ARIA-VALUSPA Full-term Report
– Public version –

Michel Valstar, Elisabeth André
Matthew Aylett, Tobias Baur
Merijn Bruijnes, Angelo Cafaro
Chloé Clavel, Eduardo Coutinho
Soumia Dermouche, Guillaume Dubuisson Duplessis
Laurent Durieu, Shashank Jaiswal
Dirk Heylen, Peter LaValle
Catherine Pelachaud, Blaise Potard
Enrique Sanchez Lozano, Björn Schuller
Mariët Theune, Johannes Wagner
Jelte van Waterschoot, Yue Zhang
April 10, 2018

Contents

1 Executive Summary

This final report describes the activities conducted during the 36-month ARIA-VALUSPA Horizon 2020 project. We will restate the main goals, describe our noteworthy outputs technical and scientific, detail our public engagement and impact activities, dwell on the challenges and unresolved issues faced during the project, and provide a task-by-task report of work-package activities.

Please let us start by restating our main goals. Task-specific AI is attaining super-human performance in an increasing number of domains. In the near future, virtual humans (VHs) will be the human-like interface for increasingly capable AI systems, in particular information retrieval systems.

However, there remains a large gap in the smoothness of the interaction between interacting with either a current VH or another human being. In ARIA-VALUSPA we aim to drastically reduce this gap.

This means first and foremost that interacting with the ARIA-agents should be engaging and entertaining. They should display interactive believable behaviour that feels real. They should be adaptive to the user at various levels, from adapting to a user's appearance, age, gender, and voice, to sudden changes in the dialogue initiated by the user.

Some particular challenges that we set ourselves in the project were to deliver a reusable framework that can be used to create Virtual Humans with different personalities, behaviours, and underpinning knowledge bases. We have done so through the *ARIA-VALUSPA Platform (AVP)*, of which the latest version is 3.0. It is described as a noteworthy output of the project in section 2. AVP is in principle independent of the language spoken by the user. We show this by delivering the assets of the system in three languages - English, French, and German.

Another important challenge that we set ourselves is to be able to deal with unexpected situations, in particular interruptions initiated by the user. This is a hard problem that has not been addressed previously. Interruption handling is integrated throughout the framework, enabling the detection of interruptions, planning new utterances when an interruption by the user has been detected, and abruptly stopping and replacing behaviour generation when necessary. The scientific breakthroughs to this are described in section 3.9, but aspects of interruptions feature in most work-package reporting.

A challenge in Text to Speech (TTS) systems is creating smooth, natural sounding voices with affect. One of our biggest achievements is the development of an affective TTS that can turn neutral speech into emotional speech using a markup language to markup the original (neutral) text. Through working closely with the visual behaviour generation team, seamless lip synchronisation has been achieved.

The behaviour analysis systems, from audio, video, or using combined audio-visual features, is truly state of the art and has progressed markedly over the period of the project. Integrated in the Social Signal Interpretation (SSI) framework, it includes the fastest and most accurate face tracker, state of the art re-trainable ASR, emotion, gender, and age estimation, and a number of other features.

In terms of impact, we have consistently engaged with the general population through a series of blog-posts on our webpage (`https://aria-agent.eu`), and through open-science events such as the London Science Museum's Lates, and the University of Nottingham's Wonder festival. In terms of academic impact, the project has led to 100 peer-reviewed academic papers, which have already attracted more than 1,000 citations between them[1]. More than 20.5 M EURO has been awarded in follow-up funding.

[1] As measured by Google Scholar on 23 February 2018

2 Noteworthy outputs

A number of outputs of the project are particularly worthwhile, and are highlighted below in some detail. Note that these outputs often bring together contributions from more than one work-package and more than one task, and as such their full value would not come to justice if read from the individual work-package report sections.

2.1 AVP: ARIA-VALUSPA Platform for Virtual Humans

The ARIA-VALUSPA Platform from Virtual Humans (AVP) is a general-purpose, modular software platform for the creation of virtual humans. It is developed in this project based on the lessons learned from the SEMAINE project and the Virtual Human Toolkit. It is free of use for non-commercial purposes, with large parts available as open source and the remainder as publicly available compiled library interfaces.

AVP is currently at version 3.0 and is available from GitHub: `https://github.com/ARIA-VALUSPA/AVP`, where you can also find all installation instructions and detailed documentation for the various modules and their configurable settings.

2.1.1 AVP Architecture

AVP is essentially an architecture for interconnected modules each with a specific functionality, which run independently but communicate through an ActiveMQ layer to update the agent's state and ultimately generate the most relevant behaviour when interacting with a user. Fig. 2.1 shows the high-level architecture of AVP. It consists of three mayor blocks: an input or *Behaviour Sensing* block, an *Agent Core* block, and a *Behaviour Generation* block. Blocks can and do consist of multiple modules. For example, the Behaviour Sensing block consists of the Automatic Speech Recognition module (ASR), the visual analysis module (eMax), and the audio-based paralinguistic analysis module (OpenSmile).

The three blocks can be briefly described as:

- The Behaviour Sensing block: responsible for collecting and processing audio-visual data about the user in terms of spoken words, recognised emotion, and estimated identity and demographics. This data would then be fed to the Agent Core.
- The Agent Core block: responsible for receiving the Input module's data, analysing it and deciding on the agent's response and behaviour. This module is also in charge of maintaining an information state that captures the agent's knowledge about the world. Finally, it is responsible for feeding back information to the Input module when incorrect information is detected.
- The Behaviour Generation block: responsible for determining fine-grained expressive behaviour, rendering the Character and playing the speech through a text to speech (TTS) component, and ensuring that the speech and character animations are synchronous and believable.

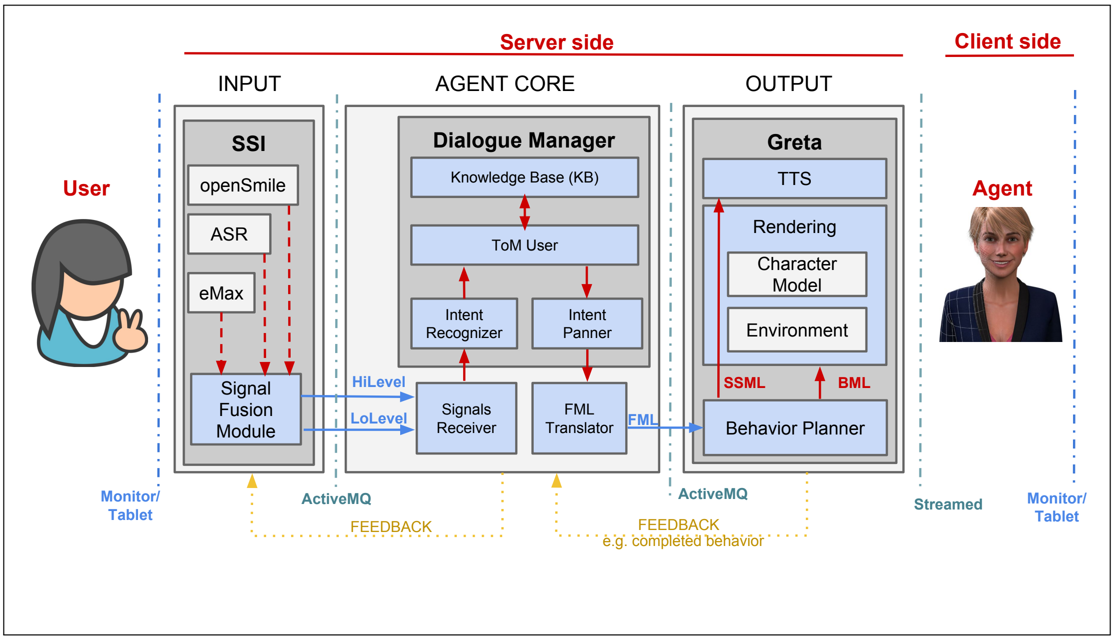


Figure 2.1: The ARIA-VALUSPA Platform (AVP) architecture.

The three blocks use ActiveMQ as a message broker for communication. This architecture allows us the flexibility of using different implementation languages and even deploy multiple machines to run the different modules, with the added overhead that all modules should interface with the ActiveMQ broker. One significant example of this modular approach is that we are actively developing three Graphics Behaviour Generation modules, that we can use to easily switch between a more polished-looking but animation-driven agent (Living Actor) with more limited behaviour or one that is more suitable for researching the minutiae of expressive behaviour (Greta). The third graphics realiser is the popular Unity3D graphics engine, which allows seamless integration of ARIA agents in games, augmented reality and virtual reality projects. Note that all three visual behaviour generation options require Greta to run to interpret FML from the Agent Core block.

In order to demonstrate how the ARIA Framework can be used to address the reality of the user's needs the project has delivered a number of instances of two use cases. One use-case is a smart and interactive book reader, and the other is an end user support system for advice on stain removal. If the first use-case - called the Book-ARIA - is an academic implementation of the ARIA Framework, the second use-case - called the Industry-ARIA - is backed by a large corporate entity. For sake of confidentiality their identity, and most details of the Industry-ARA, is removed from this public version of the final report.

2.1.2 AVP Implementation

Each of the three blocks (Behaviour Sensing, Agent Core, and Behaviour Generation) are run as separate binaries, and communicate using ActiveMQ. Below we describe the three blocks in some detail:

Behaviour Sensing: We use the Social Signal Interpretation (SSI) framework, developed at the University of Augsburg, to grab the audio-visual input of the user. This allows us to have a layer of abstraction from the raw input, but also ensures that different sub-modules will be synchronised even if they need different amounts of time to process a frame/period. SSI also collects the output of different components into a single XML file, which can be sent over ActiveMQ to the Agent Core module.

The behaviour sensing block was built by integrating three different processing components, as shown in Fig. 2.2: eMax, Kaldi (ASR), and OpenSmile. Each component is in charge of three different aspects of the interactions between humans and agents:

1. **Visual Behaviour**: eMax is used for visual analysis of a user's facial appearance and expressive behaviour. Processing consists of face detection, face recognition, facial point detection and alignment, estimation of head-pose, and finally recognition of the displays of six basic expressions and a number of facial muscle actions (FACS Action Units [18]). In addition, age and gender are estimated from the face shape and appearance. Due to processing power constraints, a fixed framerate of 5 video frames per second was used. After processing a frame, the basic expression output consists of the 6 basic emotions (anger, disgust, fear, happiness, sadness, surprise), which are further processed to provide a prediction for valence and arousal. Deep-learning based face frontalization will soon be added.

2. **Audible Behaviour**: Opensmile is used to recognise interest, gender, age, and emotions from audio. Arousal and valence are reported as values between 0 and 1, with 0.5 being neutral. Interest is reported as values between 0 and 1, 0 indicating low interest and 1 high interest. Age is reported as a probability of the user belonging to one of four categories (child, youth, adult, senior), while gender is reported as the probability of the user being either male or female.

3. **Speech Recognition**: Kaldi is used for automated speech recognition (ASR) into SSI. Unlike the rest of the framework, the ASR must run on a separate Linux server, as there is no Windows OS version available. To maintain consistency, audio is still recorded through SSI, and as a result we are opening a direct port between the machine running SSI and the machine running Kaldi, through which audio is sent and a transcript is received. The ASR outputs predictions in two versions: at word-level and sentence-level. Word-level prediction returns predictions faster but with a higher error, as it lacks the opportunity to use between-word correlation. Speech is recognised in three languages: English, French, and German.

Each 500 milliseconds a summary of the latest output of each component is added to an XML file and it this is sent through ActiveMQ for use by the Agent Core and Output modules.

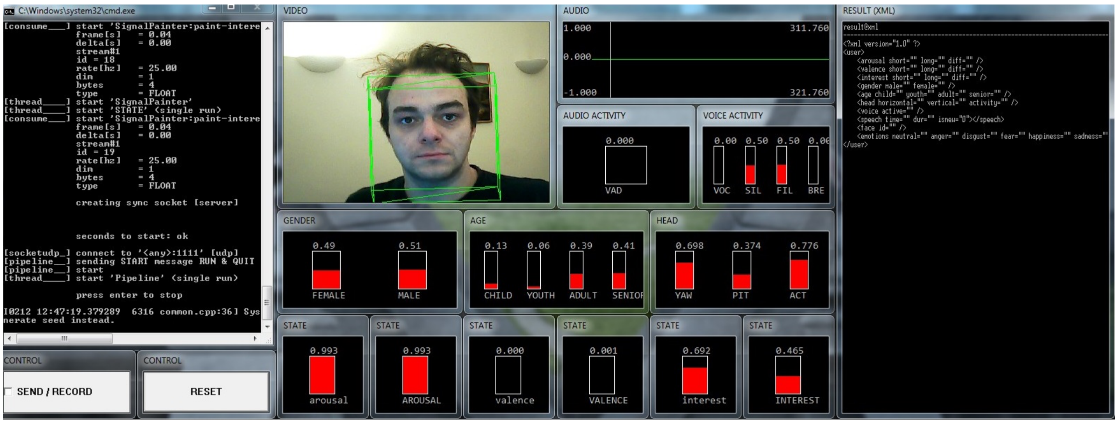


Figure 2.2: Input Module Visualisation

Agent Core: This module is implemented in Java. It contains the dialogue manager, the knowledge base, the information state, and an ActiveMQ receiver that can parse the XML files sent by the Input block. This ActiveMQ receiver converts the information contained in an XML file into objects that can be used by the dialogue manager component.

The dialogue manager was implemented using Flipper 2.0, itself based on Flipper [43], a template-based library for specifying dialogue rules for dialogue systems. Based on the user's reported emotions, gender, age, and speech transcript a response would be composed in form of a BML file, which contains the agent's behaviour and response. This is passed using ActiveMQ to the Output block where either Greta or Living Actor are responsible for interpreting and displaying the agent's response. The templates have tags that allow the agents to generate behaviours that communicate different emotions for the same utterance.

For testing purposes, a separate window was added, in which the user's speech can be written as text, was added. This text is set as the transcript to the XML file received from the Input module, as such there is no perceivable difference to a real transcript. This can also be used to monitor the conversation between the two parties. An example of this window is shown in Fig. 2.3.

Behaviour Generation: The visual behaviour is prepared by Greta, and realised (turned into graphics) by either Greta, developed by CNRS, Living Actor, developed by Cantoche, or Unity, the 3D environment development kit of choice for researchers and professionals alike. Both Greta and Living Actor instantiate a character capable of interpreting Behaviour Markup Language (BML) files. Unity uses a combination of Mpeg-4 FAP and BML, both produced by Greta. The speech synthesis is realised by CereProc. Responsibility for activating the speech synthesis and maintaining synchronisation of the visual and auditive behaviour generation (in particular lip movements) rests with both Greta and Living Actor, separately and individually.

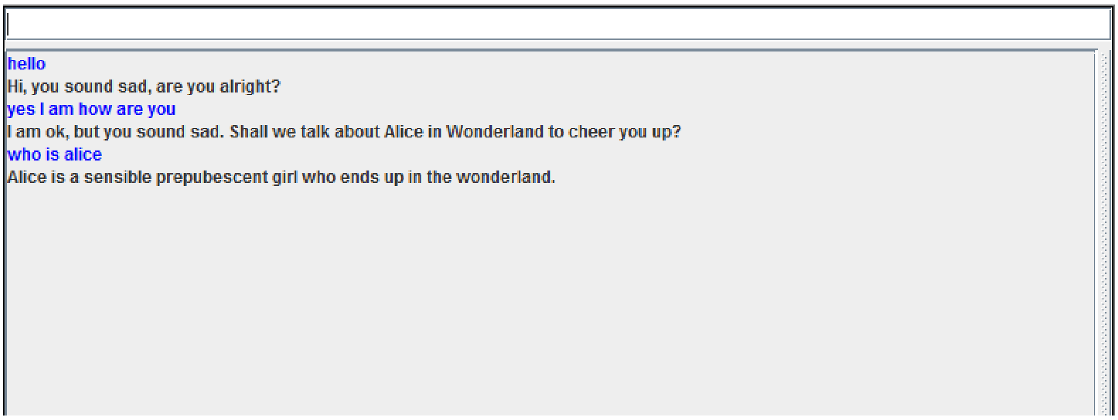


Figure 2.3: Example conversation shown in monitor.

2.1.3 System Launcher

As the system is composed of three different windows binaries (Behaviour Analysis, Dialogue Management and Behaviour Generation) and a Linux binary (ASR) that communicate using ActiveMQ, a launcher program is required to start and stop the whole system. The launcher first ensures that ActiveMQ is started and ready to use, then starts the Input block, followed by the Agent Core, followed by the Output block. It also ensures an easy way to stop the whole system.

The launcher has a small number of XML configuration files, which include the location of the binaries, which Behaviour Generation version to use (Greta or Living Actor). A separate configuration file is maintained for each block, with two variants currently available for the Output block. Having separate configuration files for each block allows us to dynamically start and stop whole blocks while the rest of the Framework continues running. Figure 2.4 shows a screenshot of the graphical launcher interface.

2.2 ARIA instantiations

One of the project's main aims is to be able to create different Virtual Humans, with different personalities, behaviours, and knowledge-bases, all based on the same framework. To proof this, we have developed the Book-ARIA and the Industry ARIA, described in the sections below. In addition, we have developed a version for the Unity 3D programmable gaming environment.

2.2.1 Book-ARIA

From the onset of the project a virtual human representing the characterization of a novel has been developed. The large number of public domain novels that can be adapted means there is an immense potential for the creation of very rich and diverse agent personalities. Moreover, the Book-ARIA is believed to have commercial value in its own right. More

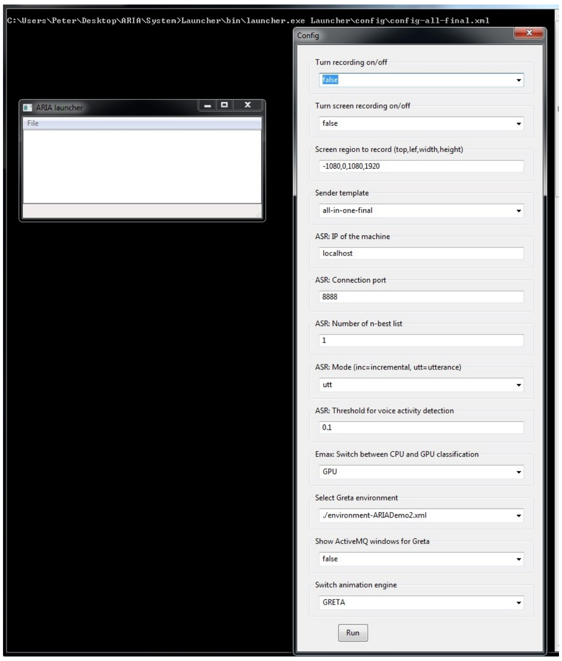


Figure 2.4: Graphical user interface for the AVP system launcher

generally, the Book-ARIA functions as a showcase of what rich personalities can be generated with ARIA-VALUSPA and how they function as interfaces for information retrieval for more complex tasks, that is, questions about the novel's content, characters, author, etc. For the purpose of this project, the novel *Alice in Wonderland* by Lewis Caroll has been selected as an illustrative example.

Alice has been created by Cantoche for WP5 to be the face of the Book-ARIA, and she quickly became the effigy of the ARIA-VALUSPA project, featuring in the project's logo and generally becoming the poster girl for all outward facing activities. Some examples of the Living-Actor version of Alice are shown in Fig. 2.5

To give life to Alice, we use the Living Actor technology which allow to convert a 3D mesh to an animated object that could be used in videos, html pages as well as 3D applications.

Proof of Concept The Book-ARIA subject has been split in two steps. First a Proof of Concept (PoC) was made using existing technologies and solutions, and this has then been followed by an implementation in the ARIA Framework. The PoC consists of an HTML page containing an animated character and an extract of the "Alice in Wonderland" novel (see Figure 2.6). The PoC is available online from `http://www.livingactor.com/clients/ARIABook/`. On the request of the user, the avatar starts reading the novel, her voice is dynamically generated by Cereproc using their text to speech (TTS) API. The user can interrupt the speech when he wants in order to ask a

Figure 2.5: Samples of Alice expressive posing.

question to the avatar about the author, a character, or a chapter, and then resume the reading where he stopped. The question answering technology used to interact with the user is set up by Cantoche.

PoC limitations In the PoC, we use an HTML5 compliant avatar. Even though the Living Actor Avatars are built in 3D, we use pre-generated 2d animations due to device technical limitations. The HTML5 Living Actor Avatar is not a video, but an animated object which can be controlled by JavaScript in order to trigger a selected animation at a specific time and for a specific duration. Even though it's sufficient for the PoC, the project has more ambitious goals that can not be reached with this kind of solution. For example, the avatar has pre-generated animations that limit its possible behaviours. Lip synchronisation with the speech, and interruptions mid-animation are also issues that are inherently difficult to fix with the HTML5 avatar.

