ARIA Valuspa

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

August, 2017

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**

(D2.3). Implementation of long-term adaptation for audio-visual communication analysis

Due date of deliverable: Month 31 Actual submission date: 13/09/2017

ARIA Valuspa

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

August, 2017

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

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

August, 2017

Table of Contents

1. Purpose of Document

The purpose of this report is to describe the methods of long-term user adaptation, the implementation and integration into the Aria system, as well as the evaluation results for the Deliverable D2.3. This work package is important regarding the "personal aspects" as denoted in the project acronym. Similar to a new friend, the Aria should get to know its user better and better with time in order to provide personal assistance with accumulated background knowledge.

2. Use Case

The goal of Task 2.3 is to endow the ARIA-VALUSPA Platform (AVP) with capabilities of learning and adapting to user characteristics with enhanced context awareness. To this end, user traits extracted from utterances in Task 2.2 (automatic speaker analysis), will allow preselection of interest and emotion models that best fit the user profile (e.g. age group, gender). Also, this meta data serves as additional high-level features for the dialogue management. Further, we can automatically adapt all models to the speaker's voice in the course of the dialogue. Novel methods for acoustic model adaptation over the state of the art developed in the project will be used for creating specific models for a single user or a user group during long-term interaction with Aria. Finally, reliable audio-visual subject identification will help to automatically select a user specific model, if a person can be identified and the accordant user profile exists in the system. Otherwise, models suitable for the user's demographics are used.

3. Methodology

From an implementation point of view, the long-term user adaptation can be realised by using semi-supervised techniques (e.g. self-training). In our use case, the user interacts with the Aria system running with pre-trained models for speaker analysis, e.g., interest and emotion recognition. The adaptation problem can be tackled by a twofold approach. First, on the input side of the Aria architecture, SSI (Johannes Wagner, 2013) detects the user's profile and performs a preselection of the trained models suitable for the specific target user(-group). In a long-term view, the system will make use of the data collected from this user by automatically labelling it and using the new samples for refining the models, thereby adapting these to the specific person.

In our experiments, we partition the selected database into speaker-disjoint training set and user-specific data. The training set contains labelled instances equal to 1/3 of the total data amount. For each target user(-group), the remaining 2/3 of the data is further split into 2/3 'unlabelled' adaptation set and 1/3 labelled test set, which is used for performance evaluation (Figure 1). In particular, for the categorisation, we considered age group (young, adult, senior), gender (male, female), and the speaker ID itself. The great benefit of this approach is that the adaptation scheme can be further extended to many other categories, e.g., different native languages.

