D4.2 Interim Report on Agent Social Interpretation Enabling

Grant Agreement nr	856879
Project acronym	PRESENT
Project start date (duration)	January 1st 2018 (36 months)
Document due:	31/08/2020
Actual delivery date	31/08/2020
Leader	University of Augsburg
Reply to	florian.lingenfelser@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	**P**hotoreal **RE**altime **S**entient **ENT**ity
Document name	Interim Report on Agent Social Interpretation Enabling
Security (distribution level)	Public
Contractual date of delivery	31/08/2020
Actual date of delivery	31/08/2020
Deliverable name	D4.2 Interim Report on Agent Social Interpretation Enabling
Type	Report
Status & version	Final Version
Number of pages	26
WP / Task responsible	WP4 / University of Augsburg
Other contributors	-
Author(s)	Florian Lingenfelser Elisabeth André
EC Project Officer	Ms. Adelina Cornelia DINU - Adelina-Cornelia.DINU@ec.europa.eu
Abstract	The PRESENT agent is challenged with a wide variety of tasks to make interaction with a user more natural and engaging. To name a few of these tasks, they include communication abilities that cater to user emotions within dialogues, adapting to user preferences in behaviour and voice based on user reactions or adapting its inner emotional simulation based on empathy to the user's emotional state. All these capabilities are enabled by a steady assessment of the users current emotional reactions. Within this document we describe the current state of the PRESENT real-time social sensing system in detail, present a first demonstrator for affective dialogue management and introduce the upcoming extensions of the sensing system.
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
v1.0 – 22/07/2020	Document created by Florian Lingenfelser
v2.0 – 14/07/2020	Version for internal review (14 days before submission date)
v3.0 – 31/08/2020	Revisions in response to review: final version submitted to Commission

Table of Contents

1 EXECUTIVE SUMMARY 5
2 BACKGROUND 5
3 INTRODUCTION 6
4 EMOTION THEORY 7
4.1 Categorical Emotion Model 7
4.2 Dimensional Emotion Model 8
5 PRESENT SOCIAL SENSING SYSTEM 9
5.1 Framework Description (SSI) 9
5.2 Modality Processing 10
5.2.1 Voice Activity Detection 11
5.2.2 Paralinguistic Analysis 12
5.2.3 Next Step: Facial Analysis 15
5.3 Multi-modal Fusion 17
6 FIRST DEMONSTRATOR 19
6.1 Demonstrated Concepts 19
6.2 Pipeline (XML) Description 21
7 CONCLUSION 26
8 REFERENCES 27

1 EXECUTIVE SUMMARY

The PRESENT agent is challenged with a wide variety of tasks to make interaction with a user more natural and engaging. To name a few of these tasks, they include communication abilities that cater to user emotions within dialogues, adapting to user preferences in behaviour and voice based on user reactions or finally, to adapt its inner emotional simulation based on empathy to the user's emotional state. All these capabilities are enabled by a steady assessment of the user's current emotional reactions.

Within this document we describe the current state of the PRESENT real-time social sensing system in detail, present a first demonstrator for affective dialogue management and introduce the upcoming extensions of the sensing system.

2 BACKGROUND

The deliverable at hand reports first advances made in WP4 towards enabling the PRESENT agent to interpret social signals of the user.

The document is a year one, second quarter deliverable, and its main function is to describe the first prototype of the agent's social sensing system. We describe the implementation of a multi-modal affect recognition system, which is able to assess social signals of the user in real-time.

Results of the presented recognition system are crucial preconditions for pursuing tasks within and beyond WP4.

3 INTRODUCTION

The PRESENT project aims at delivering an agent, which is capable of natural and believable interaction with a user as well as adaption to the user's preferences. More in detail, dialogues with an agent will be steered by considering the affective states of a user. The behaviour shown within interaction will be adapted based on implicit feedback shown by the user. Moreover, the simulated inner emotional status of the agent is among other aspects influenced by the perceived emotion of its human counterpart. In order to enable such emotional intelligence and behavioural traits, the agent needs to be capable of interpreting a user's social signals and affective states.

To this end, we realize the PRESENT social sensing system, which implements various modules for activity detection and emotion recognition from voice, face and body modalities. All components are embedded within a multi-modal framework, which guarantees synchronized sensor input and data processing in real-time. We present an XML based pipeline design, that makes it possible to reuse and rearrange all available components to quickly design custom recognition systems for the varying use cases of the PRESENT project.

The following chapters are structured as follows:

Chapter 4 describes the foundational emotion theory behind recognition approaches with respect to targeted classes and emotional dimensions to be assessed by the system, used for simulated agent emotion and published to subsequent agent modules such as dialogue management or body and face animation. In Chapter 5 we follow up with 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. Chapter 6 on the one hand shows the emotion recognition system in an application that reacts to recognized user states, on the other hand gives a detailed instruction how to design the respective processing pipeline with simple XML templates. Chapter 7 concludes the document.

4 EMOTION THEORY

Prior to the implementation of an affect recognition system, there needs to be a decision on the internal representation of the concept of user affect. Two main paradigms have found broad acceptance in literature:

- In a categorical model, expected affective states are defined as discrete classes with descriptive labels (such as *happy*, *sad*, *angry*, etc.).
- A dimensional model locates affective states within a continuous space where dimensions represent different psychological concepts (Lang 1997). The most relevant dimensions to represent affect using a dimensional model are *valence* and *arousal*.

Categorical as well as dimensional models are simplified and synthetic descriptions of human affect and are not able to cover all of the included aspects. They are however useful and needed to model emotions as concepts to be presented to a machine.

Both concepts are of course interrelated and can to a certain degree be interchanged, as a certain emotional state can be described in both models (Figure 1). This gives partners the ability to choose which representation of the current emotional user state better fits their tasks.

