D4.5 Agent Social Interpretation Enabling

Grant Agreement nr	856879
Project acronym	PRESENT
Project start date (duration)	September 1st 2019 (36 months)
Document due:	August 31st 2021
Actual delivery date	August 31st 2021
Leader	University of Augsburg (UAu)
Reply to	florian.lingenfelser@informatik.uni-augsburg.de silvan.mertes@informatik.uni-augsburg.de thomas.kiderle@informatik.uni-augsburg.de
Document status	Submission Version

Project funded by H2020 from the European Commission

Project ref. no.	856879
Project acronym	PRESENT
Project full title	Photoreal REaltime Sentient ENTity
Document name	Agent Social Interpretation Enabling
Security (distribution level)	Public
Contractual date of delivery	31/08/2021
Actual date of delivery	31/08/2021
Deliverable name	D4.5 Agent Social Interpretation Enabling
Type	Report
Status & version	Submission Version
Number of pages	25
WP / Task responsible	University of Augsburg (UAu)
Other contributors	-
Author(s)	Florian Lingenfelser, Silvan Mertes, Thomas Kiderle
EC Project Officer	Ms. Diana MJASCHKOVA-PASCUAL Diana.MJASCHKOVA-PASCUAL@ec.europa.eu
Abstract	In this report period we finalized the audiovisual emotion recognition pipeline, which forms the core of the PRESENT social sensing system. The system operates within the dimensional valence and arousal space to recognize, process and communicate affective states. It currently features analysis of the vocal modality for voice activity detection and sentence segmentation, as well as arousal recognition from paralinguistic characteristics. From the video stream we analyse facial expressions to determine valence and arousal scores. The audiovisual information is merged within a sophisticated asynchronous fusion model, delivering the final assessment.
Keywords	Social Signal Processing, Machine Learning, Real-time Emotion Recognition, Multi-modal Fusion
Sent to peer reviewer	Yes
Peer review completed	Yes
Circulated to partners	No
Read by partners	No
Mgt. Board approval	No

Document History

Version and date	Reason for Change
1.0 19-07-2021	Document created by Florian Lingenfelser
1.1 13-08-2021	Version for internal review (14 days before submission date)
1.2 30-08-2021	Revisions in response to review
1.3 31-08-2021	Final version submitted to Commission

Table of Contents

1 EXECUTIVE SUMMARY

The deliverable D4.5 - *Agent Social Interpretation Enabling* - (R, PU, M24) reports on the fully integrated PRESENT social sensing system. Since the PRESENT agent is designed to enable a natural and engaging interaction with its human user, we need to assess the affective state of the user at any point in time over the course of the interaction. Exemplary components within the whole PRESENT system that rely on steady affect recognition include the design of dialogues that take emotional reactions into account, the adaption to user preferences in behaviour and voice based on user behaviour or finally, to adapt its inner emotional simulation based on empathy to the perceived affective state.

In this report period we finalized the audiovisual emotion recognition pipeline, which forms the core of the PRESENT social sensing system. The system operates within the dimensional *valence* and *arousal* space to recognize, process and communicate affective states. It currently features analysis of the vocal modality for voice activity detection and sentence segmentation, as well as *arousal* recognition from paralinguistic characteristics. From the video stream we analyse facial expressions to determine *valence* and *arousal* scores. The audiovisual information is merged within a sophisticated asynchronous fusion model, delivering the final assessment.

2 BACKGROUND

The deliverable at hand reports the finalization of the audiovisual sensing system, which enables the PRESENT system to recognize, interpret and make use of affective states of the user. Work presented within this deliverable is part of WP4 - *Behavioural Sensitivity and Responsiveness*. Within this document, we describe and evaluate deep, transfer and multi-task learning based models for recognition of human affect and behaviour from visible and audible cues. We introduce the final version of the multi-modal fusion component and give outlooks how additional modalities and external content knowledge may be integrated in the sensing system during the remaining time of the project.

As shown by the latest presented demonstrator, the sensing system is by now fully integrated in the overall PRESENT architecture and serves as input for tasks within and beyond WP4.

3 INTRODUCTION

The PRESENT social sensing system incorporates various modules for activity detection and emotion recognition from voice, face (and optionally body and gaze) modalities. The components are embedded within a multi-modal framework, which guarantees synchronized sensor input and data processing in real-time. It is based on audiovisual input and can easily be adjusted to any of the PRESENT use-cases.

Monitoring and interpreting user affect is a crucial requirement to create an agent, which is capable of natural and believable behaviour towards the user. It should be able to consider and react to emotions within the interaction - be it in dialogues that consider affective reactions, adaptation based on implicit feedback shown by the user or during the simulation of its own inner emotional status via e,g empathic reactions. The perception of the human counterpart requires an artificial emotional intelligence. In D4.2 - *Interim Report on Agent Social Interpretation Enabling*, we introduced the theoretical concepts of discretizing emotions in order to properly handle them within the PRESENT system. The dimensional *valence* and *arousal* model was chosen as a general representation method because respective values can be used organically by subsequent modules and can be mapped onto emotional labels if needed. We use this definition throughout the sensing system and also within the PRESENT architecture. The presented work therefore enables the PRESENT agent as a whole to present itself to the user as an emotionally reactive entity.

The following chapters are structured as follows:

- Chapter 4 presents a technical description of the PRESENT social sensing system. We introduce the multi-modal framework as well as specific implementations of the machine learning components needed to automatically assess user emotion in real-time from single modalities. We describe and evaluate deep transfer and multi-task approaches for audio and video processing. Furthermore, optional modalities as well as integration of external context with potential affective information value are proposed.

- Chapter 5 explains the sophisticated asynchronous fusion approach that has been implemented to merge abstract information from differing sources with various processing rates into a continuous assessment of user emotion. We explore the parameters that influence the fusion process and determine optimal values for the given affect recognition task.

- Chapter 6 gives an overview over the system and explains close connections to other modules within the PRESENT architecture. We describe the integration and communication with the PRESENT use-case applications and explain the need for a distributed system design. Lastly the most recent demonstrator is shown.

- Chapter 7 concludes the document.

4 MODALITY PROCESSING

The PRESENT agent is meant to handle diverse roles within various applications. The unifying factor of all roles however, is that it needs to represent a trustworthy contact partner and therefore a natural interaction in which emotional expressions are considered is a must. In order to correctly include the emotional state of users in the interaction, it needs to categorize the recognized affective state in an understandable manner. We introduced the dimensional *valence- arousal* model (Russel, 1980) as representation for the affective user state in deliverable D4.2. It spans an emotional coordinate system with two axes. The *valence* axis describes the positive or negative manifestation of an emotion (e.g. pleasant versus unpleasant) while the *arousal* axis determines the degree of agitation (e.g. calm versus stressed). Discrete emotion categories can be derived from relative positions within the dimensional model. Consequently, we have decided that each unimodal recognition component within the multi-modal affect recognition component should provide its insights encoded in *valence* and/or *arousal* values and provide these to the system as a numerical value between -1 and 1.

Some of the presented components have already been covered in their initial state within deliverable D4.2 - *Interim Report on Agent Social Interpretation Enabling*. We shortly re-introduce the corresponding algorithmical approaches for the sake of a complete system description. In its core, the PRESENT sensing system is based on audiovisual input, i.e. the processing of video and audio streams coming from hardware such as webcams.