Main implementation The final implementation of the Book-ARIA is the Alice in Wonderland scenario, which allows users to chat with Alice about the book written be Lewis Carol. Two scenarios are available: a standard chat, without any particular goal for the user, and a 'quest' scenario, where the user is expected to retrieve personal information from Alice, which she will only divulge after her personal relation with the user is strong enough. See the deliverable on the system assessment (D6.4) for full details on how Alice and the user build a personal relation over time.

2.2.2 Industry-ARIA

At the end of year 1, several companies presented an application to be involved in the project. The management board has selected xxx as a partner, who presented the most

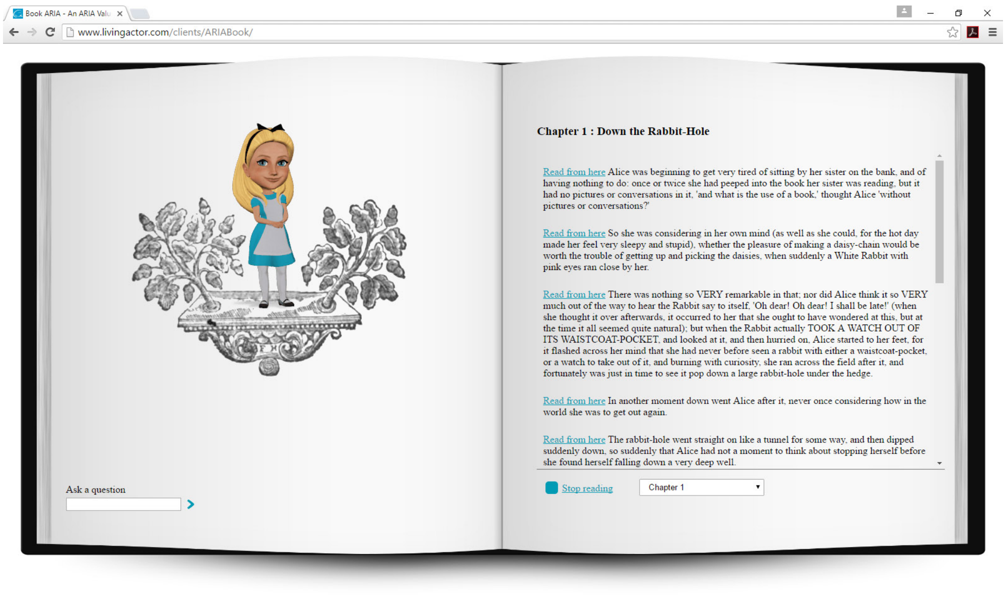


Figure 2.6: Screenshot of the Book-ARIA POC web page.

promising business case (See Deliverable D7.2 for a description of the selection process).

The Industry-ARIA is confidential and it, nor information about it, will not be made publicly available, but will be demonstrated at the final review meeting.

2.3 NoXi: Multilingual, Multimodal Database of Novice-Expert Interactions with Interruptions

An important contribution of the ARIA-VALUSPA project is the NoXi database, now consisting of two main parts: one a set of Human-Human mediated Novice-Expert interactions, and the other a set of Human-Agent interactions with the AVP system. NoXi was designed to provide spontaneous interactions with emphasis on adaptive behaviours and unexpected situations (e.g. conversational interruptions) and was recorded with the aim to be of wide use, beyond the direct goals and aims of the ARIA-VALUSPA project.

The resulting NoXi database was published and presented with a poster at the International Conference on Multimodal Interaction in November 2017, in Glasgow, UK [13]. This presentation described only the Human-Human partition. A presentation on the full database including Human-Agent interactions should follow. The poster presentation drew a lot of interest, both in the database and in the NoVa annotation tool that we used to create annotations for it (see section 2.4). At the time of writing, the NoXi database has 22 users, 12 of which are from outside the ARIA-VALUSPA consortium.

The Human-Human part of the database consists of 84 dyads recorded in 3 locations

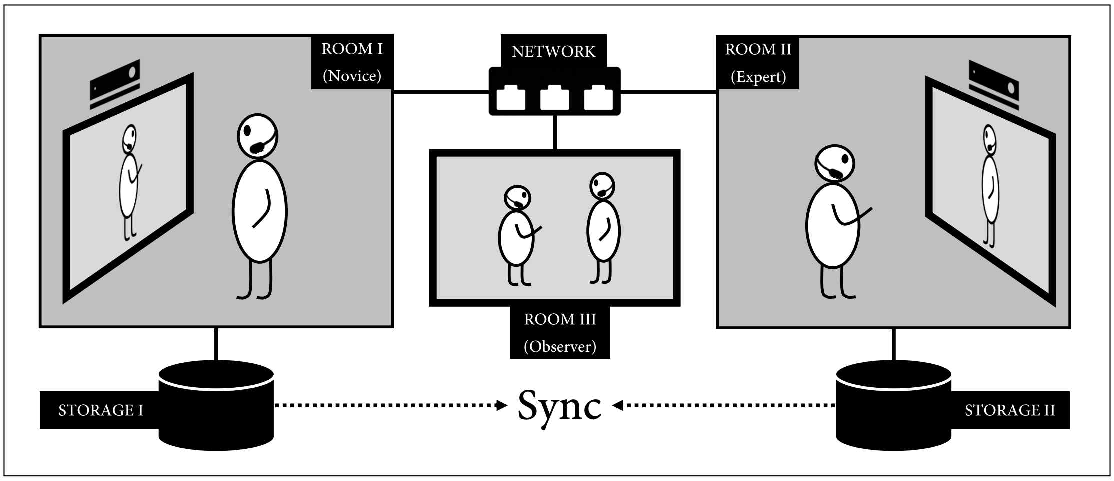


Figure 2.7: NOXI Recording setup: Novice (left) and expert (right) having a screen-mediated conversation. The interaction is monitored (middle) and recorded in sync.

(Paris, Nottingham, and Augsburg) spoken in 7 languages (English, French, German, Spanish, Indonesian, Arabic and Italian). Expert/Novice pairs discussed 58 wildly different topics and more than 25 hours of synchronized audio, video, and depth data was collected. Efforts have been made and are currently ongoing to add semi-automatic annotations to this data. See D6.2 for a full description of the Human-Human part of the NoXi database.

The Human-Agent part of the database consists of ... recorded in Nottingham, spoken in English. The same room and sensor setup was used as for the Human-Human interactions. See D6.4 for a full description of the Human-Human part of the NoXi database.

We have also added a collection of early Wizard of Oz interactions between a human and a agent controlled by a wizard, which was used to explore interactions with agents in the Book-ARIA domain (described in detail in D6.1 and below in section 6.3.5). This data, called the HAI data, is not structured in the same way as the NoXi Human-Human and Human-Agent partitions. Instead, a link to a single archive file is provided on the NoXi database website.

NoXi is made freely available to the research community and for non-commercial uses. It is available through a web interface at: https://noxi.aria-agent.eu/ after users create an account and sign the EULA.

Figure 2.7 shows a sketch of the NoXi recording setup for the Human-Human and Human-Agent interactions. We see that participants are located in separate rooms having a screen-mediated conversation. To this end, audiovisual data is streamed from one room to the other and replayed on a screen (a third end-point was added to silently monitor the recorded streams). We used Microsoft Kinect 2 to capture various signals from each participant including HD video, depth data, skeleton and face tracking. In addition we used high-quality head mounted microphones to obtain clean speech recordings. This summed up to a bandwidth of 9.3 GB per minute and user (1.4 GB after compression).

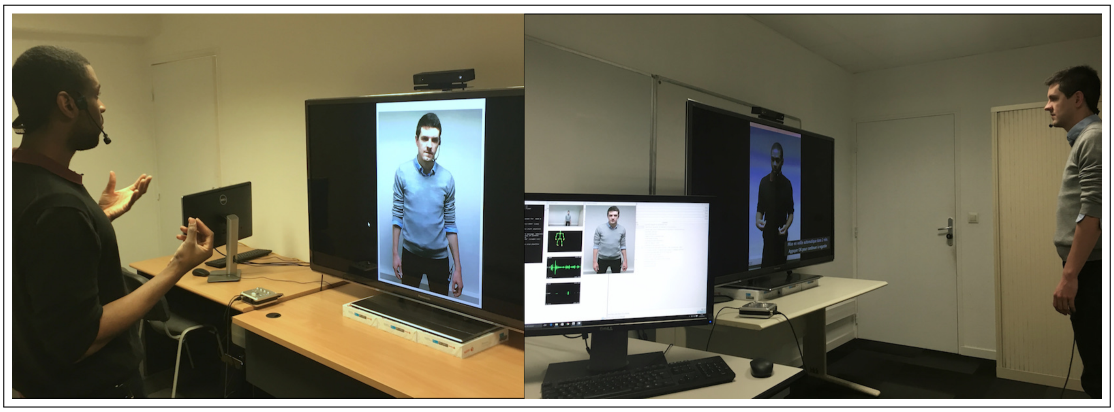


Figure 2.8: Snapshots of a novice-expert dyad in a recording session.

To keep recorded signals in sync we implemented the system with the Social Signal Interpretations (SSI) framework developed at the Augsburg University [50]. Figure 2.8 shows a snapshot of the recording.

However, NoXi not only contains a massive amount of raw interaction, but also comes with a large number of annotations. Some of the annotations have been created completely manually, while other were derived in fully or at least semi-automated way. Annotations accomplished so far range from speech and filler transcriptions, over body movements and facial features, to affective dimensions and interest scores. In the following section we will introduce a tool that we have developed to accomplish this task.

2.4 NOVA: Multimedia Annotation Tool with Integrated Machine Learning

To handle the vast amount of data in the NoXi database within the limited time span of the project, we opted for a collaborative and semi-automated workflow. Since no tools were available that would suit our needs, we decided to implement a novel annotation tool: NOVA ((Non)Verbal Annotation) for the Windows OS under the GPL 3.0 license. This tool has now been made public to the general research community, and is available on GitHub here: `https://github.com/hcmlab/nova`. A paper describing it is currently under peer review, but while presenting NoXi at ICMI 2017 in Glasgow, there was already a lot of interest in NoVa.

The main features of NOVA are:

1. Support for multiple annotation schemes (e.g. discrete labels vs. continuous scores).

2. Support for viewing content beyond audiovisual media (e.g. visualisation of tracking information).

3. Database back-end to centrally store annotations and access data from multiple sites.

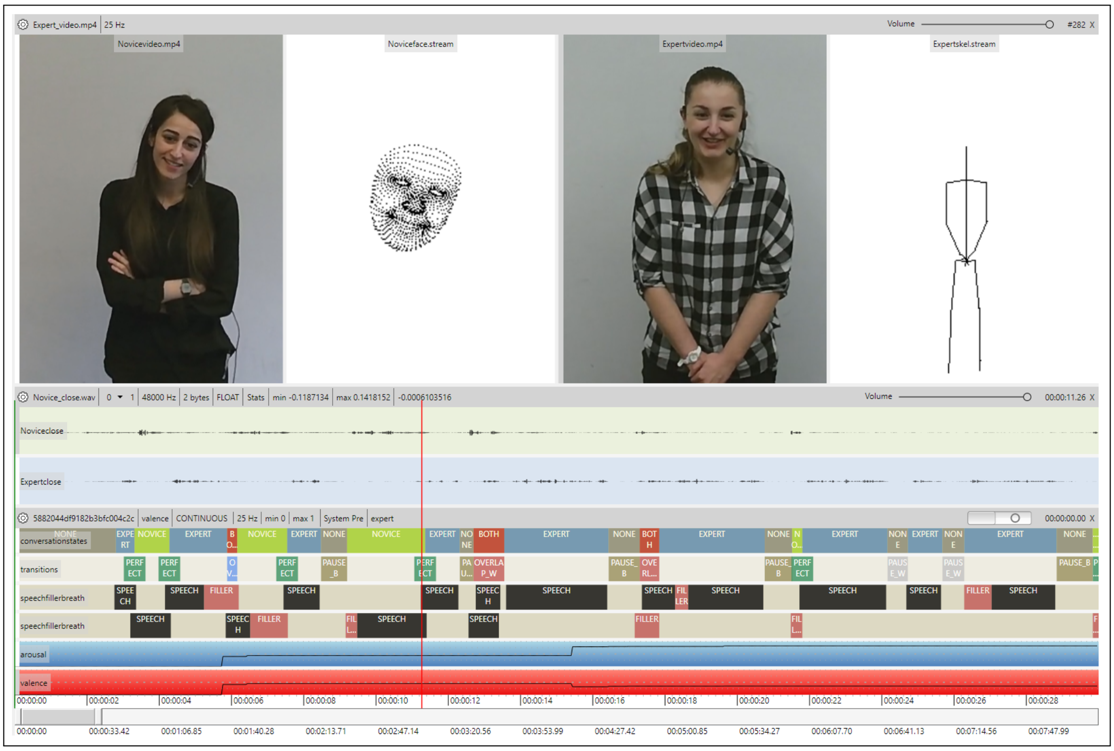


Figure 2.9: NOVA allows it to visualise various media and signal types and supports different annotation schemes. From top down: full-body videos along with skeleton and face tracking, and audio streams of two persons during an interaction. In the lower part several discrete and continuous annotation tiers are displayed. Annotations can be edited on a static fraction of the recording or interactively during playback.

4. Advanced user management to share annotation tasks among multiple raters (including strategies to combine annotations of several users).

5. Access to machine learning tools to create semi- and fully-automated annotations on the fly.

Figure 2.9 shows the main interface of NOVA. We can see that apart from audiovisual content, also facial and body tracking data can be displayed. NOVA is also not limited to a specific annotation scheme, but supports time-discrete and time-continuous annotations either based on a set of pre-defined labels or within a value range. There is no limitation on the number of data and annotation tracks that can be loaded and edited.

The real power of NOVA, however, is the full support of a cooperative machine-learning work-flow as shown in Figure 2.10. In fact, the meaning of *cooperative* is two-fold. On the one hand, NOVA allows multiple annotators to work on the same database. User rights are centrally managed and allow users to display (and sometimes even edit) annotations of other users. If multiple annotations are available for the same content they can be

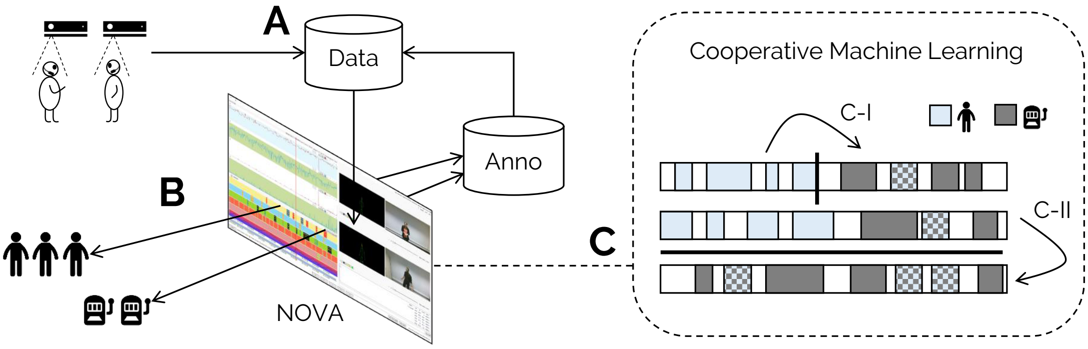


Figure 2.10: CML integration in NOVA: (A) A database is populated with recordings of human interaction. (B) NOVA functions as interface to the data and provides a database to distribute and accomplish annotation tasks among human annotators. (C) At times, CML is applied to automatically complete unfinished fractions of the database: (C-I) A session-dependent model is trained on a partly annotated session and applied to complete it. (C-II) A pool of annotated sessions is used to train a session-independent model and predict labels for the remaining sessions. In both cases, confidence values guide the revision of predicted segments (here marked with a pattern).

merged to establish a gold standard. On the other hand, users can draw on machine-aided predictions, too. For instance, a rater can ask NOVA to automatically complete a partially finished annotations. The tasks of first extracting features from the raw media files and afterwards learning a classification model are automatically handled by NOVA. A user only needs to choose from a set of available feature extraction and learning algorithms. Hence, using these tools in NOVA does not require a signal processing or machine learning background. Yet, skilled users can add their own feature extraction methods and extend NOVA with new learning algorithms.

NOVA has seen tremendous progress during ARIA-VALUSPA, with no less than 810 commits and 82 releases made since December 2017. It is now at version 1.0.1.8.

2.5 Idlak Tangle: a free DNN-based Text-To-Speech Toolkit

Statistical parametric speech synthesis based on Hidden Markov Models (HMMs) has become a common method for generating highly intelligible, flexible speech output. The dominant system, HTS [56], has been developed for over a decade, and led the way in developing parametric synthesis approaches and algorithms.

More recently, spurred on by the success of Deep Neural Networks (DNNs) in speech recognition [24], significant research has been carried out investigating the use of DNNs in parametric speech synthesis [31].

Idlak is a project to build an end-to-end parametric synthesis system within Kaldi [37], a liberally licensed Automatic Speech Recognition (ASR) toolkit. As part of Idlak, a front-end that generates full-context models compatible with HTS has been developed [4]. This front-end performed well in an evaluation against Festival, a standard front-end used by HTS. We have now released a system based on one of Kaldi's DNN frameworks as an alternative to the standard HTS/HTK modelling framework. We have called this end-to-end TTS-DNN system *Tangle*. Although other open source systems are available none offer a single framework with text normalisation and back-end processing in the same environment.

Idlak Tangle first uses Kaldi to carry out a phoneme alignment on a single-speaker corpus. This alignment is then used to train two cascading DNNs: one to predict unit durations, and a second for predicting acoustic output. Also incorporated are analysis and synthesis tools to perform MLSA vocoding with mixed excitation [54] and a simple recipe to encourage other research groups to reproduce our results. All the necessary code can be downloaded from the Kaldi-Idlak repository `https://github.com/bpotard/idlak`[2] allowing our results to be reproduced. Tangle only depends on tools that use either BSD (SPTK, expat, PCRE), Apache (Kaldi, openfst), or MIT (pugixml) licenses, allowing the use of Tangle for both commercial or academic applications. Tangle and Idlak are both released under the Apache license.

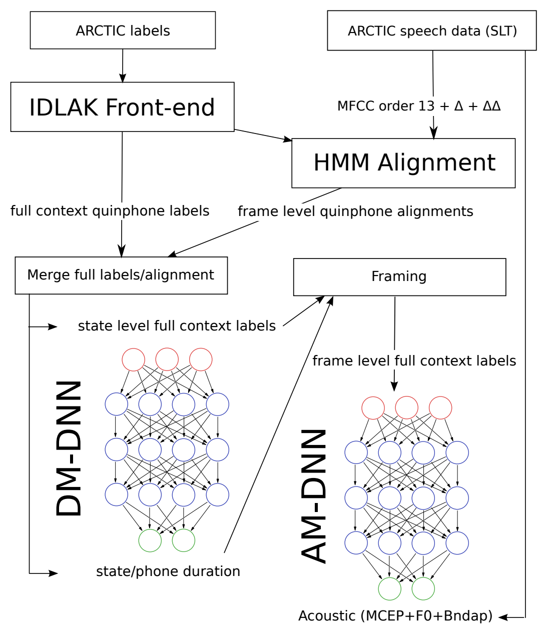


Figure 2.11: Tangle DNN training architecture

The primary motivation for our work can be summarised as follows:

[2]Currently the Idlak branch of Kaldi can be installed with `git clone https://github.com/bpotard/idlak.git`

1. It is part of a long-term goal to produce a Kaldi-based end-to-end parametric speech synthesis system. HTS suffers from licensing restrictions that prevent a standard open-source model. In addition, many new approaches in ASR are already implemented within Kaldi, such as sophisticated DNNs [49] and sub-space modelling [36]. This would allow current and future ASR developments to be directly incorporated into speech synthesis as they become available and vice versa.

2. High quality vocoders are a requirement for a good sounding parametric TTS system; however most of the available ones are either low quality or suffer from licensing restrictions that make them unsuitable to be included directly into an open source project with a liberal license. By re-implementing state-of-the-art vocoding techniques into Kaldi, we hope to bridge that gap and offer the first free high quality parametric Text-to-Speech system.

3. By making our system openly available, together with the tests we describe, we offer a useful test harness and a better sounding baseline than HTS-demo to the community.

3 Noteworthy Scientific and Technical Breakthroughs

Like the noteworthy outputs described in section 2, a number of scientific and technical breakthroughs are worth highlighting here separately.

