompassing networking, time synchronisation, real-time access, and optionally centralised data collection, viewing, and disk recording.

For a full description see Appendix 1

3.3. Mood estimation (CTL)

For mood estimation the CTL team chose to develop, train and validate a model based on Facial Emotion Recognition (FER). Since MuseIT is concerned with inclusiveness and involves VR Technologies, they also decided to focus on developing a parallel version of the algorithm that could be employed for faces occluded by VR headsets or other eyewear that are used by the visually impaired.
The team collected 50,000 online images of emotionally expressive faces. Because of inevitable imbalances in representations of individual emotions, the team chose to focus on three emotions: “happy”, “sad” and “neutral”.

For the development of the Mood Estimation Algorithm (MEA) the team chose the Mini-Xception deep learning model which combines prediction accuracy with negligible inference latency and makes use of residual modules and depth-wise separable convolutions. It also has limited parameters (54,000 trainable parameters overall) and the final model’s size is only a few megabytes.

For a full description see Appendix 2

3.4 Stress estimation (CTL)

The Catalink team also worked with estimation of stress levels of individuals using an electrocardiogram (ECG) signal, based on the inter-beat-intervals (IBIs) and the HRV metric. The extracted features were used to train and evaluate a Machine Learning (ML) model to accurately predict the stress levels of the individuals concerned.

For a full description see Appendix 3

3.5 Neurophysiological prediction (XSL)

X-System is a computational model of the musical brain that can predict the neurophysiological effects of music, and how moment-by-moment the music will activate the autonomic nervous system, endocrine system, auditory cortex, motor cortex and brainstem by XSL. It also calculates arousal and valence and plots values on the same circle as the CERTH system, with the addition of colour coding. As opposed to inducing emotion or estimating mood in the co-creators. X-System predicts these values in the music itself. This provides a very direct feed of emotional information between co-creators and a simple basis for avatar generation.

For a full description see Appendix 4

3.6 EEG Audification (XSL)

XSL has also developed a way of audifying EEG. Which will have two functions in the platform: to audify individual's EEG as a form of self-expression, as the "music of the brain" itself, and to use X-System to search the world repertoire for existing music closest to the electrical activity of the co-creator's brain. This music can then become rich material for AI assisted creativity. The EEG is processed through wavelet correlations and ridge extraction, and the resulting "score" transposed to the domain of audition.

For a full description see Appendix 5

4. User-engagement and prototype testing

The aim of this participatory workshop was:

1. to work with potential users making use of heart rate sensors to evaluate to what extent heart rate signals may be of use in the communication of states of mind and body between co-creators both in proximity and remote.
2. To work with users to evaluate the effectiveness and comfort of haptic signals in the communication of heart rate and other vibrational information.
3. To begin the process of designing “avatars” which will be visual representations of human states of mind and body to assist in emotional communication within the process of remote co-creation.

The use of heart beat audifications proved to be very effective in the communication of emotions and states of mind and body between co-creators. Participants identified the emotions embodied in heart beats and reacted in creative and inventive ways. There were strong creative and emotional reactions to a heart beat made audible in the room. The single haptic actuator was very successful. Participants felt that they were in “intimate” contact with their co-creators. Good progress was made on the design of avatars.

For a full description see Appendix 6

5. Next steps and future work

To summarise, we are on schedule, and have made good progress in relation to the correlations with other Work Packages, Deliverables etc. This initial phase of development has involved work on individual layers "in parallel". Now we are moving into integration and the articulation of the architecture as a whole. Next steps include amongst other things:

- Sensor data encoding and a proof-of-concept for sending this data through JackTrip has been done, further work is needed on improved encoding, phase locking and latency reduction.

- CTL is working on further improving their mood estimation model as well as to evaluate its effectiveness during actual use-case scenarios.

- The integration of CTL algorithms into XSL's technologies has started. A packaged version of the Mood Estimation Algorithm has been delivered to XSL who are working to run and test it.

- CTL aim to collect more data during our experiment sessions, with the PolarH10 sensor. The purpose is to incorporate some of the PolarH10 records into the training dataset, aiming to improve the performance accuracy of our classifier.

- CTL will experiment with more ML algorithms models, in order to find the best performing model.

- CERTH aim to extend the research toward emotion recognition by incorporating sensor signals into the pipeline. In particular, a multimodal approach is feasible by integrating these diverse sensor modalities, thereby enhancing the accuracy and robustness of the emotion recognition system.

- CERTH will examine each modality to determine which is better suited for detecting specific emotional states and explore methods for effectively combining their outputs. This comprehensive analysis will contribute to refining the approach and optimising the performance of the emotion recognition system.

- CERTH's emotion recognition analysis (ACF-Music service) will be integrated into the WP5 dashboard.

- More discussion and decisions on EEG will take place in the coming months, by autumn we hope to be able to have started integrations of EEG and have a prototype EEG layer up and running.

- In April we aim to have another participatory session where we can engage users to evaluate and further design the developments.

- Planning has started for the pilot demonstrations and artistic ideas are being brainstormed.

- Integration of data into the Repository (T6.4) is being discussed.

APPENDIX 1 - Affective Computing Framework service for Music (ACF-Music) (CERTH)

A1.1 Summary

The sensor diagnostics layer includes an emotion induction system designed by CERTH, whereby data from a variety of sensors is combined to induce the mood of the user, and ultimately, a) to convey this emotional information to co-creators, as well as b) to support individual creative self-expression. In both of these functions CERTH diagnostics will be combined with CTL and XSL diagnostics. The induction in itself requires multiple layers within layers, with sensor inputs including HR, HRV, BVP, EEG, temperature, accelerometer etc.

A1.2 Background

Emotions play a significant role in decision-making mechanisms and perception of individuals. Affective computing and emotion recognition technologies encompass a diverse array of devices and systems designed to perceive, understand, and respond to human emotions.[4] These technologies employ a variety of methods, including physiological signals monitoring, facial expression analysis, speech recognition, and natural language processing, to discern the emotional states of individuals, revolutionizing the way we engage and understand human emotions.[5] Physiological signals, such as EEG, GSR, and BVP, have been widely used in the area of human emotion recognition as they are directly influenced by the autonomic nervous system (ANS), which responds to emotional stimuli.[6]

Music is recognized across cultures as a powerful stimulus for evoking emotions and influencing mood. In particular, its impact on brain structures associated with emotion regulation reveals a close connection between human emotions and music.[7] These connections may even offer promising paths for therapeutic interventions in psychiatric and neurological disorders. Interestingly, given that music influences physiological reactions, it has a profound impact on emotional contagion. For example, happy music triggers.

the zygomatic muscle responsible for smiling, with an increase in skin conductance and breathing rate, while sad music activates the corrugator muscle.[8] Toward this end, the detection of emotions evoked during co-creation performances in the MuseIT project will play a vital role in the development of interactive musical experiences tailored to individual emotional states and facilitating meaningful interactions among co-creators. This integration of emotion recognition technology within the MuseIT project not only enables real-time monitoring and understanding of participants' emotional states but also contributes to the development of personalized and emotionally engaging co-creative experiences. By leveraging the insights gained from the sensor diagnostics layer, including the emotion recognition system by CERTH, a dynamic and responsive environment will be created where emotions serve as valuable cues for guiding the creative direction.

AI algorithms for emotion recognition

Research in the field of emotion recognition with physiological signals is focused on exploring the connection between various physiological signals and emotions, selecting the appropriate stimuli to induce several emotional states, and developing AI algorithms for extracting, selecting, and classifying

[4] Picard, R. W. (2000). *Affective computing*. MIT press

[5] Alheeti, A. A. M., Salih, M. M. M., Mohammed, A. H., Hamood, M. A., Khudhair, N. R., & Shakir, A. T. (2023, November). Emotion Recognition of Humans using modern technology of AI: A Survey. In 2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS) (pp. 1-10). IEEE.

[6] Egger, M., Ley, M., & Hanke, S. (2019). Emotion recognition from physiological signal analysis: A review. Electronic Notes in Theoretical Computer Science, 343, 35-55.

[7] Koelsch, S. (2014). Brain correlates of music-evoked emotions. Nature Reviews Neuroscience, 15(3), 170-180

[8] Schaefer, H. E. (2017). Music-evoked emotions—Current studies. *Frontiers in neuroscience*, *11*, 600.

representative signal features.[9] Russell's two-dimensional valence-arousal space model provides a quantitative framework for understanding emotions.[10] Typically, valence is plotted along the horizontal axis, ranging from positive to negative, while arousal is represented on the vertical axis, ranging from low to high (Figure 2). Valence reflects the degree of pleasantness or unpleasantness, while arousal indicates the level of activation. For the scope of the MuseIT project, AI algorithms for recognizing emotions will be developed, mapping the physiological reactions of participants to the valence-arousal space. Furthermore, a process is established to gather physiological data from sensors, analyze the data, and deliver real-time quantitative indicators of emotional states. We will refer to this system as the Affective Computing Framework service for Music (ACF-Music). While the emphasis of this demonstration lies on the GSR signal and arousal detection, the procedural steps for handling the physiological signals targeted for inclusion in the MuseIT project remain consistent. The steps for conducting this online analysis are outlined below.

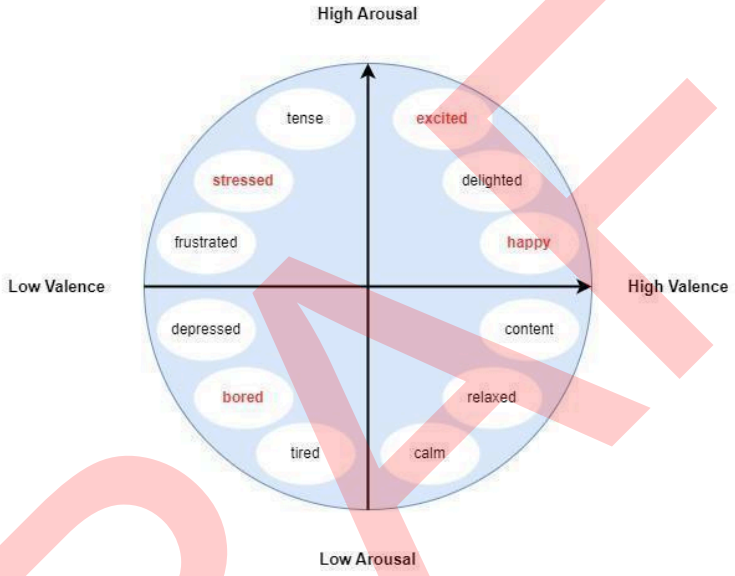


Figure 2: The 2-D valence-arousal space model for emotions

A1.3 GSR signal pre-processing and feature extraction

GSR, or galvanic skin response, refers to the measurement of skin's electrical conductivity, which fluctuates in response to changes in sweat gland activity regulated by the ANS. Research indicates a direct correlation between emotional arousal and an increase in skin conductivity, as demonstrated in previous studies.[11] GSR signals from 15 subjects in the WESAD benchmark dataset recorded with the wearable Empatica E4 are utilized in this implementation.[12] Our decision was influenced by the lightweight and unobtrusive nature of Empatica, a device we have also used for our data collection experiments (see also D4.2). In particular, we employ a median filter for eliminating artifacts generated mainly from subjects' movements, and the minmax normalization is used for each subject in the dataset to account for individual differences providing subject-independent generalized results. Next, GSR time series are grouped into 30-second windows with 50% overlap. Based on, 15 features

[9] Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*, 103-126.

[10] Basu, S., Jana, N., Bag, A., Mahadevappa, M., Mukherjee, J., Kumar, S., & Guha, R. (2015). Emotion recognition based on physiological signals using valence-arousal model. In 2015 Third International Conference on Image Information Processing (ICIIP) (pp. 50-55). IEEE.

[11] Domínguez-Jiménez, J. A., Campo-Landines, K. C., Martínez-Santos, J. C., Delahoz, E. J., & Contreras-Ortiz, S. H. (2020). A machine learning model for emotion recognition from physiological signals. Biomedical signal processing and control, 55, 101646.

[12] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018, October). Introducing wesad, a multimodal dataset for wearable stress and affect detection. In Proceedings of the 20th ACM international conference on multimodal interaction (pp. 400-408).

are extracted: statistical metrics like the mean value (SCL_mean and SCR_mean), the standard deviation (SCL_std and SCR_std), the minimum (SCL_min and SCR_min), the maximum (SCL_max and SCR_max), linear combination of these ((SCL_max-SCL_min) and (SCR_max-SCR_ min)), mean value of the first and second derivative of the SCL (respectively SCL_dot and SCL_ddot) and the number and amplitude of the SCR signal peaks.[13] We employ a machine learning algorithm, a Support Vector Machine with a RBF kernel, for classifying the features extracted from GSR into levels of arousal estimation, in a subject-independent manner by using Leave-One Subject-Out Cross-Validation (LOOCV) during training. We opted for SVM primarily because of the relatively modest size of the dataset. SVMs are recognized for their strong performance with small sample sizes and their reduced susceptibility to overfitting compared to alternative classification algorithms. Furthermore, SVMs inherently operate as binary classifiers, aligning well with the nature of our arousal classification task. The model achieved 93.22% in terms of accuracy in the prediction of the binary stress classification task. Finally, the trained model is developed and employed for real-time analysis.