4.1 Affective State Recognition from Facial Expressions

In its first iteration, the PRESENT social sensing system was solely relying on analysis of a user's voice within the human-agent interaction to deduce information about the user's current affective state (i.e. the extent of expressed *arousal*). The final system however implements a fully integrated facial analysis module to interpret the observed facial expressions with respect to their emotional content.

4.1.1 Dataset

Figure 1: Excerpt from the AffectNet (Baltrušaitis 2015) database. High quality emotional annotations (i.e. continuous valence-arousal dimensions) are more commonly available for facial expressions than for multi-modal data. In contrast to databases containing multi-modal emotional expressions (i.e. including high quality spoken audio), well annotated data exclusively featuring facial expressions are more broadly available.

The field of image processing is by far the most advanced research topic concerning deep learning approaches. Classification problems are accessible due to their visual nature and the number of adequately annotated databases for a wide range of recognition tasks is relatively high. However, the majority of huge databases are focused on object recognition. Though there is a selection of publicly available databases, affect recognition (primarily from facial expressions) is not the most common image processing topic. Our need for high quality annotations in the *valence* - *arousal* space further limits the selection, as databases featuring only categorical emotion labels are more common. Therefore our current model is again based on transfer learning techniques and is for now mainly trained on a single comprehensive affective corpus called AffectNet (Baltrušaitis, 2015). The collection of facial expressions consists of close-up shots of human faces taken from publicly available video and picture collections (reference). Annotations contain categorical emotion labels as well as continuous *valence* - *arousal* scores.

4.1.2 Recognition Approach

The basis of our recognition model is formed by a pre-trained MobileNetV2 (Sandler, 2018) network architecture, trained for object recognition on the comprehensive ImageNet (Deng, 2009) database. Since the AffectNet database features not only valence - arousal scores, but also associated categorical emotion labels, we can refine the preceding transfer learning approaches with a multi-task learning strategy. Figure (reference) illustrates the architecture: The pre-trained base model shares layers with two model heads. One model head is learning the *valence* - *arousal* recognition task, the second one is trained in classification of eight categorical emotional classes. Both model heads can backward propagate into the shared base model layers. Hereby, the second task (categorical emotions) is used to stabilize and improve performance of the initial learning task (*valence*/*arousal*). The second, redundant model head is discarded after training. The base model (further trained by the transfer and multi-task steps) together with the *valence* - *arousal* model head forms the conclusive recognition model.

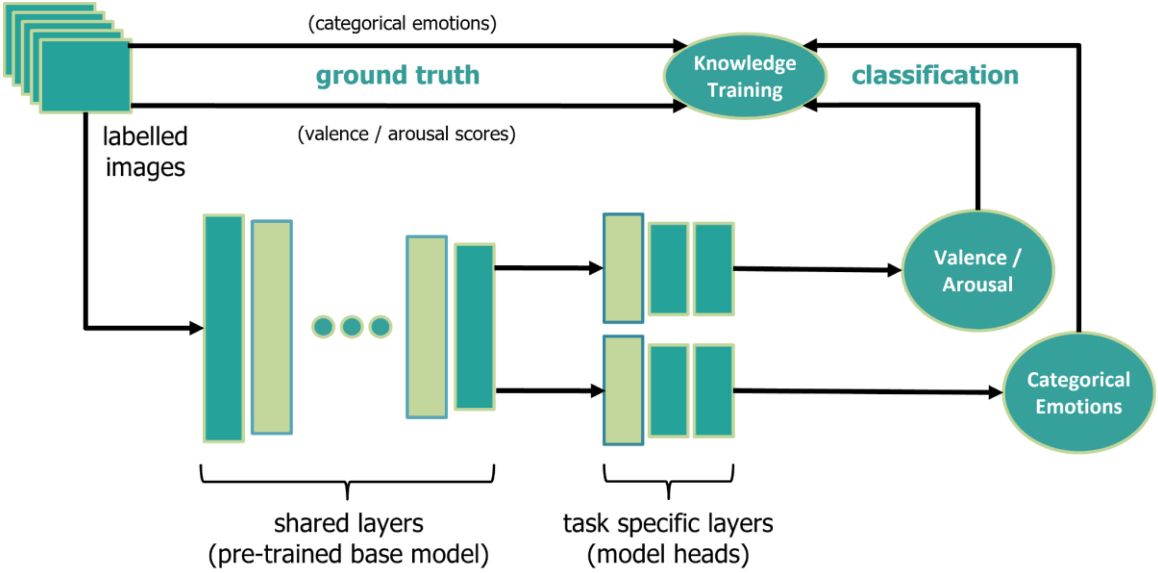


Figure 2: The pre-trained base model shares layers with two model heads, of which one trains the valence - arousal recognition task and the other classification of eight categorical emotional classes. Backward propagation from both model heads into the shared base model layers forms the base model with additional information and stabilizes the valence - arousal recognition performance.

4.1.3 Evaluation

The accuracy of the described *valence* - *arousal* recognition system is on par with the baseline presented by authors in (Baltrušaitis 2015) (Table reference). However, the deep convolutional neural network (AlexNet) used is based on a very complex architecture and results in a large (with respect to disk space) and slow (with respect to computational effort) model to be deployed in the recognition system. The additional efforts of transfer and multi-task learning pay off in achieving a comparable result with the lightweight MobileNetV2 architecture - giving us the possibility to implement a computationally efficient recognition module that can even be run on mobile devices.

	C2NN Baseline (AlexNet)		**Transfer & Multi-task Learning (MobileNetV2)**	
	Valence	**Arousal**	**Valence**	**Arousal**
RMSE	0.37	0.43	0.40	0.37
CORR	0.66	0.54	0.60	0.52

Table 1: As we deal with continuous assessment curves rather than single samples, performance is measured in RMSE (Root Mean Square Error - the predicted error aggregated over all sample points within the predicted and actual curves, where an error of 0 would indicate perfect prediction) and CORR (Pearson Correlation Coefficient - the linear correlation between trends in the predicted and actual curve, where a value of 1 would indicate perfect correlation). We achieve a comparable performance on both measurements for both valence and arousal with the practical benefit of employing a significantly slimmed model.

4.1.4 Tool for Automated Annotation of Affective Faces

In order to be able to generate visual models that refer to certain emotions, we provided project partners with a tool to automatically annotate recorded videos with respect to their affective content. The tool was implemented as a command line tool, including all necessary emotion sensing components to enable a reliable offline analysis of visual video input. The annotation process includes the analysis of both *valence* and *arousal* for every single frame of the input video by using the aforementioned recognition approach.

