European Union's Horizon 2020 research and innovation programme 645378, ARIA-VALUSPA
November, 2015

Artificial Retrieval of Information Assistants – Virtual Agents with Linguistic Understanding, Social skills, and Personalised Aspects

Collaborative Project

Start date of project: 01/01/2015 Duration: 36 months

(D2.1). Implementation of cross-domain, context- sensitive speech analysis

Due date of deliverable: Month 11 Actual submission date: 1/12/2015

ARIA Valuspa

Project co-funded by the European Commission		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

STATUS: [DRAFT]

Deliverable Nature		
R	Report	
P	Prototype	
D	Demonstrator	X
O	Other	

Participant Number	Participant organization name	Participant org. short name	Country
Coordinator			
1	University of Nottingham, Mixed Reality/Computer Vision Lab, School of Computer Science	UN	U.K.
Other Beneficiaries			
2	Imperial College of Science, Technology and Medicine	IC	U.K.
3	Centre National de la Recherche Scientifique, Télécom ParisTech	CNRS-PT	France
4	Universitat Augsburg	UA	Germany
5	Universiteit Twente	UT	The Netherlands
6	Cereproc LTD	CEREPROC	U.K.
7	La Cantoche Production SA	CANTOCHE	France

Table of Contents

1. Purpose of Document

The purpose of this report is to accompany the actual demonstrator code, describing the implementation and functionalities of the cross-domain, context-sensitive speech analysis of ARIA-VALUSPA. This deliverable pertains to WP2 (Multi-lingual audio-visual-modal speech and affect recognition). The specific demonstator associated with this deliverable are also described.

Our central aims for this deliverable were twofold:

1) To create an audio-visual affect recognition system capable of generating relevant information about the user's stable states (age, gender, personality) and transient affective and cognitive states (emotion and interest) from speech and facial expressions.
2) The creation of the first instance of the Automatic Speech Recognition (ASR) module for the English language.

These two systems are meant to underpin the generation the Avatar's verbal (dialogue management) and nonverbal behaviours (e.g., expressiveness, stance taking). Both are essential components to endowing the system with fundamental skills for the analysis of social signals, and the creation of basic emotional intelligence and behavioural strategies, including socially competent induction of positive affective states in the user.

2. Nonverbal social signal processing

The implementation of the nonverbal signal processing modules was implemented with the Social Signal Interpretation (SSI) framework[1] used in the whole ARIA-VALUSPA project. SSI supports a large range of sensor devices, filters and feature algorithms, as well as machine learning and pattern recognition plugins (cf. D5.1, section 2.2.4 for more details on SSI). Where necessary, new components were also developed to fulfil the requirements of ARIA-VALUSPA.

2.1 Speech

From an implementation point of view, our goal was to create an SSI pipeline for online speech analysis and recognition of various user's states and traits. This pipeline should capture the user's speech through a standard computer microphone, extract relevant acoustic features, feed the adequate features to the respective recognition modules (pre-trained offline), and output the predicted class or value.

In this phase of the project, we decided to develop standard, state-of-the-art models for the recognition of users' stable traits (age, gender, personality), and transient states (emotional state and interest). This selection is adequate to address the goal of user characterization and adaptation by generating information for other modules (namely

[1] Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., André, E. (2013). The social signal interpretation (SSI)

dialogue management and Agent behaviour generation), and will provide a baseline for comparison with future developments.

2.1.1 Databases

The databases used to train the speech models were taken from the Computational Paralinguistics Challenge (ComParE) series[2]. In particular, we used the following databases:

- **ComParE 2010**[3]: The *aGender* and the *TUM AVIC* corpora used in this challenge were adopted for training our age, gender and interest recognition modules. The first consists of 46 hours of telephone speech, stemming from 954 speakers, and serves to evaluate features and algorithms for the detection of speaker age and gender. In relation to age, the target labels are *children* (7-14 y.o.), *youth* (15-24 y.o.), *adults* (25-54 y.o.) and *seniors* (> 54 y.o.). In relation to gender, the classes are *female*, *male*, and *children* (male or female and less than 14 y.o.). The second database features 2 hours of human conversational speech recording (21 subjects), annotated in 5 different levels of interest (in another's speaker speech; LoI). The LoI was measured using an ordinal scale ranging from -2 (*Uninterested*) to 2 (*Curious*) and the golden standard computed as the arithmetic mean across all raters, therefore regression is used for this task.
- **ComParE 2012**[4]: we used the Speaker Personality Corpus from this challenge, which comprises 2 hours of French speech from 330 different speakers. Eleven judges were then asked to infer the personality of the speaker by completing a standardised personality assessment tests - the BFI-10[5] – which quantifies personality traits in terms of the "Big Five" dimensions - openness, conscientiousness, extraversion, agreeableness, and neuroticism[6].
- **ComParE 2013**[7]: we used the dataset from the 2013 *Emotion Sub-Challenge* – the "Geneva Multimodal Emotion Portrayals" (GEMEP) [25]. It contains 1.2 k instances of emotional speech from ten professional actors (five female) in 18 categories. The GEMEP database contains prompted speech comprising sustained vowel phonations, as well as two 'nonsensical' phrases with two different intended sentence modalities, each expressed by each actor in various degrees of regulation (emotional intensity) ranging from 'high' to 'masked' (hiding the true emotion). For the purposes of our project, emotions categories were mapped into binary Arousal (high or low) and Valence (positive or negative) classes.