A1.4 Real-time emotion monitoring

To facilitate real-time monitoring, the Lab Streaming Layer (LSL) system is used.[14] LSL is a comprehensive system designed for gathering measured time series data in research settings, encompassing networking, time synchronization, real-time access, and optionally centralized data collection, viewing, and disk recording. The liblsl library [15] offers abstractions for client programs, including Resolvers to identify available streams on the lab network, Outlets to make time series data streams accessible, and Inlets to receive data from subscribed Outlets. Information about the stream is transmitted as XML data along with the raw data. LabRecorder, the default recording software bundled with LSL, facilitates recording multiple streams from the lab network into a single file while ensuring time synchronization.

In the real-time emotion recognition system for MuseIT, the participants will wear lightweight, unobtrusive sensors for capturing their physiological reactions. CERTH employs the Empatica E4 wearable wristband in order to record the GSR data (Figure 3). LSL library is responsible for streaming data from Empatica via Bluetooth connection to a local computer. The signals are captured every 15 seconds, stored in a 30-second buffer, and subjected to a median filter with a 5-second kernel. Following this, the signals are normalized, and a feature vector comprising the aforementioned 15 features is generated. The trained subject-independent model developed predicts the levels of arousal state by computing probabilities of possible SVM outcomes in the [0 1] range. This indicates that the closer the outcome is to 1, the higher the arousal level of the individual. Finally, the buffer is updated with new 15-second samples from the stream, the oldest are discarded and an arousal state is predicted. The pipeline is depicted in Figure 4.

Figure 3: Empatica E4 wearable

[13] Cittadini, R., Tamantini, C., Scotto di Luzio, F., Lauretti, C., Zollo, L., & Cordella, F. (2023). Affective state estimation based on Russell's model and physiological measurements. Scientific Reports, 13(1), 9786.

[14] https://labstreaminglayer.readthedocs.io/info/intro.html Accessed: 07/02/2024

[15] https://github.com/labstreaminglayer/pylsl

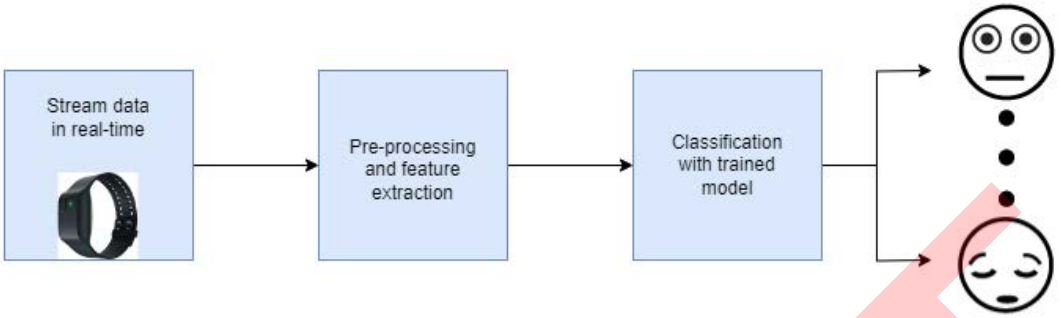


Figure 4: The pipeline for real-time arousal levels detection

APPENDIX 2 - Mood Estimation - Catalink

A2.1 Literature and overview

The area of Computer Vision (CV) and its applications around mood estimation have gained significant attention over the years. The Cohn-Kanade database, introduced in 2000, kickstarted the Automatic Facial Expression Recognition algorithms development.[16] Initially, emotion recognition was mostly based on the rule-based methodology of Facial Action Coding System (FACS), which used specific facial muscle movements, called Action Units, to identify emotions.[17] However, these early rule-based methods were limited in accuracy, due to their inability to capture the richness and complexity of human facial expressions.

Subsequently, traditional machine learning techniques, involving face detection, facial landmark extraction, and feature engineering, gained prominence. Researchers like Matthew Day, have used various Machine Learning (ML) methods like Support Vector Machines (SVM) and Gradient Boosting for automatic emotion classification.[18]

Despite the major improvements in accuracy, the above-mentioned methods required labor-intensive feature design, often leading to bias and inefficiency. Deep Learning then emerged, offering end-to-end processing of facial images, with automatic feature extraction which is learned through the training on large volumes of annotated images. Consequently, the advent of deep learning models significantly reduced time and effort in model design and training. Techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have even surpassed human-level accuracy in emotion recognition.

For example, Demochkina et al. proposed a video-based emotion recognition using MobileNet and SVM.[19] Youyi Chai et al.[20] and Yin Fan et al.[21] combined CNNs and RNNs for video emotion recognition, while M. S. Hossain et al.[22] used 2D and 3D CNNs for audio-visual emotion detection, also applying a 3D CNN in healthcare for monitoring patient emotions. These developments represent significant strides in computer vision-based mood estimation.

For our work, based on the task's requirements, a model appropriate for performing Facial Emotion Recognition (FER) should be developed, trained, and validated. Foremost, we assessed the possibility of reusing a pre-trained model for FER. But bearing in mind the mixed nature and variety of sources of our dataset, as well as the intention to refining the model by modifying the data and adjusting the model's weights for further fine-tuning in distracting scenarios such as partial occlusion, a custom-developed model was deemed more appropriate to meet our specific requirements and objectives. In addition, since the use-cases and requirements of the project involve inclusiveness and VR technologies, we decided to also focus on developing another version of the algorithm that could be employed for faces occluded by VR headsets or other eyewear that are used by visually impaired individuals. Specifically, the requirements for the proposed models are to be lightweight, fast during

[16] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 ieee computer society conference on computer vision and pattern recognition-workshops (pp. 94-101). IEEE.

[17] Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

[18] Anderson, K., & McOwan, P. W. (2006). A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *36*(1), 96-105.

[19] Demochkina, P., & Savchenko, A. V. (2021). MobileEmotiFace: Efficient facial image representations in video-based emotion recognition on mobile devices. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V* (pp. 266-274). Springer International Publishing.

[20] Cai, Y., Zheng, W., Zhang, T., Li, Q., Cui, Z., & Ye, J. (2016). Video based emotion recognition using CNN and BRNN. In *Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings, Part II 7* (pp. 679-691). Springer Singapore.

[21] Fan, Yin, et al. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks." *Proceedings of the 18th ACM international conference on multimodal interaction*. 2016.

[22] Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, *49*, 69-78.

inference, and demonstrate good performance on mood estimation tasks, including cases of partially occluded faces.

A2.2 Dataset and Simulated occlusion

As previously mentioned, FER is known for its complex nature, due to the human face's capability to create thousands of expressions using 43 different facial muscles. This complexity is compounded by individual differences in facial characteristics and expression styles. Additionally, the effectiveness of a Computer Vision algorithm is heavily influenced by its training dataset. Obtaining a representative training set typically requires gathering many thousands of images, a process that is both labor-intensive and computationally demanding.

Data collection for Facial Emotion Recognition

For constructing our dataset, we included multiple and various facial expressions within each emotion category, in order to create a dataset that well-represents the different human facial expressions. Facial images are difficult to find available online, due to the strict copyright licences. For that reason, our images were gathered from online resources that provided copyright-free images, such as Kaggle (FER 2013)[23], dataset Jafar Hussain Human emotions[24] dataset and other open-source databases such as Unsplash,[25] Pexels, [26]and Pixabay. [27]

By amalgamating images from these diverse datasets, our initial image collection comprises roughly 50,000 images of facial expressions, which are categorized into seven emotion classes ('angry', 'disgusted', 'scared', 'happy', 'sad', 'surprised', 'neutral'). Some examples for the different emotion classes are depicted in Figure 1. The categories have unequal amounts of instances, making the dataset highly imbalanced. Due to MuseIT's purposes, we decided to focus on the three most basic emotions, i.e. 'happy', 'sad', and 'neutral', and for that reason, we grouped the rest categories into a fourth class, named 'other'.

Figure 5: Examples of the FER-2023

Data collection for Facial Emotion recognition for partially occluded faces

For the version of the algorithm that is designed to work for partially occluded faces, we developed an alternated version of the dataset. To obtain representative image instances which are identical to occluded faces, a preprocessing procedure was performed. Analytically, the collected images were adjusted to our new task, by occluding the upper part of the face (i.e. the eyes and parts of the forehead and nose), inspired by the methodology originally proposed by Rodrigues et al.[28] Initially, the preprocessing algorithm uses a Multi-task Cascade Convolutional Neural Network (MTCNN) to detect five facial landmarks (two for the center of each eye, one for the nose centre and two for the

[23] https://www.kaggle.com/datasets/msambare/fer2013
[24]https://www.kaggle.com/jafarhussain786/dataset
[25] https://unsplash.com
[26] https://www.pexels.com/search/fac
[27] https://pixabay.com/vectors
[28] Rodrigues, A. S. F., Lopes, J. C., Lopes, R. P., & Teixeira, L. F. (2022, October). Classification of facial expressions under partial occlusion for VR games. In *International Conference on Optimization, Learning Algorithms and Applications* (pp. 804-819). Cham: Springer International Publishing.

right and left side of the mouth).[29] Based on the detected eye and nose landmarks, as well as the distances specified by the algorithm suggested by Rodrigues et al. (2022) a rectangle is drawn on top of each image. Therefore, the upper part of the faces is hidden, simulating in such a way the inclusion of VR headsets. An example of a pair that consists of an image and its occluded version is elucidated in Figure 6. More information regarding the Occlusion process can be found in CTL's work.[30]

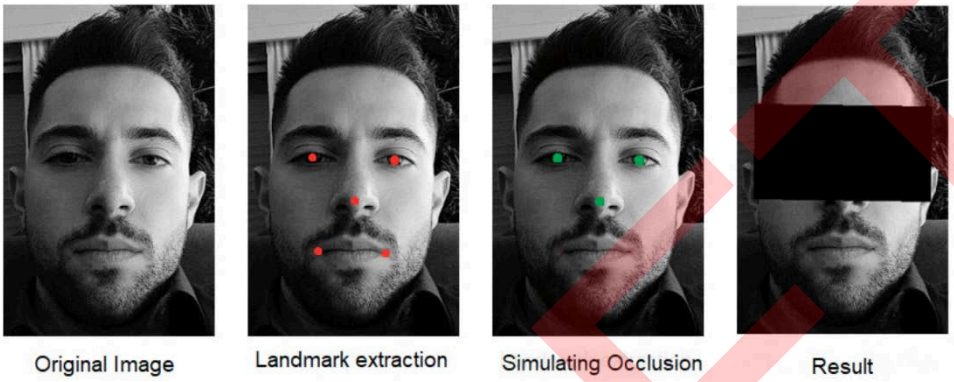


Figure 6: Occlusion Process

A2.3 Experiments and Model's Architecture

For the development of the Mood Estimation Algorithm (MEA), we analyzed various state-of-the-art deep learning models, focusing on those suitable for lightweight and embedded vision applications. Our experiments involved models like MobileNetV2,[31] MobileNetV3,[32] and mini-Xception.[33] *S*everal different architectures and hyper-parameter combinations have been evaluated and assessed with regards to both their prediction accuracy and latency for real-time inferences.

Mini-Xception emerged as the most appropriate since it demonstrates great prediction accuracy and negligible inference latency for real-time applications. Its success lies in two main features: the use of residual modules and depth-wise separable convolutions. Due to the characteristics and architecture of mini-Xception, the number of parameters is significantly reduced, ending up with an overall of ~54,000 trainable parameters. Lastly, the final model's size is only a few megabytes, less than a MB in size, so it can seamlessly be deployed and run even on some hardware-constrained devices. More information regarding our chosen model can be found in CTL work.[34]

The architecture of mini-Xception starts with two Convolution layers (which are followed by Batch Normalization and ReLU layer), followed by four residualblocks. Each block contains a convolution layer on the skip connection side, and the other side consists of two separable convolutions followed by a Max Pooling layer. All convolutional layers are followed by Batch Normalization and ReLu layers. Finally, follows a convolutional layer, a Global Average pooling layer and the final classification takes place at the SoftMax layer. A brief illustration of the architecture is depicted in Figure 7.

[29] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, *23*(10), 1499-1503.