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

August, 2017

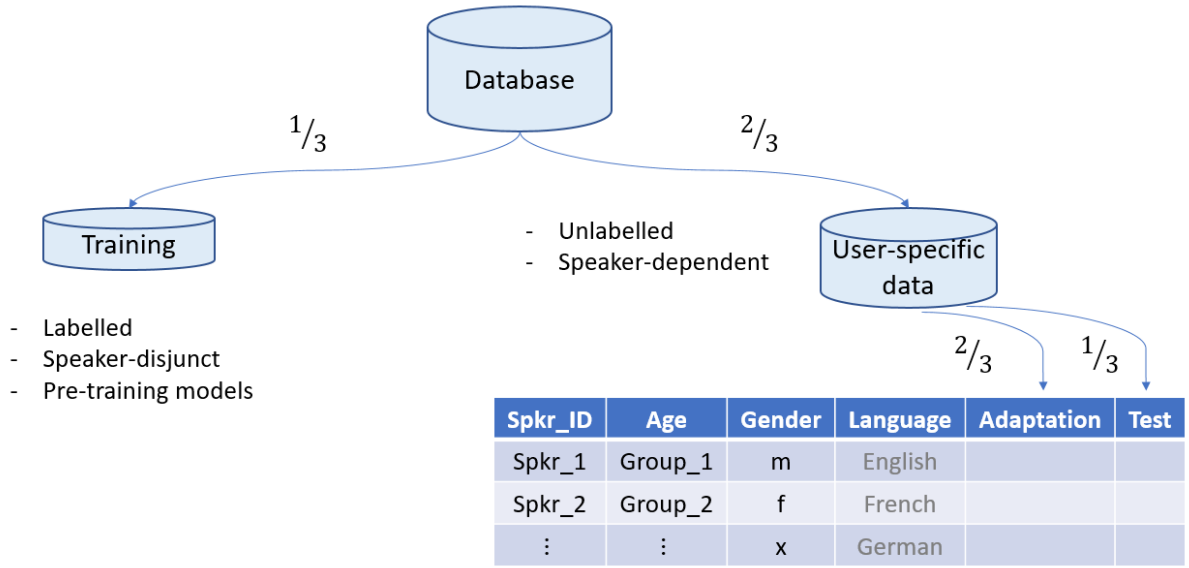

Figure 1: *Data partitioning for simulating long-term user adaptation using self-training techniques.*

In the long-term adaptation process, we aim to exploit unlabelled data by using machine predictions without any human intervention as it is unlikely that annotation of the user data can be done manually in the considered scenario. In this way, we can iteratively retrain the model adding new machine labelled instances to the original training set, thus enabling data aggregation and label enrichment.

3.1 Implementation (IC)

In the Aria project, we have developed a generic human-machine annotation framework based on Dynamic Cooperative Learning (DCL) techniques (Figure 2). It is able to fully automatically distribute the annotation workload between humans and machines by combining uncertainty sampling (AL) and self-training (SSL). This algorithm has been integrated into NOVA (Automated Analysis of Nonverbal Signals in Social Interactions) (Baur, 2013) by our partners from UA for the annotation of the NOXI database at a minimum of human labelling effort.

As the adaptation process should be unattended, we select the SSL path in the learning framework (marked as green in Figure 2). Self-Training is a well-known SSL method that trains a classifier on a small labelled data set and re-trains the model iteratively with the **most confident** machine predictions for an unlabelled data pool (Rosenberg, 2005). To enable the framework's applicability for both classification and regression tasks, we use a deep-learning based confidence measure to assess model uncertainty, which has the great advantage over state-of-the-art methods that it is applicable to both classification and regression tasks (Gal, 2016). The implementation of the DAL method is written in the Python language. For DNN training and computing of uncertainty measures, the choice fell on Keras, an open source deep learning library built on top of Theano and TensorFlow.

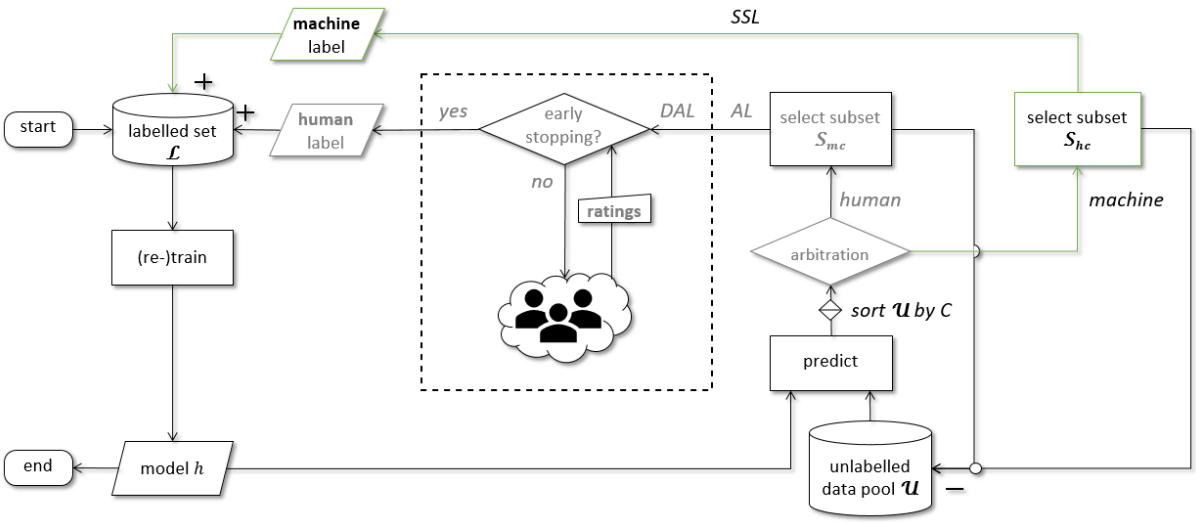

Figure 2: Flowchart of the human-machine annotation framework with support for various learning paradigms including semi-supervised learning (SSL), active learning (AL), dynamic active learning (DAL), cooperative learning (CL), and dynamic cooperative learning (DCL); the instances predicted with high model confidence (hc) are directly tagged with machine labels, whereas the instances with relatively low model confidence (mc) are subject to human inspection. Taken from the Manusscript (Yue Zhang, 2017) submitted to IEEE Transactions on Cybernetics.