[2] Schuller, B. (2012). The computational paralinguistics challenge [social sciences]. *Signal Processing Magazine, IEEE* 29(4), p. 97-101.

[3] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A., and Narayanan, S.S. (2010). The INTERSPEECH 2010 paralinguistic challenge. In *INTERSPEECH*, p. 2794-2797.

[4] Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., Van Son, R. et al. (2012). The INTERSPEECH 2012 Speaker Trait Challenge. In *INTERSPEECH*.

[5] Rammstedt, B. and John, O. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German, *Journal of Research in Personality*, 41, p. 203–212.

[6] Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, p. 26-34.

[7] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., et al. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism.

2.1.2 Features

For all user characterization tasks we use the official feature set of ComParE 2013, as it has been extensively demonstrated to be suitable for a wide range of paralinguistic tasks. During model training, features were extracted on a per-chunk level. The set includes energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs) as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. Altogether, the feature set contains 6 373 features[8]. All features were extracted using the open source feature extractor openSMILE[9], which is also included in SSI and ready to operate in real-time.

2.1.3 Model development and data enrichment

We opted for using Support Vector Machines, a standard classification technique, similar to the baseline method used in the ComParE challenges. The goals were to create a baseline model for comparison with the developments planned for subsequent phases of the project, and guaranteeing the essential functionalities of the nonverbal analysis modules. The following classification tasks (respective labels also indicated) were implemented:

- Age: *children*, *youth*, *adults* and *seniors*;
- Gender: *children*, *female*, and *male*;
- Emotion: Arousal (*high*, *low*) and Valence (*positive*, *negative*)
- Interest: *uninterested*, *neutral*, *interested*

The *aGender* and *TUM AVIC* database were used as provided by the challenge organisers, whereas the other databases used were enriched to enlarge the amount of data available for training with the aim of improving the models' performance. To do so, we developed a novel approach for large-scale data enrichment. The central concept of our approach is to join existing data resources into one single holistic database with a multi-dimensional label space by using semi-supervised learning techniques to predict missing labels. Our *Cross-Task Labelling* (CTL; article submitted to ICASSP 2016) method consists of training a model on the labelled training data of the selected databases for each individual task. Then, the trained classifiers are used for the cross-labelling of databases among each other. Our results show that CTL lays the foundation for holistic speech analysis by semi-autonomously annotating existing databases, and expanding the multi-target label space at the same time, while achieving higher accuracy as the baseline performance of the challenges.

2.1.4 Real-time workflow

Figure 1 depicts the real-time framework for nonverbal speech analysis and recognition of user states and traits. A standard microphone captures sound. SSI takes this live audio input, and sends it to a Voice Activity Detector (VAD), which will chunk the continuous inputs into segments. Typically, these segments contain a single sentence, although,

[8] Weninger, F., Eyben, F., Schuller, B.W., Mortillaro, M. and Scherer, K.R. (2013). On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology* 4.

[9] Eyben, F., Wöllmer, M. and Schuller, B.W. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor." In *Proceedings of the international conference on Multimedia*, p. 1459-1462.

depending on the amount of pauses, can contain also shorter segment (or even single words). For each chunk a feature set is extracted using a predetermined configuration file (that defines which features to be extracted; see 2.1.2). To this end, a new openSMILE plugin has been developed, which allows it to use SSI as a sink/source in a configuration file, i.e. data is efficiently exchanged between the two components in memory. The output of the feature extraction is a single vector containing the features relevant for the following recognition task(s). Once features are extracted, the feature vector is sent to the recognition modules where the features are normalized (to match the scaling used during the development stage), and then feed into the various models (at this stage a single feature set is being used for all models, but tailored features sets can also be extracted if necessary). The output of each model is then collected, displayed on the screen (for debug purposes), and sent to other modules that use information about speakers' states and traits for further processing. All models run in parallel, and outputs made available continuously.