[30] Petrou, N., Christodoulou, G., Avgerinakis, K., & Kosmides, P. (2023, July). Lightweight Mood Estimation Algorithm For Faces Under Partial Occlusion. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 402-407).

[31] Sandler, M., Howa rd, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).

[32] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1314-1324).

[33] Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.0755*

[34] Petrou, N., Christodoulou, G., Avgerinakis, K., & Kosmides, P. (2023, July). Lightweight Mood Estimation Algorithm For Faces Under Partial Occlusion. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 402-407).

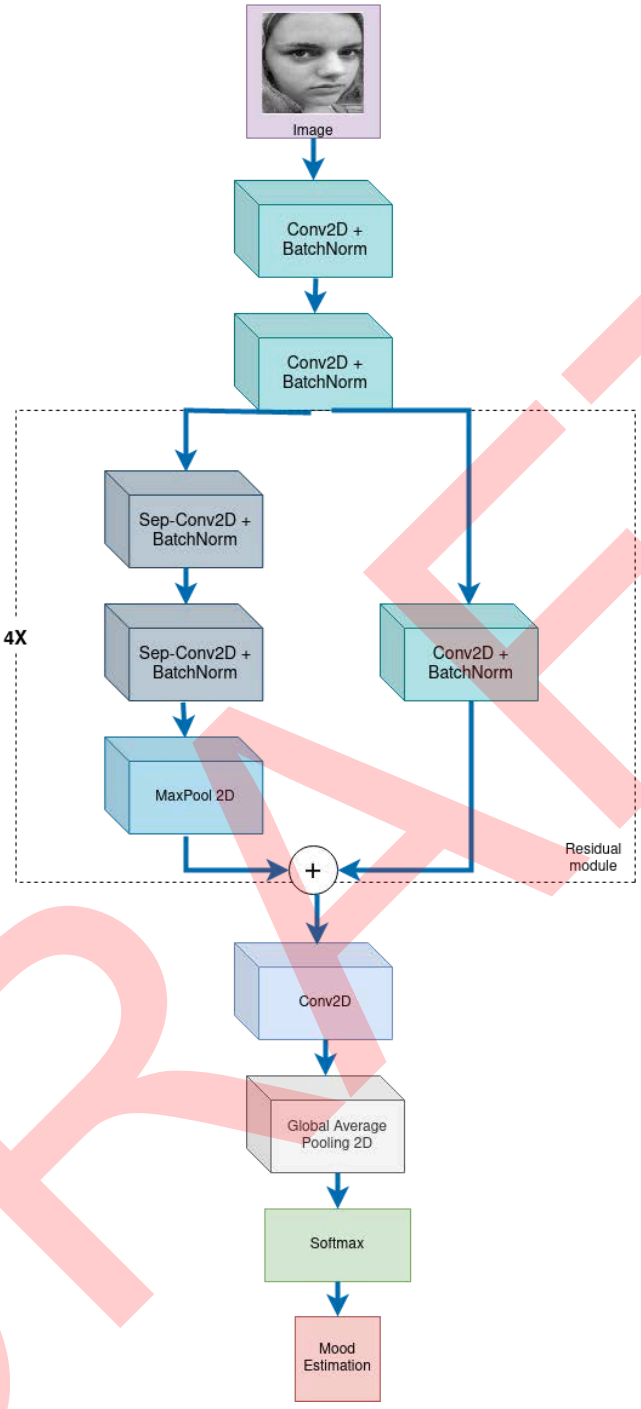


Figure 7: Architecture of mini-Xception model

A2.4 Model training and Experimental Results

Facial Emotion Recognition in full faces

Regarding the scenario that includes full faces, without any part of the face being occluded, a Mini-Xception model was trained from scratch on our data collection. The best performing model was trained with Adam optimizer, using an initial Learning rate of 1e-3, batch size 64 for 300 epochs. In addition, the learning rate was gradually reduced based on the Reduce Learning Rate on Plateau technique. Lastly, we dealt with the class imbalance problem by using weighted loss function, while to mitigate overfitting we applied L2 regularization.

For the preprocessing pipeline, the model takes video frames as input, transforming them into 64x64 images. Moreover, to increase the diversity of our training data, we apply data augmentation techniques, such as rotation, width or height shift, flip and shear transformation.

The model scored overall accuracy and F1-score equal to 0.71 on our test data. In Figure 8, we present the confusion matrix on our test data, which summarizes how the model classified the data into the emotion classes (MODEL_1). We observe that the *'happy'* class scores the highest accuracy rate (86%) among the rest, followed by the *'other'* class (74%). We also notice that the *'neutral'* images are sometimes incorrectly classified as 'sad' (16%), and sometimes as 'other' (8%), while the *'sad'* emotions are sometimes identified as either 'neutral' or 'other'. Regarding the overall performance, there is room for improvement, especially regarding the classes 'negative' and 'neutral'. However, it is quite reasonable that we do not have the perfect emotion recognition accuracy, especially on such tasks, due to the subjective nature of emotion perception. Sometimes it is challenging even for humans to distinguish similar facial expressions, such as a neutral, from a sad face, due to the subtle differences between such expressions. Thus, it is far more difficult to transfer that knowledge to a machine learning model.

Facial Emotion Recognition on partially occluded faces

For classifying facial expressions under occlusion, we chose to utilize the same mini-Xception model of our previous work, pretrained on our original data collection (that includes full faces, without any part of the face being occluded) but with some further tuning. The process aimed in utilizing the already learned knowledge of the pre-trained network, to reduce the training time as well as to improve the overall classification performance for the occluded scenario. In order to provide a fruitful comparison and applicable empirical results during our experiments, we focused on the experimentation and evaluation of four different settings for the occluded dataset:

- MODEL_1: Baseline evaluation using a the pre-trained model from the non-occluded faces setting (baseline model) dataset.
- MODEL_2: Transfer learning by freezing all parameters except those in the last convolutional layer for feature extraction.
- MODEL_3: Transfer learning with parameter initialization based on the baseline model, but without freezing any parameters during training.
- MODEL_4: Training the mini-Xception architecture from scratch on the occluded dataset, with parameter initialization based on Xavier uniform initializer.[35]

A brief summarization of the above-mentioned model settings is available in Table 1. Regarding the choice of hyperparameters, training options and other preprocessing, the same choices were used for all the models, as it was also used in the work of the non-occluded scenario.

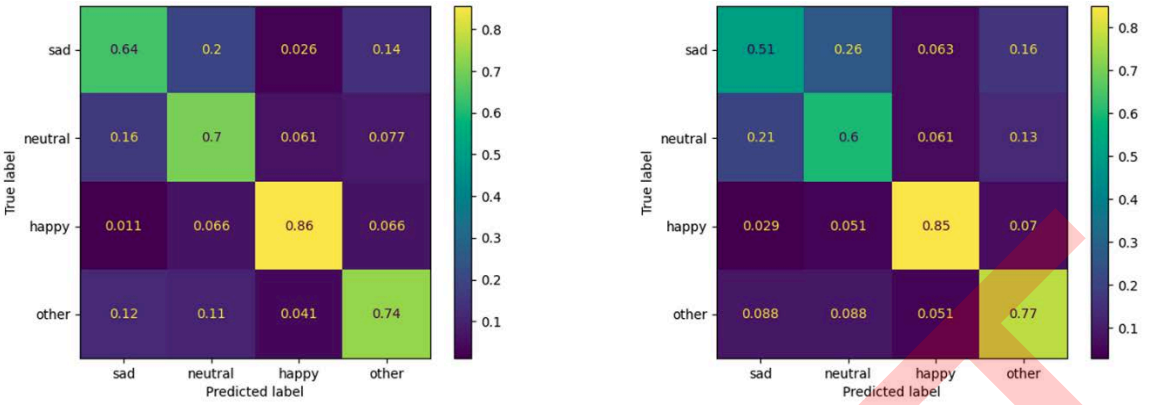
[35] Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256). JMLR Workshop and Conference Proceedings.

Table 1: Mini Exception's Experimental Settings & Results for Occlusion

ID	MODEL SETTINGS	# OF NEW TRAINABLE PARAMETERS	DESCRIPTION	TEST ACCURACY	TEST F1-MACRO	TEST F1-WEIGHTED
MODEL_1	Pre-trained for non-occluded	0	No additional training involved	0.49	0.46	0.49
MODEL_2	Pre-trained for non-occluded & Unfreeze Last Layer	4,612	All parameters apart from the last convolutional layer were frozen during training on the occluded dataset	0.63	0.61	0.63
MODEL_3	Pre-trained for non-occluded & Unfreeze All Layers	53,636	Parameters initialized based on MODEL_1 and continued training on the occluded dataset	**0.69**	**0.68**	**0.69**
MODEL_4	Trained from Scratch	53,636	Parameters were reinitialized	0.68	0.67	0.68

The best-performing model, MODEL_3, was the pre-trained model, fine-tuned for the occluded task. It is worth to note that, in general, building a model from the ground up usually results in better performance. But in our situation, there is a minor improvement of 1% in the transfer learning setting. This can be explained by the fact the original dataset and model we worked on, utilized slightly more data. That is, since the artificial process that performs the occlusion, had resulted in a dataset with almost 10% less images. That happened since in some instances the facial landmarks (including the eyes) could not be identified by the MTCNN algorithm, thus those images were not used in the training of the partial occlusion scenario. Finally, setting aside the aspect of performance, the transfer learning case and the initialization of the weights in MODEL_3 allowed for satisfactory loss and accuracy even after a few epochs, contrasting the MODEL_4 which was trained from scratch, that required 40 to 50 epochs for similar performance.. Comparing the performance between our best models for the non-occluded and occluded cases, it was noticed that the overall performance was only reduced by a small amount of 4% when occlusion was introduced. Furthermore, by comparing the performance diminishment between the two above-mentioned scenarios for the different classes, it was observed that numerous misclassifications had risen for the classes "sad" and "neutral" Figure 8. It is believed that this is due to the fact that apart from having the lip corners pulled down, people often express their sadness by crying or by raising their inner corners of eye-brows raised and eyelids loose.[36] Therefore, this information is hard to be utilized under partial or severe occlusion.

[36] Reed, L. I., & DeScioli, P. (2017). The communicative function of sad facial expressions. *Evolutionary Psychology*, *15*(1), 1474704917700418.



(a) MODEL_1 results for non-occluded test set

(b) MODEL_3 results for occluded test set

Figure 8: Confusion Matrices for unseen data

Based on the results of the above experiments, it was indicated that FER under partial occlusion is still possible. Furthermore, the results confirm that the exploitation of transfer learning as well as the simulation techniques for synthetic occlusion can lead to a respectable model that produces results that keep pace with frameworks that utilize information from the periocular area and eyes.

Evaluation under Real Conditions

In our latest phase of experimentation, we conducted some internal tests closely aligned with real-world scenarios relevant to our project's objectives. These included utilizing both 480p and full-HD webcams, simulating conditions where users wore VR headsets causing occlusion of the upper face, as well as scenarios with fully exposed faces (Figure 9). Our participants engaged with varied content under diverse lighting conditions. Overall, the outcomes were promising and aligned with user feedback collected post-session. However, we observed that under lower lighting conditions, performance slightly declined, occasionally resulting in mismatches, particularly with the recognition of sad emotions, as suggested by our previous evaluation findings.

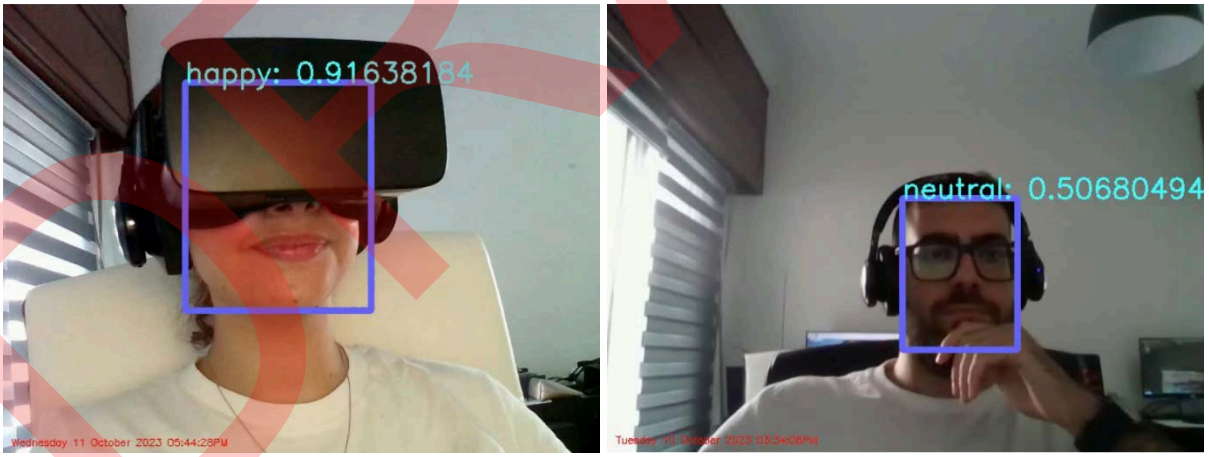


Figure 9: Experimental sessions for model testing under real conditions

APPENDIX 3 - Stress Estimation - Catalink

A3.1 Literature and overview

Stress estimation using wearable devices has emerged as a powerful tool, offering unique insights into human emotions and reactions. Equipped with advanced sensors measuring parameters like heart rate and skin conductance, these devices provide real-time data that can unveil moments of heightened stress, anger, or intense emotions. Beyond traditional health applications, this technology delves into the realm of emotional intelligence, enabling users to understand and share their responses in various situations.