The deep rectifier network consists of four hidden layers with 1,000; 750; 500; and 150 units as well as an output layer with a single neuron. The loss function is given by the mean squared error. Optimisation is done via statistical gradient descent (SGD). The initial network is trained for 200 epochs on the labelled. Re-training after each labelling step is done for 30 epochs (Yue Zhang, 2017).

3.2 DATABASES

3.2.1 AVIC

For interest recognition, we selected the Audiovisual Interest Corpus recorded at the Technische Universitat München ("TUM AVIC") as described in (B. Schuller, 2009). In the scenario setup, an experimenter and a subject were sitting on opposite sides of a desk. The experimenter played the role of a product presenter and led the subject through a commercial presentation. The subject's role was to listen to explanations and topic presentations of the experimenter, ask several questions of her/his interest, and actively interact with the experimenter. The subject was explicitly asked not to worry about being polite to the experimenter, e. g., by always showing a certain level of 'polite' attention. Visual and voice data was recorded by a camera and two microphones, one headset and one far-field microphone, at 44.1 kHz, 16 bit. In total, 21 subjects took part in the recordings, three of them Asian, the remaining European.

Table 1: *Details on subjects contained in the TUM AVIC database.*

Group	# subjects	mean age	rec. time [h]
All	21	29.9	10:22:30
Male	11	29.7	5:14:30
Female	10	30.1	5:08:00
Age <30	11	23.4	5:13:10
Age 30–40	7	32.0	3:37:50
Age >40	3	47.7	1:31:30

The language of the experiments was English, and all subjects were non-native, yet very experienced English speakers. More details on the subjects are summarised in Table 1. The level of interest (LOI) was annotated for every such sub-speaker turn. The original LOI scale reaching from LOI-2 to LOI+2 was mapped to [−1, 1] by division by 2. In the Challenge, the 21 speakers (and 3 880 subspeaker-turns) were partitioned into speaker-independent, gender, age, and ethnicity balanced sets for Train (1 512 sub-speakerturns in 51:44 minutes of speech of 4 female, 4 male speakers), Develop (1 161 sub-speaker-turns in 43:07 minutes of speech of 3 female, 3 male speakers), and Test (1 207 sub-speaker-turns in 42:44 minutes of speech of 3 female, 4 male speakers).

3.2.2 GEMEP

For emotion recognition, we selected the "Geneva Multimodal Emotion Portrayals" (GEMEP) (Bänziger, 2012). It contains 1.2 k instances of emotional speech from ten professional actors (5 f, 5 m). Using the same heuristic approach as in the INTERSPEECH 2013 ComParE Emotion sub-challenge, the 18 emotion categories were mapped to the two dimensions 'arousal' and 'valence' (binary tasks) such as to obtain a balanced distribution of positive / negative arousal and valence classes.

Table 2: *GEMAP by (pos(itive) / neg(ative); arousal (A) and valence (V).* Partitioning into training, development and test sets.

Sub-Task	#	Train	Devel	Test	Σ
AROUSAL	pos	280	100	220	600
	neg	282	108	210	600
VALENCE	pos	280	104	216	600
	neg	282	104	214	600

3.3 FEATURE SET

For feature extraction, we retain the choice of the ComParE set of supra-segmental acoustic features, which is a well-evolved set for automatic recognition of paralinguistic speech phenomena, as used for the baseline of the INTERSPEECH ComParE series. It contains 6,373 static features resulting from the computation of various functionals over

low-level descriptor (LLD) contours. The configuration file is included in the 2.1 public release of openSMILE (Eyben, 2010).