3.1 Incremental Cascaded Continuous Regression for Real-time Face Tracking

The core of the visual part of eMax is the Face Tracking system, in which an indexed set of 66 points, representing the location of key parts, such as the nose, the eyes, the mouth, or the contour, are tracked along the video sequence. All the other visual blocks build on accurately localising these points. The field of Face Tracking has long been an active research topic, aiming to develop fast and accurate methods. The state of the art method for Face Tracking, capable of working in real-time, is the Cascaded Regression of Xiong and De la Torre [52], known as Supervised Descent Method (SDM), in which an initial guess or estimate of the shape (the facial points), is passed through a cascade of linear regressors, each taking as input some local information extracted from the image around the given points (the features), and outputting a displacement towards the target locations. These regressors are fixed, and are learnt from a set of images for which the points have been manually annotated (the training set). To train the first regressor, for each of the images the initial shape guess is computed, given as random variations on the ground-truth data according to some statistics modelling how shapes vary within consecutive frames. Then, the features are collected for the training images in the given initial shapes, and a regressor is learnt through minimising the least-squares error. This regressor is then used to update the initial shapes, and the process is repeated, until the incremental improvement is too small (typically after 4 or 5 iterations).

The main problem of this approach is that the training needs to be done sequentially, and the models cannot adapt to the user once they have been trained, as this would require re-training the models again, which can take up to several hours. It has been shown that models trained to track specific and known faces are more robust than generic models, trained on a fairly big training set of images. Given that it is infeasible to train person-specific models, it is a desired target to incorporate the current information from the user to the models as the tracking is ongoing. This is known as incremental learning. The sequential learning procedure of SDM impedes its application. However, it has been shown that an SDM can be trained in parallel [3], bringing the possibility of incremental learning for the first time, at a cost of 4 seconds per frame, still far from real-time performance.

Thus, as part of the research conducted in ARIA, we have developed a novel approach to learn each of the regressors, resulting in a very fast learning method that also enables the use of incremental learning in real time. Our new method, coined incremental Cascaded Continuous Regression (iCCR), applies a Functional Regression approach to the least squares problem, assuming the target variables (the points) to be part of a continuous space, and considering an infinite set of perturbations when training the model.

More specifically, instead of generating samples at the perturbed locations, Continuous Regression approximates all of them by a first-order Taylor expansion, effectively marginalising the perturbations from the feature extraction process. This expansion linearises the features with respect to the perturbations, and enables the possibility of integrating over an infinite set of perturbations, yielding a closed-form solution. Moreover, contrary to previous works on Functional Regression, the proposed formulation integrates the space of perturbations over a non-Euclidean manifold, in which the correlation between variables are considered, thus avoiding non plausible perturbations.

They key aspect of the solution resides in the fact that it only needs the features to be extracted at the ground-truth positions, as well as that it depends on the statistics of the displacements that would be applied in the sampling-based approach. This implies that all regressors in the cascade can be trained just by replacing the statistics, thus making the effective training time to be reduced to seconds. Once the features are extracted at the annotated positions, the training process is carried out very fast.

The fact that features need to be extracted only once, at the ground-truth solution, also enables the use of incremental learning in real-time. During tracking, once a frame has been correctly fitted, we need to extract the features at the estimated points; we don't need to collect samples as in [3]. This, along with a re-arrange in the recursive least-squares update, reduces the complexity of incremental learning in one order of magnitude with respect to the most expensive operation. This improvement brings the time complexity for incremental learning down from the 4 seconds per frame in [3] to 0.15 seconds in a Matlab prototype, allowing for the first time the use of incremental learning in real time. Its implementation in the eMax C++ library incorporates the incremental learning in parallel threading, keeping a faster than real time speed.

The new proposed iCCR generated a scientific impact with top-tier publications: one at the 14th European Conference on Computer Vision (ECCV, [40]), and one in the IEEE

Transactions on Pattern Analysis and Machine Intelligence journal (TPAMI, [39], IF 8.23). To validate the proposed approach, the iCCR implementation was tested in the most extensive benchmark that exists to date for Face Tracking, the 300VW dataset ([42]). The results attained by our tracker surpassed state of the art methods, thanks to the incremental learning approach. These results are summarised in Table 3.1, in which the Area Under the Curve (AUC) for the error is reported (the higher the AUC, the better the performance), and compared against the top performance methods on the benchmark.

Method	Category 1	Category 2	Category 3
Yang et al. [53]	**0.5981**	0.6025	0.4996
Xiao et al. [51]	0.5814	**0.6093**	0.4865
iCCR	0.5978	0.5918	**0.5141**
CCR	0.5657	0.5539	0.4410

Table 3.1: AUC for 49 points configuration for the different categories.

This fast tracker is deployed as part of eMax, and has been implemented to run on an iPhone 6 at less than 20ms per frame, as part of a royalty bearing license agreement with MeoGraph, who are releasing their Augmented Reality messaging system based on our tracker mid-February 2018.

Code for iCCR is available from `http://www.cs.nott.ac.uk/~psxes1/`.

3.2 Dynamic Convolutional Neural Networks for Facial Expression Recognition

Facial expression recognition is a core part of human-human interaction. Replicating this ability is essential for building a fully functional human-machine interface. The Facial Action Coding System (FACS) developed by Ekman and Friesen [18], provides a systematic and objective way to study any kind of facial expression, by representing them as a combination of individual facial muscle actions known as Action Units (AU). However, automatic recognition of facial AUs resulting from spontaneous facial expressions is a hard task. It depends on multiple factors including shape, appearance and dynamics of the facial features, all of which are adversely affected by environmental noise and low intensity signals typical of such conditions.

As a part of the ARIA project, we developed a dynamic deep learning framework for facial AU recognition, to the best of our knowledge the first Deep method for AU detection. The framework uses deep Convolutional Neural Networks (CNN) to learn models of facial Action Units (AU). It is aimed at incorporating the three important characteristics which distinguishes one AU from another: shape, appearance and dynamics. The appearance is modelled through local image regions relevant to each AU. Shape is encoded using binary masks computed from automatically detected facial landmarks. This enables us to learn the relevant shape features instead of using hand-crafted geometric features. Dynamics is

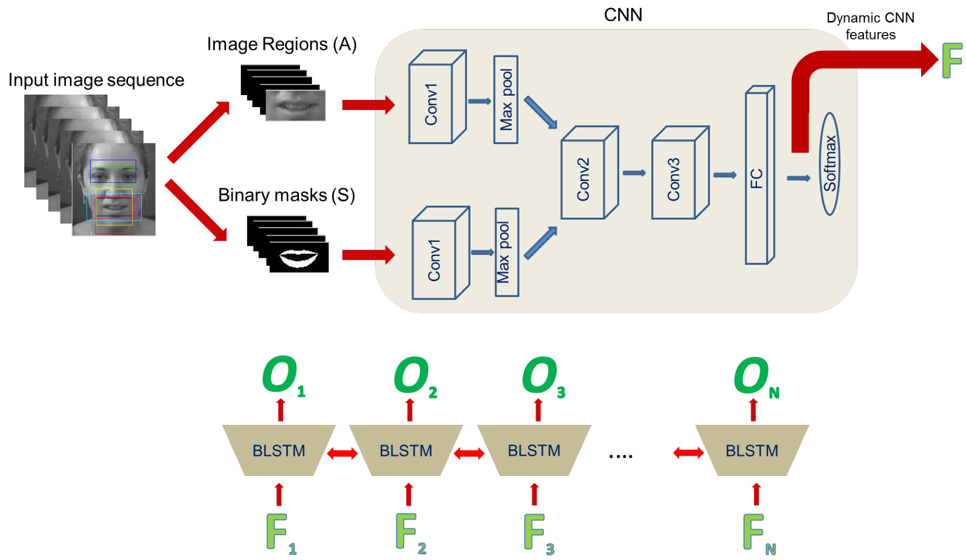


Figure 3.1: A graphical overview of our training pipeline: The colored rectangles in the input image sequence shows the different image regions selected for different AUs. Here we show the extraction of image regions (A) and binary masks (S) for AU 12. These are used as input to the train the CNN. Features extracted from the trained CNN (at the fully connected layer) denoted here as F are used to train a BLSTM network to get final output prediction values O.

modelled in two ways. Short term dynamics is encoded using a short sequence of images as input to CNN. For modelling long term dynamics, the system employs Bi-directional Long Short-Term Memory (BLSTM) recurrent neural networks.

Fig. 3.1 shows a graphical overview of our system. The system uses a CNN consisting of two input streams. The first input streams takes a transformed sequence of image regions as input (for modelling appearance). The second stream takes a transformed sequence of binary masks as input (for modelling shape). The output of the CNN after the fully connected layer (FC) is used for learning a BLSTM (for modelling long term dynamics). In contrast to previous approaches, our system learns all the key features (appearance, shape and dynamics) jointly using a deep learning framework.

The method was evaluated on a number of databases (SEMAINE, BP4D and DISFA) showing state-of-the-art performance on AU detection task. Fig. 3.2 shows the average performance (F1 measures) for AU occurrence detection on the FERA2015 Challenge dataset which is a combination of SEMAINE and BP4D datasets. The performance (2AFC scores) for occurrence detection on the DISFA dataset are shown in Fig. 3.3. Performance per AU is shown in Fig. 3.4. The proposed methodology was presented at the IEEE Winter Conference on Applications of Computer Vision (WACV) 2016 [26].

Code for the Dynamic CNN for Facial Expression Recognition method can be found from `http://www.cs.nott.ac.uk/~psxsj3/`.

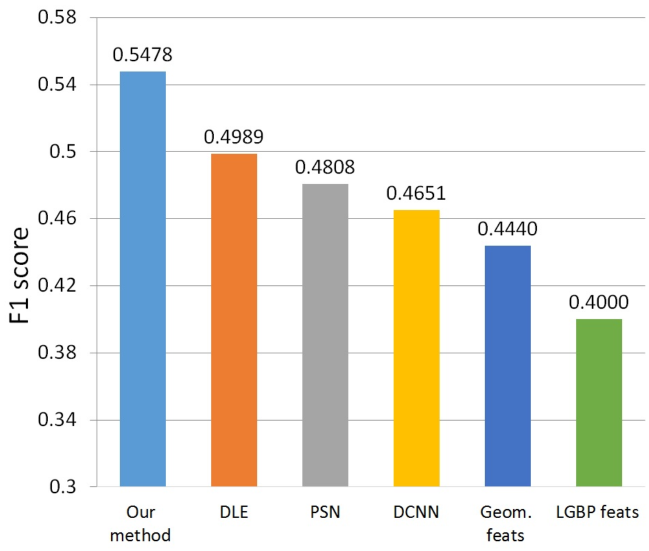


Figure 3.2: Weighted average performance on FERA-2015 test set (BP4D and SEMAINE) for AU occurrence. Our method is compared against DLE [55], PSN [6], DCNN [22] and Geometric and LGBP feats [48].

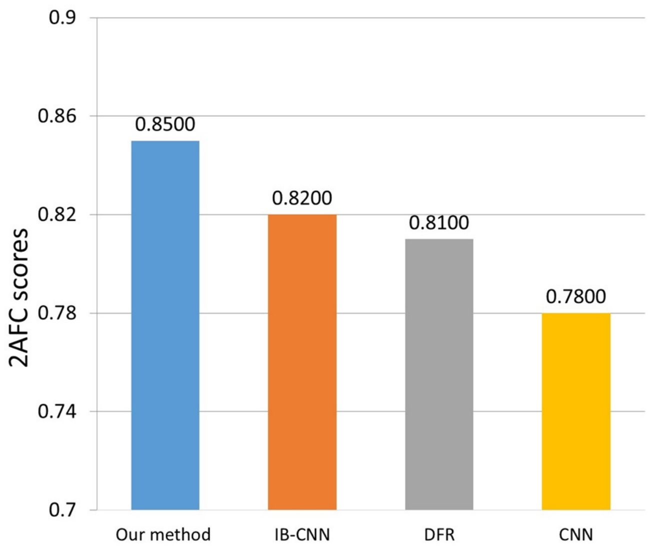


Figure 3.3: Average performance (2AFC scores) comparison on SEMAINE, BP4D, and DISFA databases for AU occurrence detection task (using 5 fold cross validation). The proposed approach is compared against CNN based approach [20], DFR [27] and IB-CNN [23]

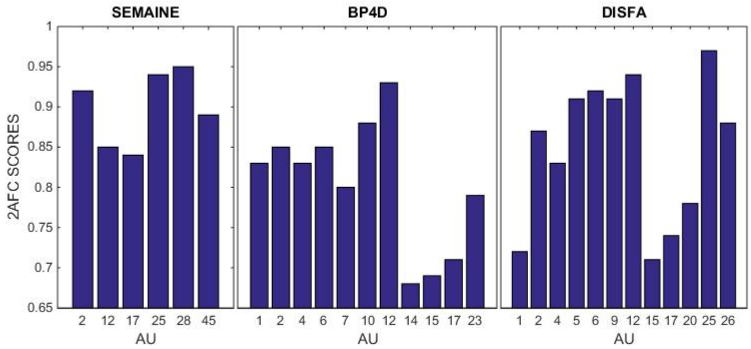


Figure 3.4: Per-AU performance (2AFC scores) comparison on SEMAINE, BP4D, and DISFA databases for AU occurrence detection task (using 5 fold cross validation).

3.3 Expressive and Reactive Text To Speech

Recently, the quality of speech synthesis has greatly improved. In many cases this speech is impossible to tell apart from real human speech. Two approaches dominate the field in terms of creating such synthetic speech, and both are based on large corpora of pre-recorded natural speech:

1. Unit Selection: New speech is generated by taking segments (or units) of these recordings, cutting them up and sticking them back together in a different order [25, 15, 28].

2. Parametric Synthesis: A statistical model, typically using hidden Markov models or deep neural nets, is created from the recorded speech. At synthesis, the model generates parameters for creating speech using vocoder techniques [56, 57].

Current systems are acceptable for reading neutral material such as bank balances, but sound unacceptable if used to read longer texts or more personal information, and in contexts that require responsive communication such as human-computer dialogue. This is a critical problem for applications where maintaining engagement with the user is a key requirement, for example in affective information retrieval applications where maintaining user motivation is crucial. In such use cases we require speech synthesis that can mimic the emotional, reactive and expressive dimension of human speech. Within this challenge ARIA has resulted in two breakthrough technologies: 1) The development of a reactive speech synthesis API which has been integrated into the ARIA-VALUSPA framework (See also section 3.9), and 2) an algorithmic approach to altering voice quality which allows emotion to be added to a previously neutral speech synthesis voice without recording additional data.

3.3.1 Reactive Speech Synthesis

The ability to be interrupted and react in a realistic manner is a key requirement for interactive speech interfaces. While previous speech synthesis systems have long implemented techniques such as 'barge in' where speech output can be halted at word or phrase boundaries, less work has explored how to mimic human speech output responses to real-time events like interruptions which require a reaction from the system. Unlike previous work which has focused on incremental production, we developed a novel re-planning approach. The system is versatile and offers a large range of possible ways to react.

Previous work has made the implicit assumption that reactive synthesis has to be incremental e.g [7, 10]. This is not the case, it just needs to be stoppable. To be reactive, the synthesis has to be fast enough to re-plan content (re-planning) and insert it (splicing). It is true that incremental systems offer locations for insertion but, given that any system has full timings described, such insertion points can be chosen without the need for incremental generation.

The re-planning and splicing approach is as follows: given a required latency, e.g. 200 ms, the system must operate fast enough to resynthesise the current chunk of speech with an alternative ending within that time. The initial part of the synthesis must match exactly the initial part of the original chunk. The new audio can then be seamlessly re-spliced into the audio stream replacing the original planned output. This requires tight control of audio playback, but has the advantage of being agnostic to the type of synthesis system you are using.

CereProc's SDK [5] synthesises on a phrase-by-phrase basis, firing a callback between phrases. During the callback a special audio buffer is available which contains the audio of the phrase as well as some metadata. This buffer is queued for playback. We created new functionality in the SDK that takes as input one of these buffers, a minimum interruption time, t_r, and an interruption type, and returns a new buffer. In this buffer the audio up to t_r is guaranteed to be identical to the original buffer. After that it will be interrupted at $t_i \geq t_r$. t_i will be a natural point for interruption, i.e. a syllable nucleus or boundary. Once this buffer is available the agent can seamlessly swap the audio buffer that is being played at some point $t_s < t_r$. By setting this time slightly in the future of when the interruption is needed some latency for processing can be added. This is illustrated diagrammatically in Figure 3.5.

Depending on the call the system has multiple strategies for finishing the phrase:

- stopping immediately,
- tailing off over a few words (a polite turn pass),
- adding Lombard effects for a few words (an angry turn pass),
- completing the original phrase with Lombard effects added.

The system can then add additional speech before returning to the original queue if appropriate. Otherwise it may need to drop some phrases that have been resynthesised

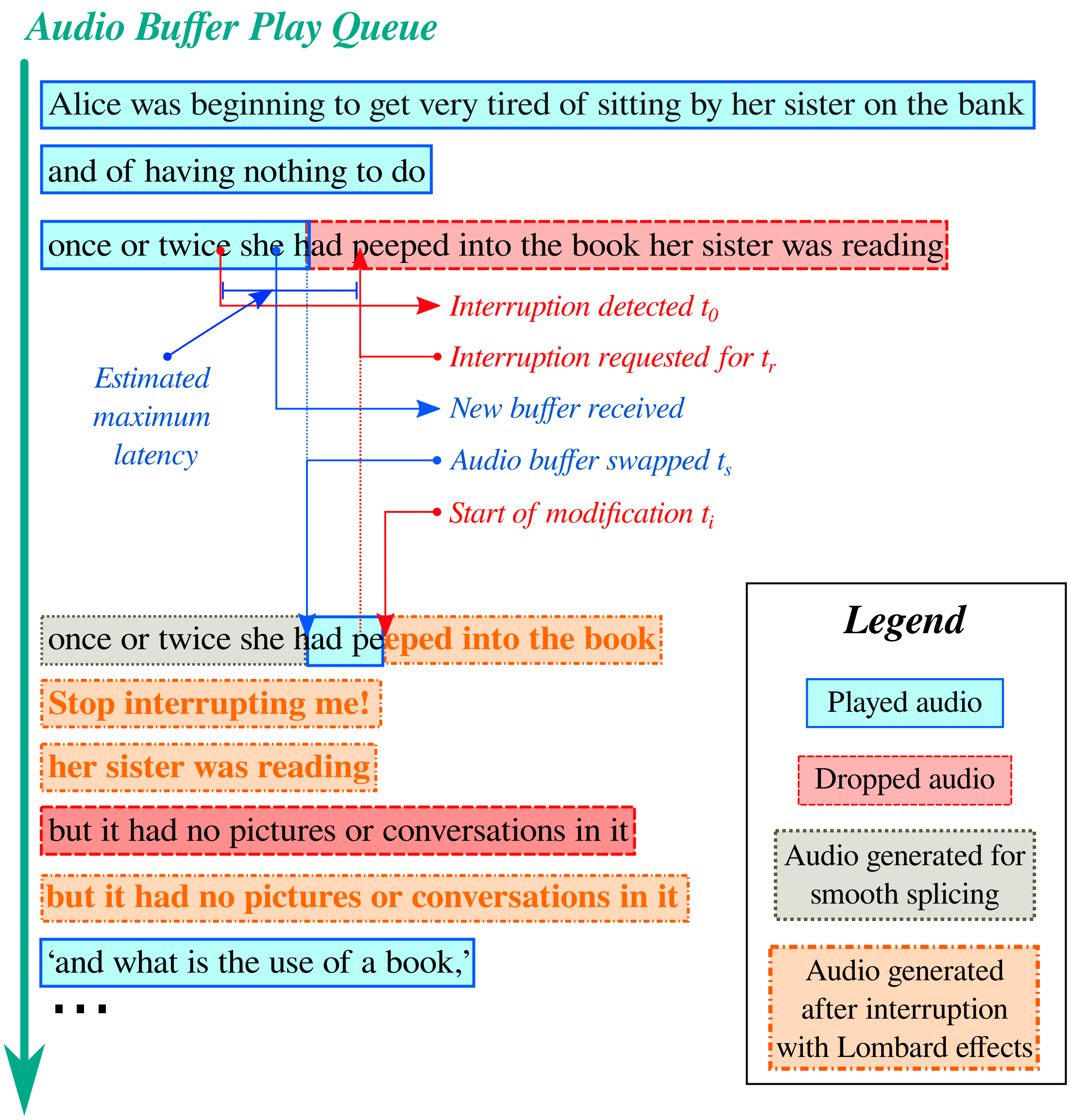


Figure 3.5: Example of the use of the interruption API, showing the changes in audio buffers. Final played audio is in blue and orange boxes, red and grey boxes are dropped. Note that $t_r - t_0$ must be larger than the maximum system latency.

differently, or empty the queue entirely, depending on the application. The approach we have developed forms the basis of a patent application (No. GB1713273.9).

3.3.2 Algorithmic Modification of Voice Quality

The CereVoice speech synthesis system uses a distinct set of sub-corpora containing different voice qualities to achieve a more subtle change in the perceived emotion in an utterance. Voice quality is an important factor in the perception of emotion in speech. However, unlike speech rate and pitch, which can be modified relatively easily using digital signal processing techniques such as PSOLA, modifying voice quality is more difficult, especially if it is important to retain naturalness. Being able to modify the voice quality, just as we modify pitch and duration, would dramatically improve the quality of emotional voices, reduce their memory footprint and allow us to alter neutral voices to make them sound more emotional.