By capturing subtle physiological changes, information can be invaluable in personal and professional contexts, helping individuals navigate social interactions, and improve communication. Additionally, the insights derived from stress estimation through wearables can be seamlessly integrated into a co-creation music service, enhancing the overall experience for users.[37] For instance, in a scenario where users not only understand their stress levels but collaboratively contribute to creating music that dynamically reflects their emotional states. This novel approach transforms stress monitoring into a shared, creative endeavor, allowing individuals to collectively shape a personalized experience that resonates with their emotional landscape. In this way, wearables not only serve as tools for self-awareness but also contribute to a collaborative and enriched emotional journey through the medium of music and art. More specifically, the exploitation of Heart Rate Variability (HRV) in wearable stress monitoring represents a significant advancement in our understanding of human psychological states. HRV, as an indicator of the variation in heartbeat intervals, provides direct insight into the autonomic nervous system's response to stress and emotional arousal.

Schmidt et al.[38] took the initiative and conducted an innovative study on stress and affect detection. The authors collected and published the WESAD (Wearable Stress and Affect Detection) dataset, a multimodal dataset that demonstrates the effectiveness of accurately predicting stress using HRV along with other physiological data. Nkurikiyeyezu et al.[39] study the impact of person-specific biometrics for stress prediction and they prove that individualized models indicate improved performance accuracy. Furthermore, Koldijk et al. [40] echoed this approach with their work, which introduced the SWELL dataset to improve stress and user modeling through personalized data.Moreover, Bobade and Vani[41] utilized deep learning and machine learning algorithms to analyze multimodal physiological data to identify stress, highlighting the potential of sophisticated computational techniques to decipher the intricate patterns of HRV and other physiological parameters.

---

[37] Turchet, L., & Barthet, M. (2018). Co-design of Musical Haptic Wearables for electronic music performer's communication. *IEEE Transactions on Human-Machine Systems*, *49*(2), 183-193.
Chen, C. C., Chen, Y., Tang, L. C., & Chieng, W. H. (2022). Effects of interactive music tempo with heart rate feedback on physio-psychological responses of basketball players. *International journal of environmental research and public health, 19(8), 4810.*

[38] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018, October). Introducing wesad, a multimodal dataset for wearable stress and affect detection. In Proceedings of the 20th ACM international conference on multimodal interaction (pp. 400-408).

[39] Nkurikiyeyezu, Kizito, Anna Yokokubo, and Guillaume Lopez. "The effect of person-specific biometrics in improving generic stress predictive models." arXiv preprint arXiv:1910.01770 (2019).

[40] Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014, November). The swell knowledge work dataset for stress and user modeling research. In Proceedings of the 16th international conference on multimodal interaction (pp. 291-298).

[41] Bobade, P., & Vani, M. (2020, July). Stress detection with machine learning and deep learning using multimodal physiological data. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 51-57). IEEE.

Within the MuseIT project, for the task of stress estimation, we explored the scenario of estimating the stress levels of an individual, using an electrocardiogram (ECG) signal. More specifically, we focused on extracting the inter-beat-intervals (IBIs) and the HRV metrics. In particular, HRV is derived from the ECG signals and captures the changes in the time intervals between consecutive heartbeats and reflects how adaptable our body can be to different environmental and psychological changes. HRV is a significant health indicator, since it can provide insights regarding overall health and wellbeing, and most significantly, it can help to uncover the mental tension of a person, since it is a strong indicator of stress. To quantify HRV, we can extract Time domain and Frequency domain features as explained in the next sections. In our application these features are extracted from an ECG signal, captured from a PolarH10[42] chest strap. Then, the extracted features are used to train and evaluate a Machine Learning (ML) model to accurately predict the stress level of an individual.

A3.2 Training data

A quality dataset is vital for training a stress detection algorithm, providing the foundation for pattern recognition and adaptability across diverse scenarios, ultimately ensuring precision and reliability in stress assessment. To this end, the WESAD dataset[35] was utilized[35]. WESAD consists of multivariate data, gathered from 15 subjects during a stress-affect lab study, while wearing physiological and motion sensors. The devices used for data collection were a chest-worn device, the RespiBAN [43]and a wrist-worn device, the Empatica E4[44]. The following sensor modalities are included: BVP, ECG, electrodermal activity, electromyogram, respiration, body temperature, and three-axis acceleration.
Specifically for our work, the data that we exploited are the ECG signals recorded from the RespiBAN. Moreover, the dataset contains three different affective states ('neutral', 'stress', 'amusement'). In addition, self-reports of the subjects, which were obtained using several established questionnaires, are contained in the dataset. Details can be found in the dataset's readme-file, as well as in WESAD's official website [45].

Data collection and processing

Data collection from chest-strap sensor

In order to test our stress detection models on realistic data, we decided to collect our own measurements, with the help of the chest-strap sensor PolarH10[46]. PolarH10 is a supremely precise heart rate sensor, providing top-quality heart rate measurements. In addition, it is considered one of the most accurate heart rate sensors by many sources. Some of its features that make it stand out from the rest are following:

- Chest straps are the gold standard, validated against the clinical ECG, sitting around 99% accurate[47]
- Polar H10 is one of the most accurate heart rate sensors currently available on the market.
- Used for medical research and sports science.[48]
- Connects with Bluetooth, ANT+ and 5 kHz.
- Several connections can be active simultaneously.
- Built-in memory for a session.
- Easy for subjects to wear and stress-free, as shown in Figure 10**.**
- Easily found in the market, with a relatively low price.[49]

[42] https://www.polar.com/en/sensors/h10-heart-rate-sensor
[43] http://www.biosignalsplux.com/en/respiban-professional
[44] https://www.empatica.com/en-gb/research/e4/
[45] https://archive.ics.uci.edu/dataset/465/wesad+wearable+stress+and+affect+detection
[46] https://www.polar.com/en/sensors/h10-heart-rate-sensor
[47] https://nesswell.com/best-chest-strap-heart-rate-monitors/
[48] Schaffarczyk, M., Rogers, B., Reer, R., & Gronwald, T. (2022). Validity of the polar H10 sensor for heart rate variability analysis during resting state and incremental exercise in recreational men and women. *Sensors*, *22*(17), 6536.
[49] https://www.polar.com/en/sensors/h10#:~:text=Heart%20Rate%20Sensor,with%20Bluetooth%C2%AE%20and%20ANT%2B. I

Figure 10: Wearable Polar H10 Sensor

Signal Pre-processing Pipeline

To build a pipeline that consists of the data collection and preprocessing steps during a session, the following procedure was automated. Firstly, when a session initiates, a data recording is performed to collect an ECG signal from an individual through PolarH10. Then the recorded signals are preprocessed and filtered, to fill in any missing values and remove the outliers. Afterwards, the IBIs are estimated from the ECG, and finally, the temporal and frequency features were extracted from the HRV, as detailed in the next subsection (Table 2). To compute the HRV features, signal segmentation is necessary. A typical segmentation for such a task involves 60-second window frames with 15 seconds overlap, achieving continuity and high sensitivity for detecting stress-induced physiological changes. The whole procedure and pipeline are depicted in Figure 11.

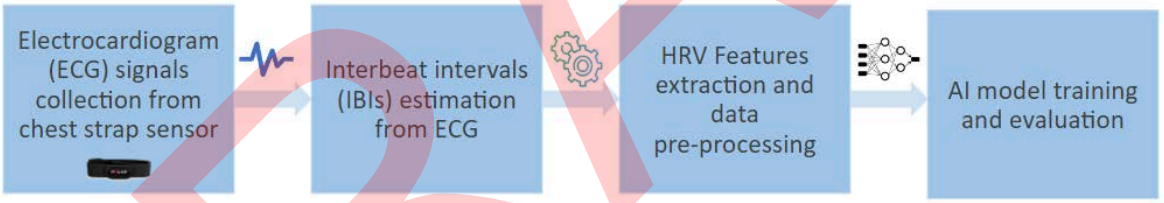


Figure 11: Data collection and Pre-processing Pipeline

A3.3 Feature Extraction

A specific feature extraction strategy was followed in order to extract aggregated and meaningful features from the raw signals contained in the training set. The different HRV metrics are categorized into *Time Domain, Frequency Domain,* and *Nonlinear* categories, providing features like the mean and standard deviation of IBIs. Finally, these direct characteristics can be used as independent variables during the learning process of an ML algorithm directly. The 33 features extracted are briefly explained in Table 2.

Table 2: Heart Rate Variability (HRV) Features OverviewModel raining and experimental results

Time Domain Features	**Frequency Domain Features**	**Non-linear Features**
Average of RR intervals	Very low frequency	Poincaré plot standard deviation perpendicular to the line of identity
Median of RR intervals	VLF power as percentage of total power	Poincaré plot standard deviation along the line of identity
Standard deviation of RR intervals	Low frequency	Kurtosis of RR intervals
Root mean square of successive RR interval differences	LF power as percentage of total power	Skewness of RR intervals
Standard deviation of successive RR interval differences	LF power in normalized units	Mean of relative RR intervals
Ratio of SDRR to RMSSD	High frequency	Median of relative RR intervals
Heart rate	HF power as percentage of total power	Standard deviation of relative RR intervals
Percentage of differences between adjacent NNs over 25 ms	HF power in normalized units	Root mean square of successive relative RR interval differences
Percentage of differences between adjacent NNs over 50 ms	Total power of RR intervals	Standard deviation of successive relative RR interval differences
-	Ratio of LF to HF power	Ratio of SDRR_REL_RR to RMSSD_REL_RR
-	Ratio of HF to LF power	Kurtosis of relative RR intervals
-	-	Skewness of relative RR intervals
-	-	Sample entropy

HRV offers vast potential for just-in-time interventions, behavioural modification, and training guidance. A high degree of methodological rigor highlights HRV's importance in creating personalized and adaptive stress management systems, paving the way for new innovations in wearable technology and stress intervention strategies.

A3.4 Feature selection

Prior to training the different ML models, we tried different feature selection techniques, to reduce the dimensionality of our data and keep only the most informative features. To this end, we applied separately the ANOVA (Analysis of Variance) and the Principal Component Analysis (PCA) method on our features. We experimented with keeping 32, 25, 12 and 10 features/components. The extracted features have been used in order to train the models which will finally detect whether a person is in a stressful state or not.

A3.5 Model Selection

The goal was to create a classifier capable of categorizing each signal segment into either a stress or non-stress state, thus enabling the inference of whether a subject is experiencing stress or not. Experimentations with different ML models were performed, to find the one that scores the highest performance for our task.
The algorithms that were tried out are the following:

- Support Vector Machine (SVM) with Linear and Radial Basis Function kernels.
- Gradient Boosting (XGBoost)
- Random Forest Trees
- Simple Multi-Layer Perceptron (MLP) Neural Network

In the following table, we provide some of the experiments conducted with different models. For each trial we provide the results (F1-scores) on the WESAD test dataset and on the data collected from the PolarH10 sensor. More details on the results are provided on the next subsection.

Model	Architecture	Feature selection techniques	number of features/components	F1-Score (WESAD)	F1-Score (PolarH10)
Neural Network	Input (12) - Hidden (24) -BatchNorm- Output (2)	ANOVA	12	0.90	0.60
XGB	200 estimators	PCA	12	0.85	0.58
XGB	500 estimators	ANOVA	12	0.88	0.55

Table 3: For each experiment, we provide the type of model used and its architecture, the feature selection technique applied to select the most informative features (or number of components, in case of PCA) for the model training, the number of features/components kept and the F1 macro scores on the WESAD testing data and on the data collected from our experiments using the chest strap sensor, PolarH10.

As it is clearly seen from the table, the model that scored the highest performance accuracy was a simple MLP network, consisting of a single hidden layer of 24 neurons, which was trained with the 12 most informative features obtained from ANOVA. The model was trained with an Adam optimizer, for 60 epochs, with a batch size of 32 and with a learning rate of 1e-3. Furthermore, during the training and model selection, the Leave-one-subject-out cross validation (LOOCV) method was utilised, namely we splitted the subjects' data into training and validation sets and afterwards we applied data scaling on each subject separately.