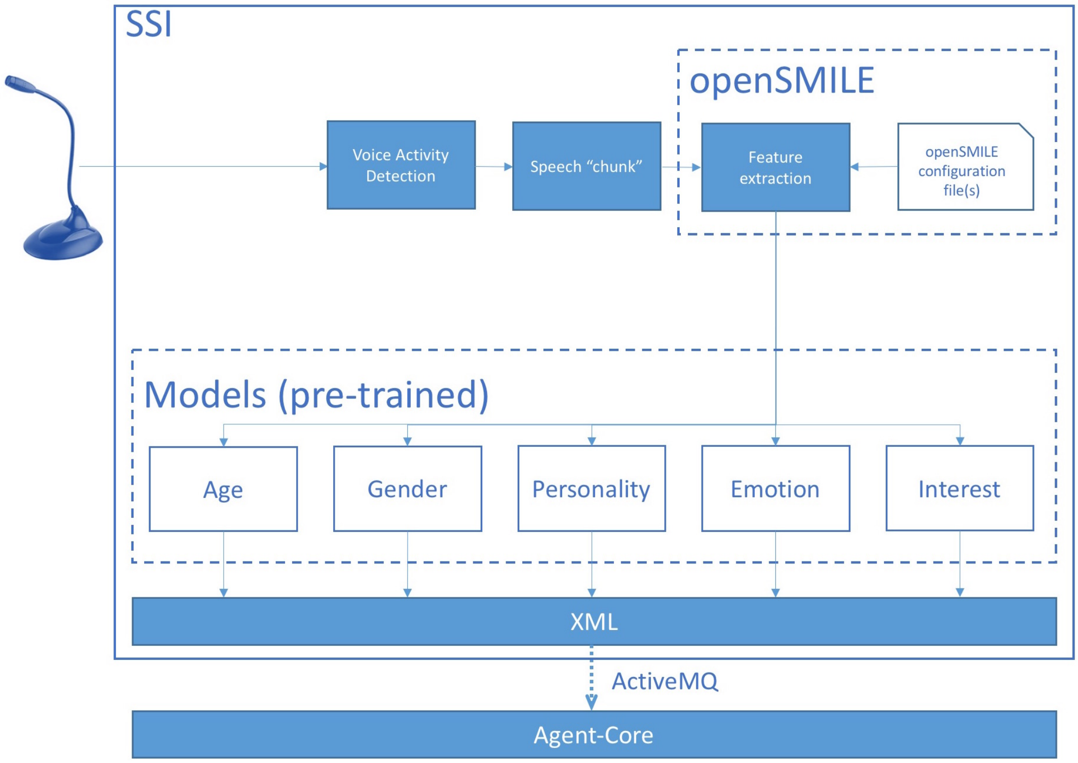


Figure 1. Overview of the nonverbal speech recognition module.

2.2 VIDEO

The visual Facial Expression Recognition (FER) module has been trained to recognise the six prototypical emotions (anger, disgust, fear, happiness, sadness, and surprise). The output of this module is a per-frame and per-label score indicating the confidene of each of the possible labels. This includes the absence of any emotion, i.e., the presence of a neutral face, leading to 7 real-valued scores.

2.2.1 Databases

- **Posed FER databases (MMI and CK):** Both the MMI (Part I-III)[10] and CK database[11] contain predominantly recordings acquired under controlled lab conditions, including fixed frontal-looking head pose and frontal illumination. The subjects were asked to pose the facial displays. While training FER for naturalistic videos using posed expressions is not ideal, these databases have been included due to the sortage of annotated in-the-wild training data.
- **Spontaneous FER database (CK+):** The extended Cohn-Kanade dataset[12] (CK+ database) is a second release of the CK database. This release consisted of spontaneous data. While the recordings were also captured under controlled lab conditions, the facial expression displays are naturalistic and thus more representative for the in-the-wild case.
- **In-the-wild FER database (BBC100):** The BBC100 database is an in-house database (thus not publicly available) developed in conjunction between the University of Nottingham and CrowdEmotion (an industry associate within the project). The BBC100 database is constantly being expanded as part of a continuous effort to tackle the training data sortage affecting FER models for naturalistic scenarios. The dataset currently includes around 150 sequences, of which 73 were used to train the current version of the models. Each sequence corresponds to a different subject, and contains one event of facial expression. The subjects are recorded interacting through a webcam. No constraints in terms of illumination, head movement, or identity have been imposed, and they are thus truly in-the-wild sequences.

2.2.2 Features

In order to extract meaningful features, it is first necessary to perform face alignment for every frame analysed. We proceed by performing face detection and facial landmark detection at every frame analysed rather than performing facial landmark tracking. This is due to the number of processes running in parallel, and the resulting requirement of robustness against frames being dropped.

We employ the face detection algorithm provided with the publicly-available dlib library[13]. This is due to its efficient multi-platform implementation, its ability to detect multiple faces simultaneously, and the precission of the localisation. The face detection bounding box is used both for maintaining a subject id (see section 2.2.4) and initialising the facial landmark detector. The algorithm used for facial landmarking is the Project-

[10] Pantic, M., Valstar, M.F., Rademaker, R. and Maat, L. (2005). Web-based database for facial expression analysis. In *Int'l Conf. on Multimedia & Expo*, p. 317–321.

[11] Kanade, T., Cohn, J.F. and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition*, .p 46–53.

[12] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J. and Ambadar, Z. (2010). The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Comp. Vision and Pattern Recognition – Workshop*, p. 94–101.

[13] http://dlib.net/

Out Cascaded Regression (PO-CR)[14]. The PO-CR algorithm is a state-of-the-art algorithm for facial landmark detection, and is capable of dealing with non-frontal head poses. We use an optimised C++ implementation to avoid having a computational overhead. The output consists on 68 facial landmarks, which are then used to register the face into a pre-defined coordinate system. This step effectively aligns all faces together, i.e., eliminates translation, scale and in-plane rotation variation between the images.

Feature extraction is applied over the registered images. We use Local Gabor Binary Pattern features (LGBP)[15] as they have shown excellent performance for a number of fine-grained face analysis algorithms, including FER.

2.2.3 Model Training

The training procedure is defined as 7 "one-vs-all" binary classification tasks. In order to guarantee real-time performance, a linear SVM classifier is used for each of the classes. Parameter optimisation is performed using a validation set. We use a portion of the BBC100 database not included during training (and subject-independent). We use this database as we are interested in "in-the-wild" performance. The training routine is totally automated, being run at regular intervals to create models incorporating the latest examples added to the BBC100 database.