Key to this approach is work carried out on closed phase LPC vocoding and reported in the ARIA-VALUSPA project's mid-term report. By decomposing the speech into a voicing component and a filter component we can attempt to modify the voicing and put the resulting elements back together. Although LPC vocoding is not a new technique, until recently it was used mostly to compress speech for efficient transmission and storage. The use of LPC analysis for decomposing and modelling elements of the approach are less explored and a subject of current international research.

We term our process Voice Modification via Glottal Signal Modelling (VMGSM). VMGSM is a novel voice transformation technique that relies on human speech source - filter uncoupling to specifically model and modify the glottal signal, which contain most of the "voice quality", which is one of the main components of affective speech.

The possibility to explicitly model the glottal part of the speech signal has many advantages, as this can allow the generation of expressive and emotional speech from neutral speech, thus allowing us to add emotional characteristics to any voice. On the other hand, the decomposition of speech is a difficult problem, relying on simplifying assumptions on the form of the voice signal, which inevitably creates artefacts in the resulting modified speech.

While it introduces artefacts, it allows one to deal better with low coverage of unit selections databases, as it contribute to reduce some of the artefacts in unit selection joints, and at the same time can produce large artificial sub-genres for emotional synthesis.

The artefacts introduced by VMGSM make the resulting output currently unsuitable as a generic replacement for the synthesis of neutral speech, however it proves its usability as a possible replacement for specialised use, such as emotional speech synthesis.

Listening tests run internally (and illustrated below) show that while VMGSM is not as effective at generating expressive speech as specially recorded sub-genres, the gap in quality is not as wide as with neutral speech. In addition, the possibility to create fully artificial sub-genres allowed us to add emotional speech capabilities to the "Alice" voice distributed freely with the AVP toolkit. This voice is built from the freely available "SLT" database from CMU, which only contains neutral speech from a single female speaker.

Voice Modification via Glottal Signal Modelling (VMGSM) was inspired by [41]. We

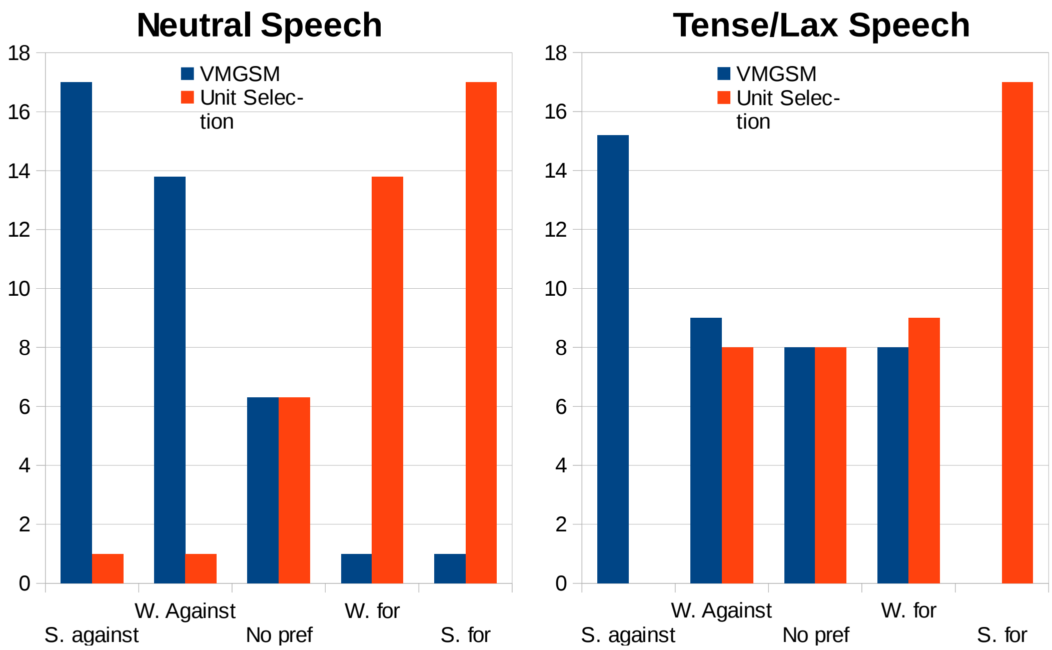


Figure 3.6: Left: Neutral speech (VMGSM modified - red vs. conventional unit selection - blue) Right: Emotional speech (VMGSM modified -red vs. conventional unit selection - blue). Figures show results of an A-B comparison tests where subjects indicated their preference on a 5-point Likert scale.

start with Pitch-Synchronous Iterative Adaptive Inverse Filtering [1], an algorithm widely used and accepted as a method of extracting a glottal signal from recorded speech. Whilst we cannot fully verify the validity of the signal output by this algorithm (indeed there is no ground-truth benchmark), there are certain attributes of a glottal signal we can expect. Spectral flatness, or a familiar pulse-like appearance for example. A more quantitative method of assessing these is via cross-correlation coefficients. The need for this measure is actually two-fold; not only does it help verify the signal, but it also assists in the optimal parameterization of the PSIAIF process for a particular speaker.

Keeping in mind the overarching goal here is voicing modification, manipulating a raw glottal signal estimate that PSIAIF produces can be a somewhat meaningless procedure; to do anything meaningful we need reliable notions of glottal characteristics. Looking at a raw glottal signal, one could guess where these regions are and tweak these durations, but there is no guarantee the modified version will retain any perceptual characteristics we consider to be voice-specific. Instead we can start with a well-formed glottal pulse whose perturbations result in another well-formed glottal pulse. An LF pulse parameterized in the traditional way [30] is an ideal candidate to perform some sort of fitting routine - and a modified version of VOICEBOX's glotlf.m function [8] takes 4 parameters to fully describe an LF-pulse within a pitch-synchronously derived frame of speech. These parameters can be estimated from a raw glottal signal to produce an LF pulse that *fits* the speech frame being analysed.

Going through this procedure per frame yielded glitchy output speech however, as pulses can vary wildly across adjacent frames. Using simple static averaging of parameters within stable regions across an entire spurt of speech, we can obtain a *default* pulse. In order to get this as faithfully close to a speaker's glottal signal as possible, we optimize PSIAIF parametrization via brute-force search, picking values that yield the highest cross-correlation score. Using optimized PSIAIF to yield a reasonable glottal signal, we then fit a modelled LF signal, enabling us to manipulate glottal signal and achieve effects such as the *stressed* and *relaxed* outputs demonstrated here; instead of an averaged LF pulse being fed into the analysis, an LF pulse derived from *irritated* and *happy* genres respectively was used. Analysis was run across all spurts available in each genre within the voice's corpus for a particular set of phonemes, alongside some additional tweaking to LF parameters that follow accepted notions of heightening or lowering vocal effort. This approach assumes that genre data was present in the first instance; for a limited corpus in which genre data is not available, we rely only on theoretical knowledge of vocal effort and glottal stress to modify the speaker-specific LF pulse. For a 7-level VMGSM voice, 6 candidate pulses of varying stress were derived from the default neutral pulse, fed into the synthesis stage of an LPC-based vocoder whose output was then added to the original corpus, expanding its emotional coverage. Other auxiliary effects were then used to augment the desired emotion (such as post-synthesis warmth and clarity filtering as well as global pitch modification).

The 7-level emotional voice using VMGSM is freely available and distributed with the ARIA VALUSPA Platform (AVP) via Github. See D7.6 for full details on how to download, install, and customise the AVP.

3.4 End-to-end Audio-Visual Emotion Recognition using Deep Neural Networks

Following a recent trend in machine learning that aims at building intermediate representations of raw input signals in order to extract task-specific information (which usually leads to a better performance on the recognition task), we used Convolutional Neural Networks (CNNs) to extract features from the speech, and a deep residual network (ResNet) of 50 layers for video. These models were then combined with a Long Short-Term Memory (LSTM) network, and trained in an end-to-end fashion where - by taking advantage of the correlations of each of the streams - we manage to significantly outperform the traditional approaches based on auditory and visual hand-crafted features for the prediction of spontaneous emotional displays.

For the visual domain, we used a deep residual network (ResNet) of 50 using the pixel intensities from the cropped faces of the subjectâĂŹs video. The first layer of ResNet-50 is a 7x7 convolutional layer with 64 feature maps, followed by a max pooling layer of size 3x3. The rest of the network comprises of 4 bottleneck architectures, where after these architectures a shortcut connection was added. These architectures contain 3 convolutional layers of sizes 1x1, 3x3, and 1x1, for each residual function. After the last bottleneck architecture an average pooling layer is inserted (see Figure 3.7).

In relation to speech, we learn the feature extraction and regression steps in one jointly trained model for predicting the emotion. The input to the model is a 6 s long segment of the raw waveform to sampled at 16 kHz (this corresponds to a 96000-dimensional input vector). Inputs were normalised to have zero mean and unit variance to account for variations in different levels of loudness between the speakers. Then, we include a temporal convolution layer with F = 20 space time finite impulse filters with a 5ms window in order to extract fine-grained spectral information from the input signal. The output of this layer is then pooled across time. The impulse response of each filter is passed through a half-wave rectifier (analogous to the cochlear transduction step in the human ear) and then downsampled to 8 kHz by pooling each impulse response with a pool size = 2. It follows a temporal convolution, for which we have used M = 40 space time finite impulse filters of 500ms window. These filters are used to extract more long-term characteristics of the speech and the roughness of the speech signal. Finally, we performed max-pooling across the channel domain with a pool size of 10 to reduce the dimensionality of the signal while preserving the necessary statistics of the convolved signal.

The final time-continuous models for the prediction of spontaneous and natural emotions (arousal and valence) were developed on the audio-visual database RECOLA. The dataset was split in three partitions - train (16 subjects), validation (15 subjects) and test (15 subjects) by stratifying (i.e., balancing) the gender and the age of the speakers. For training the models we utilised the Adam optimisation method, and a fixed learning rate of 10-4 throughout all experiments. For the audio model we used a mini-batch of 25 samples. Also, for regularisation of the network, we used dropout with p = 0.5 for all layers except the recurrent ones. This step is important as our models have a large amount of parameters ($\approx$ 1.5M) and not regularising the network makes it prone on over-fitting on the training data. For the video model, the image size used was 96×96 with mini-batch

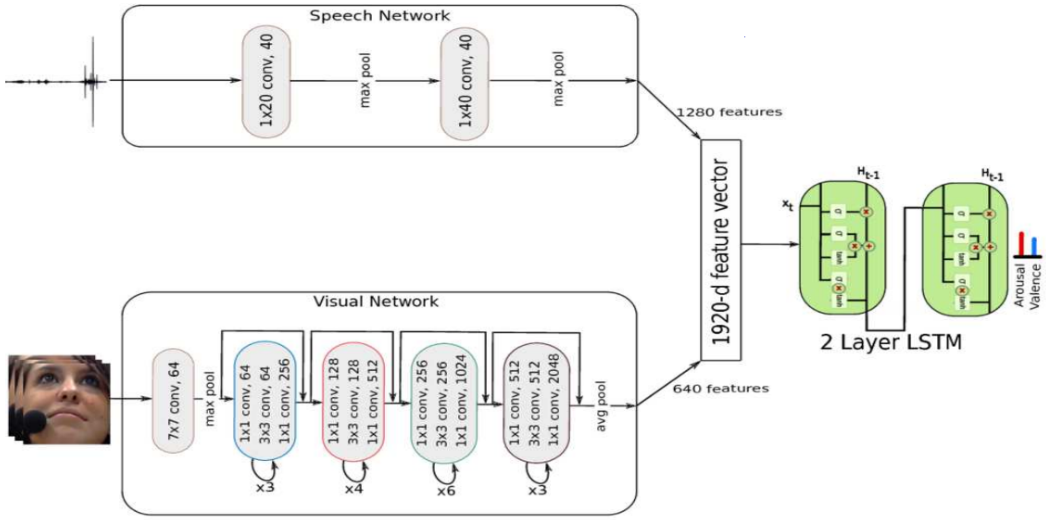


Figure 3.7: The network comprises of two parts: the multimodal feature extraction part and the RNN part. The multimodal part extracts features from raw speech and visual signals. The extracted features are concatenated and used to feed 2 LSTM layers. These are used to capture the contextual information in the data.

of size 2. Small mini-batch is selected because of hardware limitations. The data were augmented by resizing the image to size 110×110 and randomly cropping it to equal its original size. This produces a scale invariant model. In addition, colour augmentation is used by introducing random brightness and saturation to the image. Finally, we conducted a series of experiments using a chain of post-processing methods applied to the predictions obtained on the development set: (i) median filtering (with size of window ranging from 0.4 s to 20 s), (ii) centring (by computing the bias between gold-standard and prediction), (iii) scaling (using the ratio of standard-deviation of gold-standard and prediction as scaling factor), and (iv) time-shifting (by shifting the prediction forward in time with values ranging from 0.04 s to 10 s), to compensate for delays in the ratings. Post-processing steps were kept when an improvement was observed on the model performance on the validation set, and applied then with the same configuration on the test partition.

Our experiments included comparisons between single- and multi-modal models. Results show that our the present multimodal approach models achieves significantly better performance in the test set in comparison to state-of-the-art models using the RECOLA database (including those submitted to the AVEC2016 challenge). This is particularly evident for the Valence dimension. A full description of the results and approach can be found in [47].

3.5 Cooperative Learning

Like many other data-driven fields, paralinguistic speaker analysis substantially depends on the availability of labelled data, which are difficult and expensive to obtain. Within our project, a novel generic annotation framework has been developed, with the aim to achieve the optimal trade-off between label reliability and cost reduction by efficiently distributing the labelling work amongst human and machine. For the purpose of arbitration, a deep-learning based uncertainty measure is used to pass the most informative instances (with high prediction uncertainty) to human assessment, whereas those instances in a database predicted with high model confidence are labelled by machines. Further, an inter-rater agreement threshold serves as an early stopping criterion to terminate the annotation process when enough ratings have been obtained to determine each instance's gold-standard label. The efficacy of this approach is demonstrated on the "Degree of Nativeness" task of the INTERSPEECH Computational Paralinguistics Challenge. In the result, the novel dynamic cooperative learning algorithm yields .424 Spearman's correlation coefficient compared to .413 with passive learning, while reducing the number of human annotations by 74 %. For the annotation of NoXi, the proposed framework has been integrated into the social signal interpreter (SSI) and the nonverbal behaviour analyser (NOVA).

3.6 Alignment

In order to characterize verbal alignment processes for improving virtual agent communicative capabilities, we propose a framework to quantify the verbal alignment interactive process and the self-repetition behaviour of dialogue partners in dyadic textual dialogues [17]. This framework focuses on lexical patterns appearing in dialogue utterances. The code of the framework is available at `https://github.com/ARIA-VALUSPA/ARIA-DialogueManagement/tree/NLG/ARIA-NLG`. It distinguishes two main types of such patterns. The first type is *shared* lexical patterns between dialogue partners (DP) i.e. patterns that are initiated (or primed) by a speaker, subsequently adopted by the other speaker and possibly reused during dialogue by any speaker. These patterns are directly related to the verbal alignment interactive process, a particular type of on-the-fly linguistic adaptation. They can be seen as shared dialogue routines at the lexical level. They are a way to verbally align and ultimately share a common language to improve understanding, collaboration and the social connection to a conversational partner. The second type is lexical *self*-repetition. Contrary to the previous type which considers patterns that are shared between DPs, self-repetition considers each DP in isolation. Self-repetitions are lexical patterns appearing at least two times in the dialogue utterances of a given DP, independently of the other DP's utterances. Self-repetitions are directly related to the self-consistency of the linguistic production of a given DP.

The general idea of the framework is depicted in Figure 3.8. The main concept behind our model is the automatically built *lexicon*. Lexicons keep track of lexical patterns (shared ones and self-repetitions) as well as valuable features of these patterns (e.g., frequency, turns in which they appear). Lexicons can be built automatically for an entire

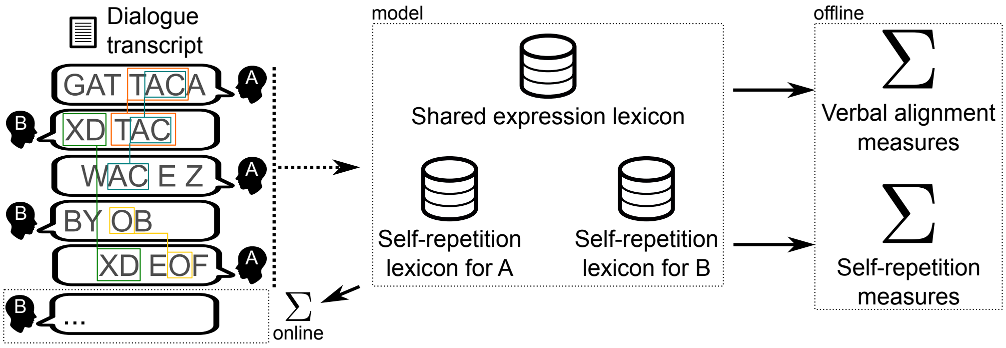


Figure 3.8: Proposed framework: automatic building of the shared expression lexicon and the self-repetition lexicons to derive offline and online measures of verbal alignment and self-repetition behaviour. Shared lexical patterns are shown on the dialogue transcript.

dialogue (i.e. offline) or incrementally for a given dialogue history (i.e. online). Our model considers three lexicons: a shared expression lexicon for shared lexical patterns, and two self-repetition lexicons (one for each DP). Lexicons and the dialogue transcript are leveraged by deriving offline and online measures to quantify aspects of the verbal alignment process and the self-repetition behaviour of DPs. Offline measures are intended to be used for past dialogue interactions (e.g., corpus studies) while online measures are intended for use in a dialogue system.

3.7 Context-sensitive analysis of complex multi-modal social signals

For the recognition of social attitudes, such as the engagement/interest of a person towards the agent, we consider Dynamic Bayesian Networks (DBNs) [34] as modelling approach. DBNs are probabilistic models that allow to design correlation between nodes in a network, but also between a node and it's state earlier in time. Even-tough the probabilities for such nodes, and even the overall network structure can be learned with machine learning techniques, it allows to retrace the decisions the DBN makes for each node or layer of nodes. One could think about using alternative models, such as deep end-to-end learning with artificial neural networks [47]. While such approaches deliver promising results on audio-visual data they only give little insights on *how* and *why* they predict behaviours the way they do. Especially in scenarios where it is essential to know why a person is interpreted as e.g. "strongly disengaged", often the idea is to identify cues that led to this interpretation, providing an additional abstraction layer. While a DBN's structure may be modelled based on a theory, our framework offers the possibility to prepare data in a way, so that the DBN learn correlations between the parallel appearance of behaviours, context and complex phenomena.

To this End, the previously introduced NOVA tool exports parallel annotations form the Annotation Database, so that a DBN may learn temporal correlations between multiple cues. In the ARIA-Valuspa Project we focused on the detection of the complex behaviour 'Engagement' (respectively the sub-concept Interest).

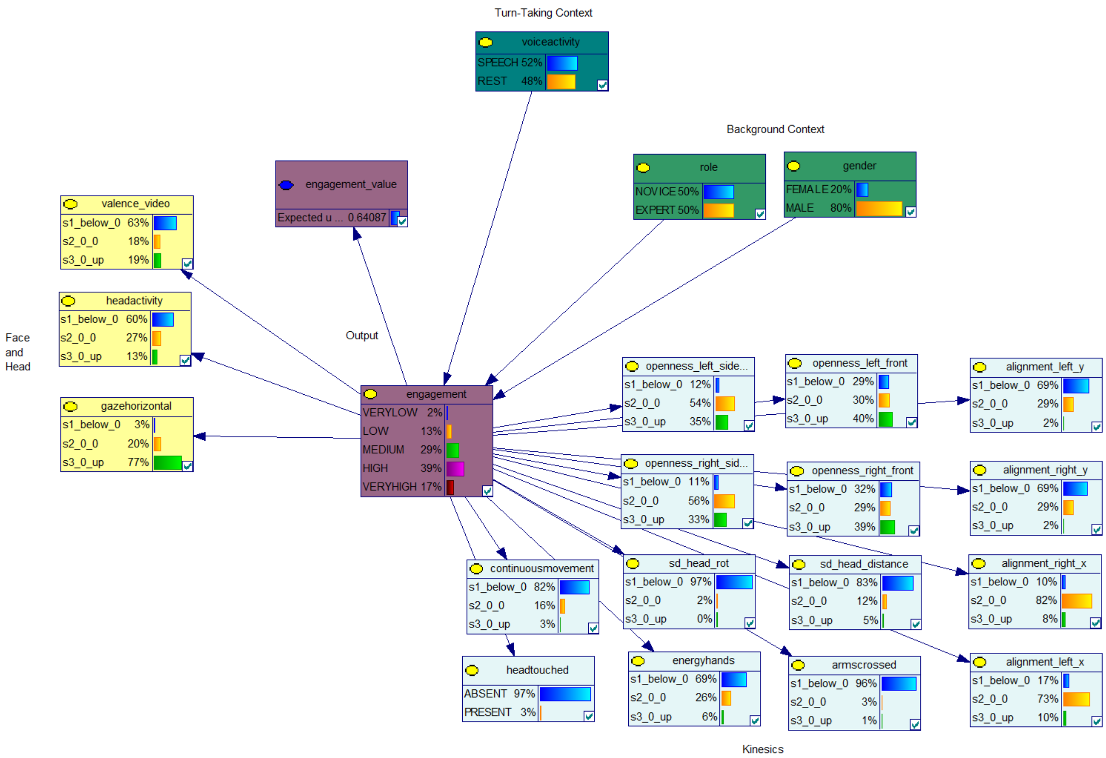


Figure 3.9: The structure of the engagement network used in the AVP

The NOVA tool, introduced in section 2.4 allows learning of parameters of a (Dynamic) Bayesian Network to fuse multiple observations for a prediction model. As training data, annotations from various annotators are combined. In Figure 3.10 we see the engagement annotation the Gold Standard user on the upper tier. The Gold Standard annotations are created by combining annotations from multiple raters to gain the ground truth. As this will be the node of interest, no evidence will be delivered later during runtime, but rather the degree of engagement will be inferred from other observations. Other annotations, such as the arms-openness, facial expressions or the amount of hand movement have been automatically created using existing models or other recognizers. Finally, manual or semi-automated annotations might be added to the model. In our case, speech/filler/breath/(silence) annotations have been added from a human annotator, that have been created using the Cooperative Machine Learning techniques (see D6.4).