A3.5 Results under real conditions

To collect some data samples for our own data through PolarH10, we designed and implemented an experiment, where participants were accomplishing some predefined tasks. During this session, the individuals were wearing a Polar H10 chest strap sensor that was recording their ECG signal. After the sessions were over, the signals were annotated and kept anonymous to evaluate our models on realistic data.

While our best model achieved an impressive F1-score of 0.90 for the LOOCV, its performance on our collected data was rather poor, scoring an F1-score of 0.60, highlighting the need for further enhancements. This discrepancy in performance can likely be attributed to the sensor differences and discrepancies; the WESAD dataset (used for training and validation) employed the RespiBAN sensor, while we used the PolarH10 sensor during our experiments. In fact, each sensor introduces varying levels of noise and sensitivity, necessitating different data pre-processing approaches for data collected from different sensors.

APPENDIX 4 - Neurophysiological Prediction - X-System

Neurophysiological prediction (XSL) X-System is a computational model of the musical brain capable of predicting the neurophysiological effects of music. It is distinct from the CERTH and CTL systems, which are respectively focused on diagnosing the emotion and mood of users. X-System is focused on predicting the neurophysiological effect of the music itself. This means it can be complementary to the CERTH and CTL systems in communicating neurophysiological information between co-creators and in the process of composition.

The demonstration below shows an X-System predictive analysis of a song by one of SHMU's (anonymised) musicians. The system models the principal areas and networks of the brain involved in processing music. Brain stem responses to sounds of primal evolutionary/survival value - for example startling, rapidly approaching or very high sounds[50] - are modelled by volume, turbulence and sharpness algorithms, as are related ascending pathways by way of the inferior colliculus to the amygdala.[51] The responses of the basal ganglia, cerebellum, premotor and motor cortex[52] are modelled by rhythmicity algorithms, detecting the power, salience and density of periodic spectral turbulence ;[53] this forms part of a complex loop with processing and retention of patterns in the auditory cortex, including the right anterior secondary cortex[54] modelled by autocorrelation and related to tempo and metrical structures. There are algorithms that as far as possible replicate basic pitch detection in the auditory brain stem as well as more complex modelling of Heschl's gyrus. Here, chroma and pitch height are detected,[55] as well as fundamentals and spectra.[56] Important outputs of these models are indicators of levels of harmonicity (how close the spectrum is to the harmonic series) and the resulting activation of limbic and paralimbic systems.[57] These are measures of

[50] Sivaramakrishnan S, et al (2004) GABA (A) synapses shape neuronal responses to sound intensity in the Inferior Colliculus *Journal of Neuroscience* 26;24(21)5031-43

Osborne, N. (2009b) Towards a Chronobiology of Musical Rhythm in Communicative Musicality Editors: S. Malloch & C. Trevarthen. ISSN 0077-8923. (Oxford, UK and New York, USA) 545-564

Erlich N, Lipp OV, Slaughter V (2013) Of hissing snakes and angry voices: human infants are differently responsive to evolutionary fear-relevant sounds *Developmental Science* 16;6 894-904

Frankland PW et al (1997) Activation of amygdala cholecystokinin B receptors potentiates the acoustic startle response in rats *The Journal of Neuroscience* 17(5) 1838-47

Panksepp, J. & C. Trevarthen. 2009. The neuroscience of emotion in music. In *Communicative Musicality*. S. Malloch, C. Trevarthen, Eds.: 105–146. OUP.

[51] Jorris PX, Schreiner CE, Rees A (2004) Neural processing of amplitude- modulated sounds *Physiological Reviews* 84 641-577

Heldt, SA, Falls, WA (2003) Destruction of the Inferior Colliculus disrupts the production and inhibition of fear conditioned to an acoustic stimulus *Behavioural Brain Research* 144 175-185

Marsh RA, et al (2002) Projection to the Inferior Colliculus from the Basal Nucleus of the Amygdala *The Journal of Neuroscience* 22/23 10449-10460

[52] Panksepp, J. (1998) Affective Neuroscience OUP Oxford *passim*

[53] Osborne, N. (2009b) Towards a Chronobiology of Musical Rhythm in Communicative Musicality Editors: S. Malloch & C. Trevarthen. ISSN 0077-8923. (Oxford, UK and New York, USA) 545-564

[54] Penhune VB, Zatorre RJ and Feindel WH (1999). The role of auditory cortex in retention of rhythmic patterns as studied in patients with temporal lobe removals including Heschl's gyrus. *Neuropsychologia,* 37(3), 215–231.

Peretz I (2001). Listen to the brain: the biological perspective on musical emotions. In P Juslin and J Sloboda, eds, *Music and emotion: Theory and research,* pp. 105–134. Oxford University Press, London.

Peretz I and Kolinsky R (1993). Boundaries of separability between rhythm in music discrimination: A neuropsychological perspective. *The Quarterly Journal of Experimental Psychology,* 46(2), 301–325.

[55] Griffiths TD, Buchel C, Frackowiak RS, Patterson RD (1998) Analysis of temporal structure in sound by the human brain. *Nature Neuroscience* 1:422–427.

Warren, J.D. et al (2003) Separating pitch chroma and pitch height in the human brain Proceedings of the National Academy of Sciences USA100 (17) 10038-10042

[56] Schneider, P. et al (2002) Structural, functional, and perceptual differences in Heschl's gyrus and musical instrument preference. *Annals of the New York Academy of Sciences*, 1060, 387-94

Menon, V. Et al (2002) Neural correlates of timbre change in harmonic sounds Neuroimage 17 (4), 1742-1754

[57] Peretz, I, Aube W, Armony, J.L. (2013) Towards a biology of musical emotions in The Evolution of Emotional Communication: From Sounds in Nonhuman mammals to Speech and Music in Man ed Altenmuller E, Schmidt S, Zimmerman E OUP

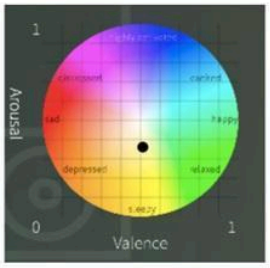
McDermott JH, Lehr AJ, Oxenham AJ (2010) Individual differences reveal the basis of consonance *Current Biology* 20 1035-1041

Koelsch S, Fritz T Schlaug G (2008) Amygdala activity can be modulated by unexpected chord functions during music listening *Neuroreport* 9(18):1815-9.

"vertical" harmonicity, but In pathways to emotional centres, for example the amygdala, "linear" harmonicity, or how notes and chords follow one another, is also significant, and modelled by a linear harmonic cost algorithm.[58] The values are plotted on an emotion colour circle - low arousal towards the bottom of the circle, high arousal towards the top, negative valence left, positive valence right.[59] The coordinates offer approximate locations for tracks in zones of emotion, mood and feeling within the circle.

The colour circle plots autonomic arousal (y axis, top to bottom) against vagal power, or positive feeling (x axis, left to right). All human emotions can be located within this circle Figure 12 below also shows XSL analysis graphs corresponding to predictions of activity in the autonomic nervous system, endocrine system, auditory cortex, motor cortex and brain stem.

The colour circle may act as a generator of avatars. X-System analyses also offer feedback to co-creators, as for example to the feedback in below figure. Figure 12: Example of XSL analysis



This is a lovely song with a beautiful simplicity, using primarily chords of C, F and G major, sometimes with a repeated C-G on top, sometimes moving to A minor. It has high **harmonicity**, sending warm messages to the emotional part of the brain, but with a little dissonance and reserve, There is a gentle **rhythmicity** and the voice has a beautiful smooth quality with an engaging "huskiness" which is captured by 50Hz **turbulence**. The position of the song on the colour circle suggests it is **between low and moderate autonomic arousal**, and on the positive side of emotions, but phlegmatic and thoughtful rather than overtly joyous

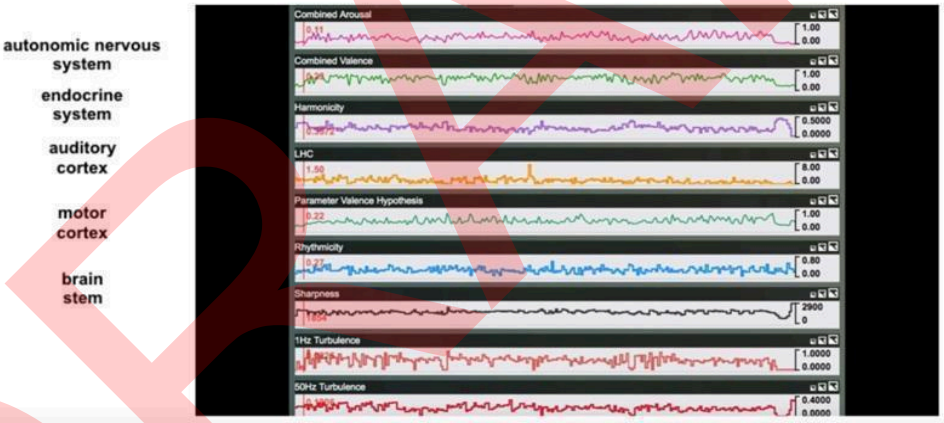


Figure 12: Example of XSL analysis

Stein MB, Simmons AN, Feinstein JS, Paulus MP.(2007) Increased amygdala and insula activation during emotion processing in anxiety-prone subjects. *Am J Psychiatry* 164(2): 318-27

Baumgartner T, Lutz K, Schmidt CF and Jancke L (2006). The emotional power of music: How music enhances the feeling of affective pictures. *Brain Research,* 1075 (1), 151–164.

Eldar E, et al (2007) Feeling the real world: limbic response to music depends on related content. *Cereb Cortex* 17(12):2828-40.

Blood AJ and Zatorre RJ (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences USA,* 98(20), 11818–11823.

[58] Koelsch S, Fritz T Schlaug G (2008) Amygdala activity can be modulated by unexpected chord functions during music listening *Neuroreport* 9(18):1815-9.

[59] a development and revision of

Scherer, K.R., Shuman, V., Fontaine, J.R.J, & Soriano, C. (2013). The GRID meets the Wheel: Assessing emotional feeling via self-report. In Johnny R.J. Fontaine, Klaus R. Scherer & C. Soriano (Eds.), Components of Emotional Meaning: A sourcebook *(pp. 281-298)*. Oxford: Oxford University Press.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44(4)*, 693-727.

Russell JA. (1980) A circumplex model of affect. *Journal of Personality and Social Psychology.* 39:1161–1178.

APPENDIX 5 - EEG Audification - X-System

The demonstration shows how X-System uses Multi-Oscillatory Dynamic Analysis to "audify" users' EEG, and then uses its model of the musical brain to search for the music that is closest to the users' EEG. EEG Audification can provide users with an audio representation of what is happening in their own brain, or the brains of others. The demonstration shows how XSL uses Multi-Oscillatory Dynamic Analysis to "audify" users' EEG, and then uses its model of the musical brain to search for the music that is closest to the users' EEG.

While CERTH is using EEG primarily diagnostically to help with emotional induction, to facilitate body-and-mind communication and to support creative self-expression, the XSL approach is designed to "audify" the user's brain to provide musical material directly from the user's mind, but also to search the world's repertoire to find the music closest in frequency profile behaviour to the user's brain. Both of these approaches lend themselves to rich compositional AI. It means that people with no movement or communication can still create music from their minds and bodies.

XSL has developed two ways in which an EEG signal can be transformed into music, both of which may serve as starting points for the development of co-creation specific technology.

The first of these methods is direct audification[60]. Brain wave activity is separated into different frequency bands, for instance delta waves (between around 0.8-4hz) are active during deep sleep, while beta waves (12-30Hz) signify concentration. Each of these waves can be thought of as a different instrument, which has its specific pitch range but at a certain moment in time plays one defined pitch in that range. X-system orchestrates these brain waves by first transposing them into the audible spectrum (50hz-20kHz) and additionally arranging the brainwaves detecting from different regions in the brain into harmonics of the brain-instrument.

A second technique employed by XSL is using these direct audifications, along with XSL's INRM, to find existing music that 'matches' the brain. This is accomplished by analysing both the direct audifications and a library of music in terms of X-system parameters, such as harmonicity, rhythmicity, linear harmonic cost, etc, which are designed to mimic how the human brain responds to musical signals in various different brain regions (brain stem, amygdala, motor cortices, auditory cortex, etc). Once this analysis is complete, both the music and the direct EEG audification are decomposed into a series of parameters, at which point it is possible to match existing music and EEG based on those parameters.

The diagrams below show icons of the process of recording EEG (Figure 13), the process of wavelet transform, filtering and prioritising the EEG (Figure 14) and the process of audification, turning the EEG signal to audible sound (Figure 15).