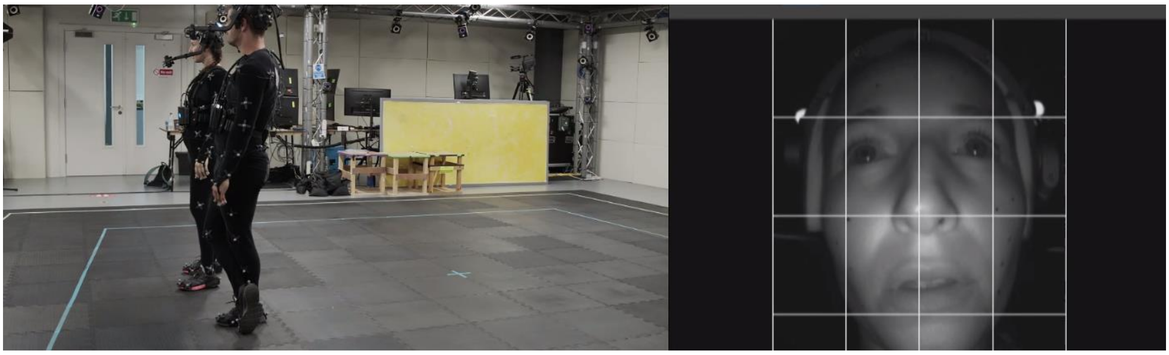


Figure 3: PRESENT recording session with professional actors. Head-mounted cameras recorded facial expressions and in order to quickly generate a rough annotation, an offline version of the

aforementioned facial analysis component was provided to partners as an easy-to-use command line tool.

The tool can be used to analyse recordings of the PRESENT actors (Figure 3) regarding their *valence* and *arousal* values. Due to the fact that this data is recorded with a head mounted camera, the recording setting is very consistent. To take advantage of this, we enhanced the offline annotation tool with an additional preprocessing stage. The implemented pre-processing algorithms take care of only taking relevant areas of the visual input into account. Thus, a face cropping algorithm that is exactly adapted to the specific recording scenario has been deployed.

4.2 Affective State Recognition from Voice

In addition to facial expression analysis, the audiovisual input constantly provides the user's voice during interaction. Intensity in the voice modality is suspected to be highly correlated with the emotional activation of a person (Lassalle, 2019) and is consequently well suited for estimating the *arousal* dimension, as loudness, pitch or speech rate carry descriptive information for determining user arousal. To further our insight about the current user state, we can analyse the raw audio signal gained from any available microphone sensor for emotional content. Traditionally, vocal characteristics were extracted from the voice signal as a feature vector via a feature extraction component (e.g. EmoVoice (Vogt, 2008)) to an extracted spoken sentence. These were then classified by a statistical classification model. This approach is limited by the characteristic that feature calculation works most reliably when applied to longer sequences of audio data (i.e. whole sentences). To achieve a continuous and reactive analysis we use neural networks that have the ability to bypass the feature calculation step and instead take small chunks of the raw audio signal as input.

In order to achieve high quality continuous arousal classification, a lot of training data - i.e. well annotated audio data with continuous arousal scores - is theoretically needed. In practice, this kind of affective data is not widely available. To counter this problem, we apply a transfer learning approach with a pre-trained convolutional network. Hereby, transfer learning describes a machine learning approach, in which the aim is to transfer knowledge gained in having solved one problem to a new but related problem.

4.2.1 Dataset

Transfer-learning is done with an excerpt of the SEMAINE (McKeown, 2013) database, which is one of the few sources to obtain high quality arousal annotations for audio data. The excerpt consists of around 30 sessions each containing 5 to 15 minutes of recording. The SEMAINE project[1] was one of the first endeavours to enable natural conversations with human-like avatars. The interaction - i.e. the dialogue - was steered from outside with a *Wizard-of-Oz* setup and users were recorded interacting with the avatar. Application of a transfer learning approach enables us to use the (relatively small but well annotated) corpus to circumvent the problem of sparse data availability and to nevertheless implement a dedicated deep neural network within the PRESENT social sensing system.

4.2.2 Recognition Approach

As mentioned above, the goal of our online classification system is a continuous and reactive analysis of the audio signal. The processing of raw audio with neural networks can process very short signal

[1] http://www.cs.nott.ac.uk/~pszmv/page5/page8/page8.html

segments (i.e. 50 milliseconds) and enables classification at a very high sample rate – resulting in the targeted smooth and reactive assessment of the current arousal level.

The base model (SoundNet[2]) is trained to recognise scenes based on present sounds in the audio channel. Of course, scene classification is not the target of our emotional analysis. However, the output of the last pooling layer of the pre-trained model represents a compact and focused distillation of the audio signal that is well suited for further processing. We feed the output of the last pooling layer (pool 5) of the SoundNet architecture into two additional fully connected layers which are meant to map the representation to the final arousal score.

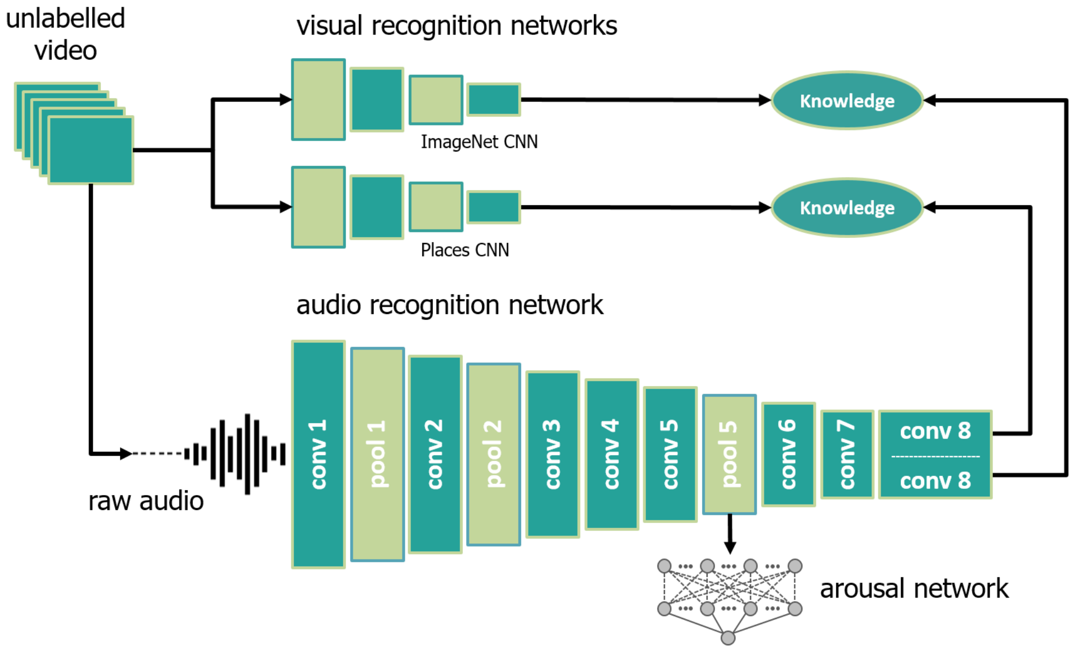


Figure 4: SoundNet (Aytar, 2016) is a deep convolutional network for natural sound recognition. The network is trained by transferring discriminative knowledge from visual recognition networks into sound networks. The approach capitalizes on the synchronization of vision and sound in video.

4.2.3 Evaluation

For Evaluation of the system we again use the excerpt of the SEMAINE database and carry out a leave-one-out cross validation. The RMSE and CORR for *arousal* recognition are on a comparable level to the performance of facial expression analysis (Table 2). The *valence* analysis however underperforms in comparison to arousal detection. This shows especially in the case of the very low CORR score, which means the signal is often trending in the wrong direction. This behaviour is of course very misleading for the succeeding fusion algorithm, therefore we only consider *arousal* assessments for following processing steps.