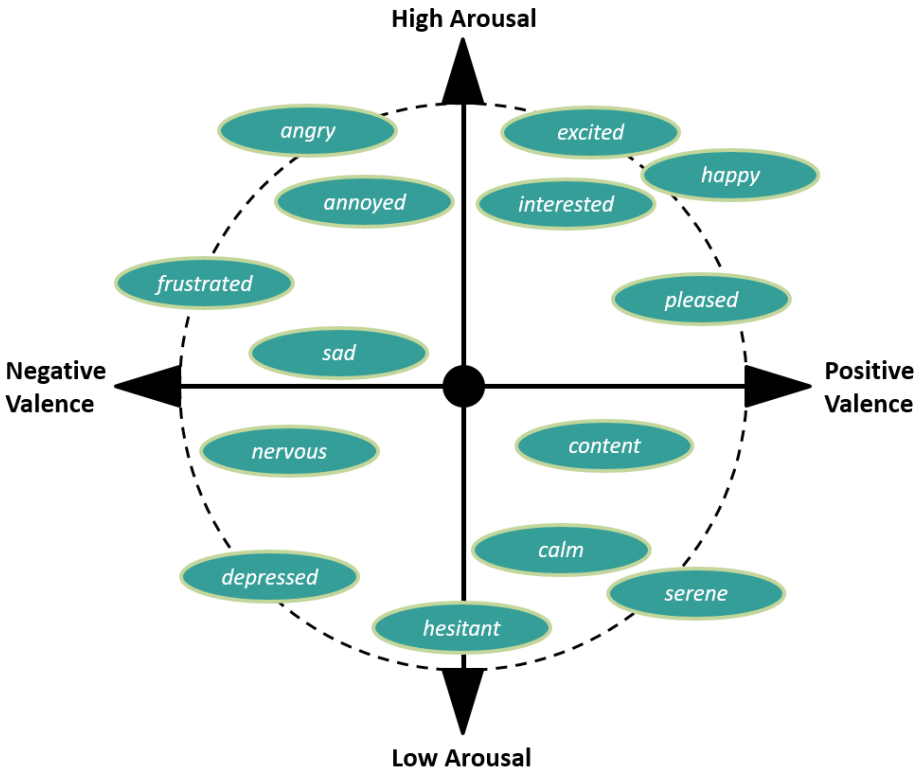

Figure 1: Interrelation of categorical emotion classes with the valence-arousal dimensional model.

4.1 Categorical Emotion Model

At first glance, categorization of emotions into a finite set of classes seems a reasonable approach. We are very used to the common naming conventions for certain affective states such as *anger* or *joy* and have general conception of the meaning of these labels. However, the category-based paradigm is limited for two main reasons:

Firstly, affective states in every-day life are too complex to be well represented by a limited number of discrete categories. Unfortunately, augmenting the number of possible labels complicates the annotation process while creating a database and increases the complexity of the recognition system that will classify them.

Furthermore, a category based model is by default not able to blend between labels. Recognition classes are considered to be independent and there is no definition or notion for distance measures between them. Both mentioned drawbacks can however be relieved by the dimensional approach.

4.2 Dimensional Emotion Model

Russell's valence-arousal scale is a widely used concept in research on affect to quantitatively describe emotions (Russell 1980). In this scale, each emotional state can be projected on a two-dimensional plane with arousal and valence as the vertical and horizontal axes respectively.

Hereby, valence refers to how pleasant or unpleasant is an affective state while arousal indicates the activation or deactivation level (Russell 1980). This approach features several advantages over a categorical model:

- In contrast to discrete classes, blended emotions can naturally be represented within a dimensional model, as affective states share a common set of continuous dimensions.

- If an application requires labels for emotions (e.g. for visualisation purposes), a simple mapping from valence / arousal values to discrete classes is possible (Figure 2). The mapping is however not loss-free and therefore only unambiguous in this direction, as exact values cannot be inferred from emotional labels.

- Not all considered modalities (e.g. facial expressions, vocal characteristics, body language) are able to express the whole range of an emotional category, but can well characterize certain aspects of an emotional state. For example, a high pitch in the voice could be related to either a happy or an angry state, however, it is in any case pointing to a highly activated and aroused condition.

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 2: Example for a mapping from emotional dimensions to categorical classes based on predefined intervals on the valence and arousal axes.

5 PRESENT SOCIAL SENSING SYSTEM

Automatic detection and interpretation of social signals can be carried by voice, gestures, mimics, etc. Meaningful insights in the current emotional user state plays a key-role in the PRESENT project. Given the highly naturalistic requirements and standards of the targeted PRESENT applications, we require these recognition tasks to be performed close to real time, which of course is a non-trivial task.

To tackle this challenge, we utilize the versatile Social Signal Interpretation framework (SSI) (Wagner 2013). Besides its suitability for recording and integration purposes, SSI also offers helpful tools for analysing and interpreting any given input from multiple sensors in real time. The following chapter describes those capabilities in detail first. Afterwards the extensions which are currently developed to adapt the framework to the specific needs of the project are explained.

5.1 Framework Description (SSI)

The Social Signal Interpretation (SSI) framework[1] offers tools to record, analyse and recognize human behaviour in real-time, such as gestures, mimics, head nods, and emotional speech. Following a patch-based design pipelines are set up from autonomic components and allow the parallel and synchronized processing of sensor data from multiple input devices.

[1] https://hcm-lab.de/projects/ssi/

In particular, SSI supports the machine learning pipeline in its full length and offers a graphical interface that assists a user to collect own training corpora and obtain personalized models. In addition to a large set of built-in components SSI also encourages developers to extend available tools with new functions. For inexperienced users an easy-to-use XML editor is available to draft and run pipelines without special programming skills. SSI is written in C++ and optimized to run on computer systems with multiple CPUs. Binaries and source code are freely available under GPL.