Figure 13: EEG of system user

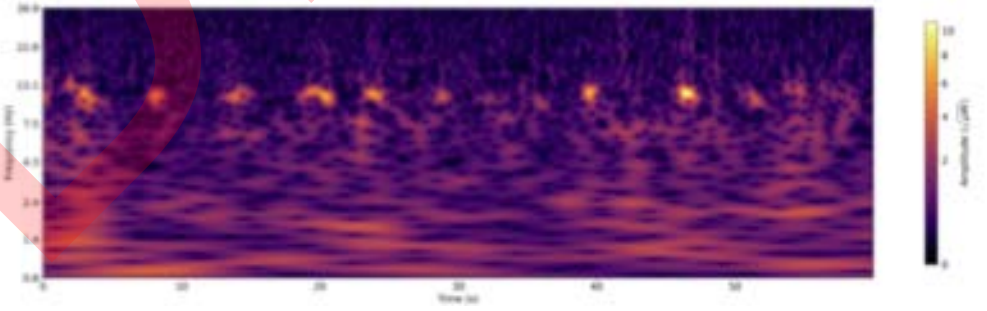


Figure 14: Wavelet transform.

60 https://en.wikipedia.org/wiki/Audification#:~:text=By%20definition%2C%20it%20is%20described,mapped%20to%20sound%20pressure%20levels.

Figure 15: Audification of EEG signal

There is the second approach - which involves routing the audification to X-System, which uses its model of the musical brain to search for the music in the existing world repertoire closest to the electrical activity of the user's brain (Figure 16). This provides very personal musical material for AI development and modulation.

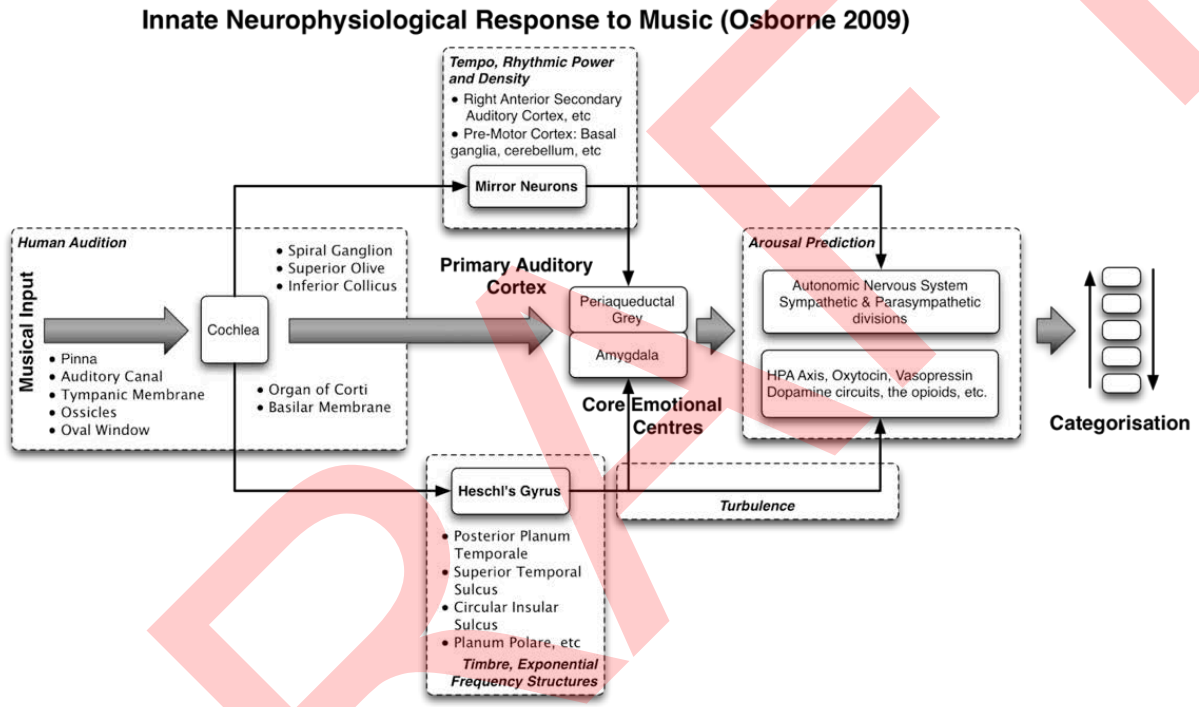


Figure 16: Neurophysiological response to music

APPENDIX 6 - Participatory Session - Share Music & X-System

A6.1 Workshop Details

Time and Location

Participatory Session 3
Visual Arena, Lindholmen, Gothenburg, Sweden
Monday 11th March 10:00 AM to 3.30 PM (DAY 1)
Tuesday 12th March 10:00 AM to 3.30 PM (DAY 2)

Aim

The aims of this participatory workshop were threefold:

1. HEART RATE

To work with potential users making use of heart rate sensors to evaluate to what extent heart rate signals may be of use in the communication of states of mind and body between remote co-creators (for the purposes of this workshop, the co-creators will be in proximity).

2. HAPTICS
   To work with users to evaluate the effectiveness and comfort of haptic signals in the communication of heart rate and other vibrational information.
3. VISUAL REPRESENTATION
   To begin the process of designing "avatars" which will be visual representations of human states of mind and body to assist in emotional communication within the process of remote co-creation.

Each module of the workshop had its own specific objectives.

Contributors

The session consisted of 7 participants, with different kinds of disabilities, from different genders and ranging from 20 to 44 in age contributing to the workshop artistic input and expertise on user needs. The session was led by Nigel Osborne (SHMU). 4 other members of the SHMU team participated, as did 2 members from XSL and 3 from HB. 5 personal assistants for participants were also present in the room.

Instruments

Participants were given the option to use a variety of electronic musical instruments, and/or microphones for the voice.
Nigel Osborne (SHMU) led the workshop with guitar / violin and voice, with Jonathan Walton (XSL) on trumpet and voice.
Instruments available for participants were:

- Software keyboard (iPad)
- Software keyboard (iPad)
- Korg Synthesizer with vocoder
- Moog Theremin
- Drum machine
- Amplified metal percussion

Set-up

HEART RATE

- Python script to read live streaming heart rate data from Polar H10 sensor
- Python script to transform live streaming heart rate data to audio
- Python script to send and receive data over Jack Trip

HAPTICS

- Python script to transmit live heart rate data as haptic output, using the Actronica HSD mk.2 board and 2 HapCoil Plus actuators

VISUAL REPRESENTATION

- XSL analysis of examples of favourite pieces of music supplied by our participants. This enabled the team to talk to each participant in terms of neurophysiological predictions, including such things as emotions and behaviour of the heart and it introduced XSL's colour circles, a possible generator of avatars and/or aspects of avatars.

- Collecting different versions of avatars from games / unicode / popular culture. Examples were emoji-like, figurative, abstract human images, abstract art, colour systems (e.g. related to XSL) and combinations of the same.

A6.2 Module A - Recorded heartbeat exploration

Duration: 2 hours with a 15-minute break

Objectives

- To familiarise participants with playing music with heart beats; to work co-creatively exploring how emotions and states of mind and body may be communicated through heart beats. This work was focussed on creative materials and heart beats featured in two existing pieces of music - classic rock from Pink Floyd, classical romantic from Gustav Mahler.

- To explore different speeds of heart beat, different heart rate variabilities and the different emotions and states of mind and body they are associated with.

Exercises

1. Listen to Pink Floyd's Breathe and comment on how it feels to hear a heartbeat in a song
2. Co-improvise on the structure of the song over the original recorded heart beat
3. Short discussion about the autonomic nervous system, arousal and heart rate
4. Co-improvise on a sound file of a slow male heart beat recorded during meditation.
5. Co-improvise on a sound file of a fast female heart beat recorded after intensive exercise.
6. Short discussion about heart rate variability, vagal power and valence
7. Improvisation on fast heart beat with high heart rate variability (positive valence, joyful)
8. Co-improvise on a sound file of a fast male heart beat with high variability, probably associated with a very strong, positive feeling of joy - with a free choice of pitches.
9. Co-improvise on a sound file of a rigid techno beat with low variability in order to illustrate the difference between positive and negative valence.

Animateurs supported participants in all improvisation exercises. Software instruments were set to a particular scale (E Dorian / D Pentatonic) to make sure that participants did not need specifically musical knowledge in order to be able to co-improvise harmoniously with the group.

Research Questions

- What does it feel like playing with a heartbeat?
- What, if anything, does the heartbeat give to you in terms of personal/emotional information or stimulation?
- What is the difference between playing with a slow heartbeat and a fast heartbeat?
- What is the difference between playing with a rigid heartbeat and a variable heartbeat?

Findings

In Module A the group co-improvised with an existing piece of music based on heart beats - Pink Floyd's Breathe - and on anonymous recordings of heart beats related to different emotions and states of mind and body - a man meditating, a woman after intensive exercise, a person very activated and joyful and a rigid echo track based on a fast heart beat.
Participants reported that they found the exercises interesting and enjoyable to work with. They identified the Pink Floyd heartbeat as maybe "bored", "long-suffering" or even "slightly fearful" and

“searching for communication” which is exactly how Roger Waters of Pink Floyd describes the track: “Speak to me and Breathe together highlight the mundane and futile elements of being alive, but also the importance of living one's own life – and, crucially "Don't be afraid to care”” The consensus of the group was that the Pink Floyd heart beat communicated these emotions, and that this was helpful and informative in the process of co-improvisation.

The purpose of the two examples that followed was to investigate heart beats as indicative of autonomic arousal. The slow, meditative (c40bpm) heart beat produced a calm and spacious co-improvisation. The fast heart beat after intensive exercise produced high energy and fun.

The fast joyful heart beat and rigid techno beat were intended as invitations to the group to explore high vagal power (high heart rate variability, associated with positive feelings) and low vagal power (low heart rate variability, associated with negative feelings. Some members of the group found the heart beat with high variability more difficult to “perform” with, which was indicative of a very important point: many participants were using the heartbeat as a rhythmic cue, as much as, and possibly more than as an emotional cue. It is clear that using heart beats is potentially musically “invasive” or rhythmically “obliging” as well as emotionally informative. The group indicated that this was not necessarily a bad thing.

An interesting conversation followed the exercise with techno beats. Surprisingly the group found that the beats in some ways communicated negative emotion, which is what the experience of neurophysiology would suggest: high autonomic arousal combined with low heart rate variability and low vagal power is associated with anger or distress. A conversation followed about how we can sometimes enjoy elements of negativity or “danger” in music.

A6.3 Module B - Live heartbeat co-creation

Duration: 2 hours with a 30-minute break

Objectives

To begin co-improvisation with the live sound of the heart of one or two of the participants.

To work with haptic transmission of heart beats between participants, and to explore and assess how effective this may be in communicating states of mind and body and supporting remote co-creation

Exercises

1. Listening live to heart rate of a single participant
2. Group co-improvisation with live heart-rate data
3. Passing around the haptic actuators of two participants for reaction
4. Asking participants to respond to where on their body it feels most comfortable to place haptic actuators
5. Group co-improvisation with heart beats of two members of the group using haptic actuators to convey the sensation of the heart beat.

One of the participants wore the Polar H10 sensor, and the heart rate data was played live as audio to the room. The group as a whole then co-improvised with these individual heart beats, as a musical-emotional reaction to the quality of heartbeat. The chosen participant decided from among the improvisations in Module A which pitch material they wanted to use with their heart beat.