2.2.4 Real-time workflow

The video data is captured through a standard webcam. Face detection is first run, and different processes are started for each of the faces detected (to avoid excessive computational overhead for now we constrain the maximum number of faces to be two). For each new frame, a face ID is assigned to each face detected based on the overlap with previous face detections. Then, facial landmark detection is performed for each face, followed by face registration and feature extraction. Each of the classifiers trained are run on the same features, yielding a real-valued score. This produces a per-frame label that is then fed into the SSI. In order to avoid excessive fluctuation of the frame rate, we run at approximately 5 fps. The process is summarised in Figure 2.

[14] Tzimiropoulos, G. (2015). Project-out cascaded regression with an application to face alignment, *IEEE Conf. on Computer Vision and Pattern Recognition*.

[15] Zhang, W. Shan, S., Gao, W., Chen, X., and Zhang, H. (2005). Local Gabor binary pattern histogram sequence (lgbphs): a novel nonstatistical model for face representation and recognition, *Int'l Conf. on Computer Vision*.

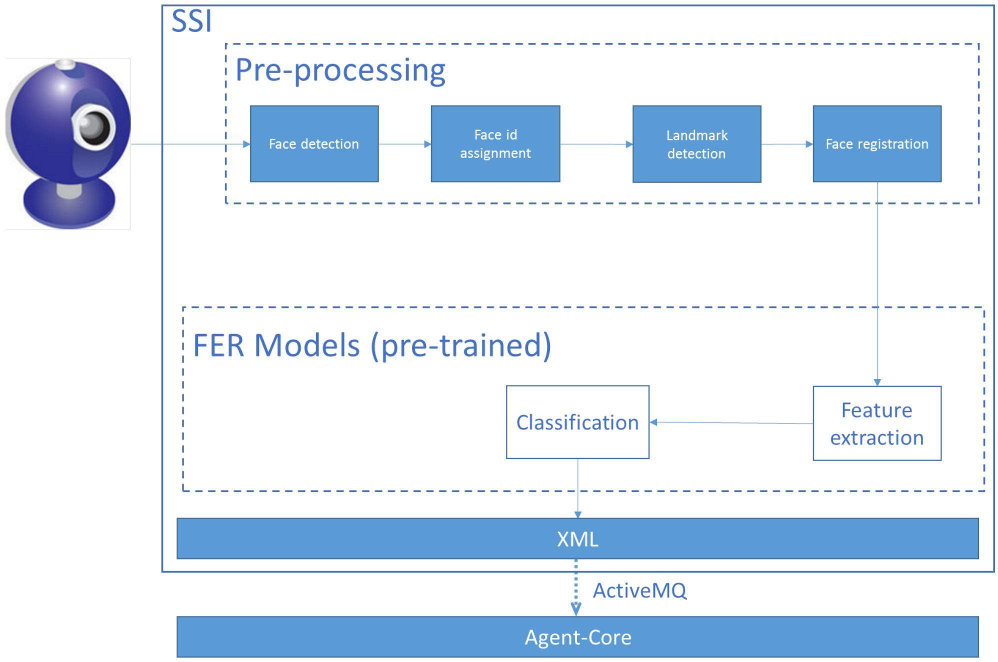


Figure 2. Overview of the video-based Facial Expression Recognition module.

3. Analysis of verbal content: Automatic Speech Recognition (ASR)

3.1 Databases

The English ASR module was created using a recently released corpus of read English speech called *LibriSpeech*[16], which is suitable for training and evaluating speech recognition systems and is freely available for download. The corpus was derived from audiobooks that are part of the *LibriVox* project. It contains approximately 1000 hours of speech (2338 speakers: 1128 females, 1210 males). Alongside the corpus, pre-built language models are also available, and were trained on approximately 14,500 public domain books taken from *Project Gutenberg* with around 803 million running words and 900 000 unique words. The lexicon contains the 200 000 most frequent words of the text corpus. The pronunciations of one third of these words are taken from the publicly available CMU pronunciation dictionary. Pronunciations of missing words are generated using the Sequitur Grapheme-to-Phoneme (G2P) converter toolkit[17].

In addition to the *LibriSpeech* corpus, we automatically segmented and manually aligned five audiobooks of *Alice's Adventures in Wonderland* to further train the acoustic and language models of the ASR with a vocabulary specific to our case scenario. Finally, for testing purposes and performance estimation in a typical target scenario, we created a hypothetical dialogue between Alice and a testing user, and asked eleven people (7

[16] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). LibriSpeech: an ASR corpus based on public domain audio books, *ICASSP*.

[17] Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion, *Speech Communication*, 50(5), p. 434-451,.

male) to read the user parts and record them using a standard PC microphone in realistic conditions. The dialogue script is as follows:

User: *Hi! How are you?*
Alice: *Hello! I am fine thanks. What is your name?*

User: *My name is [user's name]. And yours?*
Alice: *Hi [user's name]! My name is Alice.*

User: *Alice? Like Alice in Wonderland?*
Alice: *Exactly. I love that book. I read it a million times.*

User: *Then you should know a lot about it.*
Alice: *Well, I know the story by heart ... and some curiosities about the book as well.*

User: *That is nice to know. Would you mind if I ask you a few questions?*
Alice: *Of course not.*

User: *Who wrote Alice in Wonderland?*
Alice: *That's easy. It was Lewis Carol. Actually, his real name was Charles Lutwidge Dodgson.*