Once the data sheet is created, it is combined with a Bayesian network to automatically map the nodes of the network with annotator/scheme combinations from NOVA. The network used to infer engagement in the AVP is shown in Figure 3.9

To learn parameters in the model the expectation-maximization (EM) algorithm [32] is applied.

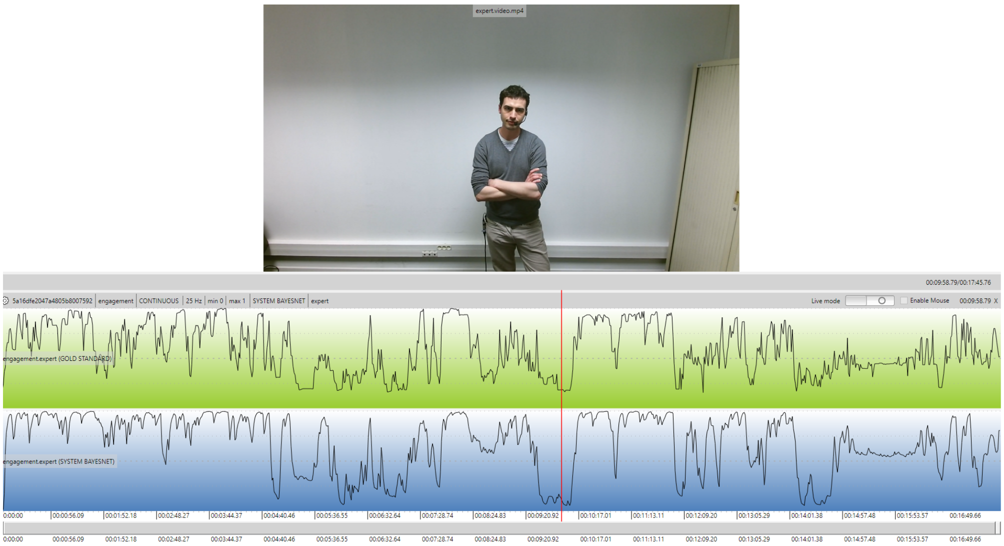


Figure 3.10: The gold standard annotation for engagement (top tier) and the prediction of the DBN (lower tier) shown in the NOVA tool

For training models with the NOVA tool, annotations are used to generate data sheets to learn the probabilities in the network. There has to be a trade-off between using sheer manual annotations, which represent the "ground truth" and deliver a perfect foundation for creating the models, and on the other hand automatically created annotations that represent outputs that our classifiers are actually able to predict. By using the cooperative machine learning we are able to adapt the outcomes of the machine already during the annotation process, so that the models reach a state where we can "trust" them to output annotations just like a human annotator would do. That means in conclusion, for our model to work as expected, we need to find social cues that are recognized reliably well, and that represent the problem at hand. Of course, the complex behaviour needs manual annotations to represent the ground truth. As complex behaviours, such as our use-case "engagement" are not straight-forward in terms of interpretation, it's preferable (one could say necessary) to have multiple raters for the given problem. In the Aria-Valuspa Platform we constantly infer the user's engagement/interest based on observations of social cues in multiple modalities, while at the same time considering background context such as the role, gender and turn-taking.

Once we learned the parameters of our DBN, we may use it either for statistical prediction purposes, or in a real-time scenario by updating the nodes with evidences received from our Social Signal Interpretation component, as well as external sources such ARIA's dialogue management system. We receive these evidences by using the network in a SSI [50] pipeline, updating evidences with observations from multiple social signal recognizers, but optionally also external information. Figure 3.10 shows and example instance of the NOVA tool, showing the gold standard annotation for engagement on the upper tier (green), as well as the prediction on the lower tier (blue).

3.8 Situation-Driven Dialogue Management

Dialogue management requires knowledge about the domain that the agent should be knowledgeable about. This means that domain experts are often called upon when the behaviours of an agent are crafted. However, domain experts often have limited knowledge about dialogue systems making it difficult for them to provide the necessary information. That is why we designed our dialogue management system in a situation-driven manner. This means that domain experts with limited programming knowledge can specify the dialogues an agent should be able to have, by describing situations they expect the agent to be in and describing what the agent should do in those situations. Within dialogue management we distinguish three concepts:

- **Dialogue Manager**. This deals with how the agent behaves in an interaction. This is specified in Management templates: the rules that govern the considerations that the agent can have and all the behaviours that the agent can decide to do. These templates are the abstract framework of the dialogue manager. For example, these templates govern what the situation is when the agent speaks and the user also starts speaking (that would be an interruption).
- **Dialogues**. These describe the scenario that the agent knows about and can converse about. They are defined in a Dialogue Structure consisting of Move templates (see below). Scenarios consisting of Move templates are what a domain expert has to write to create their agent. These templates contain the behaviour that an agent can do and the rules that define when this behaviour is appropriate. For example, when the agent is talking about rabbits and the user interrupts with a question about other animals, an appropriate response can be talking about other animals that the character Alice encountered.
- **Dialogue Engine**. This is the component that performs the underlying tasks that are necessary to create a dialogue manager. The dialogue engine we developed is called Flipper 2.0[3]: an extended and improved version of the original Flipper engine described in [44].

The Dialogue Manager takes a scenario and situation- driven approach to creating dialogue structures based on conversational acts, and is therefore called Situation-driven Dialogue Manager (SDDM). It shares some properties with current tools for the development of dialogues, such as the use of dialogue trees from DISCO [38] and the use of question-answer matching for information retrieval from the NPC Editor of the Virtual Human Toolkit [29]. Similar to the FLoReS dialogue manager of Morbini et al. [33], the SDDM has been set up to facilitate the creation of structured dialogues with the use of domain experts.

Dialogues for the ARIA system are defined in terms of hierarchical dialogue acts in a Dialogue Structure, see Figure 3.11:

- **Dialogue structure**: this is the root. The name should refer to the name of the character for which the dialogues are defined.

[3] `https://github.com/hmi-utwente/Flipper-2.0`

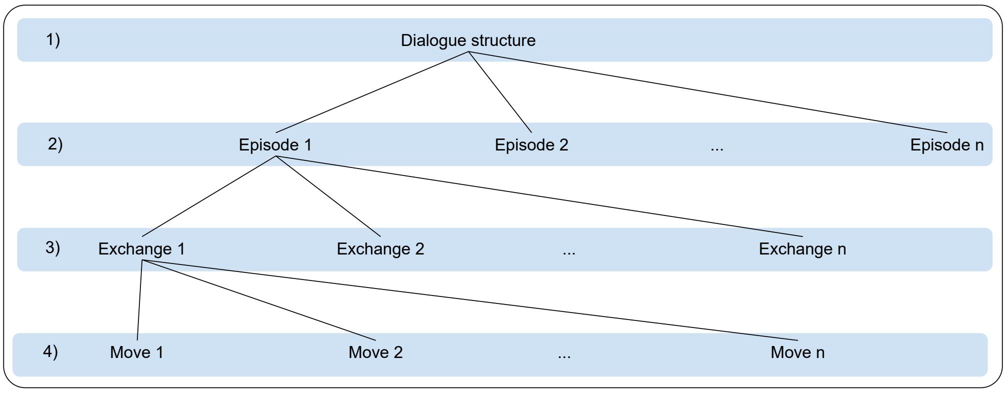


Figure 3.11: The hierarchical dialogue structure.

- **Episode**: an episode covers a phase of a conversation (e.g. social, Q&A, reading along with the agent, unexpected situations).
- **Episode.Exchanges**: episodes are made up of exchanges that are related to a topic in the episode (e.g. greeting, farewell, return_greeting).
- **Episode.Exchanges.Moves**: an exchange is made up of several moves (conversational acts, based on DIT++ [9]). A move can be realized with an utterance, nonverbal behaviour, or a combination of the two. A move has a goal that can be achieved by the behaviour and a status for that goal. Moves are selected for execution based on how relevant they are to the situation in the conversation. Rules define when a move becomes relevant.

Moves are the atomic units (dialogue items) in the Dialogue Manager. A move can refer to a dialogue act of the user or the agent. A move can be listening or speaking, depending on its DIT++ category. We distinguish three types of moves, depending on the information they carry:

- Content/Dialogue Act (C): C-Moves contain the content of the agent or user's utterances, for example if the user asked the question "What can you tell me about Alice in Wonderland?". This is the main move type, see also the next paragraph.
- Interaction Management (I): I-Moves contain information about the interaction. These moves are descriptors for the state of the floor and the agent uses these to decide when to speak. For example, the information that the agent and user are both speaking. This information is obtained from SSI (for the user) and from feedback from the embodiment (for the agent).
- Socio Emotional (S): S-Moves contain information that has to do with the social and emotional state of the agent or user. For example the valence of the user's emotional state as obtained from SSI. A computational emotional model can generate these moves for the agent (currently this task is done by the DM).

The content moves (C-Moves) *for the agent* are further subdivided depending on the type of content that is contained within the move. This division helps with planning the agent moves (for example combining content with an opinion) and helps a dialogue scenario author to keep track of what purpose a move has:

- Content (C-tag): For the Alice character, this is the type of C-Move that contains factual information from the book, for example which events happened and whom Alice met. An example is: "When I saw the White Rabbit, I chased him into a rabbit hole."

- Opinion (O-tag): A C-Move with an opinion tag contains an opinion, for example Alice's opinion about events or characters from the book, such as: "I thought it was rather strange to see a white rabbit with a watch."

- Meta information (M-tag): A C-Move with a meta information tag contains information on a meta level about the interaction. Using such moves, the agent can talk about the interaction, for example during a lull in the conversation the agent might say "Do you want to continue talking about this?".

The Dialogue Engine (Flipper 2.0) stores all information the agent knows in the Information State. Information comes from various sources and is represented in the form of Moves. During an interaction, the moves *of the user* are created by the system via the Input Understanding component. Some examples of user moves are:

- The user has made an utterance, and the automatic speech recognition (ASR) outputs a string of words: a C-type user move is generated holding the ASR output.

- The user has started speaking, and this is detected by the Voice Activity Detection: an I-type user move is generated stating that the user has started speaking (and when the agent was also speaking that this is an interruption).

- The user has started smiling, and the SSI updates the valence of the user: an S-type user move is generated holding the user valence.

Additionally, information about the agent's actions is received as input (i.e. feedback) from the behaviour realiser. The behaviour realiser (e.g. Greta) sends continuous feedback about what behaviour has been carried out. Feedback can be:

- BML Callbacks: the BML realiser sends information about which behaviour (BML block) has started, ended, or has been stopped.

- Time Markers Callbacks: during the agent behaviour, the realiser sends feedback on the exact timing of each behaviour that is executed. This is done using time-markers (see section 3.9). For agent utterances this is done on word level.

The feedback allows the Input Understanding to keep track of the floor (i.e. turn-taking) and the completion of the goals of the agent. For example, knowing from the feedback when the agent has stopped speaking and knowing from the ASR when the user has started speaking allows us to determine whether there is overlapping speech and thus whether the user has interrupted the agent. Additionally, the time markers allow us to know what part of the agent utterance has been said uninterrupted (and thus was heard by the user) and what part was not heard because it was interrupted. The agent can concatenate moves, for example a C-tagged C-move ("The White Rabbit had a watch") with an O-tagged C-move ("I liked that watch"). Time marker feedback is used to determine whether the goal of the agent's move was accomplished: if the agent was interrupted before it could complete the utterance, the goal of the move (e.g. of conveying this information) is not accomplished. This may lead the agent to repeat the move.

Moves have rules that determine when the move becomes relevant. The move an agent carries out is selected based on its relevance. We view relevance as the utility value of a move, where the agent is trying to maximize the utilities of moves and plans the moves with the highest relevance above a certain threshold. This threshold is dynamic and decreases when for a certain amount of time no move has been performed and increases when the agent is speaking.

The relevance of a move gets updated by the Agent Move Updater. Relevance is based on the rules in the move. When a rule (akin to a precondition) is met, the relevance of the move increases. Additionally, when the user has said something, this utterance can be compared to utterances predefined in the move to which this move would be an appropriate response. This is an extension of the QA Matching approach. Furthermore, relevance of a move increases if closely related moves (e.g. moves in the same exchange) become more relevant. We use Management templates in Flipper to update the relevance of the dialogue structure.

The Agent Move Selector keeps track of the relevance of all the moves in the Dialogue Structure. Once it has found a move with relevance above the threshold, it selects this move and sends this move to the Move Planner. Additionally, it sends the selected move to the Move Generator for execution by the agent embodiment. The Move Planner keeps track of the current agent move and the planned agent move. It gets information from the move selector and the Input Understanding modules for observed and predicted user moves.

Once an agent move has been selected and put in the move planner, this move is translated to FML-APML. First, the agent's verbal utterance (if present) is extracted from the selected move, and time markers are added to it. Secondly, the emotion of the agent is set, based on the current emotional state of the agent stored in the information state. Furthermore, additional parameters (e.g. backchannel, stance) in the move can be used to fill the placeholders in the FML-templates. Finally, the agent can align the verbal content of its move to the user's word choice by taking the dialogue history into account.

3.9 Interruptions

Interruptions are phenomena that frequently occur in human interaction. To study and model interruptions, we first defined a taxonomy of interruptions and measured how different types of interruptions may affect the perception of the social attitude of the interrupter and her level of engagement in the interaction (see D6.1). Interruption may be characterized in two broad types: disruptive or cooperative [12]. They correspond to different strategies which are expressed by the interrupter through different dialogue acts (i.e. communicative functions). Interruptions may also be distinguished by their temporal relation with the interlocutor's speech. We consider overlaps, silence and replanning.

We dealt with three aspects:

1. Studying interruptions and their meaning and effects during the interaction;
2. Detecting when a user's interruption occurs;
3. Reacting appropriately (i.e. agent) when such interruptions occur.

In regards to the first aspect, we conducted a web study aimed at investigating the effects of interruption strategies and types, in agent-agent interactions, on human perception of both agents' interpersonal attitudes (dominance and friendliness) towards each other, of their level of engagement, and involvement in the interaction (See D6.1). We considered a dyadic agent interaction as it allowed us a complete systematic control of both the interrupter and the interruptee's behaviour. Our next step was to study human-human interruptions. With SSI developed within WP2, we annotated the NoXi corpus [13] to extract information when interruption occurs (see D6.2). This annotation was done, in most part, automatically. We looked at the reactions deployed in response to an interruption at the behaviours level and at the strategy level (see D4.2). Regarding the last aspect, we model different interruption types for the agent. That is we model an agent stopping and holding its gesture; continuing speaking, thus overlapping with user's speech, while marking an increase of emphasis; and replanning what to say and what to gesture. Regarding the third aspect, to characterize precisely the behaviour of the virtual character, we conducted an experimental study where users chose interactively the video of the virtual agents (see D4.3). Videos were ordered following a genetic algorithm. We have identified a large number of parameters that described the virtual agent's behaviours. Manipulating these parameters one per one is very cumbersome. So we proposed to use a web study where participants could visualize four videos of an agent reacting to an interruption. The animations of the virtual characters within the four videos are computed on the fly using genetic algorithm (See D6.4).

3.10 Modelling social attitudes

Interpersonal attitudes and emotions can be characterized by "multimodal non-verbal sequences". We have proposed a model of social attitude as sequences of behaviour. To develop this model, we rely on a sequence-mining method to extract, for each attitude type, (1) the most relevant quantitative timing, and (2) the sequential non-verbal behaviour representing this attitude.

In a first step, we relied on the annotation of an existing database [14]. The annotation is done at two levels: non-verbal behaviour and expressed attitudes (see D6.1). For the non-verbal behaviour annotation we consider the following modalities: gesture (e.g., communicative gestures), hand position (e.g., hands together), posture (e.g., leaning backwards), head movement and direction (e.g., nods), gaze (e.g., looking at the interlocutor), facial expression (e.g., eyebrow). For social-attitude representation, we use Argyle's bi-dimensional model of attitudes [2], with an affiliation dimension ranging from hostile to friendly, and a status dimension ranging from submissive to dominant. We applied temporal sequence mining so as to find temporally frequent sequences called patterns from a sequence database [16]. We built four datasets of non-verbal signal sequences representing the four attitude variations: dominance increase, dominance decrease, friendliness increase, and friendliness decrease (see D4.2).

Our next step was to evaluate the extracted sequences and measure if they convey a given attitude variation. To this aim we conducted an evaluation study where the behaviour of a virtual character reproduced the annotated sequences (see D6.4). We have used the Greta/VIB platform to generate videos of a virtual agent displaying non-verbal patterns. As our model only considers nonverbal behaviour, we left aside the content of the speech. For this, each non-verbal pattern was shown while the agent spoke the same nonsense speech. For each video, participants rated their perceived attitudes of the agents along 16 adjectives following Leary's model [45].

To model a virtual agent communicating with different social attitudes, we have developed a behaviour planner, called Sequential Attitude Planner (see D4.3). It takes as input an FML file (utterance to be said by the agent) and the attitude variation that the agent will express toward the user. Four steps defined the Sequential Attitude Planner: 1) FML-sequence generation: generation of a sequence of non-verbal signals that expresses the communicative intentions. 2) Attitude-sequence selection: from the sequence dataset that represents the attitude variation that the agent will express, the algorithm selects the most similar sequence to the FML-sequence. 3) FML-sequence enrichment: all signals in the attitude-sequence that do not appear in the FML-sequence are considered to be added to the FML-sequence. 4) Priority signals selection: we designed a Bayesian Network (BN) to model the occurrence probability of non-verbal signals for each attitude.

We have integrated the Sequential Attitude Planner into the Greta/VIB agent platform. Finally we have evaluated this last model. As for the first evaluation study, participants had to evaluate the perceived attitude changes in videos of agents (see D6.4). We found significant differences for two variations, dominance increase and friendliness decrease. We also highlighted correlations between both attitude dimensions, friendliness and dominance, as well as asymmetric correlations between the two extremities of the friendliness axis, friendly vs hostile.

4 Public engagement

Below we list the various ways in which the ARIA-VALUSPA project engaged with the public.

4.1 Blogs

We established the website http://aria-agent.eu that contains blog entries and news from all partners of the project (see Fig. 4.1.) It further features a list of publications and demo videos, as well as direct download links of the AVP system, the NOVA tool and the NoXi corpus.

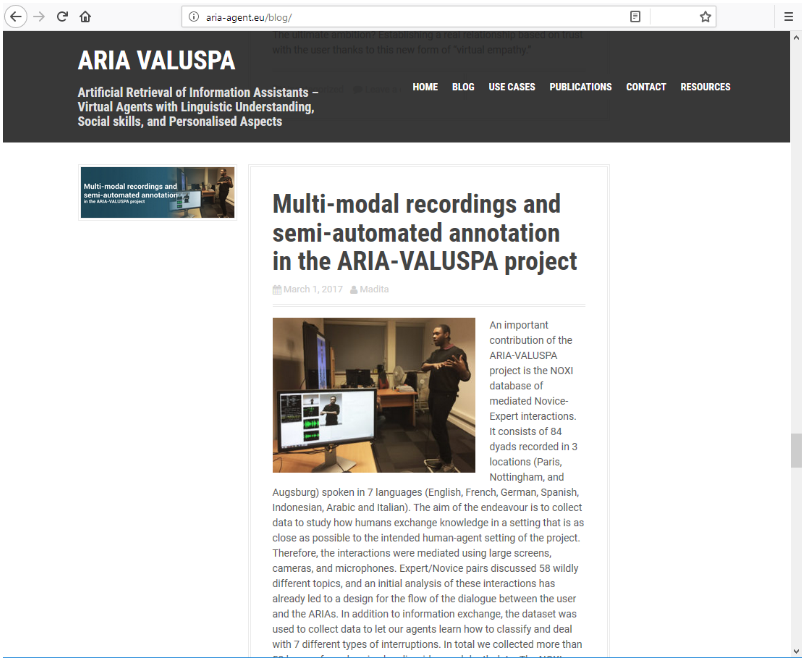


Figure 4.1: The Aria Valuspa Blog.