3.3 Evaluation on AVIC

For our empirical evaluations, we used the AVIC database as it contains the speakers' demographic data such as age group (20s, 30s, 40s) and gender. The evaluation metric is the Spearman's rank correlation coefficient (CC). As can be seen in Figure 3-4, the model performance curves for the different target groups consistently improve by adding new machine labelled instances to the training set. This substantiates our assumptions that first, the most reliable predictions are selected and thus the applied confidence measure is effective, and second, data aggregation and label enrichment can lead to enhanced recognition accuracy on the test data of the specific user group, thereby achieving the goal of long-term user adaptation. The different length of the learning curves is due to the varying size of the adaptation data sets. Further, it is noted that a different test set is used for each user group, thus although the initial model is saved and reused for each learning curve, the same starting point is not possible. Finally, as can be seen in Figure 5, it is more challenging to adapt the model to a single user although a slight raising trend can be observed for most of the users. One reason for the curves' instability and high variation among the users can be the ambigous nature of interest recognition, which is a typical subjective task in paralinguistics. The task of user identification is a challenge in itself and will be addressed in Section 5.

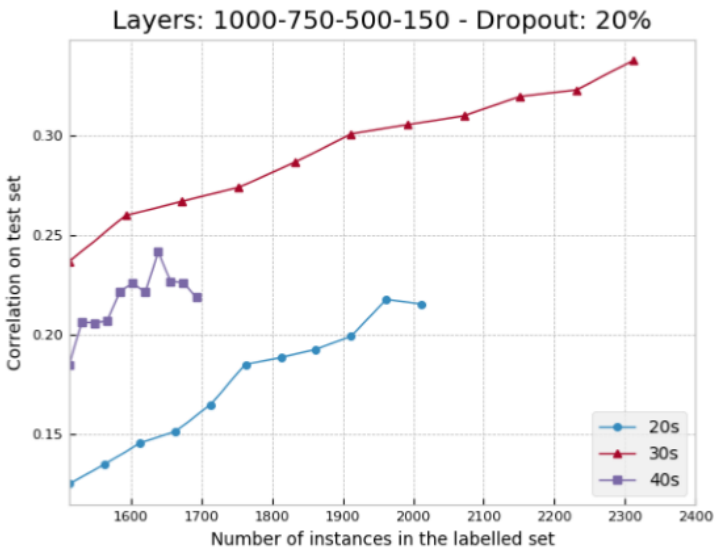

Figure 3: *Performance curve of the SSL-based adaptation technique to the different age groups on the AVIC database.*

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

August, 2017

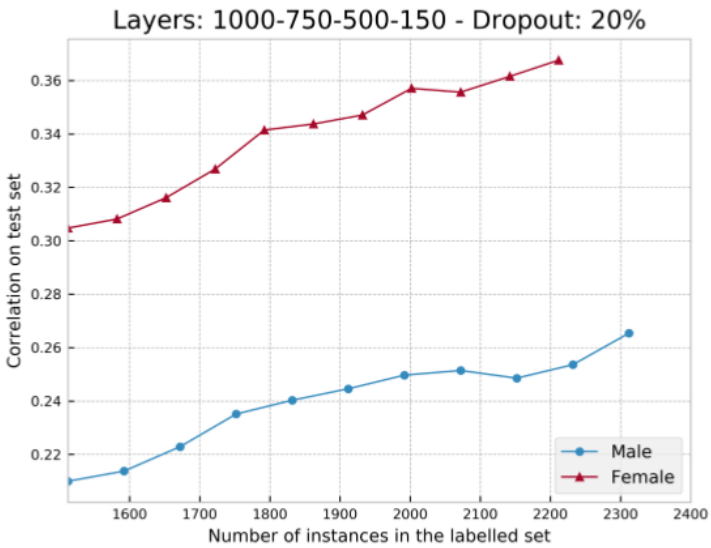

Figure 4: *Performance curve for adaptation to the gender groups on the AVIC database.*