User: *He also wrote Through the Looking Glass, right?*
Alice: *Yes, he did.*

User: *I have forgotten quite a lot of things about the book, but I remember a long and sad tale. That is in Alice in Wonderland, correct?*
Alice: *Yes, it is in Chapter three if I remember well - a Caucus-Race and a Long Tale.*

User: *Do you remember who told the tale to Alice?*
Alice: *Sure. It was the Mouse. Shall I read you a passage?*

User: *Yes, please!*
Alice: *[`You promised to tell me your history, you know,' said Alice, `and why it is you hate--C and D,' she added in a whisper, half afraid that it would be offended again.*
`Mine is a long and a sad tale!' said the Mouse, turning to Alice, and sighing. `It IS a long tail, certainly,' said Alice, looking down with wonder at the Mouse's tail; `but why do you call it sad?']

User: *There were also a character in the book who kept appearing and disappearing. What was its name?*
Alice: *You mean the Cheshire Cat?*

User: *Yes! The Cheshire cat. What a curious character ... I also recall the mad Hatter, the one with a card on his hat showing '10/6'. Do you know what does that mean?*
Alice: *The card is a price tag in old English money. It means ten shillings and six pence. This meaning is explained in a shorter version of the book called The Nursery Alice, also by Lewis Carroll.*

User: *That is fantastic! Thank you so much. I always wanted to know that!*
Alice: *You are very welcome.*

User: *It was really nice to talk to you and recall the book. It was also nice to meet you, Alice.*
Alice: *Same here. I also enjoyed our conversation. Hope to see you again. I could read the book for you if you want.*

User: *That would be nice. See you soon then.*
Alice: *Bye, bye!*

3.2 ARCHITECTURE

The general architecture of our ASR system is depicted in Figure 3.

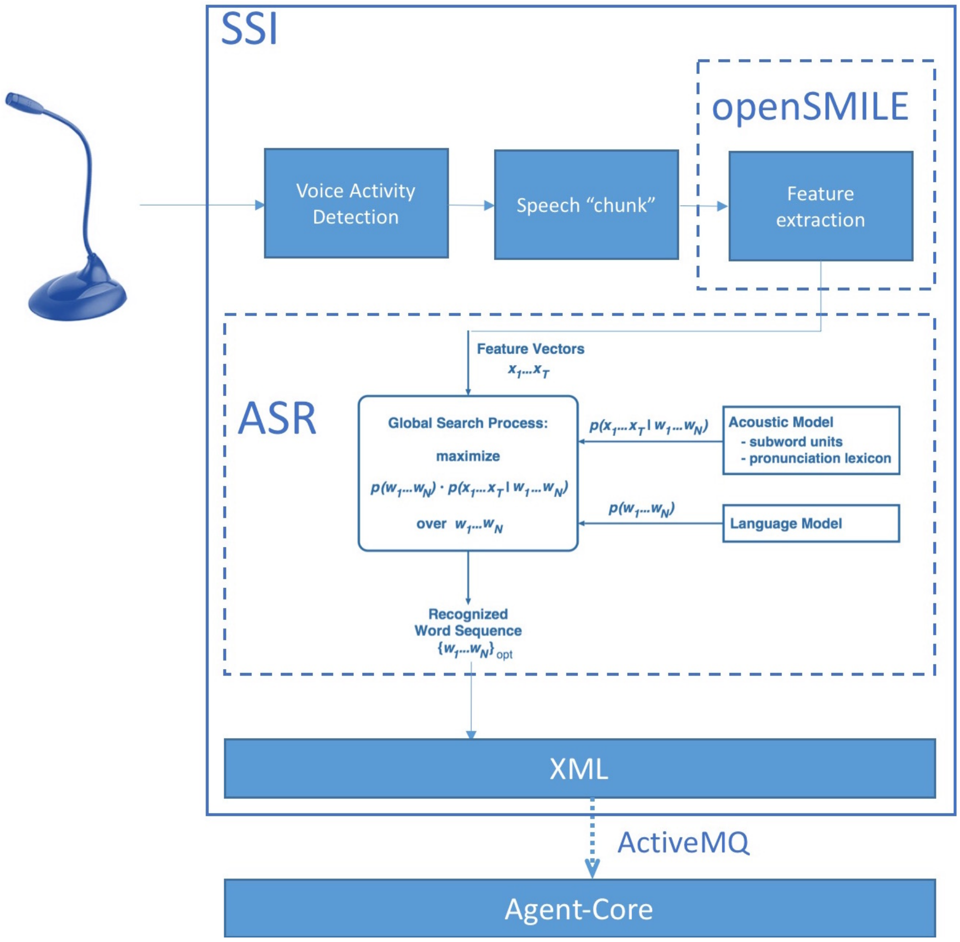


Figure 3. Overview of the Automatic Speech Recognition module.

The system is implemented using state-of-the-art approaches. Feature extraction is based on Mel-frequency Cepstral Coefficients (MFCCs) with additional delta and acceleration coefficients. The language model is a modified Kneser-Ney smoothed backoff 4-gram language model. Acoustic models are context-dependent triphone models trained using a hybrid Deep Neural Network – Hidden Markov Models (DNN-HMM) setup. The DNN part is based on deep maxout neural networks with p-norm pooling strategy[18]. These neural networks are trained to predict the posterior probabilities of each context-dependent state, which are then divided by the corresponding state prior probability to provide a "pseudo-likelihood" that is used in place of the state emission probabilities in the triphone HMMs. The neural network training is performed on top of feature space maximum a posteriori linear regression

[18] Zhang, X., Trmal, J., Povey, D. and Khudanpur, S (2014). Improving Deep Neural Network Acoustic Models using Generalized Maxout Networks, ICASSP.