Further we disseminated blog entries on the Social Media Platforms Facebook [4] and twitter [5].

4.2 Science Museum Lates

A team from CereProc supported by Dr Eduardo Manuel De Brito Lima Ferreira Coutinho from Imperial College demonstrated various aspects of speech synthesis at the Science Museum in London, as part of the Royal Society's 'The Next Big Thing' project.

[4]https://www.facebook.com/ariavaluspa/
[5]https://twitter.com/ariavaluspa

Lates is a prestigious free public event held once a month where adults take over the Science Museum. Every event has a different theme, covering a wide range of topics, from climate change to alcohol, from childhood to robots. These showcases have turned out to be extremely popular and attract around 5000 visitors per night.

Not surprisingly then, CereProc / ARIA team was kept busy all night. Our main activity was 'Bot or Not' – a quiz that lets you test your ability to recognise a synthetic voice and learn about speech synthesis in the process. Everyone who took part was added to the leader board and received a personalised message from Donald Trump (totally fake of course, generated using CereProc's prototype Trump voice).

Feedback showed that most players found it a lot more difficult than they thought it would be, and no one has yet reached the perfect score of 20/20.

We also introduced visitors to (the voice of) Roger who gets very cross if you try to interrupt him while he's speaking. The interruption demo was created as a test harness for the reactive speech synthesis framework, and used a set of different strategies to respond to user's and demonstrator's interruptions. These varied from gracefully ceding the floor and restarting, to holding the floor with tense voice quality (see section 3.3 for more details).

In addition, Dr Coutinho presented his work on sentiment analysis by demonstrating how to tell if a politician is being sincere when giving a speech. Once again, Mr Trump took a centre stage! Visitors also got a chance to record and analyse their own speech for signs of dis-ingenuousness.

4.3 Book deal

This section has been redacted from the public version of this report.

4.4 Public panels, talks, and keynotes

A large number of public panels, academic talks and invited keynotes were delivered that featured (research conducted in) ARIA-VALUSPA. Below is a full list of such talks:

- Matthew P. Aylett: Bot or Not? Exploring the perception of acted, modified and synthetic speech. Keynote, workshop "Investigating Social Interactions with Artificial Agents", ICMI November 2017
- Matthew P. Aylett: Delighting the User with Speech Synthesis. Invited talk. Symposium on speech synthesis with special emphasis on non-verbal control KTH, January 2017
- Matthew P. Aylett: Delighting the User With Speech Synthesis. Glasgow Social Robotics, invited talk. November 2015
- Dirk Heylen: Keynote. On Labels, Theory, Methodology, Practice. Boston. Society for Affective Science, April 2017

- Dirk Heylen: Keynote. Reflections on Data Collection and Annotation. IEEE SMCS Technical Committee on Computational Psychphysiology. Beijing. May 2017.
- Dirk Heylen: Keynote. Engagement in Conversation Revisited. ACII Workshop on User Engagement and Interaction. San Antonio. October 2017.
- Catherine Pelachaud: Keynote, workshop "Investigating Social Interactions with Artificial Agents", ICMI November 2017
- Catherine Pelachaud: Keynote, 3rd Global Conference on Artificial Intelligence, Miami, USA, October 2017.
- Catherine Pelachaud: Keynote, ACM Symposium on Applied Perception, SAP'17, Cottbus, Germany, Sept 17
- Catherine Pelachaud: Keynote, Robotics and Emotions, International Robotics Festival, Pisa, September 2017
- Catherine Pelachaud: Keynote, Interaction with Agents and Robots: Different Embodiments, Common Challenges, satellite workshop of IVA'17, Stockholm, August 2017
- Catherine Pelachaud: Keynote, 3rd Workshop on virtual social interaction, Bielefled, Allemagne, July 2017
- Catherine Pelachaud: Keynote, Interspeech, Stockholm, August 2017
- Björn Schuller: Keynote "LP in Tomorrow's Profiling – Words May Fail You", 14th International Conference on Natural Language Processing (ICON 2017), Kolkata, India, 18.-21.12.2017.
- Björn Schuller: Keynote "Mental Health Monitoring in the Pocket as a Life Changer? The AI View.", ISRII 9th Scientific Meeting, International Society for Research on Internet Interventions (ISRII), Elsevier, Berlin, Germany, 12.-14.10.2017.
- Björn Schuller: Keynote "Big Data, Deep Learning – At the Edge of X-Ray Speaker Analysis", 19th Conference on Speech and Computer (SPECOM 2017) jointly with 2nd International Conference on Interactive Collaborative Robotics (ICR 2017), Hatfield, UK, 12.-16.09.2017.
- Björn Schuller: Keynote "Automatic Speaker Analysis 2.0: Hearing the Bigger Picture", The 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD 2017), IEEE/EURASIP, Bucharest, Romania, 06.-09.07.2017.
- Björn Schuller: Keynote "24/7 Computational Psychophysiology", IEEE SMCS Technical Committee Workshop on Computational Psychophysiology, IEEE, Beijing, P.R. China, 22.-23.05.2017.

- Björn Schuller: Opening Plenary “Artificial Emotional Intelligence – A Game Changer for AI and Society?”, Annual Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB), AISB, Bath, UK, 19.-21.04.2017.

- Björn Schuller: Keynote “Reading the Author: A Holistic Approach on Assessing What is in one’s Words”, 18th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Springer, Budapest, Hungary, 17.-23.04.2017.

- Björn Schuller: Plenary Keynote “Tiefes Lernen und die breiten Ḿoglichkeiten Anwendungen invited talk, MobileTechCon, Munich, Germany, 13.-16.03.2017.

- Björn Schuller: Keynote “Engage to Empower: Emotionally Intelligent Computer Games & Robots for Autistic Children”, Conference on âĂIJThe world innovations combining medicine, and technology in autism diagnosis and therapy”, SOLIS RADIUS, Rzeszow, Poland, 29.09.2016.

- Björn Schuller: Keynote “Intelligent Diagnosis and Monitoring of Autism”, Conference on âĂIJThe world innovations combining medicine, and technology in autism diagnosis and therapy”, SOLIS RADIUS, Rzeszow, Poland, 29.09.2016.

- Björn Schuller: Keynote “Computational Paralinguistics in Everyday Environments”, The 4th International Workshop on Speech Processing in Everyday Environments (CHiME 2016 Workshop), San Francisco, CA, 13.09.2016.

- Björn Schuller: Keynote “7 Essential Principles to Make Multimodal Sentiment Analysis Work in the Wild”, 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016 Workshop, IJCAI/AAAI, New York, NY, 10.07.2016.

- Björn Schuller: Keynote “Say no more – the computer already deeply knows you?”, SWS 2016 Speech Signal Processing Workshop, ACL/ACLCLP, National Taiwan University, Taipei, Taiwan, 18.03.2016.

- Björn Schuller: Keynote “‘Less Input’: Cooperative Learning for Emotionally Intelligent Systems”, 4th Machine Learning for Interactive Systems Workshop (MLIS 2015) held at the International Conference on Machine Learning (ICML’15), Lille, France, 11.07. 2015.

- Björn Schuller: Keynote “Computational Paralinguistics: Breaking the Voice”, UK SPEECH 2015 – 4th Meeting of the UK and Irish Speech Science and Technology Research Community, Norwich, UK, 01.-02.07.2015.

- Björn Schuller: Keynote “Speech Analysis in the Big Data Era”, 18th International Conference on Text, and Dialogue (TSD 2015), Springer LNCS/LNAI, Plzen, Czech Republic, 14.-17.09.2015.

5 Economic Impact

Below we give an estimate of the economic impact, for as far as we are able to judge this.

5.1 Industrial Impact

Economic impact on our industrial partners was assessed but has been redacted from this public version of the report.

5.2 New funded activities directly following ARIA work

A number of projects and activities are the direct result of research, technology, and know-how developed in ARIA-VALUSPA. The total worth of these new projects is more than 14.5 Million Euros.

5.2.1 Affective Language

The University of Twente has received a 236,663 EURO award (total funding: 733,800 EURO) from the Dutch NWO funder, for the project 'Affective Language'. This project aims to investigate (1) how the emotional states of speakers influence the language they produce and (2) how the influence of emotion on language production can be modeled in computational tools for affective natural language generation.

5.2.2 ALTCAI

The 338,000 Euro ALTCAI project funded by the UK's DIFFID department has been awarded to Dr Valstar of the University of Nottingham to research using ARIA agents to deliver health-care advice to people in sub-Saharan Africa.

The University of Nottingham's 5-year, 26,600,000.- EURO Biomedical Research Centre has adopted the ARIA-VALUSPA code-base, ensuring access and improvements to AVP until at least April 2022. Valstar is the deputy director of the Mental Health theme making up 20% of the total centre's activities, and will use ARIA agents to create diagnostic, monitoring, and treatment tools for people with major depression disorder.

5.2.3 COUCH

The University of Twente and UPMC-CNRS have received 3,704,000 EURO, (UPMC-CNRS 657,500 EURO, U Twente 777,357 EURO) for a Horizon 2020 project called COUCH: Council of Coaches. In COUCH, multiple virtual coaches form a personal council that supports the user in their health and well-being. Individual coaches have their own area of expertise, personality, and style of coaching. Join a council meeting! Give the council your thoughts, or listen and observe how the individual coaches exchange their views on your health behavior. Take what you've learned into your daily life, and if the need arises, contact any of the coaches anytime, anywhere.

5.2.4 EVA

The University of Augsburg has won funding from the German Science Foundation (DFG) to support two research associates, student researchers, travel for EVA: How to Win Arguments - Empowering Virtual Agents to Improve their Persuasiveness . The EVA project, we will simulate argumentation dialogues between humans through embodied conversational agents. We will rely on Reinforcement Learning (RL) to optimize the agents' argumentation strategies in an interaction with a simulated opponent.

5.2.5 EMMA

The University of Augsburg has won funding from the German Federal Ministry of Education and Research (BMBF) worth ca. 1 M Euro for EMMA: EmmA âĂŞ Emotionaler mobiler Assistent. In the EMMA project, Augsburg will investigate the use of a socio-emotional assistance system for improving the psychological health of people at work.

5.2.6 FIODSpraak

The University of Twente received a 207,349 EURO award from the FIOD - the Dutch Fiscal Information and Investigation Service, to realise an infrastructure (based on a set of software tools, available data and protocols) to speed up and improve the analysis of recorded conversations.

5.2.7 GrassrootWavelengths

CereProc's involvement in the GrassrootWavelengths project is a direct result of our success in working with academics and publishing innovative work on expressive speech synthesis - a key area of innovation in the ARIA project.

The Grassroot Wavelengths is a 2.17M EURO project that will create a game-changing network of inclusive digital platforms for citizen engagement, community deliberation, and the free flow of information within, into, and out of discrete geographic communities by piloting solutions for connected, inexpensive, community owned and operated radio across Europe. See `https://cordis.europa.eu/project/rcn/213180_en.html` for more details.

5.2.8 R3D3

The University of Twente has received 110,250 EURO (total funding: 247,500 EURO) for R3D3: Rolling Receptionist Robot with Double Dutch Dialogue from COMMIT - Zwaluw Project. The R3D3 project investigated situated natural language dialogue systems that combine limited natural language understanding with the understanding of non-verbal behaviour of the users in a real-life context.

5.2.9 VIVA

The University of Augsburg has won funding from the German Federal Ministry of Education and Research (BMBF) worth ca. 2 M for VIVA - Vertrauen und Sympathie schaffender lebendiger sozialer Roboter. The objective of the VIVA project is the development of a hardware and software platform as a basis for the creation of lively social robots that establish trust and empathy.

6 Summary of Technical Effort per Work-Package

This section provides a brief summary of all work done, structured by work-package and task. Rather than repeat reports, we reference heavily to relevant previous deliverables or sections of this final report. For the technical work-packages, we also report on the progress made after July 2017, which is not captured by any dedicated work-packages.

6.1 WP1: System design and realisation for web and mobile device environments

6.1.1 1.1 System integration

Task 1.1 is all about continuously integrating the various processing components. By having a dedicated scientific programmer at Nottingham, this was ensured. Only when this post changed from the original member of staff to a new member of staff was there a period of about 3 months where continuous integration was paused. We employed an early integration approach, and had a fully functioning minimal viable product at the end of year 1 (Milestone 1, or AVP 1.0). The aim was to continuously keep the components created by all partners in synchrony to have a working system all the time even while new functionality is added. This was mostly successful, although there were invariably times where progress in one module broke functionality of another, requiring us to temporarily use an older version of the offending module while the problem was resolved.

The task was kick-started using the SEMAINE system, insofar that we used the same messaging protocol (ActiveMQ) and messaging formats, as well as the general system divide of behaviour analysis, dialogue management, and behaviour generation. But in AVP, the three blocks are themselves larger integrated entities and most ActiveMQ communication is between these three blocks, rather than a very large number of small modules all communicating through ActiveMQ.

Where possible, we adhered to software engineering best practices, such as continuous integration, and relatively short release cycles. AVP was released as versions 1.0 (12 months), 2.0 (24 months), 2.1 (26 months), 2.2 (28 months), 2.3 (30 months), 2.4 (32 months), and 3.0 (36 months). Clearly, as all partners were closer to integrating their research into software, there was a greater need and opportunity for new releases. This was particularly the case for the Dialogue Management, on which work did not seriously start until month 15 of the project.

All partners received training on software during two courses, one in Augsburg (month 12) and the other in Twente (month 19), as part of face to face consortium meetings.

6.1.2 Task 1.2 End-to-end system realisation on web and smartphone technology

The original proposal envisaged the realisation of ARIAs on mobile devices using web browser technology. However, after the mid-term review the official advice was not to pursue this goal.

Nevertheless, some progress has been made. The iCCR face-tracker was deployed to the iOS platform, running on a native platform in less than 20 ms/frame, i.e. at 50 frames per second, allowing some additional processing on top of the face tracker while maintaining a 30 frames per second throughput. The AVP allows sending of audio and video data using webRTP from a chrome browser to a server running AVP. Cantoche have made a streaming server to send video of the generated Living Actor from this AVP server back to the browser. So, in theory a fully reactive web-enabled system has been delivered, however it isn't seamlessly integrated with the main AVP codebase, and has not been extensively tested.

6.1.3 Task 1.3 Realisation of a real-time distributed system

This task focuses on delivering a real-time distributed system. Thankfully, by employing the SEMAINE Active-MQ system and existing components such as eMax, SSI, and Greta, such a system came virtually for free from Milestone 1 (AVP 1.0) onwards. AVP runs seamlessly on multiple machines. Because the ASR is the only module that requires Linux as the operating system, this either runs in a virtual machine, but in most of our experiments on a separate physical machine. eMax works best on machines with a decent Graphics card, so it can utilise the GPU. In practice, we often use two or three machines to run ARIA interactions.

6.1.4 Task 1.4 Support for user-profiles

The user adaptation capabilities in WP 2-4 crucially depend on the system's ability to represent and remember a user's profile. To do so, we use a simple but effective face recognition system based on the facial point tracking and appearance of the face, and integrate this in eMax as part of the visual behaviour analysis. The face recognition performs well for well-lit, frontal-view faces. The user ID utilised in the eMax face recognition module is sent by SSI over ActiveMQ to the Dialogue manager, which links it to its in-built user profile system. This system stores the agent's belief of a user's age, gender, and preferred language. Unfortunately, recognising and pronouncing people's names is a very hard problem, so we don't store or use people's names. Instead, the DM can say "Hi again!" when we recognise someone as being in the database, and "Hi, we haven't met before, have we? It's nice to meet you." when the user-id is new. It also allows us to count the number of interactions with a person, which can be used to build e.g. longer term personal relations. Combined with storing the interaction history, that is the dialogue history together with the interest level at different times, allows the agent to learn longer term what a user finds interesting and what not. This long-term adaptation remains future work.

6.1.5 Task 1.5 Implementation of standards

Several web standards that will be used by us are in the process of formal specification. Many of these are highly relevant for the future commercial success of ARIA-VALUSPA technology. Among them are the W3C HTML Speech Incubator Group and the W3C Emotion Markup Language. We have adhered to these standards wherever possible.

6.2 WP2: Multi-lingual audio-visual-modal speech and affect recognition

6.2.1 Task 2.1 Cross-domain, audio-visual multi-lingual detection of verbal and non-verbal cues

This task is focused on recognition of naturalistic speech in three languages, as well as affect and interest recognition from nonverbal cues from prosodic and facial dynamics.

Speech recognition We have developed a fully Automatic Speech Recognition system for English, French and German languages that is real-time capable with very low latency. The objective of this module is to recognise the verbal content of user's voice and send that information to the dialogue system for further processing and the generation of the verbal interaction. The basic architecture of our ASR system was created using the Kaldi toolkit, a very well-known open-source ASR toolkit that is well tested, optimised and actively maintained by many researchers to support the state-of-the techniques (and therefore can easily be extended and further developed). Deliverable 2.1 includes a complete technical description of the ASR system. The final versions of our ASR models have the following Word Error Rates: 39.0% (English), 28.8% (German) and 40.2% (French). These values were estimated on a subset of the NoXi database. See Deliverable 2.2 for further details on the ASR implementation and performance.

Recognition of the user affective state and interest level In the context of this task, we started our work by developing single modality classifiers for affect – Arousal (high, low) and Valence (positive, negative) – and interest – uninterested, neutral, interested – recognition systems. For arousal, we obtained an Unweighted Average Recall (UAR) of 68.9 %, for valence, 61.6 %. We developed well-established models (Support Vector Machines) and engineered features sets (using the openSMILE audio feature extractor). Arousal and Valence from video was established based on the 6-basic emotion prediction. Full details of this initial is presented in Deliverable 2.1.

6.2.2 Task 2.2 Automatic Audio-visual User Profiling

User traits estimated from vocal patterns We implemented models for the recognition (classification tasks) of the following traits for creating the user profiles; class labels and the classifier performances (UAR) are also given:

- Age: children, youth, adults and seniors; (UAR = 48.91 %)
- Gender: children, female, and male; (UAR = 81.21 %)

These long-term traits are used to profile the users. Furthermore, according to the WP, we have conducted research on the recognition of the big five personality traits (mean

UAR = 71.4 %), native language (L1) (UAR = 47.5 % for eleven classes), and health condition (UAR = 70.2 % for cold vs non-cold), as well as drowsy state induced by alcohol intoxication or sleep deprivation (UAR = 69.0 % for drowsy vs non-drowsy). The obtained results are consistently significant above chance. In addition, we proposed a multi-task learning method based on deep neural network with shared hidden layers for universal speech emotion recognition. Thus, our system is able to predict various emotion representations based on categorical, dimensional, and appraisal modelling conceptions. Full details about these models are described in Deliverable 2.1 and in our recent publications as listed in Section 7.3.

6.2.3 Task 2.3 Adaptation to user, context, and environment

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.

In relation to demographic information, user traits extracted from utterances in Task 2.2 (automatic user analysis) were used for

- preselecting interest and emotion models that best fit the user profile (e.g. age group, gender, Speaker ID);
- as high-level features for adapting the dialogue management system;
- automatically adapting emotion and interest recognition models to the speaker's voice in the course of the dialogue.

We achieved these outcomes through the creation of a novel learning framework for acoustic model adaptation that allowed us to train customised models for a single user and user groups during long-term interaction with ARIA. In order to identify the user at each moment and initiate user adaptation strategies, we implemented reliable face identification algorithms (see Fig. 6.1). When a specific, known user is identified, the respective user profile is retrieved and previously adapted models are initiated. Otherwise, other models that are as close as possible to the new user profile are used (see Fig. 6.2 for a visualisation of the adaptation process). Full details on these features are provided on Delivery 2.3.

We also endowed our ASR systems with the capacity to adapt to a specific user voice in real-time. This is achieved through the generation of speaker dependent acoustic features that provide the acoustic models with a vector of speaker-dependent features (iVector) in addition to the standard un-adapted and non-normalised MFCC features. The iVector in our system include 100 features that provide the acoustic models with sufficient knowledge about the speaker characteristics. This vector is estimated in a left-to-right way, i.e., at a certain time t, it sees input from time zero (beginning of the segment) to t (present moment). It also receives information from previous utterances of the current speaker, i.e, from the beginning of the session and will therefore improve over time.