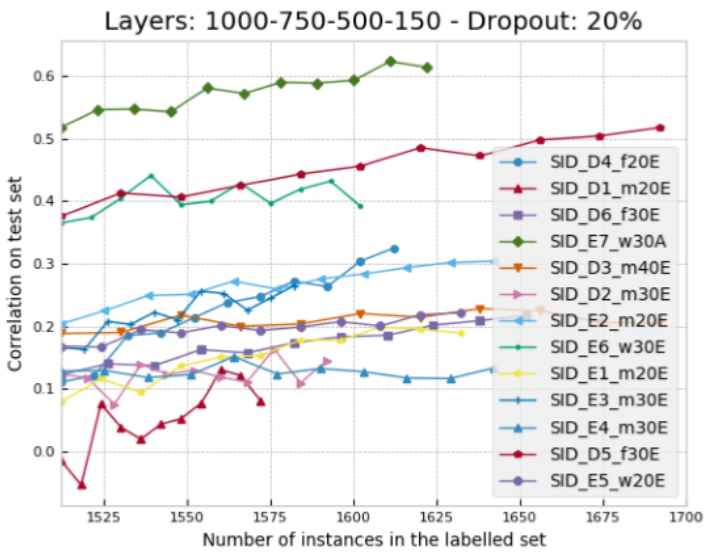

Figure 5: *Performance curve for adaptation to the speaker ID on the AVIC database.*

4. Real-Time Implementation (UA)

The real-time detection system of ARIA is implemented with the Social Signal Interpretation (SSI) framework (Johannes Wagner, 2013). For a detailed description of the system please refer to D.2.1. To enable user adaptation in real-time, the classification module of SSI had to be adjusted. So far, each classifier of a prediction pipeline was a single

model, which was loaded at start-up. At run-time, it was possible to halt and resume prediction by turning a classifier off and on, but not to switch between models on-the-fly.

The new implementation now allows it to assign to a classifier a list of models, where each model is tagged with a unique identifier. By default, the first model in the list will be used for prediction. However, at any time a message with name of desired model can be sent to the classifier to switch to another model of the pool. This can be either done manually, or by using the output of another classifier. In the latter case, SSI tries to match the name of the predicted label to one of the loaded models. If a fitting model is found that does not match with the current one, the models will be switched. The following figure illustrates the data flow:

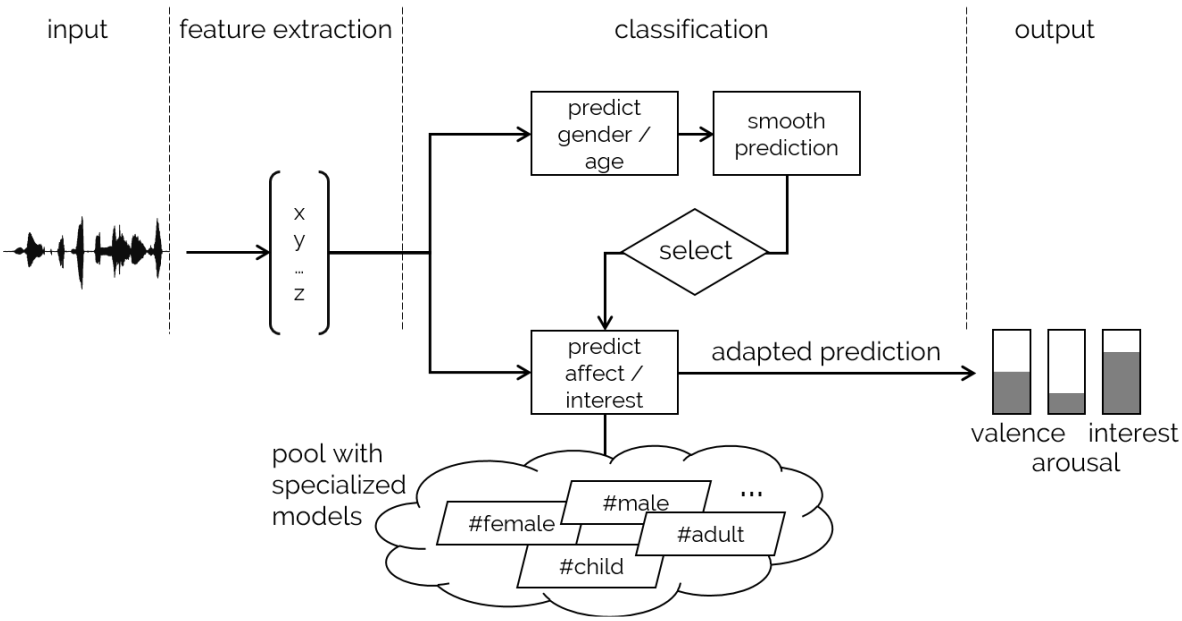

Figure 6: *Real-time Implementation of user adaptation into SSI.*

We can see that first audio chunks are first transformed into a compact feature representation. The feature set is shared by two types of classifiers: on the top, we implement conventional classifiers, which use a single model to predict static personality traits (here the gender and the age of the user). The predictions of the classifiers are received by a second type of classifiers (bottom), which operate on a pool of models. These models have been specialized for the gender (male, female) and age (child, youth, adult, senior) classes. If, for instance, a male voice is detected, but currently the female model is loaded, the classifier switches to the model specialized for male speakers. If we expect that users will interact with the system for a longer period (e. g. several minutes), an optional smoothing step can be added to avoid switching models back and forth in case of false predictions.

In a SSI pipeline, we implement the new behaviour by populating the path option of a classifier with a list of tagged models and setting it to receive the events of another classifier predicting the according classes. The following xml snippet shows an excerpt of the pipeline, where a classifier predicting the gender of the speakers is used to guide the selection of an arousal detector specialised on gender groups (changed spots are highlighted):

```
...
<!-- feature extraction -->

<consumer create="TupleEventSender" address="feature@audio">
       <input pin="audio8k" address="vad@audio" state="nonzerodur">
              <transformer create="OSWrapper" configFile="IS13.conf"/>
       </input>
</consumer>
<!-- static classifier
outputs 'female' or 'male'
-->
<object create="Classifier" path="gender" address="gender@audio">
       <listen address="feature@audio"/>
</object>
<!-- adaptive classifiers
switches between the specialized models 'arousal-f' and 'arousal-m'
-->
<object create="Classifier" path="female:arousal-f;male:arousal-m">
       <listen address="feature,gender@audio"/>
</object>
...
```

The same trigger mechanism we use to switch to a model specialized on a group of people, can also be used to adapt to a particular user (see Section 3.4). Of course, this requires that specialised models are available for the detected user. However, in the future we could implement a system that dynamically adapts to the user during an interaction. Therefore, we buffer the predictions (along with the feature vectors) and from time to time use them to retrain the models in the background. After the retraining is finished, we can use the existing interface to switch to the adapted models on-the-fly. This way, the system can even adapt to users that are not yet in the database. Of course, we have to make sure to reset the models when a new user enters the floor. Also, we have to be careful when selecting the instances used for retraining. Ideally, only instances predicted with a high confidence should be kept as otherwise noise will be introduced in the training process.

5. Visual User Identification (UN)

The visual system of ARIA incorporates a face tracking system (E. Sánchez-Lozano, 2016), described in ARIA D2.2, Section 3.2). The tracking system basically estimates, for each frame taken from the video stream, a sparse set of 66 fiducial points, meant to meaningfully locate key parts of the face, such as the mouth, the contour, the eyes or the nose. These points are used as input for the emotion classification system (described in ARIA D2.2, Section 3.2).

The User Identification system builds on the work of (Déniz, 2011), and takes as input both the current image and the tracked points, and then extracts some local information within local patches surrounding each of the points. The local information is a high-level representation of the image features, meant to be robust to changes in rotation, scale, and illumination, and it is based on a Histogram of Oriented Gradients (Dalal, 2005), described in ARIA D2.2, Section 3.2). The HOG features are concatenated to form a column vector, resulting in the representation of a face for a given frame.

These features are extracted for each frame, and stored in a sliding window of 50 frames. For each sliding window, a mean feature vector, and a covariance matrix, are stored, describing the temporal features for the current user.

The incremental CCR algorithm (iCCR) includes tracking lost/failure detection, which is activated each time the quality of the estimated landmarks is under a certain threshold. The lost detection is activated either when the fitting (the landmark estimate) is poor due to a large movement, or when the user is no longer within the camera's range of visibility.