(fMLLR) speaker adapted features. The decoder is based on Weighted Finite State Transducers (WFSTs).

3.3 Performance (preliminary evaluation)

The ASR trained on the *LibriSpeech* corpus was evaluated on the five manually aligned audiobooks of the "Alice in Wonderland", as well as on the dialogue test scenario. The Word Error Rates (WER) are displayed in Table 1.

Dataset	WER (%)
Audiobook 1	29.22
Audiobook 2	40.33
Audiobook 3	23.40
Audiobook 4	21.31
Audiobook 5	9.98
Scripted dialogue	29.48

Table 1. Performance of the ASR-English system on the target scenario vocabulary.

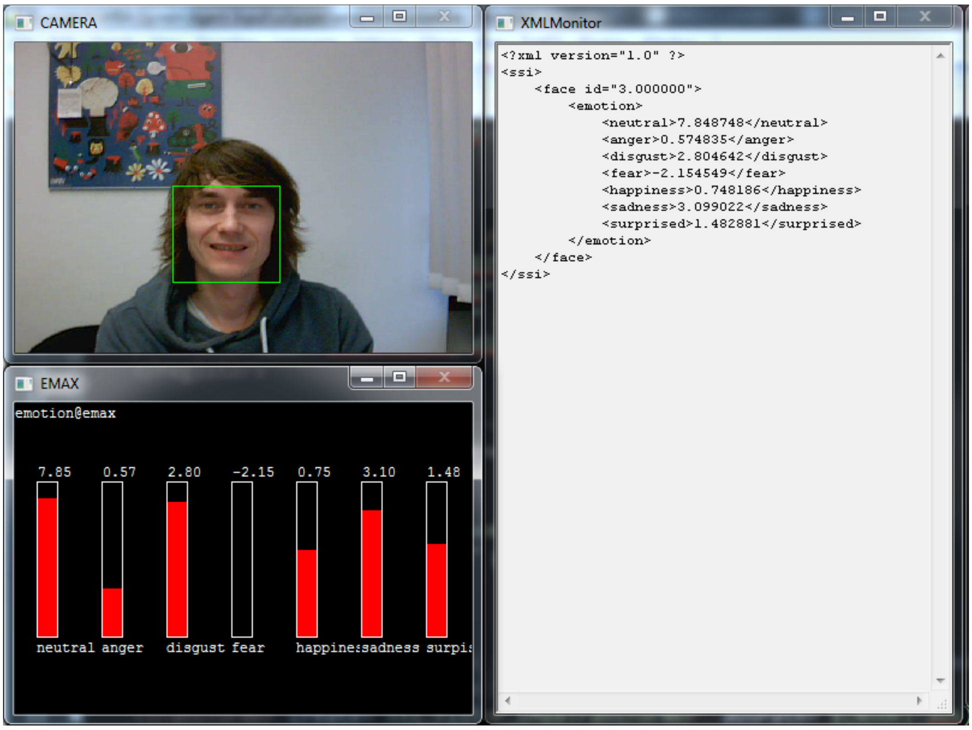
4. Case scenario and real-time demonstration

We created a scenario to demonstrate the use of nonverbal and verbal analysis in the context of the Book.-ARIA The code is publicly available from https://github.com/ARIA-VALUSPA/ARIA-System. The particular demonstrator for D2.1 is run on Windows machines by double-clicking the file "RUN-Emotion-Mimic.bat". For a full system demonstration, including dialogue management and behaviour generation, you can run "RUN-All.bat", but please note that this is essentially deliverable D1.1, which is not finalised until end of December 2015.

The scenario consists of mimicking the user's emotional state, by mirroring the emotions estimated from speech and video signals. Throughout the interaction we keep track to the user's emotional state at each moment in the interaction, as well as an historic of the sequence of emotional states expressed in order to estimate the mood of the user. The estimated mood is then used by the Agent to read the story with a tone of voice and facial (expressive) behaviours that match the user's emotional state. The output of the ASR is displayed on the GUI console.

As a first step, we integrated eMAX into the SSI framework. To capture the user's face we use a webcam and forward every image frame to the eMAX component (currently ~5 fps at 640x480 RGB). The result is a vector with the position of the face (if a face is found) and scores indicating the estimated evidence of detecting seven basic emotions: 'neutral', 'anger', 'disgust', 'fear', 'happiness', 'sadness', and 'surprised'. A higher score expresses a higher probability that the current frame expresses the according emotion. Results are visualized in a bar plot to give an immediate feedback of the recognition (see Figure 4. Real-time facial expression recognition). eMAX also provides a discrete estimate of the Valence and Arousal[19], where the possible output labels are positive or negative valence and low or high arousal. To share the detected user state with Dialogue Manager an XML template has been defined. At run-time the template is filled with the current values and published through an ActiveMQ port.