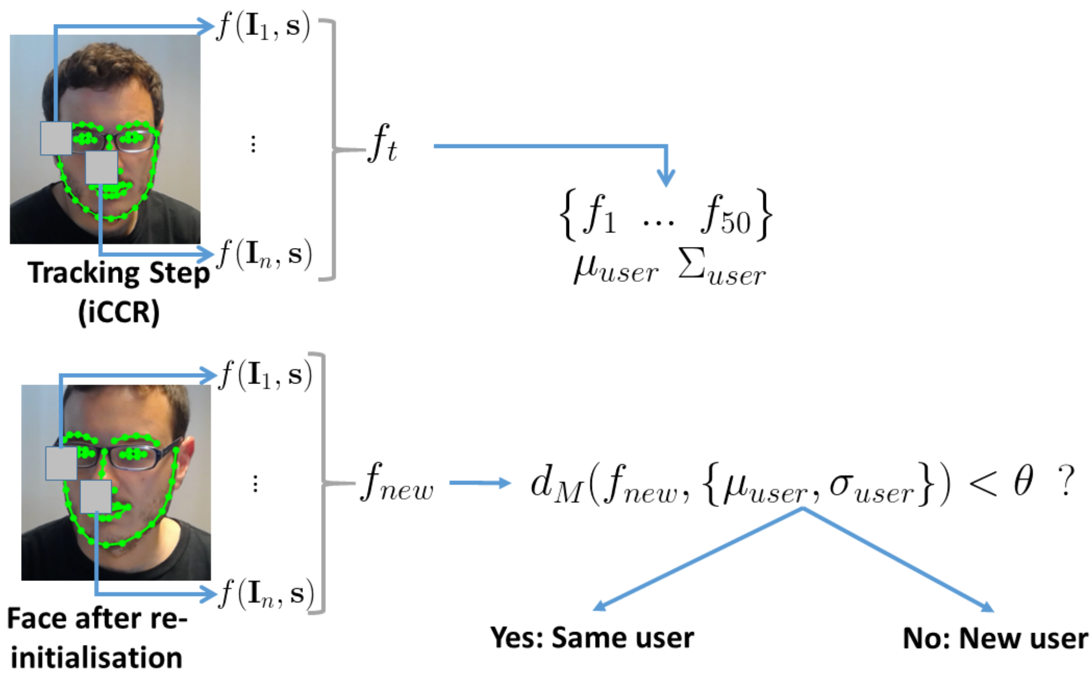


Figure 6.1: User re-identification procedure: While the tracking is ongoing with no failure, the user-specific features are stored in a sliding window of 50 frames. When the tracker needs to be reinitialised, the new features are first compared to the statistics stored for the previous user (mean and covariance of the features). When the distance is higher than a threshold, the system detects a new user, and the data collection re-starts.

6.2.4 Task 2.4 Audio-visual Fusion for Social and Emotional Skill Enhancement

Human emotions are expressed non-verbally in multiple ways, being the face and the voice two of the most important modalities. Not only information from both these modalities is important, but the interplay between the expression of emotion in both face and voice are relevant for the communication of particular emotions. In order to create a more robust emotion recognition system, we developed a novel framework for time-continuous audio-visual emotion recognition – Dynamic Difficulty Awareness Training (DDAT). The DDAT framework consists of a multi-task learning scenario to train a Deep Neural Network whereby, in addition to predicting Arousal and Valence affective dimensions, the uncertainty level of these predictions based on the agreement level between annotators is also predicted. The emotion prediction uncertainty is used in this context as a proxy for the "difficulty level" of the task, and together with audio-visual features, is fed to the input layer of the DNN at the following time step (see Fig. 6.3a, i.e., the estimated "difficulty" of the task is regarded as a complementary descriptor of the audio-visual

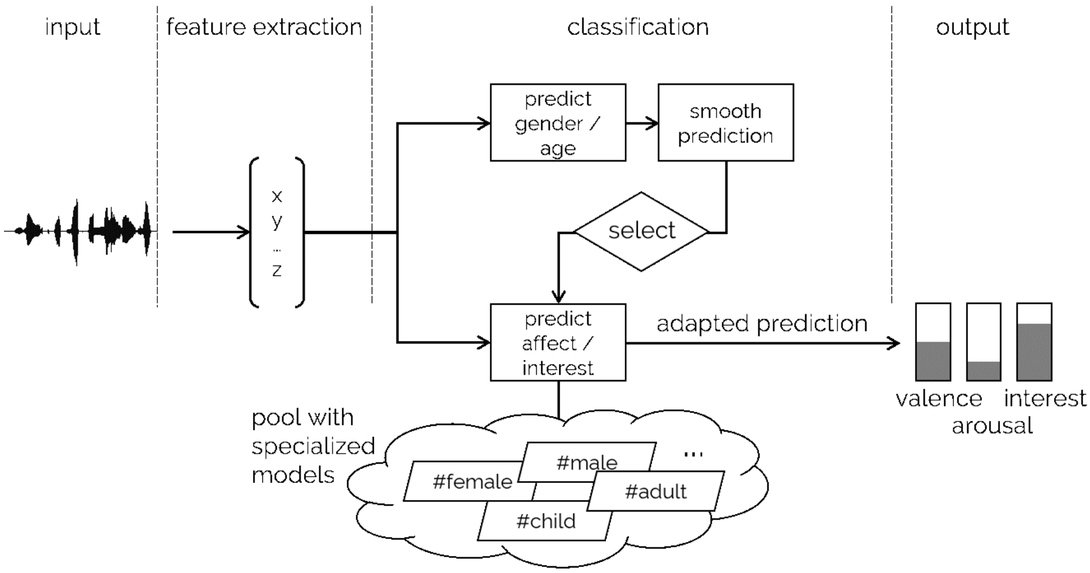


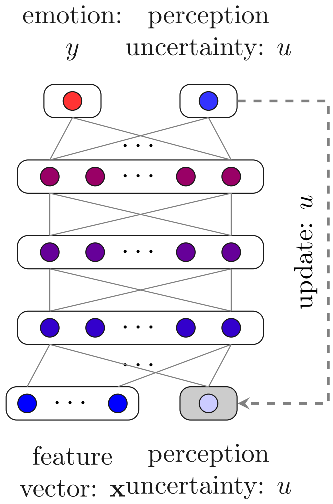
Figure 6.2: Real-time workflow for user adaptation in ARIA.

features and combined with them to form an extended feature set for emotion recognition (both during training and testing phases). The goal of this approach is to allow the model to create an expectation of the difficulty level for emotion prediction at any given moment, and eventually use this knowledge to improve the learning process. This assumption is inspired by human learning processes in which levels of attention are normally higher when learning difficult or ambiguous tasks.

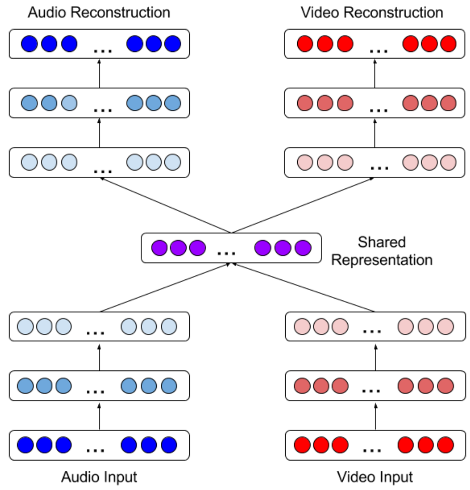
6.2.5 Advances made after July 2017

Intermediate audio-visual behaviour representation Following a recent trend in machine learning that aims at building intermediate representations of raw input signals in order to extract task-specific information (which usually leads to a better performance on the recognition task), we used Convolutional Neural Networks (CNNs) to extract features from the speech, and a deep residual network (ResNet) of 50 layers for video (see Fig. 6.4. These models were then combined with a Long Short-Term Memory (LSTM) networks, and trained in an end-to-end fashion where - by taking advantage of the correlations of the each of the streams - we manage to significantly outperform the traditional approaches based on auditory and visual hand-crafted features for the prediction of spontaneous emotional displays.

Face frontalization Another problem we worked on is the automatic frontalization of face images, i.e. creating a canonical frontal view representation of a face of arbitrary head-pose. Recent advances in deep learning have led to high performing facial expression recognition algorithms. However, their performance can still severely degrade with large variation in head pose. Rotating the face to a reference position (as a pre-processing step)



(a) Dynamic Difficulty Awareness Training (DDAT) framework. Augmented inputs are updated using emotion perception uncertainties.



(b) Bimodal stacked denoising autoencoder architecture.

before being used as input to a classifier/regressor is an effective solution to overcome this problem. However, this frontalization step becomes especially challenging when out of plane rotations are involved as that results in non-linear transformations of the face shape and appearance, and often parts of the face are self-occluded. With this goal in mind we developed a data driven approach for generating frontalized face images using deep Convolutional Neural Networks.

We trained a hourglass network [35] to generate frontalized version of the input faces. The network consists of 4 residual modules designed to capture information at multiple scales. Following these modules, there are 4 additional layers to do the up-sampling and combining features across adjacent resolutions. The input to the network is a cropped face image which can be in any pose. The network is trained to output frontalized face image using Mean Squared error loss. The method was trained and evaluated on the BP4D dataset [58] and it achieved a RMSE of 0.11 on a validation set. Some results from our proposed method are shown in Fig. 6.5. If it is found that the frontalisation is indeed helpful for further face analysis, this method will be introduced into eMax, and thus the Behaviour Analysis pipeline in due course.

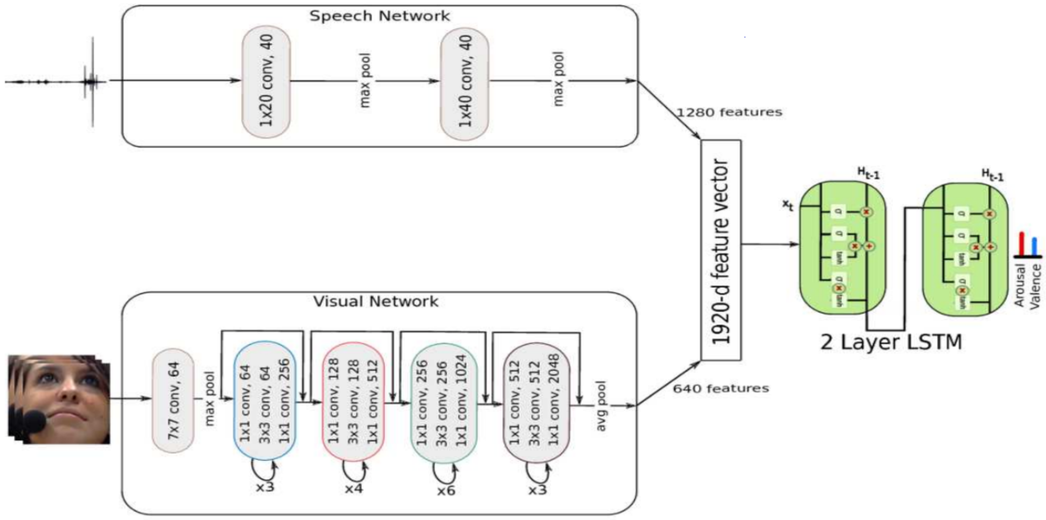


Figure 6.4: The network comprises of two parts: the multimodal feature extraction part and the RNN part. The multimodal part extracts features from raw speech and visual signals. The extracted features are concatenated and used to feed 2 LSTM layers. These are used to capture the contextual information in the data.

6.3 WP3: Multi-modal dialogue management for Information Retrieval

6.3.1 Task 3.1 Multi-lingual natural language understanding

This task concerns the way the ARIA-VALUSPA dialogue management system handles the understanding of the user's utterances in the three natural languages targeted by the project: English, French and German.

In the dialogue management system, user intents are defined independent of language, according to the DIT++ (Dynamic Interpretation Theory) taxonomy of communicative functions [9]. Intents include greeting, approach, topic introduction, inform, question, valediction, thanking and apologizing. These intents are generally applicable for all languages and only require a translation of the verbal parts.

As described in Section 6.2.1, an automatic speech recognition (ASR) system has been developed for English, French and German. For the subsequent step of handling multi-lingual user input, we have opted for a shallow understanding approach that is not dependent on the availability of language-specific resources. After the utterance recognized by the ASR has been put in the information state, and a user move has been created (see section 3.8), the Input Processing component checks if the user's utterance matches any of the predefined user utterances in the agent's moves (where a match

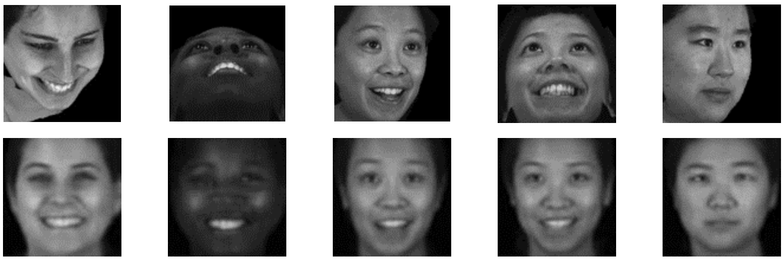


Figure 6.5: Some generalisation results from our face frontalization algorithm on a left-out test set. The top row shows the input (non-frontal pose) face images. The bottom row shows the frontalized face images generated from the network.

would make the agent move relevant). This matching of user utterances is done in a language-independent fashion by measuring unigram overlap between sentence strings; it does not involve any syntactic or semantic parsing.

The user utterances specified in the agent moves have been written in English, but simply translating them is sufficient to have dialogues in French or German as well.[6] Future users of the system can author the dialogue templates in their preferred language, provided that an ASR component and a text-to-speech component are available for that language. For non-verbal behaviour specified in the moves, we make the simplified assumption that there is no difference amongst the different languages.

6.3.2 Task 3.2 Task-oriented dialogue management

This task involves the implementation of an information state-based architecture for dialogue management that can run different dialogue management dimensions in parallel, focusing on interaction management aspects such as turn taking, repair mechanisms, and floor management.

The dialogue management system uses an information-state based approach. The information state according to Traum and Larsson represents "the information necessary to distinguish it [= the current dialogue] from other dialogues, representing the cumulative additions from previous actions in the dialogue and motivating future action" [46]. An advantage of the information-state based approach is the ability to formalize linguistic models which in turn enables easier integration in a computer system.

The dialogue engine underlying the dialogue management system is called Flipper 2.0: an extended and improved version of the Flipper engine described in [44]. Development of Flipper 2.0 involved multiple steps. We first built an extension of the original Flipper to enable a multi-agent approach and facilitate integration with other dialogue system components. We also made some other improvements to Flipper, such as making it possible to load dynamic classes which can be called. We then updated Flipper to Flipper 2.0 by using more standardizations such as JavaScript and JSON. This makes

[6] The dialogue templates used in the book personification demonstrator have already been translated to French; translation to German is work in progress.

the dialogue engine less computationally expensive and also creates more flexibility in defining dialogues, useful for creating more dynamic conversations.

Based on Flipper 2.0. we designed a dialogue manager that provides a dialogue structure based on the DIT++ standard. We have extended DIT++ with some domain-dependent functions (see next section), making it possible to specify domain-specific as well as domain-independent dialogue moves that cover multiple conversational dimensions. The dialogue manager takes a scenario and situation-driven approach to creating dialogue structures based on conversational acts; see Section 3.8 for more details.

6.3.3 Task 3.3 User-adaptive dialogue strategies

A special new feature of the ARIA-VALUSPA dialogue management is adaptation to the user. Adaptation, or alignment, in natural dialogues appears in many aspects of dialogue. This task involves the implementation of different adaptive strategies that are relevant to the application.

One simple form of user adaptation is by making use of the input provided by the SSI framework. The SSI can detect demographics of a user (e.g. the gender) and this can be used by the agent to create an appropriate of responses, for example, 'Hello Sir' versus 'Hello Madam'. Other, more advanced forms of adaptation we have focused on are adaptation in terms of turn-taking and alignment of the agent's word choice to that of the user. Here focus on the alignment.

We measured the verbal alignment between the user and the agent in a small corpus of Wizard of Oz dialogues (the HAI corpus; see D6.1 and section 6.3.5) and used this as a basis to create an algorithm that performs verbal alignment, as shown in Figure 1.

```
Data: Dialogue history, planned agent utterance
Result: List of all possible (un)aligned agent utterances
initialization;
while NPs with modifiers remaining in agent utterance do
    select NP;
    if aligning then
        remove modifiers not used by user;
        add modifiers used by user;
    else
        remove modifiers used by user;
        replace modifiers with synonyms;
    end
end
```

Algorithm 1: Verbal alignment generation

Using this algorithm makes it possible for the agent to use referring expressions that are preferred by the user (as shown by the dialogue history). An example is talking about the 'tiny golden key' found by Alice in the book Alice in Wonderland. Depending on the expression used by user, the agent can refer to it in the agent utterance as 'key', 'golden key', 'tiny key' or 'tiny golden key'. If the user uses a particular description, the agent is able to mirror this.

6.3.4 Task 3.4 Reinforcement learning based on user feedback

This task refers to training the system's adaptive dialogue strategies using reinforcement learning. Our approach to dialogue management lends itself to online learning. Our dialogue policy is based on a single value of relevance (described in section 2), used to select behaviours of the agent. The multi-modal information from the agent's mental model (the information state) could be encoded to use the relevance value as a reward. The dialogue history could be used as an input for learning as well, in addition to all possible dialogue moves the agent can make. We have collected data (WOZ) and implemented relevance values that could be learnt from. Actually performing reinforcement learning experiments on these data remains as future work.

6.3.5 Task 3.5 Dealing with unexpected situations

This task concerns enabling the ARIA agents to deal with unexpected situations that occur during an interaction. For this task, we started out by collecting Wizard-of-OZ dialogues in the Alice in Wonderland domain. The goal of the data collection was to create a corpus with unexpected situations that can occur during a conversation between a virtual agent and a user, such as misunderstandings, (accidental) false information, and interruptions by another person. In a classic WOz approach where the wizard uses a button interface, it is nearly impossible to improvise in unexpected situations. This is why we gave our wizard the freedom to choose his own words and facial expressions to respond to and initialize unexpected situations. The corpus, called HAI (Human-Agent Interaction) Alice, consists of 15 conversations and more than 900 utterances.

One type of unexpected situation occurs when the agent gives an unexpected answer due to speech recognition errors or knowledge base limitations. We carried out a few additional WOz studies to investigate the use of dialogue repair strategies in this type of situation. In the first study, we investigated which repair strategies the users employed in reaction to off-topic answers by the agent. In most cases, they reacted with a clarification question or a follow-up question about the unexpected answer. The next most frequent reaction was to simply ignore the answer and ask a different, unrelated question. This is a strategy we do not expect to see much outside the experimental context, where the participants have no real information need. Therefore, in the next experiment we investigated the effects of adding some simple repair strategies to the agent's repertoire: instead of giving an irrelevant answer, the agent asked the user to rephrase their question, or said that it could not answer the question. These repair strategies slightly improved the performance of the virtual agent (i.e. more questions were answered correctly). Ironically, they also caused the participants to perceive the agent as less intelligent, presumably because the repairs drew attention to the agent's limitations. For this reason we have not implemented such strategies in the dialogue manager.

Instead, we have implemented a strategy where the agent only initiates a repair strategy if the user explicitly signals that the agent has made an inappropriate response. If the user has a negative reaction to the agent's response, the agent will politely apologize and repeat what she thought the user said. This way the agent does not repeat itself and gives the user some feedback on what it has recognized. This gives the user an opportunity to voluntarily rephrase their earlier utterance, while maintaining the current topic of the conversation.

Finally, we have implemented a method in the dialogue manager to deal with interruptions based on the personality traits of the agent. For example, you can set the agent to be dominant and she will talk louder and finish her behaviour, even if the user tries to interrupt her. You can also make a more submissive agent that stops talking as soon as she hears the user speak, or minimizes the overlap with the user's speech. It is also possible to dynamically alter the approach to user-interruptions, making the agent less or more receptive to user interruptions.

6.3.6 Task 3.6 Generation of Dialogues for Book Personification demonstrator

For the generation of dialogues for the book personification application, we aimed to create a set of dialogue structures that covers the themes in the book and avoids open domain conversation. In preparation for this, we performed an analysis of a subset of the NoXi corpus of novice - expert conversations (see Section 2.3). Our focus was on determining the general discourse structure of these conversations, following the intentions of the interlocutors. More details can be found in Deliverable 3.2, Section 2.3.

A first version of the book personification demonstrator was described in Deliverable 3.2. For this first version, We extended DIT++ with domain-dependent functions, such as talking about a character in the book or storytelling about an event in the book. We use a simple semantic representation of utterance content, making it possible to specify the topic and type of open questions, thus refining the general category of 'setQuestion' (open question) in DIT++.

Since then, the demonstrator has been extended with question-answer (QA) matching functionality comparable to that of the VH Toolkit [21]. A question by the user is mapped to the most similar question found in a database with question-answer pairs, and the corresponding answer is returned.

In addition to the HAI Alice corpus (see above) through which question-answer pairs were collected, we carried out several other small WOz studies to obtain more question-answer pairs and to populate the initial knowledge base of our agent.