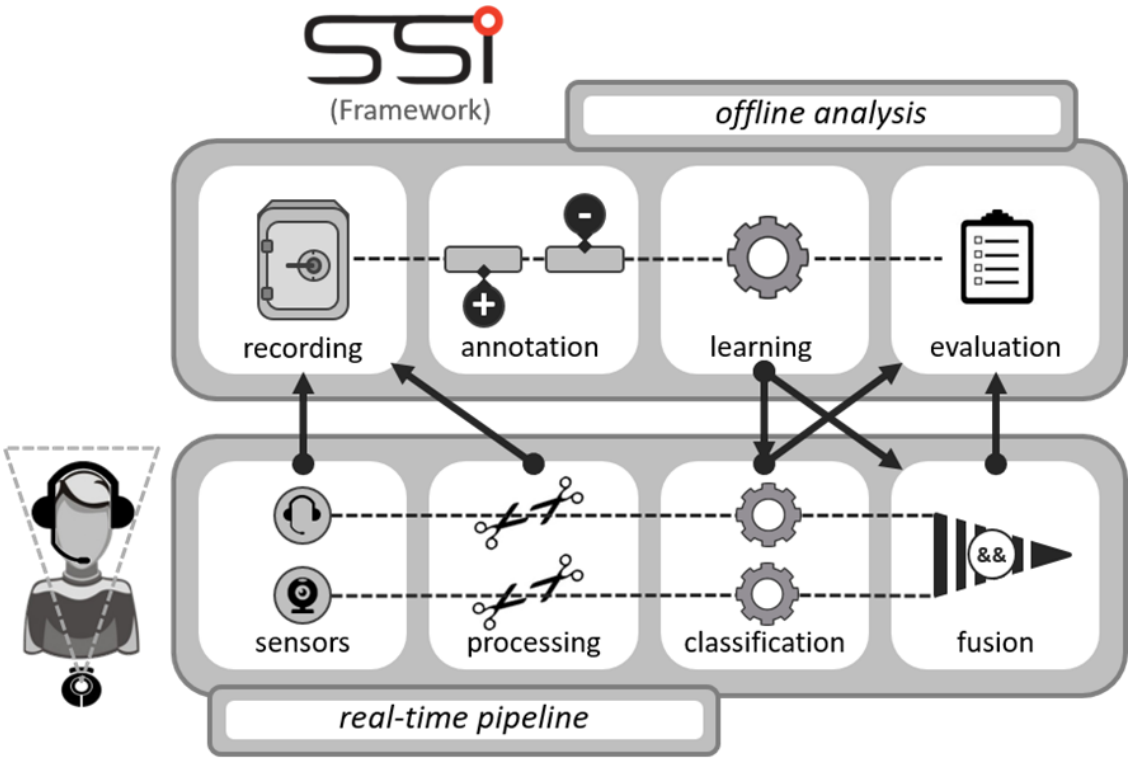

Figure 3: The SSI framework supports the real-time machine learning pipeline in its full length, from synchronised sensor input to multi-modal fusion of classification results. It additionally offers tools that assists a user to collect own training corpora and train custom models.

5.2 Modality Processing

In the following sections we will describe the actual recognition modules that have been or are currently implemented within the Social Signal Interpretation framework in order to realize the PRESENT social sensing system. This includes a sophisticated voice activity detection component to guarantee proper dialog management, emotion recognition components for multiple modalities as well as a fusion component for a coherent integration of results coming from the multi-modal classification models.

5.2.1 Voice Activity Detection

The task of voice activity detection is a crucial step for any agent system that aims to interact with the user via a dialogue system. Having correct recognition of spoken sentences is not only the basis for dialogue management, it also helps the emotion recognition module (Section 5.2.2) to identify signal parts that may carry meaningful information.

Simplistic approaches to voice activity detection, such as looking at audio signal's energy levels or signal to noise ratio may suffice in a very controlled setting (few background noise or distracting sounds, close talk microphone, etc.). However, in a more natural environment a reliable identification of voiced segments requires more effort.

In the PRESENT sensing system we follow a machine learning based approach by classifying voice in the audio signal with the deep learning model VadNet[2] (Wagner 2018). The model is trained using publicly available videos from television media libraries, featuring high quality subtitles with timestamps (
Figure 4). These videos offer a huge amount of training data in which we are able to automatically create annotations for voiced audio parts against segments containing silence or various kinds of sounds we want to identify as background noise.

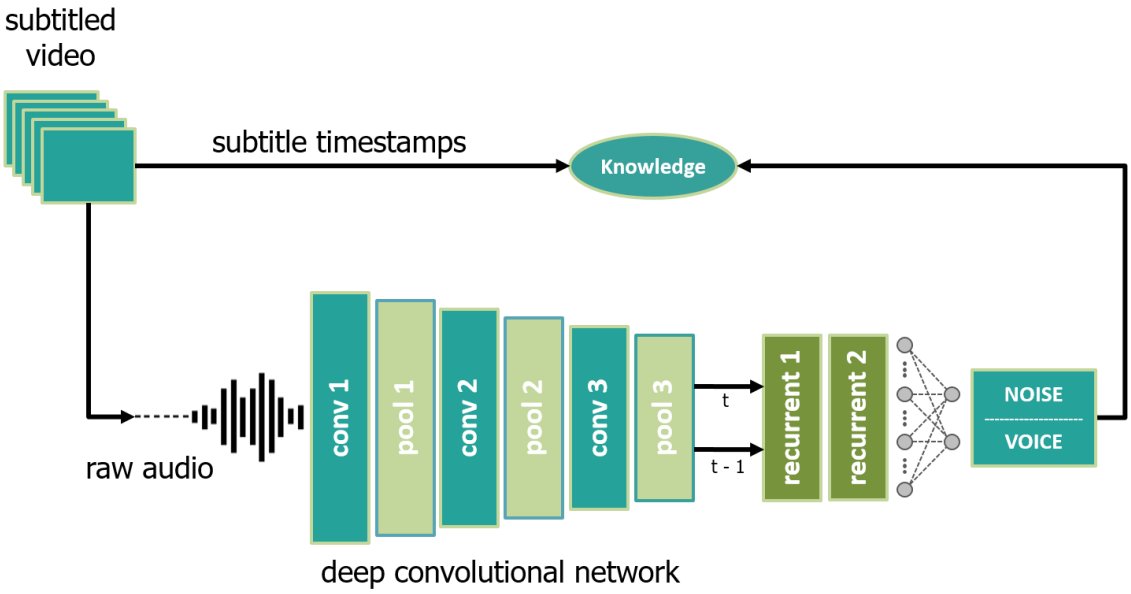

Figure 4: The VadNet model has been trained with Tensorflow (Abadi 2016). It takes as input the raw audio input and feeds it into a 3-layer convolutional network. The result of this filter operation is then processed by a 2-layer Recurrent Network containing 64 RLU cells. The final bit is a fully-connected layer, which applies a softmax and maps the input to a tuple <noise, voice> in the range [0..1].

[2] https://github.com/hcmlab/vadnet

Output of the classification model is smoothed over consecutive analysis frames and results in a continuous estimation of NOISE versus VOICE classes (Figure 5). By tracking this assessment, we are able detect the on- and offset of spoken sentences in the audio channel.

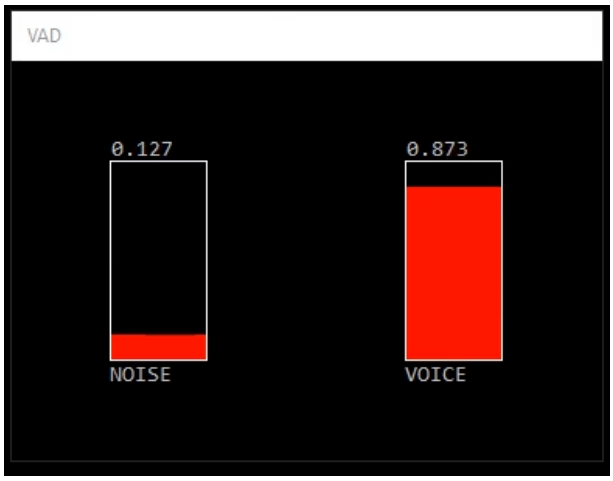

Figure 5: Continuous assessment of NOISE versus VOICE of the VadNet model.