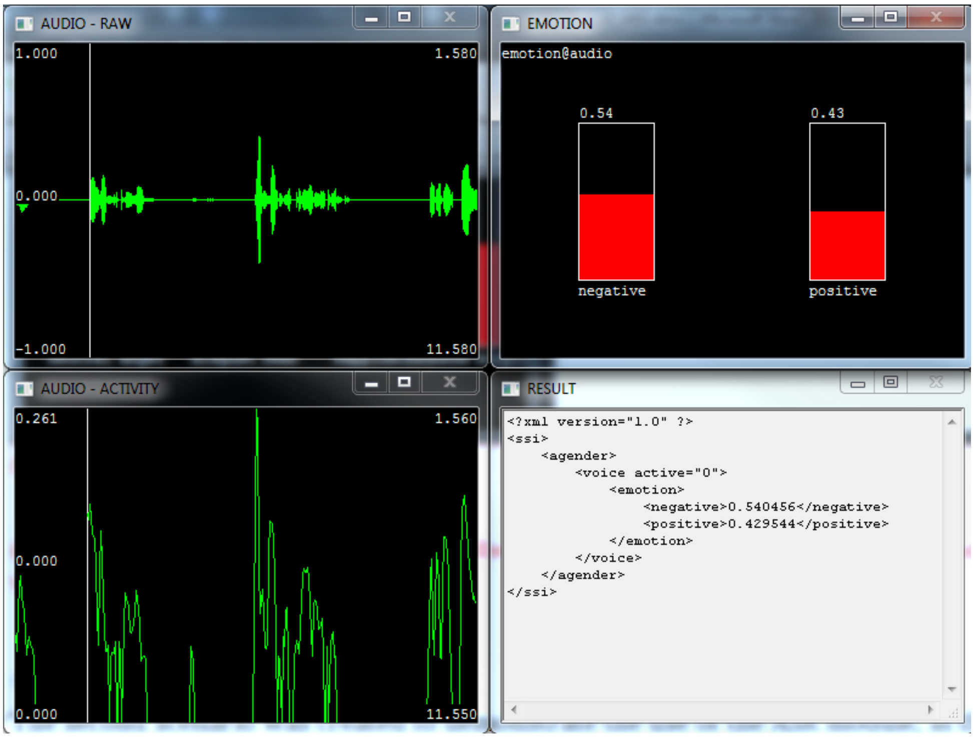
The audio is processed as described in 2.1.4. Again we give feedback of the recognition in a bar plot (see Figure 5. Real-time speech emotion recognition) and fill an according XML template. Finally, results are published through an ActiveMQ port. In favour of improved maintainability we currently keep separate pipelines, which will be integrated in a single pipeline for the final system. This will alsog allow publishing all recognition results in a single XML structure.



```
<?xml version="1.0" ?>
<ssi>
    <face id="3.000000">
        <emotion>
            <neutral>7.848748</neutral>
            <anger>0.574835</anger>
            <disgust>2.804642</disgust>
            <fear>-2.154549</fear>
            <happiness>0.748186</happiness>
            <sadness>3.099022</sadness>
            <surprised>1.482881</surprised>
        </emotion>
    </face>
</ssi>
```

Figure 4. Real-time facial expression recognition

[19] Russell, J.A. (1980). A Circumplex Model of Affect, *J. Personality and Social Psychology*, 39, p. 1161-1178.



```
<?xml version="1.0" ?>
<ssi>
    <agender>
        <voice active="0">
            <emotion>
                <negative>0.540456</negative>
                <positive>0.429544</positive>
            </emotion>
        </voice>
    </agender>
</ssi>
```

Figure 5. Real-time speech emotion recognition

5. Plans for Next Period

In our on-going work, we are improving the nonverbal audio-visual recognition modules, as well as exploring the advantages of using information from multiple modalities for improving the robustness of the whole system. Specifically, we are now working with deep neural architectures for single- and multi-modal user's states and traits recognition using, which can work when only one or both modalities are available (which is particularly important for the robustness of the system in real-time). Furthermore, we are also developing post-modelling strategies for using the consecutive outputs of the various modules to improve the stability of the system. Since states like age or gender, but also mood, will not change during interaction with the same user, we are currently exploring ways to keep track of recognition results over a longer period of time and output smoothed probabilities. In this way we are hoping to increase the stability of the recognizers. Once we approach a stable classification state we would want to suspend recognition for the remaining interaction. However, in the event that the interaction partner changes we have to be able to turn on according recognition modules once again to adjust to the new situation. Therefore, we are currently extending SSI with a mechanism that will allow us to use certain components in the pipeline on demand. In relation to the ASR, we are now training the system to recognize German speech, and soon we will start the French system. Finally, we are also assessing the part of the system that should be adapted to the user, including individual models for emotion recognition, and online collection of new material for model retraining.

6. OUTPUTS

In what follows, we indicate the outputs with pertinence to this deliverable (categorised by topics) that have been published (or are *in press*) in the first 11 months of the project.

Face alignment

Sanchez-Lozano, E., Martinez, B., and Valstar, M.F. Cascaded Regression with Sparsified Feature Covariance Matrix for Facial Landmark Detection, *Pattern Recognition Letters* (accepted for publication).

Martinez, B., and Valstar, M.F. L21-based regression and prediction accumulation across views for robust facial landmark detection, *Image and Vision Computing* (accepted for publication).