	Transfer Learning (SoundNet)

[2] http://soundnet.csail.mit.edu/

	Valence	Arousal
RMSE	0.47	0.35
CORR	0.09	0.59

Table 2: As we deal with continuous assessment curves rather than single samples, performance is measured in RMSE (Root Mean Square Error - the predicted error aggregated over all sample points within the predicted and actual curves, where an error of 0 would indicate perfect prediction) and CORR (Pearson Correlation Coefficient - the linear correlation between trends in the predicted and actual curve, where a value of 1 would indicate perfect correlation). We achieve a comparable performance on both measurements for both valence and arousal with the practical benefit of employing a significantly slimmed model.

4.3 Voice Activity Detection

In D4.2 - *Interim Report on Agent Social Interpretation Enabling*, we introduced our voice activity detection approach for the segmentation of voiced parts of the audio signal versus silence, background noises, etc. A robust voice activity detection model is needed for interactions with the user via a dialogue system. Not only is the correct recognition of spoken sentences the basis for dialogue management, it furthermore can guide the emotion recognition module to signal parts that may carry relevant information.

4.3.1 Dataset

To generate the vast amount of training data needed to train a deep neural network from scratch, we employ a sophisticated method to automatically generate annotations: We looked for publicly available videos from television media libraries that feature high quality subtitles with timestamps (reference). With these timestamps we are able to automatically segment annotations for voiced audio parts against silence or various kinds of sounds we want to identify as background noises.

4.3.2 Recognition Approach

The PRESENT sensing system implements a deep learning based approach by continuously classifying the audio signal with University of Augsburg's VadNet (Wagner, 2018). Output of the classification model is smoothed over consecutive analysis frames and results in a continuous estimation of *NOISE* versus *VOICE* classes. By tracking this assessment, we are able detect the on- and offset of spoken sentences in the audio channel. This information is for example used in dialogue tasks, where spoken sentences are extracted from the raw audio signal and transmitted as dialogue acts to the external dialogue management tool. Here, speech recognition is carried out and corresponding answers are created.

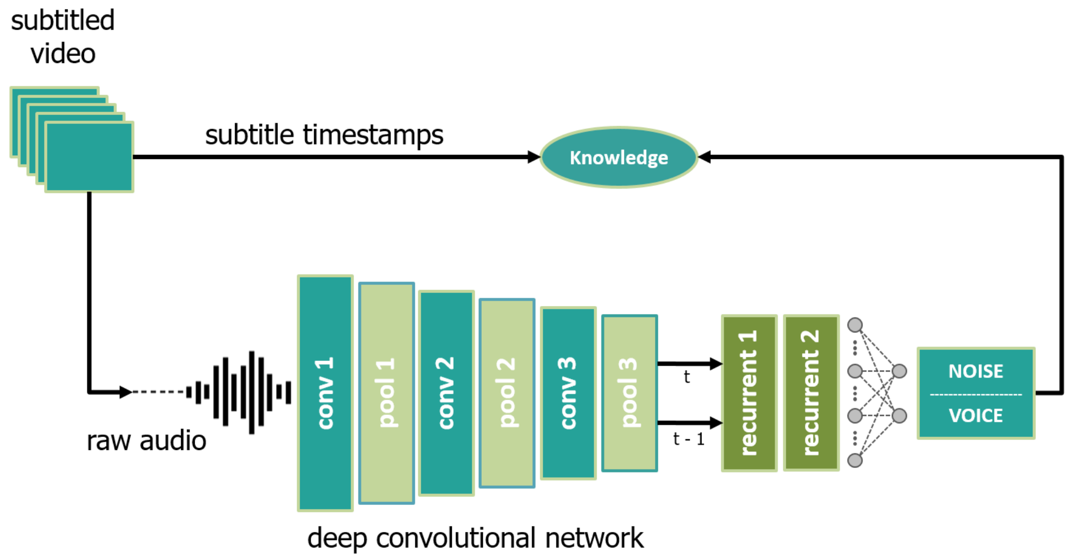


Figure 5: The VadNet model has been trained with Tensorflow (Abadi, 2016). It takes as input the raw audio input and feeds it into a 3-layer convolutional network. The result of this filter operation is then processed by a 2-layer Recurrent Network containing 64 RLU cells. The final bit is a fully-connected layer, which applies a softmax and maps the input to a tuple <noise, voice>.

4.4 Optional Modalities

The audiovisual core system is theoretically finalized, but as development of PRESENT use cases carries on, we figured out potential needs for additional modalities to be included in the sensing system. We have therefore included several software toolkits to access more information about the user. Hereby we chose open-source software solutions in order to avoid problems with distribution among partners as well as late changes in the hardware setup.

4.4.1 Speech-to-Text

In D4.2 we introduced a streaming interface to a sophisticated dialogue management tool (i.e. Google Dialogflow[3]). Partners communicated the need for a more simple approach, that could for example be used for keyword spotting and a custom lightweight dialogue system. To this end we included a Speech-to-Text (STT) system into the sensing system. As Mozilla's DeepSpeech (Hannun, 2014) system is one of the current state-of-the-art STT frameworks, we decided to use that architecture in order to allow for a high recognition accuracy.

DeepSpeech is an open-source end-to-end neural STT engine that has already been applied to various scenarios and applications in academic and non-academic environments. UAu integrated a DeepSpeech implementation that was modified and improved by Mozilla. That particular implementation consists of a recurrent neural network (RNN) that was trained on the task of generating English text transcriptions of audio data containing speech. After 5 hidden layers the

[3] https://cloud.google.com/dialogflow

resulting data is fed to a softmax before being mapped to the respective characters. A pre-trained model that is publicly available via the official DeepSpeech repositories[4] was used for the task.

4.4.2 Full Body Analysis

The analysis and synthesis of body movement and gestures will be covered in D4.7 - *Reactive Agent and Touch Enabling*. However, we have by now included a full body analysis tool within the signal processing framework. Affective concepts such as a person's current expressivity (Caridakis, 2010) or engagement in a conversation (Baur, 2016) can be reliably described by suitable movement features (Caridakis, 2010) and body orientations and postures. Therefore it might be interesting to include this information in the social sensing system.

Figure 6 shows possible ways to track user body postures. All are based on tracking the skeleton (i.e. body parts, joints, rotations) - the hardware setup however varies greatly. Boody suits provide the most precise tracking, but are of course not applicable in standard applications. Depth image based tracking as provided by the Microsoft Kinect[5] is a solid, accurate and affordable hardware solution but in comparison to a software based tracking method still needs a certain investment and setup. We have decided to integrate a software based open-source solution to the signal processing framework. OpenPose[6] offers reliable tracking of multiple people in video streams, which makes it a good choice in our audiovisual setup.