Research Questions

- How much can you sense a person's character and feelings by improvising with their heartbeat?
- Which felt better - improvising freely with heartbeat, or improvising in time with it?
- What did the participants think and feel about experiencing someone else's ECG signal haptically? Was it comfortable? Intrusive? How was the intensity of the sensation, was it too intense? Not intense enough? Did the participants notice any increase in emotional connection to the person whose ECG they could feel?
- How did these experiences relate to the accompanying music? Did the experiences compound and fit together naturally, or did they feel a disconnect or juxtaposition? Did they find the experience inspiring musically or otherwise?
- What other things would the participants want to experience haptically in the context of a musical experience? Other biological signals? The music itself (melodic, harmonic or rhythmic information)?
- What kind of device suits the participants for receiving haptic information? Is the vest comfortable? Can they imagine spending a long period of time in the vest? Does the vest stifle any movements/actions needed for creating music? Can the participants think of any other ways in which they would prefer to receive haptic information? (I'm not sure here what choices/possibilities we have for haptic devices)

Findings

The responses to "anonymous" heart beats had been positive, in the sense that emotions and states of mind and body had been to some extent communicated, and that the experience of

co-improvisation had been enjoyable - but with the important caveat that heart beats can also impose themselves as rhythmic cues.
Audifications of heart beats of those in the room, however, produced very strong reactions. Participants described the experience as “strange”, “weird” and “spooky” on the one hand and “intimate” , “strong” and “very moving” on the other. One participant, among the more musically experienced, reported that she could “feel the energy” of the person whose heart beat was being audified.
The session was organised as a series of duets between the participant whose heart beat was being audified (he was also singing), and each of the remaining participants in turn. The result was, by agreement of both participants and animateurs, among the best and most exciting work of its kind those present had ever experienced. Once again there was the issue of the musically rhythmic (as opposed to emotionally informative) effect of the heart, but this seemed to pale into insignificance in the presence of such powerful and immersive creativity.
The encouraging surprises were not over. The next stage was communicating the heart beat through a single haptic actuator ,that could be held in the hand, pressed against chosen parts of the body or worn in a sleeve. The team hooked up two participants. One of them, when he was handed the actuator, shouted out in high passion, “Oh f**k! I’m holding my heart!”. Passions were indeed high, and passing around the actuator seemed to create a less rhythmically obliging but no less emotionally informative experience for the participants.
There were several different aspects of the haptic heart beat communication to consider. The first was the choice of actuator algorithm. Three different algorithms had been developed by XSL. Participants were very clear on their choice: it was in fact a continuous vibration modulated by haptic signals from the heart. It gave the most complete and “realistic” representation of the heart beat.
Other aspects for consideration were related to where the actuator should be placed on the body. There were widely differing opinions among participants, which was perhaps to be expected among a diverse group. There were two front runners: on the upper forearm (which is serendipitously where the actuator “sleeve” is designed to sit) and on the back of the neck at the top of the spine.
It is possible that in the design of the platform the team will have to prepare for more than one actuator location on the body.

A6.4 Module C - Music and emotion

Duration: 2 hours with a 15-minute break

Objectives

To explore issues of music and emotion, including X-System analyses and issues of images and colours to prepare the way for the choice and design of avatars.

Exercises

1. A review of X-System analyses of participants' creative work or personal choices of repertoire, including discussion of autonomic, endocrine, limbic, motor and and emotional issues.
2. Each soloist chose one of the emotions listed on the colour wheel (Figure 17) and created a piece responding to a particular emotion or combination of emotions.
3. The group performed each of the pieces as a co-improvisation.

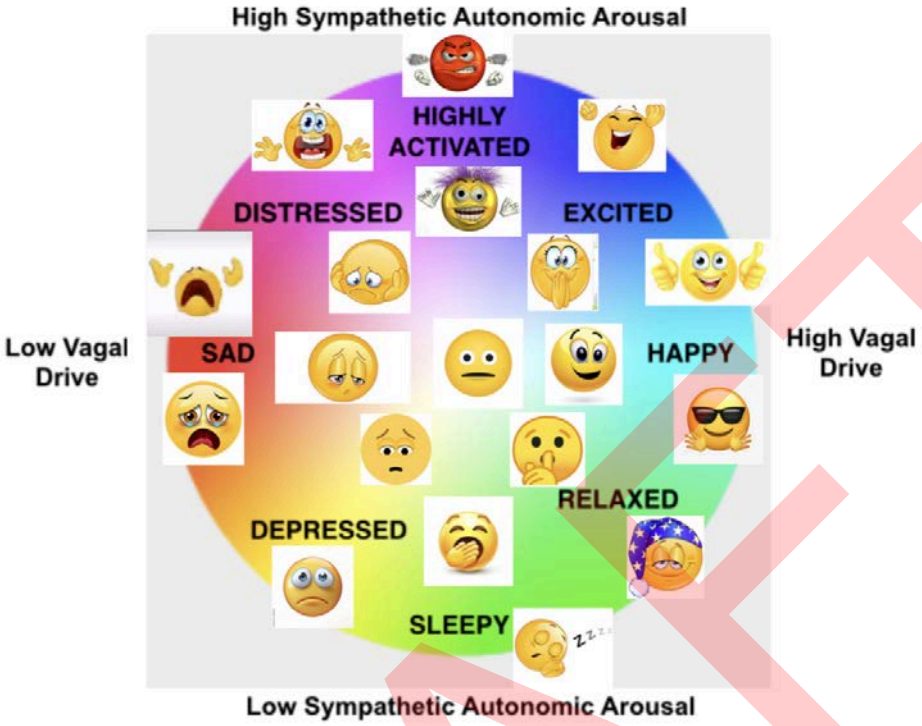


Figure 17: Emotional Colour Wheel

Research Questions

- What emotions do the participants feel when playing or listening to music?
- Can the same piece of music trigger different emotions depending on how you are feeling?

Findings

The session began with discussion of XSL analyses of “favourite” pieces of participants.

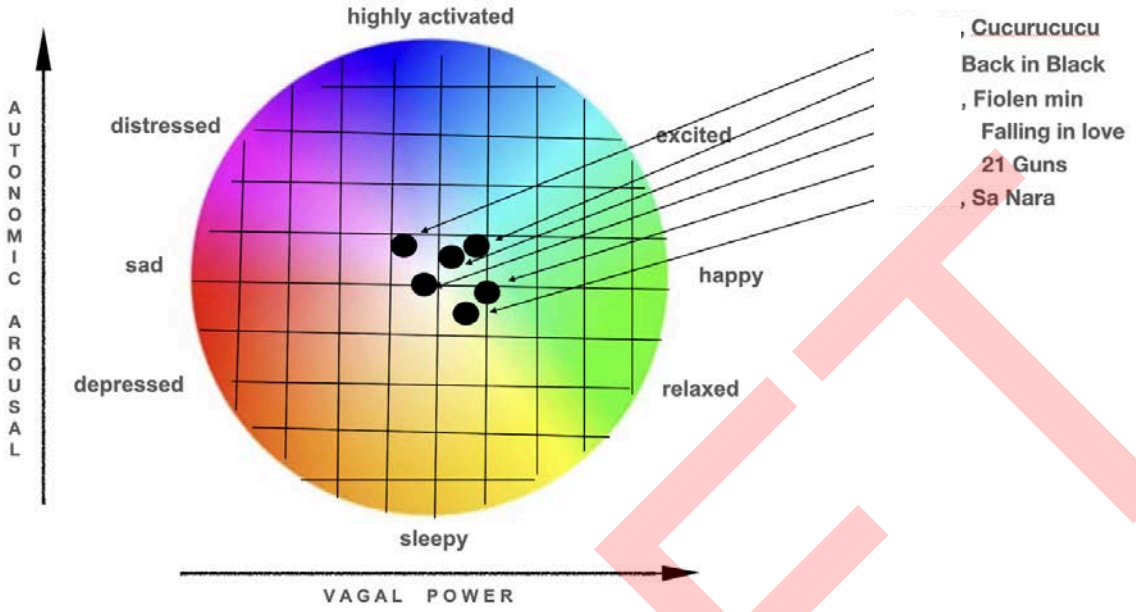


Figure 18 X-System Arousal/Vagal Power prediction

The XSL colour circle is a way of displaying emotion through predictions of arousal (related to speed of heart, the y axis) and vagal power (related to heart rate variability, the x axis). It was very interesting to see how close together the participant's musical-emotional preferences were. Even though Back in Black is an AC/DC song and 21 Guns a Green Day number, both are gentler and have greater vagal power than most heavy metal or neo-Punk rock.

The analyses served as a bridge between discussions about the heart and autonomic nervous system and more general discussions about emotion, including the body's chemistry etc.

Participants chose emotions from the colour circle/emoji graphic in the manual and developed musical material, subsequently shared with the whole group in co-improvisation. One participant chose "thrilled", which the group decided belonged somewhere between "excited" and "highly activated". All participants said they were comfortable with this way of working and with relating emotions to musical expression. Indeed, disabled musicians often seem to be more "at home" relating to emotional cues than others.

A6.5 Module D - Avatars

Duration: 2 hours with a 30-minute break

Objectives

To explore co-improvisation using a variety of avatars as stimuli, with a view to determine what kind of avatar will be most effective in communicating emotion and states of mind and body.

Creative design of avatars.

Exercises

1. Participants discussed various kinds of avatars

2. Using coloured pencils and paper, participants created designs for potential avatars.

Research Questions

- How do participants experience other people's emotions in a musical performance environment but also in general? How do they pick up on these emotions?
- How could these emotions be represented visually, haptically or otherwise? Is there a difference in how participants would like their own emotions represented and how they would like to learn about others?
- How could these avatars, or representations, respond to the beating of the heart or other information like brain waves?
- Can the participants draw or describe or otherwise illustrate these emotions? Do they have a colour, shape, or action associated with them? If they can imagine themselves represented by a small avatar, what would this avatar be doing while experiencing each emotion? How would its appearance change?

Findings

In Module D the group discussed avatars as a way of communicating states of mind and body from one remote co-creator to another. Various examples were projected onto a screen in the studio, ranging from realistic faces, to cartoon- or emoji-like images, to more abstract shapes Most participants could relate to most of the avatars, but found some of them, in particular a cartoon image of a girl's face with long eyelashes and heavy lipstick as stereotypical in a sexist way. In general participants responded most enthusiastically to expressive and richly colourful "painterly" images of human faces with aesthetic ambition, fantasy and elements of abstraction.

A crucially important point was raised by one of the participants who is a committed gamer. He pointed out that when he chose an avatar, it was because he wanted to become someone else, or more precisely someone other than himself and to feel different things, as opposed to the avatars that we were discussing, which were intended to be true representations of the emotions of co-creators.

In the next phase of Module D, participants designed avatars for themselves specifically intended to be capable of communicating their true emotions and states of mind and body, in a way that would be useful for remote co-creation.

Here are the avatars, together with the comments and explanations of their creators.

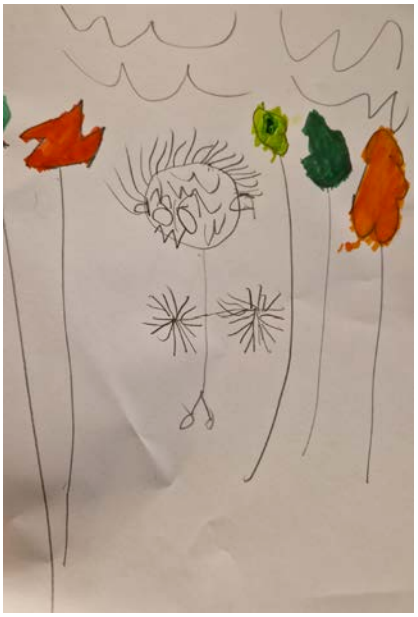


Figure 19: This was designed by one of the assistants. It represents six different avatars for emotions conveyed through different colours and dynamics and energy flow of the lines.

Figure 20: This was designed by a participant. The avatars are flowers and images of nature representing human emotion. Here the avatars are angry. This is communicated through colours and through the disturbed expression on the face of the dandelion.

Figure 21: This is a fantasy animal avatar. He is King of the Cats, proud,
confident, resilient, “in charge” and happy.

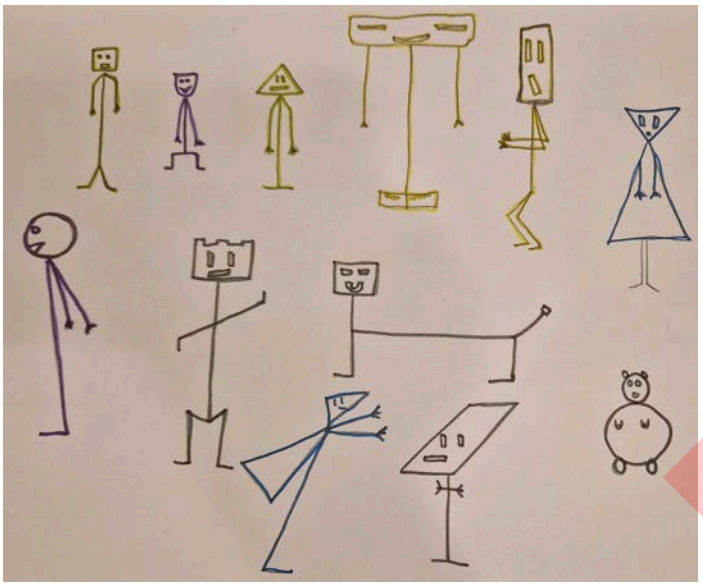


Figure 22: This was designed by an assistant and contains 12 possible models for an avatar, each of them capable of expressing emotions through facial expression and movement of limbs.