In our first demonstrator (Chapter 6), we use this information to extract the spoken sentences from the raw audio signal and transmit respective segments as dialogue acts to the external dialogue management tool. Here, speech recognition is carried out and corresponding answers are created.

5.2.2 Paralinguistic Analysis

In parallel to voice activity detection, we analyse the raw audio signal for emotional content. As described in Section 4.2 we focus on the dimensional valence-arousal model for representation of user emotions. Intensity in the voice modality is suspected to be highly correlated with the emotional arousal of a person (Lassalle 2019) and is therefore well suited for estimating the arousal dimension: Paralinguistic features such as loudness, pitch or speech rate carry descriptive information for determining user arousal. Traditional machine learning models would make use of these characteristics by applying a feature extraction step (through the appliance of an audio analysis module such as EmoVoice[3] (Vogt 2008) to an extracted spoken sentence, followed by automatic classification of the resulting feature vector.

[3] https://github.com/hcmlab/emovoice

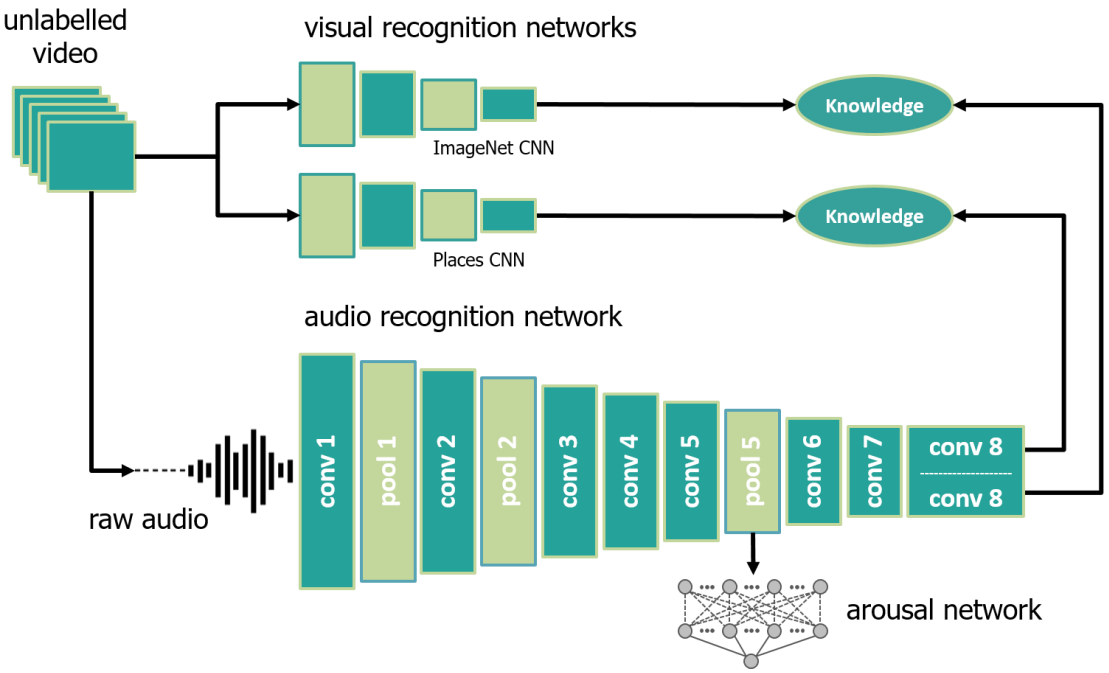

Figure 6: SoundNet (Aytar 2016) is a deep convolutional network for natural sound recognition. The network is trained by transferring discriminative knowledge from visual recognition networks into sound networks. The approach capitalizes on the synchronization of vision and sound in video.

The feature based classification method is however slightly limited in that feature calculation works most reliable when applied to longer sequences of audio data (i.e. whole sentences). Our goal is a more flexible and reactive analysis based on much shorter frames of the signal. Neural networks have the ability to bypass the feature calculation step an instead take chunks of the raw audio signal as input. These segments can be very short (i.e. 50 milliseconds) and enable classification at a very high sample rate – resulting in a smooth and reactive assessment of the current arousal level.

Figure 6 shows the architecture of the applied deep learning model for continuous arousal classification from audio. As the availability of annotated audio data with continuous arousal scores is sparse, we use a pre-trained convolutional network for the pre-processing of audio data. SoundNet[4] (Aytar 2016) exploits the leading edge of computer vision algorithms - which are in general very accurate in classifying objects and scenes in video data - to generate labelled data for audio classification. SoundNet is trained to recognise scenes based on present sounds in the audio channel.