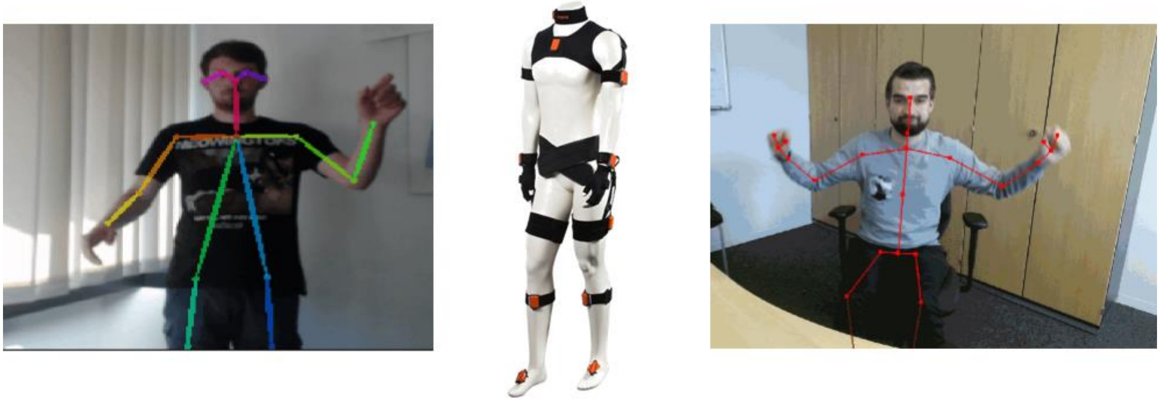


Figure 6: Affective events are generally discarded rather fast, with cues from the voice having a longer lasting effect - pointing to slower changes in the vocal modality with longer influence on the final result. No modality is weighted as 'high', most probable due to non-optimal accuracy of both modalities.

4.4.3 Gaze Analysis

The last modality we consider for possible analysis and information gain is the gaze direction of the user. Gaze behaviour can inform the application about the area of focus of the user within the visual

[4] https://github.com/mozilla/DeepSpeech/releases

[5] https://developer.microsoft.com/en-us/windows/kinect/

[6] https://github.com/CMU-Perceptual-Computing-Lab/openpose

representation. As this might be of interest for the development of PRESENT use cases, we decided to also include a possible software solution to the sensing system. Again, costly and cumbersome hardware is avoided in favor of software based open-source solutions within the audiovisual setup. The OpenFace library[7] is primarily known for detection of facial landmarks, but it also calculates an estimation of the viewport and gaze direction (Figure 7).

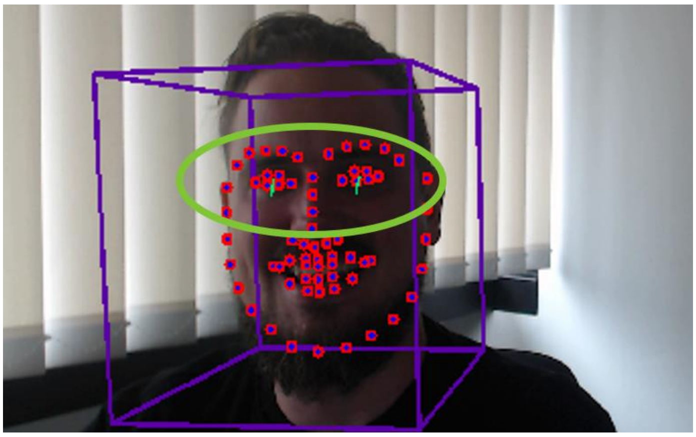


Figure 7: The OpenFace library is an open-source software tool which in addition to the detection of facial landmarks is able to estimate the viewport and gaze direction of a user based on video input only.

5 MULTI-MODAL FUSION

Given reliable recognition output for all considered modalities, we have finalized in the reporting period the asynchronous fusion system to incorporate the multi-modal assessments into a coherent prediction of a user's valence and arousal scores, which can then be utilized by subsequent PRESENT modules (e.g. dialogue management, socially aware reinforcement learning, agent emotion simulation) to generate fitting reactions to affective user states. We already used a basic version of the multi-modal fusion algorithm in the first demonstrator (reference) to handle arousal estimation from the vocal modality. We have by now adapted the algorithm in further iterations of the PRESENT social sensing system, i.e. with respect to inclusion of the facial expression modalities, proper weighting of considered modalities and their accuracy in valence and arousal recognition. The current version of the PRESENT emotion recognition system features audiovisual recognition systems as well as the implementation of the asynchronous, multi-modal fusion algorithm with optimal, modality-specific parameters determined by an evaluation study.

In D4.2 we introduced our underlying multi-modal signal processing framework SSI[8] (Wagner, 2013) and a technical perspective, we have to take into consideration that while the framework guarantees synchronization of data streams from multiple sensors, the recognition results of considered modalities will appear at differing sample rates, because of varied analysis windows for data processing and classification. This means that the fusion algorithm needs the ability to handle asynchronous input in the form of events and has to keep track of timings to calculate the amount of time a registered event has influence on the fusion result. In (Lingenfelser, 2016), we present an asynchronous and event driven fusion mechanism, which exactly matches the mentioned affordances. In short, modalities publish recognition results based on their inherent recognition rate in the form of asynchronous events. The fusion algorithm collects these outputs (in our case, *valence*

[7] https://github.com/TadasBaltrusaitis/OpenFace

[8] https://hcai.eu/projects/ssi/

and *arousal* values), remembers respective time of occurrence and offsets this timing with predefined time spans and weights, defining how long and to which extent the event should influence the fusion result until it gets discarded. The algorithm uses the modified input values to calculate a running mean over all currently active events on all considered emotional dimensions.

5.1 Fusion Approach

In D4.2 - *Interim Report on Agent Social Interpretation Enabling* we stated the need for a fitting multi-modal fusion algorithm to incorporate impressions from multiple sources (i.e. modalities) into a coherent final result within the targeted *valence* - *arousal* space which can be published to the rest of the system. As indicated in D4.2, the implemented multi-modal fusion system which processes the unimodal recognition assessments is an asynchronous and event driven approach (Lingenfelser, 2016) and features various options to adapt to the characteristics of monitored modalities.

Events are hereby defined as affective manifestations that are emitted by the involved modalities, occur asynchronously across affective channels and influence the assessment of the current affective state. They are coupled to activity detection in the respective affective channels. The assurance of meaningful activity can be as simple as guaranteeing stable tracking in frames containing video data of e.g. faces for facial analysis. More elaborate activity measures include signal processing or segmentation steps such as the introduced voice activity detection to determine parts of the audio channel that probably hold paralinguistic information. Frames that do not pass an activity check will not be qualified to generate affective events.

We have defined every contributing affect recognition modality to generate assessments in the *valence* - *arousal* space. Consequently, the fusion algorithm is also designed with this emotion paradigm in mind. The fusion algorithm handles the incoming *valence* and *arousal* events with respect to their occurrence over time. As the *valence* - *arousal* space spans two axes, we need to handle multidimensional events, so the processing of events holding two entries (one current *valence* and one current *arousal* score) is theoretically possible. However, this approach induces an interdependence between the two axes, which may be beneficial if the dimensions of a recognition problem for example mutually exclude each other (e.g. happy - unhappy). In the *valence* - *arousal* space this behaviour is not desirable, the two dimensions are generally independent. As established and shown in the system schematic (reference), the audio modality (at the moment of writing) only provides reasonable insights in the *arousal* level. If we would demand two-dimensional events, the valence entry would need to be filled, but every possible entry (e.g. zero) would suggest an unwanted influence on the fusion result. The solution is the definition of two separate event processing fusion components, one for each axis of the valence-arousal space. The single results are merged back into a two-dimensional final result to be published to the PRESENT system.