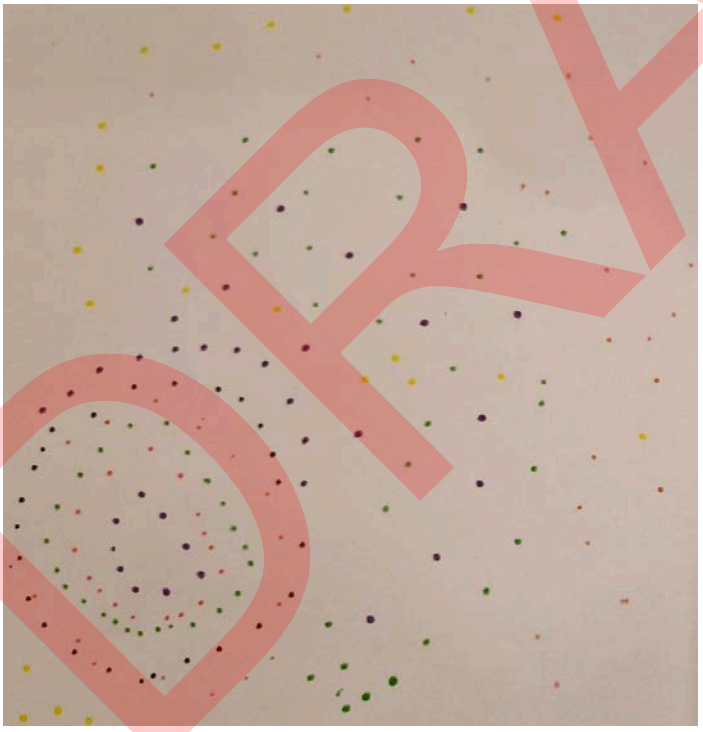


Figure 23: This is designed by a participant. The different-coloured "pixels" express emotion, and "swarm" in dynamic shapes that are also capable of expressing the energies of various emotions.

Figure 24: For this avatar, the shape is intended to remain more or less the same. The colours change to convey emotions.

Figure 25: This is an amoeba that can change shape and move in expressive ways. Emotion is also conveyed by colours. The creator generated it by arm movements and by way of clear instructions where colours should begin or end.

A6.6 Summative exercise

Objectives

To evaluate the performance of the algorithms discussed in section 3.1 in the context of the work performed in module B. In other words, the concepts and technologies explored in the session were

put more concretely into the context of a usable co-creation platform, using transmission of heart-rate data as a case study.

Exercises

1. Remote co-creation between two rooms incorporating live heart-rate and avatar design.

Research Questions

- Would direct ECG audification transmit cleanly and synchronously over Jacktrip?
- Would the platform succeed in remotely reproducing the co-creative experiences the participants had during the in-person sessions?
- Would the technology that was tested 'in-house' perform in a new and unfamiliar setting?

Findings

The participant who created the avatar in Figure 25 volunteered to be the "remote" co-creator in the final exercise of the workshop. She was isolated with her assistants in a separate room with sound attenuation, headphones, a microphone and polar heart sensor. Her avatar (above) was projected in the room with the rest of the group.

Jack Trip was used to carry the heart beat signal and the sound of her voice to the group as a whole. The group could hear her heart beat and "feel" it through the haptic actuator.

The group co-created and co-improvised on the basis of this remote musical, emotional and "state-of-mind-and-body" communication. It was a fitting summation - including remote co-creation, audifications of heart beats, haptic heart beat, emotional communication and avatars - to a productive and insightful two-day workshop.

REFERENCES

Alheeti, A. A. M., Salih, M. M. M., Mohammed, A. H., Hamood, M. A., Khudhair, N. R., & Shakir, A. T. (2023, November). Emotion Recognition of Humans using modern technology of AI: A Survey. In 2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS) (pp. 1-10). IEEE.

Anderson, K., & McOwan, P. W. (2006). A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *36*(1), 96-105.

Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*

Basu, S., Jana, N., Bag, A., Mahadevappa, M., Mukherjee, J., Kumar, S., & Guha, R. (2015). Emotion recognition based on physiological signals using valence-arousal model. In 2015 Third International Conference on Image Information Processing (ICIIP) (pp. 50-55). IEEE.

Baumgartner T, Lutz K, Schmidt CF and Jancke L (2006). The emotional power of music: How music enhances the feeling of affective pictures. *Brain Research,* 1075 (1), 151–164.

Blood AJ and Zatorre RJ (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences USA,* 98(20), 11818–11823.

Bobade, P., & Vani, M. (2020, July). Stress detection with machine learning and deep learning using multimodal physiological data. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 51-57). IEEE.

Cai, Y., Zheng, W., Zhang, T., Li, Q., Cui, Z., & Ye, J. (2016). Video based emotion recognition using CNN and BRNN. In *Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings, Part II 7* (pp. 679-691). Springer Singapore.

Chen, C. C., Chen, Y., Tang, L. C., & Chieng, W. H. (2022). Effects of interactive music tempo with heart rate feedback on physio-psychological responses of basketball players. *International journal of environmental research and public health, 19(8), 4810.*

Cittadini, R., Tamantini, C., Scotto di Luzio, F., Lauretti, C., Zollo, L., & Cordella, F. (2023). Affective state estimation based on Russell's model and physiological measurements. Scientific Reports, 13(1), 9786.

Day, M. (2016). Exploiting facial landmarks for emotion recognition in the wild. *arXiv preprint arXiv:1603.09129*.

Demochkina, P., & Savchenko, A. V. (2021). MobileEmotiFace: Efficient facial image representations in video-based emotion recognition on mobile devices. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V* (pp. 266-274). Springer International Publishing.

Domínguez-Jiménez, J. A., Campo-Landines, K. C., Martínez-Santos, J. C., Delahoz, E. J., & Contreras-Ortiz, S. H. (2020). A machine learning model for emotion recognition from physiological signals. Biomedical signal processing and control, 55, 101646.

Egger, M., Ley, M., & Hanke, S. (2019). Emotion recognition from physiological signal analysis: A review. Electronic Notes in Theoretical Computer Science, 343, 35-55.

Eldar E, et al (2007) Feeling the real world: limbic response to music depends on related content. *Cereb Cortex* 17(12):2828-40.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.
Erlich N, Lipp OV, Slaughter V (2013) Of hissing snakes and angry voices: human infants are differently responsive to evolutionary fear-relevant sounds *Developmental Science* 16;6 894-904

Fan, Yin, et al. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks." *Proceedings of the 18th ACM international conference on multimodal interaction*. 2016.
Frankland PW et al (1997) Activation of amygdala cholecystokinin B receptors potentiates the acoustic startle response in rats *The Journal of Neuroscience* 17(5) 1838-47

Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256). JMLR Workshop and Conference Proceedings.
Griffiths TD, Buchel C, Frackowiak RS, Patterson RD (1998) Analysis of temporal structure in sound by the human brain. *Nature Neuroscience* 1:422–427.

Heldt, SA, Falls, WA (2003) Destruction of the Inferior Colliculus disrupts the production and inhibition of fear conditioned to an acoustic stimulus *Behavioural Brain Research* 144 175-185

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1314-1324).
Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, *49*, 69-78.
Jorris PX, Schreiner CE, Rees A (2004) Neural processing of amplitude- modulated sounds *Physiological Reviews* 84 641-577

Koelsch, S. (2014). Brain correlates of music-evoked emotions. Nature Reviews Neuroscience, 15(3), 170-180.

Koelsch S, Fritz T Schlaug G (2008) Amygdala activity can be modulated by unexpected chord functions during music listening *Neuroreport* 9(18):1815-9.

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014, November). The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 291-298).
Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops* (pp. 94-101). IEEE.
Nkurikiyeyezu, Kizito, Anna Yokokubo, and Guillaume Lopez. "The effect of person-specific biometrics in improving generic stress predictive models." arXiv preprint arXiv:1910.01770 (2019).
Marsh RA, et al (2002) Projection to the Inferior Colliculus from the Basal Nucleus of the Amygdala *The Journal of Neuroscience* 22/23 10449-10460

McDermott JH, Lehr AJ, Oxenham AJ (2010) Individual differences reveal the basis of consonance *Current Biology* 20 1035-1041

Menon, V. Et al (2002) Neural correlates of timbre change in harmonic sounds Neuroimage 17 (4), 1742-1754

Osborne, N. (2009b) Towards a Chronobiology of Musical Rhythm in Communicative Musicality Editors: S. Malloch & C. Trevarthen. ISSN 0077-8923. (Oxford, UK and New York, USA) 545-564

Panksepp, J. & C. Trevarthen. 2009. The neuroscience of emotion in music. In *Communicative Musicality*. S. Malloch, C. Trevarthen, Eds.: 105–146. OUP.

Panksepp, J. (2003). Can anthropomorphic analyses of separation cries in other animals inform us about the emotional nature of social loss in humans? Comment on Blumberg and Sokoloff (2001). *Psychological Review, 110*(2), 376–388.

Panksepp, J. (1998) Affective Neuroscience OUP Oxford *passim*
Penhune V.B., Zatorre R.J. (2019) Rhythm and time in the premotor cortex *PLoS Biology* Sep 14;683:27-3 doi: 10.1016/j.neulet.2018.06.030 . Epub 2018 Jun 19.

Penhune VB, Zatorre RJ and Feindel WH (1999). The role of auditory cortex in retention of rhythmic patterns as studied in patients with temporal lobe removals including Heschl's gyrus. *Neuropsychologia,* 37(3), 215–231.

Peretz, I, Aube W, Armony, J.L. (2013) Towards a biology of musical emotions in The Evolution of Emotional Communication: From Sounds in Nonhuman mammals to Speech and Music in Man ed Altenmuller E, Schmidt S, Zimmerman E OUP

Peretz I (2001). Listen to the brain: the biological perspective on musical emotions. In P Juslin and J Sloboda, eds, *Music and emotion: Theory and research,* pp. 105–134. Oxford University Press, London.

Peretz I and Kolinsky R (1993). Boundaries of separability between rhythm in music discrimination: A neuropsychological perspective. *The Quarterly Journal of Experimental Psychology,* 46(2), 301–325.

Petrou, N., Christodoulou, G., Avgerinakis, K., & Kosmides, P. (2023, July). Lightweight Mood Estimation Algorithm For Faces Under Partial Occlusion. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 402-407).
Petrou, N., Christodoulou, G., Avgerinakis, K., & Kosmides, P. (2023, July). Lightweight Mood Estimation Algorithm For Faces Under Partial Occlusion. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 402-407).
Picard, R. W. (2000). *Affective computing*. MIT press
Reed, L. I., & DeScioli, P. (2017). The communicative function of sad facial expressions. *Evolutionary Psychology*, *15*(1), 1474704917700418.
Rodrigues, A. S. F., Lopes, J. C., Lopes, R. P., & Teixeira, L. F. (2022, October). Classification of facial expressions under partial occlusion for VR games. In *International Conference on Optimization, Learning Algorithms and Applications* (pp. 804-819). Cham: Springer International Publishing.
Russell JA. (1980) A circumplex model of affect. *Journal of Personality and Social Psychology.* 39:1161–1178.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).

Schaefer, H. E. (2017). Music-evoked emotions—Current studies. *Frontiers in neuroscience*, *11*, 600.

Schaffarczyk, M., Rogers, B., Reer, R., & Gronwald, T. (2022). Validity of the polar H10 sensor for heart rate variability analysis during resting state and incremental exercise in recreational men and women. *Sensors*, *22*(17), 6536.
Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44(4)*, 693-727.

Scherer, K.R., Shuman, V., Fontaine, J.R.J, & Soriano, C. (2013). The GRID meets the Wheel: Assessing emotional feeling via self-report. In Johnny R.J. Fontaine, Klaus R. Scherer & C. Soriano (Eds.), Components of Emotional Meaning: A sourcebook *(pp. 281-298)*. Oxford: Oxford University Press.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018, October). Introducing wesad, a multimodal dataset for wearable stress and affect detection. In Proceedings of the 20th ACM international conference on multimodal interaction (pp. 400-408).
Schneider, P. et al (2002) Structural, functional, and perceptual differences in Heschl's gyrus and musical instrument preference. *Annals of the New York Academy of Sciences*, 1060, 387-94

Sivaramakrishnan S, et al (2004) GABA (A) synapses shape neuronal responses to sound intensity in the Inferior Colliculus *Journal of Neuroscience* 26;24(21)5031-43

Stein MB, Simmons AN, Feinstein JS, Paulus MP.(2007) Increased amygdala and insula activation during emotion processing in anxiety- prone subjects. *Am J Psychiatry* 164(2): 318-27

Turchet, L., & Barthet, M. (2018). Co-design of Musical Haptic Wearables for electronic music performer's communication. *IEEE Transactions on Human-Machine Systems*, *49*(2), 183-193.
Warren, J.D. et al (2003) Separating pitch chroma and pitch height in the human brain Proceedings of the National Academy of Sciences USA100 (17) 10038-10042

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, *23*(10), 1499-1503