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

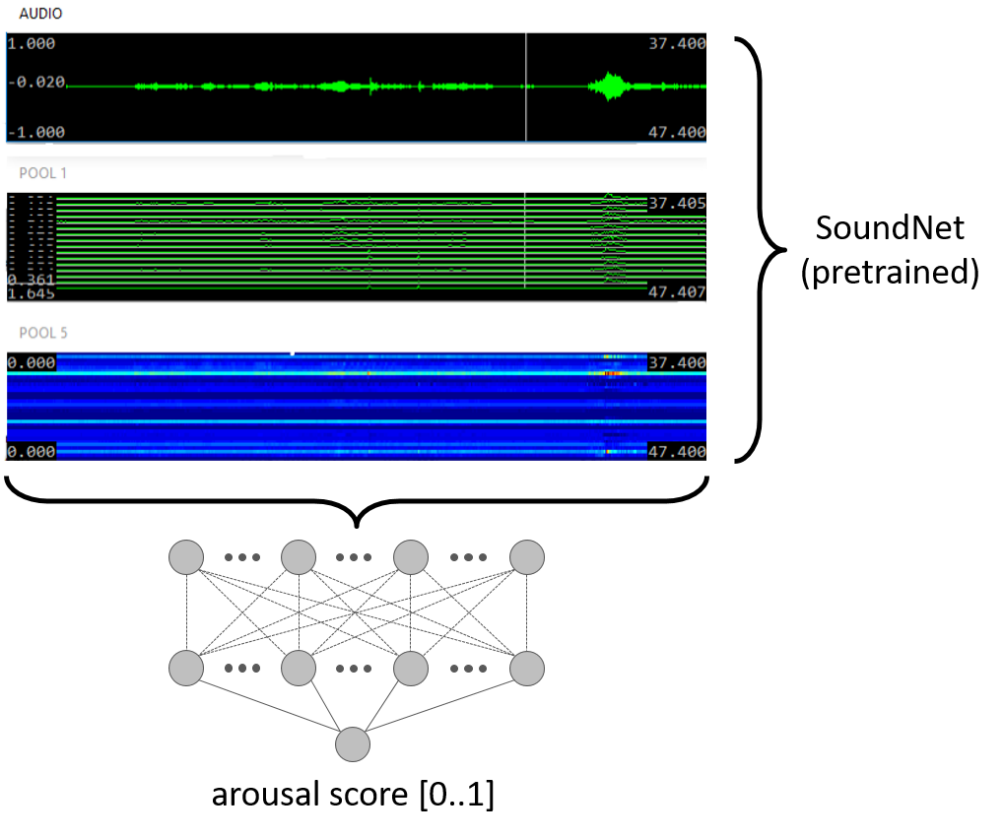

Figure 7: We use the condensed representation of the pre-trained deep convolutional SoundNet model as input for a small, fully connected network. This network sits on top of the preceding SoundNet layers and is trained to recognise continuous arousal scores.

Though scene classification is not the target of our emotional analysis, we found that the last pooling layer of the trained model has learned a compact and focused representation of the audio signal, apparently emphasizing characteristics of the initial input data that are well suited for arousal classification. Hence we feed the output of the last pooling layer (pool 5) of the SoundNet architecture into a small network featuring two fully connected layers that result in the final arousal score classification in the range [0..1].

The final arousal network is trained on an excerpt of the SEMAINE[5] database, which is one of the few sources to obtain high quality arousal annotations for audio data. Though the size of this database is probably insufficient to train a deep convolutional framework, we circumvent this problem by using the condensed representation of the pre-trained SoundNet and only train a small network on top (Figure 7).

[5] https://semaine-db.eu/

5.2.3 Next Step: Facial Analysis

The first demonstrator shown in Chapter 6 works with the arousal dimension of the user interacting with the system, based on paralinguistic analysis only. However, the goal of the PRESENT sensing system is to assess the whole valence-arousal space for emotional evaluation of the user. The facial modality offers a very robust source of information on emotions expressed by users. Automatic facial behaviour analysis classically focuses on the recognition of universal facial expressions and ActionUnits (AU) - atomic movements in the face caused by the activation of one or more facial muscles. The research is mainly motivated by well-known studies of the psychologists Ekman and Friesen (Ekman 1971). Any possible facial expression can be inferred from combinations of Action Units. Hence, we are currently implementing a facial analysis module to include in the emotion recognition system.

There is a wide array of software (partly open-source) for the automated analysis of facial expressions, which provide easy access to face tracking, facial landmark estimation and action unit assessment. An example of a facial analysis module readily integrated in the SSI framework is Openface[6] (Baltrusaitis 2018) – see FigureFigure 8 for an example of real-time face recognition and landmark tracking on webcam input.

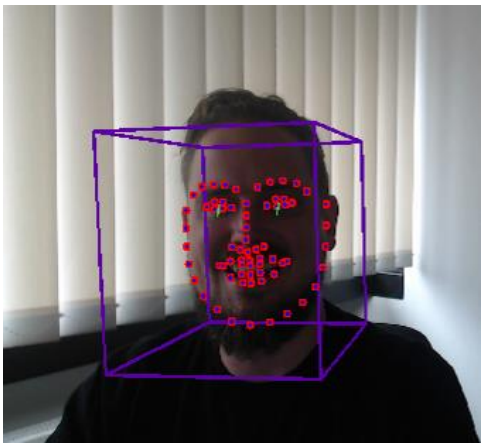

Figure 8: Example of real-time tracking of face position, estimated viewport and facial landmarks using the Openface library as module within the SSI framework.

Figure 9 shows the inferred activation of relevant ActionUnits which can be used as input for a machine learning model to automatically recognize the shown emotion.

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

Figure 9: Real-time calculation of ActionUnits. These features offer the possibility to classify emotions based on the facial expression. AU's such as *LipCornerDepressor*, *CheekRaiser* or *UpperLipRaiser* are especially well suited to detect states of valence.

ActionUnits such as *LipCornerDepressor*, *CheekRaiser* or *UpperLipRaiser* are especially well suited to detect states of valence (i.e. a positive or negative emotional state) within the facial expression of a user and our first approach will consequently focus on valence recognition based on action unit detection. In addition to the already implemented arousal recognition module from vocal properties, our multi-modal recognition system will from this point on be able to cover the whole valence-arousal space.

In addition to feature based (i.e. ActionUnits) approaches we are also looking into deep learning based recognition, which will be able to forego the computational expensive action unit calculation step and take raw frames of video as input. We are planning to satisfy the rising need for training data of a convolutional deep learning approach with publically available databases such as AffectNet[7] (Baltrušaitis 2015).

[7] http://mohammadmahoor.com/affectnet/

Figure 10: 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 (Figure 10). We therefore plan to have a fully trained deep convolutional emotion recognition model within the upcoming months.

5.3 Multi-modal Fusion

The final step needed in our multi-modal emotion recognition pipeline is a fitting multi-modal fusion algorithm. At the time of the upcoming second prototype, we will deal with a recognition module for the arousal dimension based on paralinguistic analysis of the voice and a facial expression analysis component for continuous valence estimation. In the course of the project, this will extend to valence and arousal estimations from described modalities as well as additional input modalities (e.g. body language - **WP4T4**). Consequently, we need a way to incorporate impressions from multiple sources into a coherent final result within the targeted valence-arousal space which can be published to the rest of the system.

While we finally aim for every modality to generate estimations within the whole valence-arousal space, some modalities are more reliable within a specific emotion dimension (e.g. body language is suspected to be more reliable in arousal estimation and is hardly useful for valence recognition, while facial expressions reliably convey valence related assessments). The fusion algorithm should be able to reflect these imbalances, typically through the appliance of weights to respective modality-emotion combinations.