5.2 Parameter Exploration

During the description of the event driven fusion algorithm, we have described its adjustability via a number of parameters. These of course influence the recognition quality of the multi-modal system. There are two defining parameters for each involved modality that are declared for the *valence* and *arousal* axis independently. The first is the *weight* of respective affective events which determines the proportionate impact they will have on the final fusion result. The second is the *decay-speed*, the rate at which this influence will decrease over time.

Typically, a greedy grid-search algorithm is used to determine optimal parameter configuration for the final recognition system. Given a suitable training corpus, we define a set of possible values for each parameter and run evaluations against gold standard annotations for each possible combination. These tests are evaluated against a pre-defined development set, separated from the training data. As we have no annotated data from PRESENT use cases yet, that would ideally serve as the basis for the following exploration study, we resort to a selection of audiovisual recordings featuring a communicational setting which we have gathered in former projects.

As quality criterion, we again use the root mean square (RMSE) error and the Pearson correlation (CORR) coefficient. Hereby, we give more weight to the RMSE. The reason is that the gold annotation can indeed be interpreted as a continuous annotation, but as the annotation values are chosen from a fixed set, we see a staircase annotation. The RMSE should consequently be more significant than the CORR, which mainly analyses analogue signal courses. At this point we have run a large amount of parameter combinations and the following results with indications for the currently developed system have been found:

	VALENCE		**AROUSAL**	
	Decay Speed	**Weight**	**Decay Speed**	**Weight**
Voice	-	-	*medium*	*low*
Face	*fast*	*medium*	*fast*	*medium*

Table 3: Affective events are generally discarded rather fast, with cues from the voice having a longer lasting effect - pointing to slower changes in the vocal modality with longer influence on the final result. No modality is weighted as 'high', most probable due to non-optimal accuracy of both modalities.

The *valence* score (for now) consists of events from the facial modality only. As described earlier, we do not receive affective hints with positive or negative emotional content from the vocal modality. The *arousal* score receives events from both modalities. In both cases, the fusion needs to react quickly to incoming events. Affective cues from both modalities are discarded rather fast, with cues from the voice having a longer lasting effect - pointing to slower changes in the vocal modality with longer (though leighter weighted) influence on the final result. The fact that no modality is weighted as 'high' may relate to the former unimodal evaluations showing non-optimal RMSE scores for both modalities. A more sophisticated determination of the fusion parameters for the social sensing system will be carried out, once additional modalities (e.g. body or gaze behaviour) become available together with annotated recordings of users within PRESENT use cases.

5.3 Context Integration

Asynchronous fusion on event level has proven to be robust in affect recognition scenarios (Lingenfelser, 2016) and provides an abstraction level that allows to build a highly adaptable recognition system, as modalities that contain information about a sought target class and provide events for the fusion algorithm can (from a technical point of view) be easily added or removed.

Additionally, the fusion module is not a one-way street. There is not only the possibility to add input (i.e events) to the algorithm from the sensor side, but also to integrate feedback and context from the application side into the fusion process. Such context information from the system may include the general tone of the application (e.g. dark, serious, entertaining, funny, etc.) or shifts thereof. The behavior and goals of the agent may influence the user's affective state as would the outcome (success or failure) in a competitive scenario.

Technically speaking, the event concept allows for feedback to the fusion algorithm as long as the amplitude of influence is expressed in values the fusion algorithm is designed to handle - in our case *valence* and *arousal* values. This approach is deliberately designed in a very generalistic way and is meant to give application and use-case designers an uncomplicated way to provide information whenever they see fit.

6 SOCIAL SENSING SYSTEM

We have by now defined all uni-modal affect recognition systems that have been implemented for the audiovisual social sensing platform as well as an extensible fusion algorithm for the asynchronous and event driven processing of affective events from any desired sources. The following chapter will show a fully implemented social sensing system and its structure, the parallel processes connected to it as well as its integration to the game engine in which the main PRESENT applications are seated.

6.1 System Overview

Figure 8 shows the distribution and connection of services within the social sensing system and other relevant parts of the PRESENT architecture. First off, the components shown with the D4.5 icon mark modules that are part of the current deliverable. Input is currently designed to only include audio and video sensors, so that easy setups with a webcam as only hardware requirement is possible. From here respective signals are directed to the processing components.

The audio signal is analysed by voice activity detection (Section 4.3), voiced segments (i.e. sentences) can be extracted and used for dialogue management. An open-source solution for text-to-speech analysis (Section 4.4.1) is integrated in the system and enables simple scripting with keyword spotting. For a more sophisticated approach, we established an exemplary streaming connection to external management tools (i.e. Google Dialogflow[9] - as described in D4.2). The raw audio stream is also fed into the respective affect recognition module, where events for the fusion approach are generated.

Video images are directly fed into our facial expression component (Section 4.1) where again affective events are generated. Additional possibilities of processing of the video frames include open-source based tracking and analysis of gaze behaviour and body language (Sections 4.4.2 and 4.4.3).

The sensing system is very closely related to two components presented in D4.4 - *Non-verbal Agent Behaviour Enabling*: We implemented an emotion simulation component, that continuously calculates fitting *valence* and *arousal* scores for the virtual agent in interaction with the user. These can be interpreted as a suggested reactive behaviour which can e.g. be used in non-scripted or idle states of the avatar. Reaction and empathy to user emotions and feedback is a crucial part to this simulation, so there are several connections depicted within Figure (reference) that link the sensing system with the simulation component. Second is the socially aware reinforcement learning approach (also presented in D4.4), which takes affective user reactions as input for adaptation processes of the avatar behaviour. Consequently this component also has a strong information exchange with the social sensing system.