Wang, X., Valstar, M.F., Martinez, B., Khan, M.H., Pridmore, T.P. (2015). Tracking by Regression with Incrementally Learned Cascades, *Int'l Conference on Computer Vision* (ICCV).

Facial expression analysis from video

Almaev, T., Martinez, B., and Valstar, M.F. (2015). Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection, *Int'l Conference on Computer Vision* (ICCV).

Jaiswal, S., Martinez, B., and Valstar, M.F. (2015). Learning to combine local models for Facial Action Unit detection, FERA workshop, *IEEE Conf. on Face and Gesture Recognition*.

Jaiswal, S., Valstar, M.F. (2016). Deep Learning the Dynamic Appearance and Shape of Facial Action Units, *Winter Conference on Applications of Computer Vision* (WACV), (accepted).

Martinez, B., and Valstar, M.F. (2016). Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition, Face Detection and Facial Image Analysis, M. Kawulok, E. Celebi, B. Smolka editors, Springer (In press).

Ringeval, F., Schuller, B.W., Valstar, M.F., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., Pantic, M. (2015). AV+EC 2015--The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data, *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, ACM Multimedia.

Valstar, M.F., Girard, J., Almaev, T., McKeown, G., Mehu, M., Yin, L., Pantic, M. and Cohn, J. (2015). FERA 2015 - Second Facial Expression Recognition and Analysis Challenge, *Facial Expression Recognition and Analysis Challenge workshop*, in conjunction with IEEE Int'l Conf. Face and Gesture Recognition, 2015.

Dealing with unexpected situations (acoustic novelty detection)

Marchi, E., Vesperini, F. Weninger, F., Eyben, F., Squartini, S., and Schuller, B. (2015). Non-Linear Prediction with LSTM Recurrent Neural Networks for Acoustic Novelty Detection. In *Proceedings 2015 International Joint Conference on Neural Networks (IJCNN)*, IEEE, p. 12-17.

Emotion recognition in music (context awareness)

Sagha, H., Coutinho, E., and Schuller, B.W. (2015). Exploring the Importance of Individual Differences to the Automatic Estimation of Emotions Induced by Music. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, p. 26-30.

Coutinho, E., Trigeorgis, G., Zafeiriou, S, and Schuller, B.W. (2015). Automatically Estimating Emotion Music using Long-Short Term Memory Recurrent Neural Networks. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, p. 14-15.

Sentiment analysis (inputs to dialogue management)

Schuller B., Mousa A. E., and Vryniotis V. (2015). Sentiment analysis and opinion mining: on optimal parameters and performances. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5), p. 255-263.

Schröder M., Bevacqua E., Cowie R., Eyben F., Gunes H., Heylen D., Maat M., McKeown G., Pammi S., Pantic M., Pelachaud C., Schuller B., De Sevin E., Valstar M.F., and Wöllmer M. (2015). Building Autonomous Sensitive Artificial Listeners (Extended Abstract). In *Proc. Conference on Affective Computing and Intelligent Interaction (ACII)*, p. 21-24.

Estimation of user's states

Coutinho, E. and Schuller, B.W. (2015). Automatic estimation of biosignals from the human voice. In B. Hu and J. Fan (Eds.), *Advances in Computational Psychophysiology.* Science, Special Supplement on Advances in Computational Psychophysiology, AAAS, 350(114), p. 48-50.

Gentsch, K., Coutinho, E., Eyben, E., Schuller, and Scherer, K.R. (2015). Classifying Emotion-Antecedent Appraisal in Brain Activity using Machine Learning Methods. In *Proceedings of the Bi-Annual Conference of the International Society for Research on Emotion*, Geneva, Switzerland: International Society for Research on Emotions, p. 8-10.

Azäis, L., Payan, A., Sun, T., Vidal, G., Zhang, T., Coutinho, E., Eyben, F., and Schuller, B.W. (2015). Does my Speech Rock? Automatic Assessment of Public Speaking Skills. In *Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association,* ISCA, p. 2519-2523.

Trigeorgis, G., Coutinho, E., Ringeval, F., Marchi, E., Zafeiriou, S, and Schuller, B.W. (2015). The ICL-TUM-PASSAU Approach for the MediaEval 2015 "Affective Impact of Movies" Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, p. 14-15.

Dynamic Active Learning (reduction of annotation effort and costs)

Zhang, Y., Coutinho, E., Zhang, Z., Quan, C., and Schuller, B. (2015). Agreement-based Dynamic Active Learning with Least and Medium Certainty Query Strategies. In F. Bach and D. Blei (Eds.), In *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France. JMLR: W&CP, 37, p. 6-11.

Zhang, Y., Coutinho, E., Zhang, Z., Adam, M., and Schuller, B. (2015). On Rater Reliability and Correlation Based Dynamic Active Learning. In *Proceedings 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII),* p. 21-24.

Zhang, Y., Coutinho, E., Zhang, Z., Quan, C., and Schuller, B. (2015). Dynamic Active Learning Based on Agreement and Applied to Recognition of Emotions in Spoken Interactions. In *Proceedings of the 17th ACM International Conference on Multimodal Interaction*, p. 9-13.

Big data

Schuller B. (2015). Speech Analysis in the Big Data Era. In *Proceedings of the 18th International Conference on Text, Speech and Dialogue*, 9302, p. 3-11.