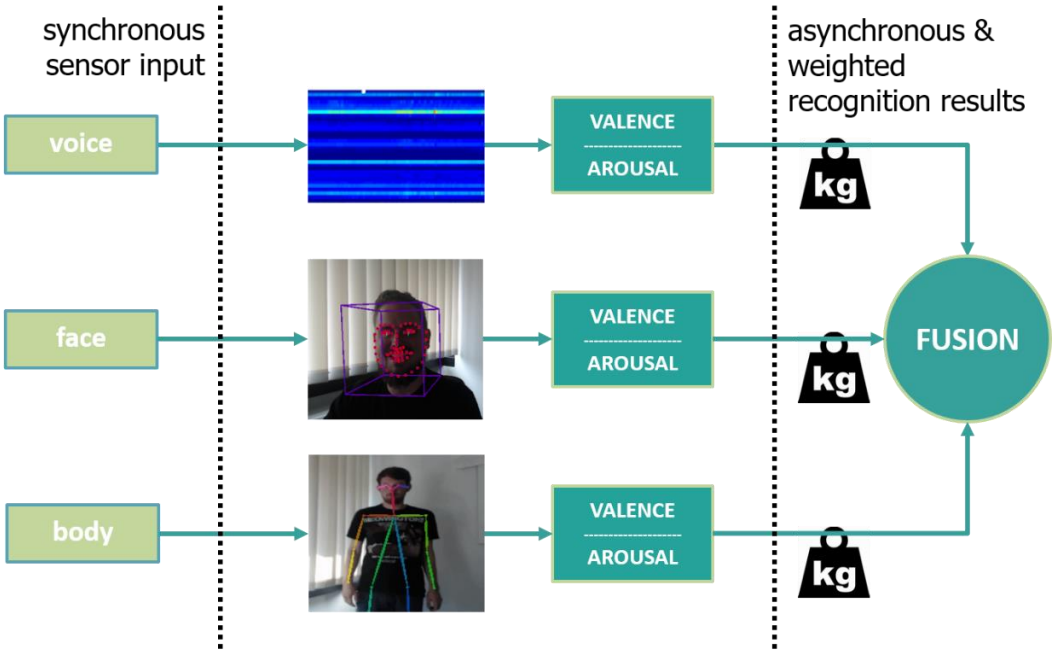

Figure 11: While the SSI framework guarantees synchronization of data streams from multiple sensors, recognition results of considered modalities may appear at differing sample rates and need to be weighted based on their accuracy in valence and arousal recognition.

From 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 (Figure 11). 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 and arousal values), remembers respective time of occurrence and offsets this timing with predefined timespans 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.

We currently use a basic version of the algorithm in the first demonstrator (Chapter 6) to handle arousal estimation from the vocal modality. We will adapt the algorithm in further iterations of the PRESENT social sensing system, i.e. with respect to proper weighting of considered modalities and their accuracy in valence and arousal recognition.

6 FIRST DEMONSTRATOR

The following first prototype[8] is meant to demonstrate the capabilities of the PRESENT agent to interpret a user's social signals and show behavioural sensitivity and responsiveness. The setting is based on a first idea for a use-case that circulated in the consortium at the very beginning of the project. The agent is meant to act as a presenter within a museum environment. While showing the user a specific piece of art, it is able to give general information on the topic as well as to spontaneously react to affective user states during the interaction.

Figure 12: Screenshot of the first demonstrator. Video available on the official YouTube PRESENT channel.

The modern piece of art in question is called *Comedian* and is chosen because of its potential to elicit aroused behaviour - the installation depicts a banana stuck to a wall with a piece of tape. The agent is capable of detecting exceptional positive or negative swings in user arousal and consequently tries to calm the user down with supporting information or encourage him to engage more in the experience of observing the installation.

6.1 Demonstrated Concepts

Several key concepts of the social capabilities of the PRESENT agent as well as their integration with other modules are meant to be demonstrated within this dialogue setting:

[8] https://www.youtube.com/watch?time_continue=2&v=4n7Q75D1AO0&feature=emb_logo

- The consortium has decided to use an external dialogue management tool (in this case Google Dialogflow[9]) to enable the simple creation of dialogues by all partners for use-cases as needed. In this prototype, we show the technical integration of the respective tools and how to communicate emotional user states to other components of the system. (**WP5T3**)

- We created an example dialogue, which is primarily meant to give an idea how to reasonably integrate the emotional state of a user into human-agent interaction. (**WP4T1**)

- From a technical perspective, the demonstrator proves the capability of the first social sensing system prototype to correctly assess relevant user states in real-time during interaction (Figure 13). Please note that the focus on a single emotional dimension (arousal) based on a single modality (audio / voice) is a temporary feature of the first prototype. The extension to the full valence-arousal emotional based on multi-modal recognition is currently being implemented. (**WP4T3**)

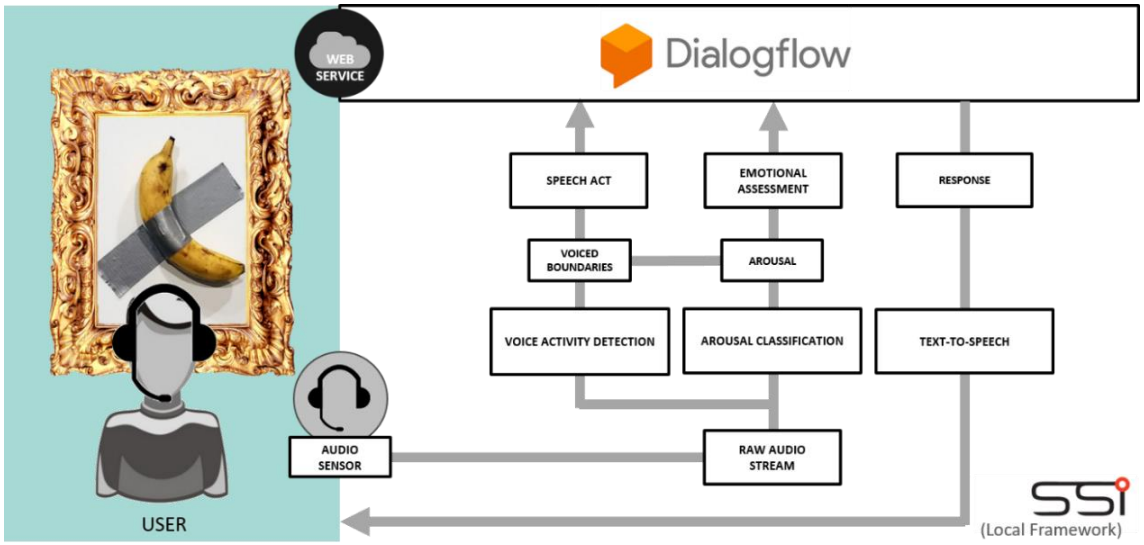

Figure 13: The first demonstrator is only based on the arousal dimension. The scenario gives an impression of the communication between the social sensing system and dialogue management as well as the affective loop between user and agent in the sense that emotional reactions trigger behaviours of the agent and vice versa.

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

6.2 Pipeline (XML) Description

To help users without deeper programming experience get started with affect recognition pipeline construction, our framework offers the possibility to define the processing chain with XML files. At runtime, the XML specifications are translated into pre-complied C++ code. Please note that while this approach is very convenient to design any signal processing pipeline with existing components, there is no way to implement new components with <XML>.