[9] https://cloud.google.com/dialogflow

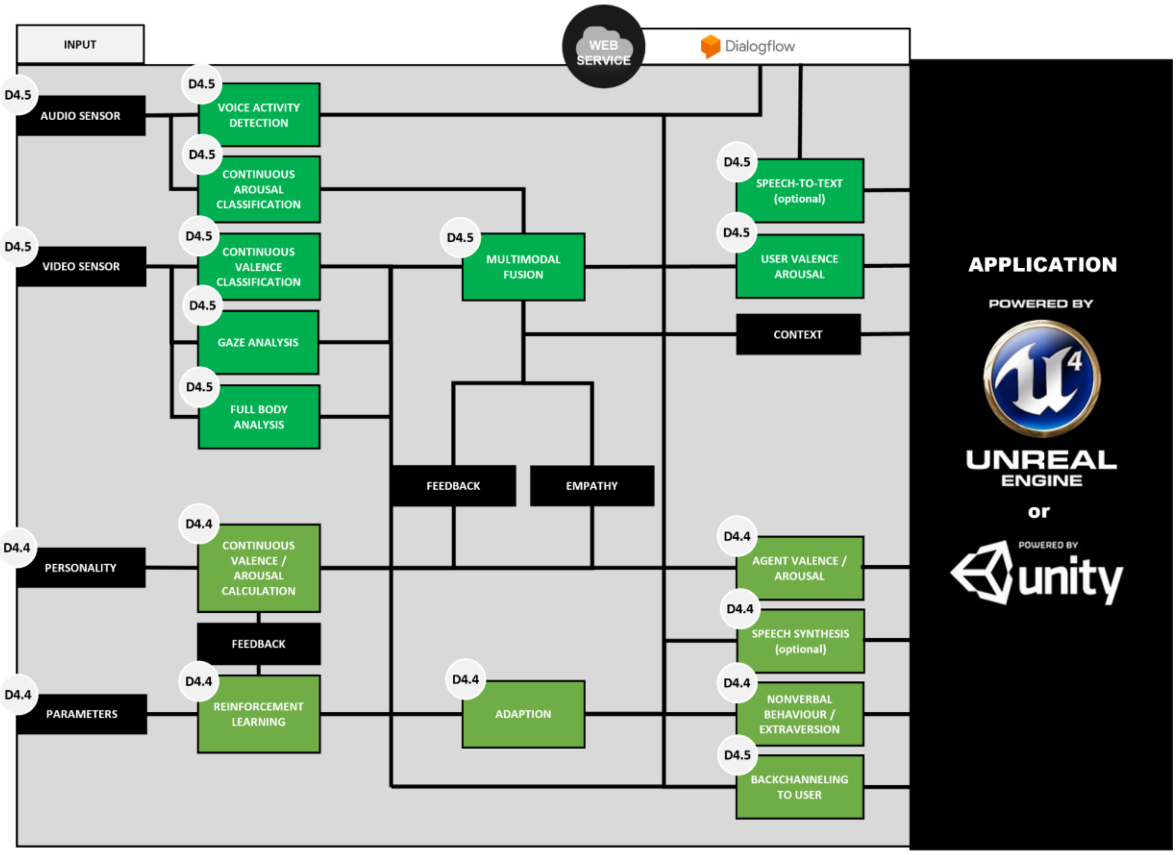


Figure 8: Schematic overview of the PRESENT social sensing system, its logical and practical connections to agent emotion simulation and socially aware adaption systems. All relevant information is streamed to an always accessible and continuously updated interface in the game engine.

6.2 Distributed System Design

On the technical side, we aim for a distributed system design. The physical separation of the sensing system and the PRESENT application is a necessary demand, as the targeted visual quality of the final high-resolution avatar is not combinable on a single hardware instance with a sensing pipeline running amongst others continuous video processing processes. The raw sensor data can also be possibly streamed from hardware devices to the sensing system, creating a separation layer between hardware and software which supports several physical setups. Lastly, we implemented new framework components that enable us to even outsource single processes within the sensing system itself to further physical or virtual environments. This is not only meaningful to distribute processing workload over several machines if needed, it also helps with conflicting system configurations needed for single components and their external dependencies.

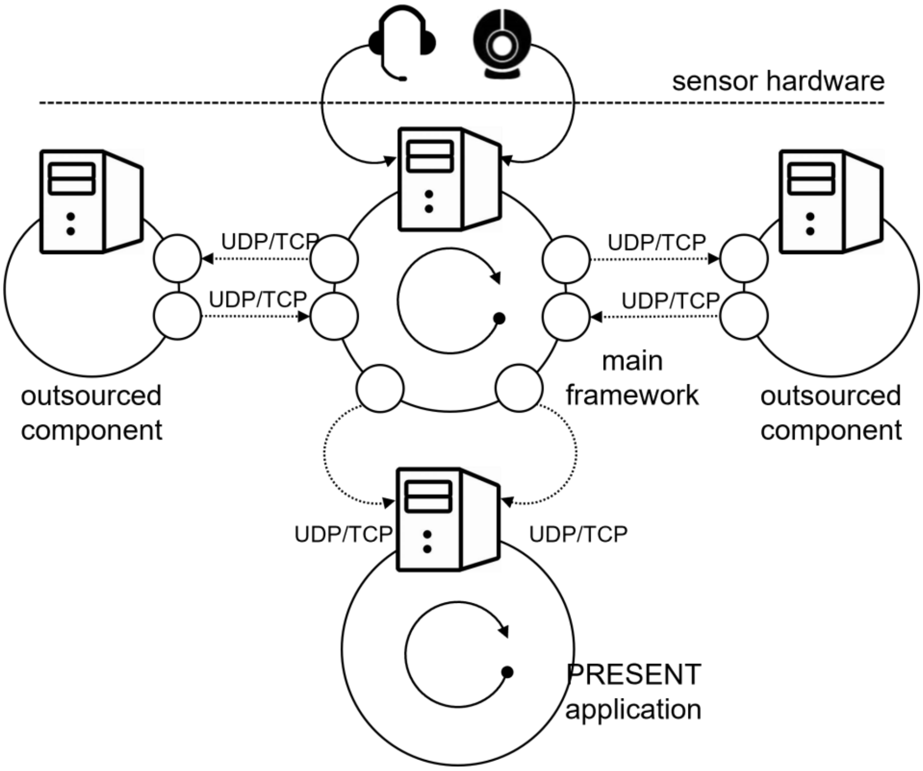


Figure 9: The dimensional valence-arousal value model spans an emotional coordinate system with two axes. Discrete emotion categories can be derived from relative positions within the dimensional model.

6.3 Game Engine Integration

The first demonstrator[10] was implemented without an interface to the game engine and was based on arousal scores from the vocal modality only. It demonstrated the design of a dialogue with the provided tools and accompanying information about affective user behaviour. Since then, the full audiovisual sensing system has been fully integrated with the Unreal Game Engine and the first agent prototypes respectively.

[10] The First Emotion Sensing Pipeline for PRESENT is working

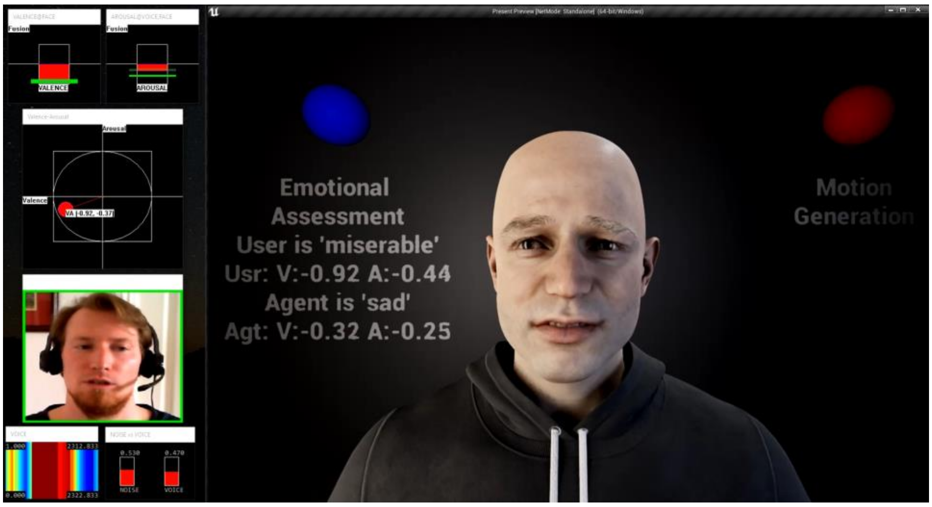
Figure: Our first prototype showed the capabilities of an affective dialogue without integration of the PRESENT agent.