In these cases, the tracker needs to be restarted, by first applying a face detection step (i.e., locate whether there is a face in the current frame, and where). When a new face is detected and the facial points are estimated again, the feature extraction process is repeated. The new set of features, collected after a restart, are now compared to the stored mean and covariance features. The comparison is done via the Mahalanobis Distance, which basically indicates the likelihood of these features to have been generated by the same Gaussian distribution than that defined by the mean vector and covariance matrix. If the distance is under a certain threshold, the features can be said to belong to the previous user. Otherwise it is assumed a new user is now considered. In such a case, the mean vector and covariance matrix belonging to the previous user are stored and referred to as the person-specific statistics for "user 1", and a new storage process starts. The process is depicted in Figure 7.

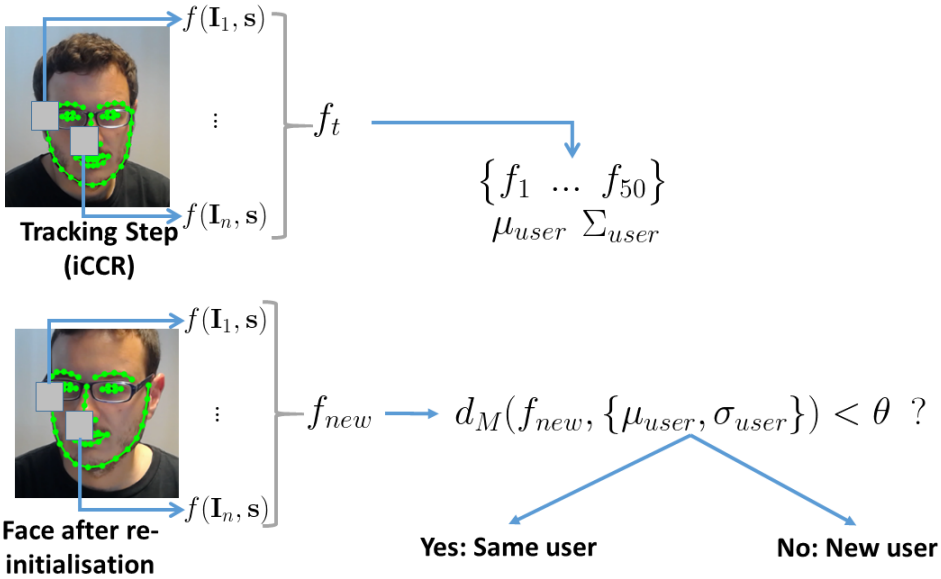

Figure 7: User re-identification procedure: While the tracking is ongoing with no failure, the user-specific features are stored in a sliding window of 50 frames. When the tracker needs to be reinitialised, the new features are first compared to the statistics stored for the previous user (mean and covariance of the features). When the distance is higher than a threshold, the system detects a new user, and the data collection re-starts.

When the tracker needs to be restarted, the identification process is repeated, although the Mahalanobis Distance is now obtained not only with respect to the previous user, but also with respect to all the previous users. That is to say, the new features are compared to all previous stored users. The user corresponding to the lowest distance is considered as the most likely to have taken over the tracking, considering the corresponding distance to be under a threshold. If none of the distances are under the threshold, a new user is added to the current session.

For long-term user identification, the system will allow the user to store the person-specific data collected during a specific session. Further to this option, it will incorporate an enrolling step, to allow the user to be fully identified each session. In this scenario, the system will not only know that a set of specific features belong to an existing user, but reach session, the system will try to identify the user from the stored database, and will assign a temporal intra-session ID to the user in case no existing user is found to be handling the system.

8. Plans for Next Period

In the next period, we aim for several achievements. First, after the adaptation approach has been validated on AVIC for regression and implemented in SSI, we will apply it to the classification task, i. e., emotion recognition on GEMEP. For ASR, we will add confidence

measures for the transcriptions and integrate the French models. Finally, to further advance multi-modal audio-visual affect recognition, we will investigate Deep Neural Networks (DNNs) with pretrained layers.

9. 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 31 months of the project.