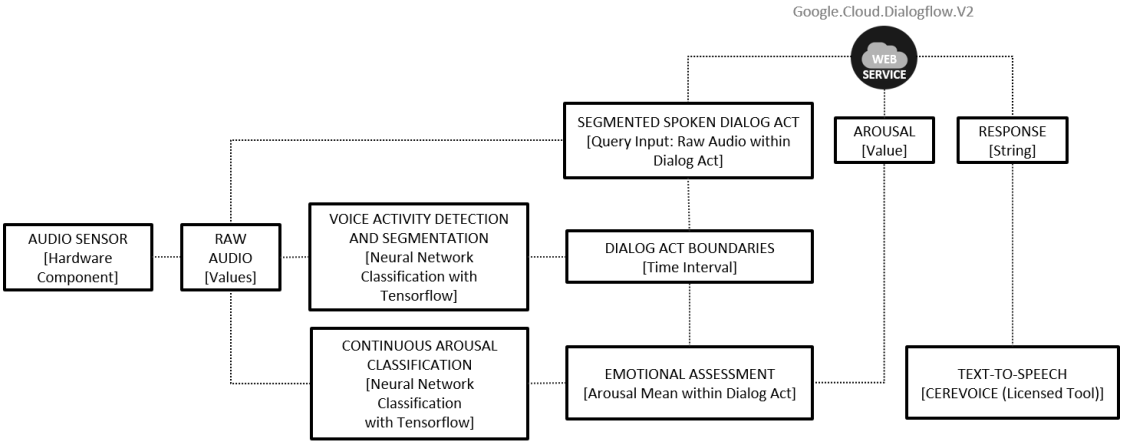

Figure 14: Schematic of the XML pipeline, which defines the processing chain of the first demonstrator.

The (slightly simplified) XML pipeline of the demonstrator (Figure 14) starts with the introduction of an audio sensor. The entry <pin="stream_audio"> defines the data stream which provides the raw audio data to the following processing components in the chain.

```
<sensor create="Audio" option="audio_options" sr="16000>
        <output pin="stream_audio"/>
</sensor>
```

Our framework is designed to be capable of handling input from multiple sensors (e.g. audio, video, physiological, etc.) as well processing of respective data streams in parallel and in synchronicity. Upcoming prototypes will include multiple modalities, for this first example we will however work with audio input only.

Next component in the chain is a feature extraction component in form of a transformer, followed by further transformers processing the preliminary results. The feature extraction component is labelled <PythonFeature> which tells the framework, that the actual extraction component is implemented in the form of the referenced Python script <script="vad.py" >. We included this possibility in addition to precompiled C++ modules in order to make the framework more accessible and well suited for rapid prototyping.

```
<transformer create="PythonFeature" script="vad.py" >
        <input pin="stream_audio" frame="3.0s" delta="1.0s"/>
        <output pin="stream_vad_raw"/>
</transformer>
```

The algorithm applied in this module is the voice activity detection approach described in Section 5.2.1. It is applied on a sliding window of raw audio data of 4 seconds with a 1 second overlap <frame=”3.0s” delta=”1.0s”>. The output of the component is reintroduced to the framework as a data stream <pin="stream_vad_raw">. More in detail, it is a two-dimensional stream, with dimensions holding the respective recognised probabilities for the *noise* or *voice* class.

In order to condense information needed in future steps, we select only the dimension holding the *voice* probabilities and consequently generate a one-dimensional data stream with help of a <Selector>. The statement <frame=”1”> tells the framework to process every consecutive available frame of the stream without timing restrictions.

```
<transformer create="Selector" indices="1">
        <input pin=" stream_vad_raw" frame="1"/>
        <output pin="stream_voice_raw"/>
</transformer>
```

As we treat the probability stream as a signal for further processing, we smooth the included results over time via a *Moving Average* calculation <MvgAvgVar>.

```
<transformer create="MvgAvgVar" format="1" method="2" win="4.0s">
        <input pin="stream_voice_raw" frame="1"/>
        <output pin="stream_voice"/>
</transformer>
```

After this processing, we end up with the stream <stream_voice>, containing the continuous probability for voiced signal parts in an interval [0..1]. We can at this point start the communication with the external dialogue management tool by generating a triggering event via a <TriggerEventSender>.

```
<consumer create="TriggerEventSender" address="voice@vad"
                thresholdIn="0.5" thresholdOut="0.5">
        <input pin="stream_voice" frame= "1.0s"/>
</consumer>
```

Here, we publish an event within the framework, containing the timings whenever the probability for voice in the audio channel crossed 0.5 and when it dropped below that mark again. This interval gives the on- and offset of a spoken sentence. Components which subscribe to <address="voice@vad"> receive this event. We will later on use this event to send the corresponding snippet of raw audio to the dialogue manager, but as we are designing an emotional reactive interaction, we first need to also calculate the accompanying user state in the form of current arousal level.

As described in Section 5.2.2, we apply the pre-trained SoundNet model to create a condensed representation of the audio stream (<soundnet_pool5_stream>), to be classified by our arousal network.

```
<transformer create="PythonFeature" script="soundnet.py">
        <input pin="audio_stream" frame="882" delta="11883"/>
        <output pin="soundnet_pool5_stream"/>
</transformer>


<consumer create="Classifier" trainer="classifier_arousal"
                address="arousal@voice">
        <input pin=" soundnet_pool5_stream " frame="1"></input>
</consumer>
```

The classifier creates a recognition result every 50 milliseconds again as an event within the framework and in order to smooth these results we collect respective events within a fusion component as introduced in Section 5.3.

```
<object create="Fusion" dimension = "1" address="arousal@fusion">
        <listen address="arousal@voice"/>
</object>
```

Of course, this fusion component will - at later stages - be able to collect and process recognition results of multiple components and handle the whole valence-arousal space. For now, we use it only for arousal level estimation from the audio channel. Results of the fusion component are again published as events and in order to further process the results, we serialize the events into a data stream <stream_arousal>.

```
<sensor create="EventToStream" sr="25">
        <listen address="arousal@fusion"/>
        <output pin="stream_arousal"/>
</sensor>
```

Based on our former analysis steps, we now have all information extracted to drive an emotionally reactive conversation with the agent. The following component evokes a Python script <script="dialogflow.py"> whenever an event from the voice activity detection module is received <address="voice@vad">. As stated earlier, the event includes timestamps for the beginning and the end of a spoken sentence, so we are able to extract the corresponding frames out the internally buffered streams containing raw audio and arousal scores (<input pin=" stream_audio; stream_arousal”>).

```
<consumer create="PythonConsumer" script="dialogflow.py"
          address="speech@dflow">
    <input pin=" stream_audio; stream_arousal"
          address="voice@vad"/>
</consumer>
```

The extract of raw audio is sent to the external natural language recognition and dialogue management tool. Together with the raw audio, the mean value of the arousal scores that were recognized during the spoken sentence are provided to dialogue management, so that the dialogue can deviate in differing branches as reaction to the current user arousal level.