In order to use the analysis insights and results on application side, we defined a continuously updated socket interface, so that the most contemporary information concerning the fused *valence* and *arousal* scores of the user as well as the latest result of the built-in text-to-speech module is always accessible to the application. The same interface is used to communicate context information of the application back to the sensing system. Outputs of D4.4 - *Non-verbal Agent Behaviour Enabling* systems (e.g. simulated agent emotion, adaption suggestions, etc.) are also handled by this interface. To further ease partners' work with the interface, we also included a mapping function, that derives discrete emotion categories from relative positions within the dimensional *valence* and *arousal* model (reference).

Valence very low Arousal very high "*angry*"	Valence low Arousal very high "*afraid*"	Valence neutral Arousal very high "*alarmed*"	Valence high Arousal very high "*excited*"	Valence very high Arousal very high "*happy*"
Valence very low Arousal high "*frustrated*"	Valence low Arousal high "*annoyed*"	Valence neutral Arousal high "*tense*"	Valence high Arousal high "*amused*"	Valence very high Arousal high "*pleased*"
Valence very low Arousal neutral "*distressed*"	Valence low Arousal neutral "*sad*"	Valence neutral Arousal neutral "*neutral*"	Valence high Arousal neutral "*content*"	Valence very high Arousal neutral "*satisfied*"
Valence very low Arousal low "*miserable*"	Valence low Arousal low "*nervous*"	Valence neutral Arousal low "*anxious*"	Valence high Arousal low "*calm*"	Valence very high Arousal low "*serene*"
Valence very low Arousal very low "*depressed*"	Valence low Arousal very low "*worried*"	Valence neutral Arousal very low "*hesitant*"	Valence high Arousal very low "*at ease*"	Valence very high Arousal very low "*relaxed*"

Figure 10: The dimensional valence-arousal value model spans an emotional coordinate system with two axes. Discrete emotion categories can be derived from relative positions within the dimensional model.

Figure 11 shows a demo (presented at the second review) with full system integration and a user emotionally expressing to the PRESENT mid-resolution agent within the Unreal engine.



Emotional Assessment
User is 'miserable'
Usr: V:-0.92 A:-0.44
Agent is 'sad'
Agt: V:-0.32 A:-0.25

Figure 11: The second prototype features the fully integrated audiovisual sensing system. Apart from emotional analysis of the user, the interface provides suggestions of the emotion simulation and socially aware adaption systems to the Unreal engine and therefore any PRESENT use case application.

7 CONCLUSION

In this document, we gave an overview concerning the current state of the PRESENT social sensing system. We have described and evaluated in detail our custom deep transfer and multi-task learning based models for voice activity detection, *arousal* classification based on live audio input and *valence* and *arousal* recognition from video streams. We have furthermore presented the multi-modal fusion system capable of weighting asynchronous recognition based on evaluation performances of recognition modules for all available modalities. Optional parameters have been explored and tuned to the affect recognition problem at hand.

While the first demonstrator was only able to show the capability of the purely audio-based social sensing system to drive an emotional reactive conversation with an agent, we have by now demonstrated the PRESENT sensing system fully integrated in the system architecture. It is able to cover the whole emotional range of a user based on pure audiovisual input. Additionally, we introduced software based solutions to kickstart full body and gaze analysis if needed, without further addition of hardware. Though not an actual PRESENT use case and only based on the arousal dimension, the scenario gives a good impression of the affective loop between user and agent in the sense that emotional reactions trigger behaviours of the agent and vice versa, resulting in a more complex and more emotionally rich interaction.

Some concepts, such as, e.g., the inclusion of context in the affective assessment process are defined and implemented as general as possible. The reason is the range of developed use case applications which are still to be finalized. We want the sensing system to be as accessible and configurable as possible for other partners, therefore we try to avoid restrictive definitions. As the social sensing system is a crucial precondition for subsequent tasks such as dialogue management, agent behaviour or agent adaption, the audiovisual core system is by now finalized but can be modified (e.g., introduced additional modality possibilities such as gaze analysis and body tracking) whenever the development of case applications demands.

8 REFERENCES

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin et al. "Tensorflow: A system for large-scale machine learning." *12th USENIX symposium on operating systems design and implementation*, 2016: 265-283.

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." *Advances in neural information processing systems*, 2016: 892-900.

Baltrusaitis, Tadas, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. "Openface 2.0: Facial behavior analysis toolkit." *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, 2018: 59-66.

Baltrušaitis, Tadas, Marwa Mahmoud, and Peter Robinson. "Cross-dataset learning and person-specific normalisation for automatic Action Unit detection." *Facial Expression Recognition and Analysis Challenge, IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

Baur, Schiller, and André, "Modeling users social attitude in a conversational system," in Emotions and Personality in Personalized Services. Springer, 2016,pp. 181–199.

Caridakis, Wagner, Raouzaiou, Curto, André, and Karpouzis, "A multimodal corpus for gesture expressivity analysis," in International Conference on Language Resources and Evaluation (LREC), Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, Malta, 2010.

Hannun, Awni, et al. "Deep speech: Scaling up end-to-end speech recognition." *arXiv preprint arXiv:1412.5567* (2014).

Deng, Dong, Socher, Li, Li, Fei-Fei: *ImageNet: A large-scale hierarchical image database*. In: *CVPR 2009 - IEEE Conference on Computer Vision and Pattern Recognition, 2009*. 2009, S. 248–255.

Lassalle, Amandine, Delia Pigat, Helen O'Reilly, Steve Berggen, Shimrit Fridenson-Hayo, Shahar Tal, Sigrid Elfström et al. "The EU-emotion voice database." *Behavior research methods*, 2019: 493-506.

Lingenfelser, Florian, Johannes Wagner, Jun Deng, Raymond Brueckner, Björn Schuller, and Elisabeth André. "Asynchronous and event-based fusion systems for affect recognition on naturalistic data in comparison to conventional approaches." *IEEE Transactions on Affective Computing*, 2016: 410-423.

Mckeown, Gary & Valstar, Michel & Cowie, Roddy & Pantic, Maja & Schroder, M.. (2013). The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. Affective Computing, IEEE Transactions on. 3. 5-17. 10.1109/T-AFFC.2011.20.

Russell, J. A. "A circumplex model of affect." *Journal of Personality and Social Psychology*, 1980: 1161-1178.

Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov and Liang-Chieh Chen. "MobileNetV2: Inverted Residuals and Linear Bottlenecks" The IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520.

Vogt, Thurid and André, Elisabeth and Bee, Nikolaus. “EmoVoice—A framework for online recognition of emotions from voice.” *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, 2008: 188-199.

Wagner, Johannes and Lingenfelser, Florian and Baur, Tobias and Damian, Ionut and Kistler, Felix and André, Elisabeth. “The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time.” *Proceedings of the 21st ACM international conference on Multimedia (MM ’13)*, 2013: 831-834.

Wagner, Johannes and Schiller, Dominik and Seiderer, Andreas and André, Elisabeth. “Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?” *Proceedings of Interspeech*, 2018: 147-151.