Machine Learning

[ML1] Yue Zhang, Yifan Liu, Felix Weninger, and Björn Schuller. Multitask deep neural network with shared hidden layers: Breaking down the wall between emotion representations. In *Proc. of ICASSP*, New Orleans, LA, March 2017. IEEE.

[ML2] Eduardo Coutinho and Björn Schuller. Shared acoustic codes underlie emotional communication in music and speech - evidence from deep transfer learning. PLOS ONE, 12(e0179289):1–24, June 2017.

Paralinguistics

[PL1] Maximilian Schmitt, Erik Marchi, Fabien Ringeval, and Björn Schuller. Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices. In *Proc. of ITG Symposium on Speech Communication (ITG SC)*, pages 264–268, Paderborn, Germany, 2016. VDE, IEEE.

[PL2] Björn Schuller, Stefan Steidl, Anton Batliner, Elika Bergelson, Jarek Krajewski, Christoph Janott, Andrei Amatuni, Marisa Casillas, Amanda Seidl, Melanie Soderstrom, Anne Warlaumont, GuillermoHidalgo, Sebastian Schnieder, Clemens Heiser, Winfried Hohenhorst, Michael Herzog, Maximilian Schmitt, Kun Qian, Yue Zhang, George Trigeorgis, Panagiotis Tzirakis, and Stefanos Zafeiriou. The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Proc. of INTERSPEECH*, Stockholm, Sweden, August 2017. ISCA.

[PL5] Yue Zhang, Felix Weninger, Boqing Liu, Maximilian Schmitt, Florian Eyben, and Björn Schuller. A paralinguistic approach to speaker diarisation. In *Proc. of ACM International Conference on Multimedia*, Moutainview, CA, October 2017. ACM. to appear.

[PL6] Yue Zhang, Felix Weninger, and Björn Schuller. Cross-domain classification of drowsiness in speech: The case of alcohol intoxication and sleep deprivation. In *Proc. of INTERSPEECH*, Stockholm, Sweden, August 2017. ISCA.

[PL7] Zixing Zhang, Felix Weninger, Martin Wöllmer, Jing Han, and Björn Schuller. Towards intoxicated speech recognition. In *Proc. of International Joint Conference on Neural Networks (IJCNN)*, pages 1555–1559. IEEE, 2017.

References

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A. and Konosu, H., 2009. Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. *Image and Vision Computing Journal, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, 1760-1774.

Bänziger, T., Mortillaro, M. and Scherer, K.R., 2012. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *American Psychological Association*, 1161–1179.

Baur, T., Damian, I., Lingenfelser, F., Wagner, J. and André, E. 2013. NOVA: Automated analysis of nonverbal signals in social interactions. *International Workshop on Human Behavior Understanding* (pp. 160-171). Springer.

Dalal, N. and Triggs, B., 2005. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition (CVPR)* (pp. 886-893). San Diego, CA, USA: IEEE.

Déniz, O., Bueno, G., Salido, J. and De la Torre, F., 2011. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 1598-1603.

Sánchez-Lozano, E., Martinez, B., Tzimiropoulos, G. and Valstar, M., 2016. Cascaded Continuous Regression for Real-time Incremental Face Tracking. *European Conf. on Computer Vision (ECCV)* (pp. 645-661). Amsterdam, The Netherlands: Springer.

Eyben, F., Wöllmer, M. and Schuller, B., (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1459--1462). Florence, Italy: ACM.

Gal, Y., 2016. *Uncertainty in deep learning.* PhD thesis, University of Cambridge.

Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F. and André, E., 2013. The social signal interpretation (SSI) framework – multimodal signal processing and recognition in real-time. *Proceedings of the 21st ACM international conference on Multimedia* (pp. 831--834). Barcelona, Spain: ACM.

Rosenberg, C., Hebert, M. and Schneiderman, H., 2005. Semi-supervised self-training of object detection models. *Proceedings of IEEE Workshop on Motion and Video Computing* (pp. 29-36). Breckenridge, CO: IEEE.

Zhang, Y., Weninger, F., Michi, A., Wagner J. André, E., Schuller, B. 2017. A Generic Human-Machine Annotation Framework Using Dynamic Cooperative Learning with a Deep Learning-based Confidence Measure. *IEEE Transactions on Cybernetics*, submitted.