The last component in the chain rounds up the dialogue setting, by receiving the answer to the last user dialogue act as a string and calling a text-to-speech[10] routine to present the agent's answer to the user.

```
<object create="PythonObject" script="cereproc.py">
    <listen address="speech@dflow"/>
</object>
```

The resulting real-time recognition pipeline (Figure 1515) is meant to run in the background of the actual PRESENT application and stream recognition outputs via UDP sockets.

[10] https://www.cereproc.com/en/home

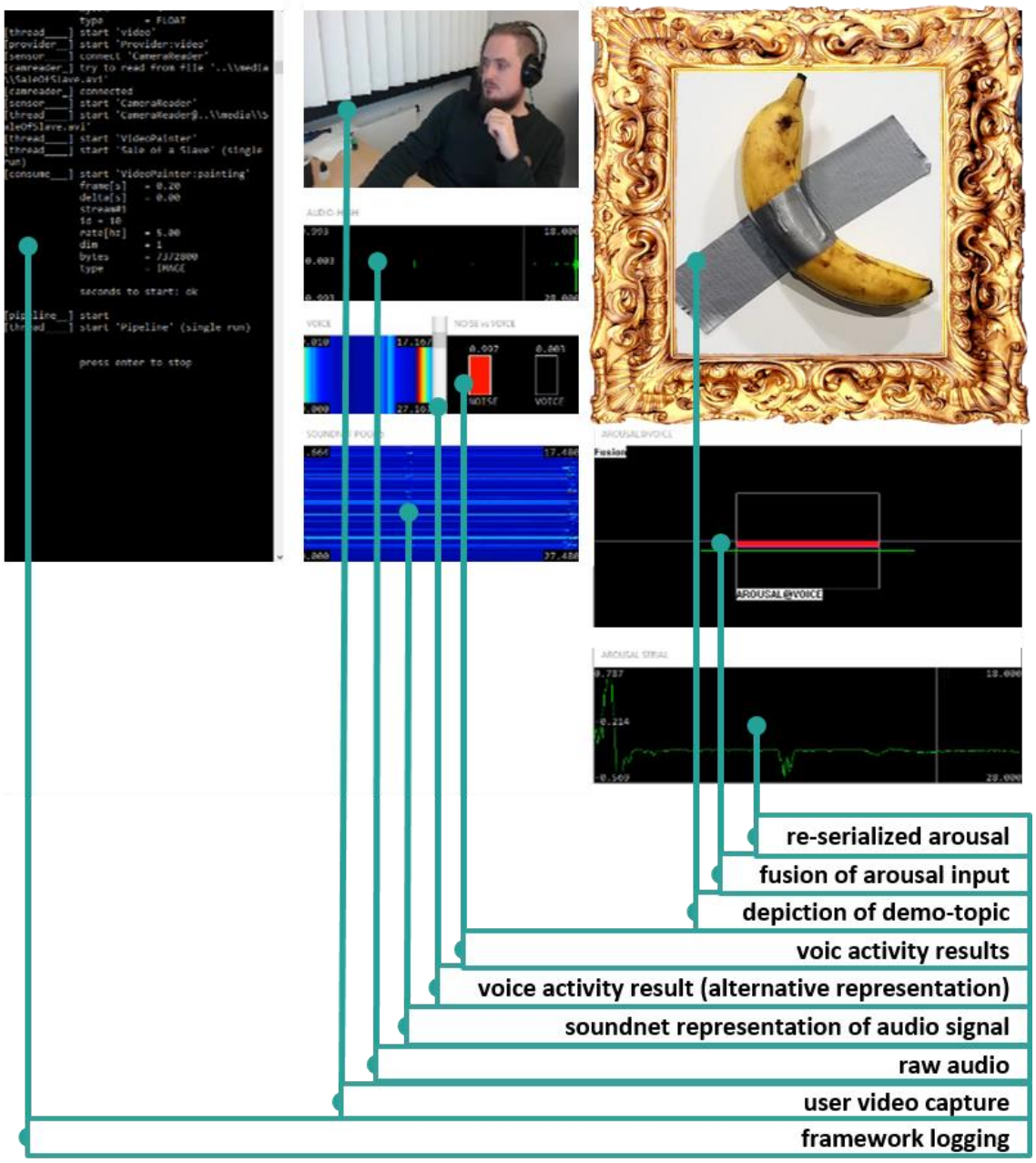

Figure 15: The recognition pipeline created by the given XML description fully drives the first demonstration dialogue, upcoming iterations will however run in the background of PRESENT applications and stream recognition results in real-time.

7 CONCLUSION

In this document, we introduced the theoretical concepts of discretizing emotions in order to properly handle them within the PRESENT system. The dimensional valence-arousal model is chosen as general representation method because respective values can be used organically by subsequent modules and can be mapped onto emotional labels if needed.

In the following, we gave an overview concerning the current state of the PRESENT social sensing system as well as current and future developments. The Social Signal Interpretation (SSI) framework offers the perfect environment to integrate our recognition modules for multi-modal user state recognition. We have described in detail our custom deep-learning based models for voice activity detection and arousal classification based on live audio input. We have furthermore outlined the ongoing effort to implement a video-based facial expression recognition module to introduce the valence dimension in the recognition system. This effort will be followed up by training all components to recognize both emotional dimensions and consequently a multi-modal fusion system capable of weighting asynchronous recognition based on evaluation performances of recognition modules for all available modalities.

The first demonstrator shows the capability of the purely audio-based social sensing system to drive an emotional reactive conversation with an agent. 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.

As the social sensing system is a crucial precondition for subsequent tasks such as dialogue management, agent behaviour or agent adaption, it is important to make the pipeline as modular, accessible and modifiable as possible for other partners. We therefore included the demonstrated ability to configure the whole recognition pipeline with XML syntax.

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.

Ekman, P. and W. Friesen. "Constants across cultures in the face and emotion." *Personality and Social Psychology*, 1971: 124–129.

Lang, P. and Bradley, M. and Cuthbert, B. "Motivated attention: Affect, activation, and action." *Psychology Press*, 1997: 97–135.

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." *EEE Transactions on Affective Computing*, 2016: 410-423.

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

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.