To easily expand the range of user questions that the Book Personification agent can answer through QA matching, we have developed a question generation system that takes text as input and generates a large number of QA pairs from it [19]. This form of question generation can be used to complement or replace data collection with human users. For example, based on the input sentence (from a summary of Alice's Adventures in Wonderland) 'Her giant tears form a pool at her feet' the following questions can be generated: 'What happens at her feet?' Or: 'What happens to her giant tears at her feet?' The answer to both questions is 'Her giant tears form a pool'. The software for

Table 6.1: An example dialogue with the Book ARIA. (Sp=Speaker, U=User, A=Agent)

#	Sp.	Utterance
1	U	Hello
2	A	Hello to you too, human!
3	U	What is your name?
4	A	I'm Alice, from the book Alice in Wonderland.
5	U	What can you do?
6	A	I can tell you about the book Alice in Wonderland.
7	U	Tell me about the book.
9	A	I followed the rabbit into a hole and then fell down a very deep well! I came upon a heap of sticks and dry leaves and found myself in a long, low hall. There, on a stool, was a tiny golden key!
10	U	What did you do with the key?
11	A	I looked around and I found a little door. The key fitted in the lock, but the door was too small for me to pass! I looked around again and found a bottle with 'drink me' on it.
12	U	Did you drink the bottle?
13	A	The 'drink me' bottle did not have poison written on it, so I drank it! Then I felt myself shrinking, until I was only 10 inches high.
14	U	What is your name?
15	A	My real name is Alice Liddell.
16	U	Do you have any siblings?
17	A	My parents had nine children. When I was born, I had an older brother and sister, and a younger sister.
18	A	Right, look at the time! I am afraid I must go at once, tea time with my cat!
19	U	Alright, goodbye.
20	A	Goodbye!

this question generation can be found on GitHub[7].

6.3.7 Task 3.7 Generation of Dialogues for Industry Associate demonstrator

This task is devoted to the Industry Associate application. It encompasses defining the exact mission of the Virtual Assistant, creating a dialogue scenario to achieve this goal, and building adaptive and task-oriented dialogues in multiple languages.

The remainder of this section has been redacted for the public version of this report.

6.3.8 Advances made after July 2017

After the last deliverable in July, the team in Twente worked to improve the system and prepare it for the final evaluation in October and November.

In September we integrated the new dialogue engine, Flipper 2.0, into the dialogue manager. We implemented most of what was described in Deliverable 3.3, Section 2.1. We defined episodes, exchanges and content, interaction and social moves in the format that the Dialogue Engine can more easily work with. We created new templates and added relevance values to each template to match up to our dialogue structure.

We implemented the GOAL markers that have been introduced in GRETA in the DM templates as well. These markers indicate whether a sentence said by the agent is completed and or accomplished. We do this by using the tags `DMBegin` and `DMEnd` for indication if the agent has completed her sentence and the tags `DMImpBegin` and `DMImpEnd` for indication of accomplishing the intent of the agent. The latter tags were used to mark the most important aspects of the utterance of the agent. With these markers we have more expressive behaviour in the agent and we can adapt our interruption strategy. For example, the agent will not repeat the sentence if the most important aspect has already been said.

We modified the knowledge base to match to the concept of dialogue moves better, so that the system can make better use of the dialogue structure we defined. Each move has a goal formatted as episode_exchange_move, making it easy to determine the current episode the conversation is in, and also to measure distance between different types of moves, based on if they are in the same exchange and/or episode.

After July we integrated the verbal alignment tool (see Section 6.3.3 in the agent. We set up a server on which the algorithm runs and included an API to perform verbal modifications. We can save the dialogue history and adapt the agent utterances in the long term (across interactions with the same user). The verbal alignment tool only works for English.

For the final evaluation of the system we have worked together with the partners at UoN to set up a game with the book demonstrator. We developed a scenario where the user is engaged in a conversation with Alice. Alice believes that her adventures in Wonderland are real and wants to talk about this. She does not like to talk about her 'real life', especially not to strangers. The user is instructed to discover the truth about

[7] `https://github.com/evania/alice-qg`

Alice and is instructed to learn more about her and eventually develop a trust bond with her. Once Alice trusts the user, Alice will tell something about her real life, such as her real name. We have augmented the dialogue moves of the agent with more FML capabilities described in [11].

The method to measure trust is based on an interest level that has been developed by our partners at the University of Augsburg. We kept track of the long-term interest level of the user and used this as the measure of trustworthiness. The interest level was based on multi-modal input from the user, namely the arousal and valence level of the user, together with the amount of head movement and gaze.

Finally, regarding Task 3.7, the prototype task-oriented demonstrator for the Industry-ARIA was implemented after July.

6.4 WP4: Context-sensitive generation of acoustic and visual agent behaviour

6.4.1 Task 4.1 Overall dynamic non-verbal communicative behaviour model

For the generation of dynamic non-verbal communicative behaviour we implemented a solution that remains SAIBA compliant and spans over different components of the ARIA system. Figure 6.6 depicts the proposed architecture of the ARIA system. We made an important change compared to common SAIBA platform. The Intent Planner is placed inside the Dialogue Manager (DM) component. The role of the DM is to produce the communicative intents and dialogue acts for the agent by choosing among several FML templates corresponding to the desired intents to communicate. FML templates are a specialized version of FML-APML containing more advanced constructs. They also offer several input parameters that support online dynamic changes in order to produce different FML that can be processed by Greta. A translator component (FML Translator in Figure 6.6) transforms a given FML template and its input parameters into a valid FML-APML script.

6.4.2 Task 4.2 Adaptive nonverbal communicative behaviour generation model

FML Templates The DM has an FML manager component in the DM is responsible of the communicative behaviour generation. This FML manager (1) selects one of the available FML-templates according to the communicative intents to accomplish, (2) fills the available placeholders (each template supports a set of parameters) with information extracted from the Information State in the Dialogue System (e.g. a subject in an utterance or an emotion) and (3) sends the template filled in with information to an FML Translator component that in turn produces a regular FML-APML script for Greta.

The FML Templates are based and categorized according to the Dynamic Interpretation Theory (DIT++) taxonomy of communicative functions [8]. The DM can pass in input

[8]Bunt, Harry. "The DIT++ taxonomy for functional dialogue markup." AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts. 2009.

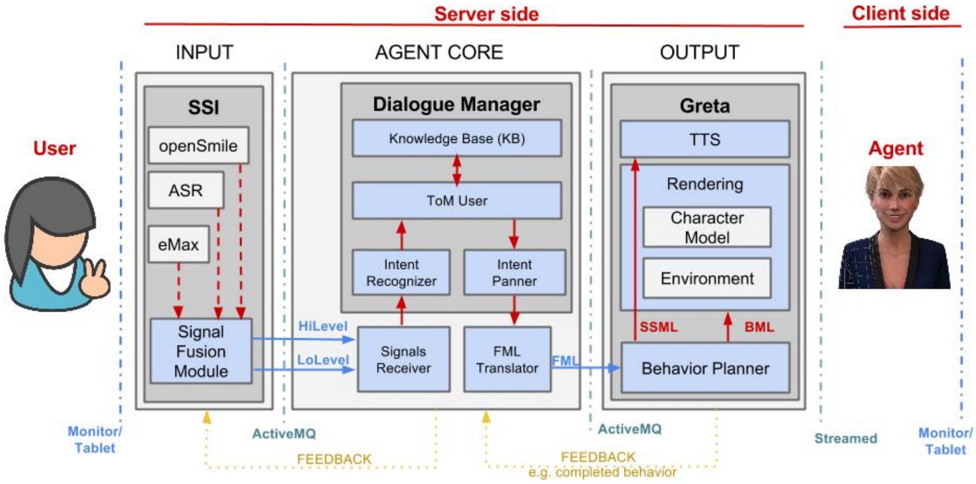


Figure 6.6: Overall architecture following SAIBA

different types of parameters such as: emotion label and its intensity, pitch accent, certainty level, different words alternative. Once the DM has selected an FML template and instantiated its parameters, it sends the template to the FML Translator. His task is to transform each FML Template and its parameters into FML-APML for Greta.

Expressing interpersonal attitudes CNRS developed a model for making the ARIA agent capable of expressing different interpersonal attitudes, for example dominant or hostile, toward the user. Attitudes are displayed through sequences of multimodal behaviours. A corpus was annotated along 2 dimensions: multimodal behaviors and social attitudes. Attitudes are represented as the 2D space (Friendly, Dominance). We were interested in looking at the behaviours that trigger a change of attitude perception. The data was segmented to identify the non-verbal behaviors that characterize a variation (increase or decrease) in attitude. We have applied HCApriori, a temporal sequence mining algorithm, to extract temporal patterns of nonverbal signals expressing the four attitude variations and two 'neutral' attitudes (neutral dominance, and neutral friendliness).

First study: Extraction of multimodal behaviors patterns

We have evaluated our algorithm that extracts patterns of multimodal behaviors in link with attitude variation. We have performed two types of evaluation: objective and subjective. For the former, we compared our algorithm against four state-of-the-art algorithms: QTIPrefixSpan-Kmeans, QTIPrefixSpan-AP, QTIApriori-Kmeans, and PESMiner. For this, we rely on two criteria: the pattern extraction accuracy and the empirical efficiency (running time). We found that our algorithm HCApriori outperforms the other algorithms and is able to achieve over 0.92 accuracy whereas the runner-up achieves 0.70.

We have also run an empirical study to investigate whether non-verbal patterns extracted with our model for a given attitude variation are perceived as conveying the same attitude variation. We evaluated eight non-verbal patterns for the perception of dominance variation (4 for dominance increase vs. 4 for dominance decrease) and eight nonverbal patterns for the perception of friendliness variation (4 for friendliness increase vs. 4 for friendliness decrease). We also evaluated 2 non-verbal patterns for the perception of neutral attitude (1 for neutral dominance and 1 for neutral friendliness). We used Greta to generate videos of ECA displaying non-verbal behaviour patterns. As our model only consider nonverbal behavior, we left aside the content of the speech. For this, each non-verbal pattern was shown while the agent spoke nonsense speech. For measuring each attitude's dimension we relied on Leary's model and used 16 variables (four for each attitude): helpful, cheerful, cooperative, warm, leaderlike, assertive, domineering, forceful, aggressive, arrogant, defiant, distant, withdrawn, timid, depend and unauthoritative. To asses if there is a significant difference between the reference video and the comparison videos, we conducted a Wilcoxon test. We found that: (1) patterns representing dominance increase are evaluated as more dominant, more hostile and less friendly compared to the reference video. (2) patterns representing dominance decrease are evaluated as more submissive compared to the reference video. (3) patterns expressing friendliness decrease are evaluated as more hostile and more dominant compared to the reference video. (4) patterns expressing friendliness increase were perceived as equivalent to the neutral expression.

Model of Sequential Attitude Planner Once our extraction model has been validated, we have implemented a new module in the ARIA system that we call **Sequential Attitude Planner** to generate the non-verbal behaviour of the ARIA agent expressing an attitude variation. The Sequential Attitude Planner takes as input an FML file (as produced by the Dialogue Manager through the FML Translator) and the attitude variation that the agent will express toward the user. These information are defined with the Functional Markup Language (FML) [3]. Our Sequential Attitude Planner is composed of four steps:

1. From FML to BML Sequence generation: The first step of our model is generating a sequence of non-verbal signals expressing the communicative intentions contained in the input FML file.

2. BML Attitude sequence selection: from a dataset of behaviours sequences expressing attitude variations, the algorithm selects the sequence that is closer to the behaviours (or sequence of behaviours) that the original behaviour planner in ARIA-Greta proposed when transforming the FML into a sequence of BML descriptors.

3. BML Sequence enrichment: all signals in the attitude-sequence previously selected that do not appear in the original sequence of behaviours generated by the ARIA-Greta behaviour planner are added to the produced sequence of behaviours that the agent will exhibit.

4. Priority signals selection: we designed a Bayesian Network to model the probability of occurrence for non-verbal signals for each attitude. Based on these probabilities,

the algorithm replaces a signal in the final behaviour sequence with a mapped signal expressing a specific interpersonal attitude, if the probability of this last signal is higher than the probability of the first one for expressing that attitude.

Second study: Sequential Attitude Planner

We report on the evaluation study we conducted on our Sequential Attitude Planner. We follow the same evaluation protocol as the first study with one difference: this time the agent did not say nonsense utterances but said meaningful utterances. So we have added four dependent variables: dominant, friendliness, submissive. The agent utters the same utterance in the reference video and in the comparison ones. The behavior of the agent in the reference video has been generation by the Greta/VIB behavior planner without using the sequential attitude model. In the reference video, the agent displays the behaviors generated by our sequential attitude planner. 64 participants took part of the study. We ran t-test to compare the results on the reference video with the comparison ones. We found two significant results: for dominance decrease and friendliness increase. The results are similar to the first study.

6.4.3 Task 4.3 Emergence of synchrony during engagement phases between ECA and User

The relationship between user and agent during the interaction can be characterized through the emergency of synchrony and, in particular, alignment. More specifically, the recognized user's emotion through facial expressions can elicit an agent's emotional response through the generation of appropriate facial expressions. The agent responds by aligning its nonverbal behaviour to the user's detected emotional state defined as empathy level (0-1). The agent's behaviour alignment with user's detected empathy level can be mapped with agent's disengagement (low user's empathy) and full engagement (high user's empathy). CNRS created a graphical tool in Greta that allows simulating the detected user's empathy and map it into animation parameters in real-time for the agent as depicted in Figure 6.7.

Regarding verbal alignment, CNRS has provided measures characterising verbal alignment processes based on repetitions between dialogue partners. To this end, CNRS has proposed a framework based on repetition at the lexical level which deals with textual dialogues (e.g., transcripts), along with automatic and generic measures indicating verbal alignment between interlocutors.

6.4.4 Task 4.4 Adaptive speech synthesis

The task comprised of two elements, reactivity and expressive and conversational speech.

1. A ground breaking reactive speech synthesis API was designed, built, tested and integrated with the ARIA-VALUSPA code base. This allows speech synthesis to be altered during production without requiring a pause, for example increasing vocal effort to hold the floor, or using conversational speech elements to gracefully cede the floor.

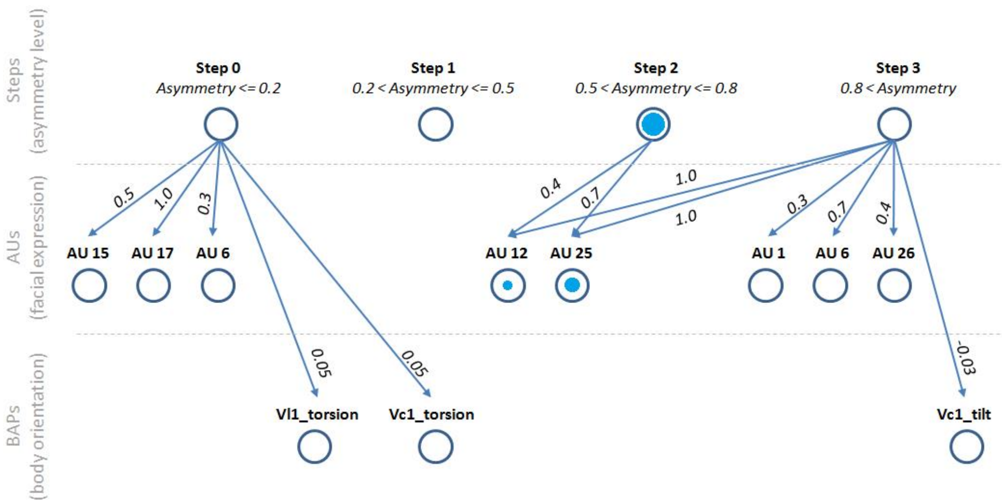


Figure 6.7: Graphical tool to define behaviours mapping

2. CereProc APIs for controlling phrasing, reduction and non-verbal cues were integrated into the ARIA-VALUSPA framework and made available to the dialog and graphic rendering modules. Conversational speech was recorded as part of Task 6.4 and integrated into a voice released to ARIA-VALUSPA as part of the Industry ARIA allowing blending of acted conversational speech with more formal read style speech. Significant work in exploring closed phase LPC vocoding techniques was carried out. This work was also extended to produce Idlak Tangle, an open source speech synthesis system based on deep neural nets (DNNs) (see section 2.5). DNN approaches were also applied to prosody modelling which resulted in higher quality synthesis, especially for more sparse genre synthesis. The vocoding approach was then used to model tense and lax voice qualities. This approach was then used to produce two voices for ARIA-VALUSPA, the first a version of the Alice voice used in the book ARIA. The second a voice created using freely available Arctic data which can be freely distributed with the project outputs. Although the quality of the resulting algorithmic modelling of voice quality was not sufficient for replacing the Alice voice, it allowed the release of the freely available emotional voice. Finally unit selection approaches to controlling emphasis were added and integrated into the ARIA-VALUSPA framework.

For more details and a focus on the breakthrough elements of this work see section 3.3.

6.4.5 Task 4.5 Synthesis-Analysis feedback loops

The ARIA agent can adapt to the user's socio-emotional state detected by the INPUT module of the ARIA system (e.g. user's presence, voice activity, speech, etc.). This adaptation can be used to create synthesis-analysis loops of adaptive audio-visual behaviours. CNRS added two adaptive features in the ARIA system that supports Task 4.5: the handling of Interaction States and the realtime Language (audio synthesis) switch.

The ARIA agent can be in 4 different states with respect to a dyadic interaction with a user:

- IDLE
- ENGAGING
- ENGAGED
- DISENGAGING

The INPUT module informs the ARIA system about the detection of multimodal signals. The Dialogue Manager keeps track of this information and informs all other components, including ARIA-Greta that synthesizes appropriate multimodal behavior according to the current interaction state of the ARIA agent.

ARIA agents are multilingual (English, French and German) and can adapt its language to their users. Therefore, in addition to the Interaction State, the Dialogue Manager holds information about the current language of the ARIA agent. CNRS implemented a dynamic language switch that can be done in real-time when the Dialogue Manager informs ARIA-Greta of a language change. This change affects Cereproc's synthesised speech because ARIA-Greta loads in the speech engine the new language modules in real-time. UAugsburg and ICL worked on the real-time detection of user's language changes (among the three system languages). This supports an adaptive analysis-synthesis loop in which a change in the language is automatically detected by the INPUT module of the ARIA system and the ARIA agent dynamically adapts to it.

6.4.6 Task 4.6 Multimodal behaviour response model to unexpected situations

At CNRS we focused on dealing with interruptions as unexpected situation that may occur during the user-agent interaction. We tackled three questions:

1. Study interruptions and their meaning and effects during the interaction;
2. Detect when a user's interruption occurs;
3. React appropriately (i.e. agent) when such interruptions occur.

For the first question, we proposed a taxonomy for modelling user's interruptions and we evaluated the impact of interrupting behaviour, based on this taxonomy, on interpersonal attitude and engagement judgements from a third person point of view[9]. The proposed taxonomy embeds two main categories: the interruption type, whether or not the speaker-switch is successful and the presence or not of simultaneous speech and the strategies underlying the interruption, namely **disruptive** or **cooperative**.

[9] A. Cafaro, N. Glas, and C. Pelachaud. 2016. The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions. In Proceedings of the International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16), 911-920.

When dealing with the second question, it was important to differentiate back-channelling behaviour (i.e. when the user gives vocal feedback while listening to the agent) and interrupting behaviour. We approached this problem by taking advantage of the data collected for the NoXi database and developed a corpus-based machine learning approach. We learned from annotated interactions whether the speaker's acoustic features (e.g. prosody) allow discerning if the user is interrupting (as opposed to back-channelling) and in this case which the employed strategy (disruptive or cooperative) is.

Regarding the third question, we first enhanced the animation capabilities of the Greta virtual agent. The current behaviour of the agent can be:

- **Ended**: the formerly specified behaviour(s) finished (i.e. has been displayed).
- **Stopped**: the formerly specified behaviour has been stopped (e.g. due to a newly created FML-APML set in replace mode).
- **Aborted**: the formerly specified behaviour failed to be displayed for some error(s).

And a new behaviour can be:

- **Started**: the formerly specified behaviour(s) started (i.e. is being displayed).

It is the dialogue manager that re-plans the communicative intentions as reaction to the interruption. It sends these new intentions to Greta using the FML template mechanism, but in **replace** mode. The dialogue manager makes use of two parameters, namely **reaction type** and **reaction duration**, to specify how the agent reacts to the interruption.

Finally, we conducted an experiment to determine the multimodal behaviour of the agent when reacting to different interruption types. From a careful analysis of human data using the NoXi database, we extracted multimodal behaviours that are commonly displayed during an interruption. To validate these behaviours on the virtual agent, we defined an interface with four videos of the agent. In these videos the agent's animation is obtained by manipulating the different multimodal behaviour using a genetic algorithm. We asked human participants to choose the videos that correspond best to the reactive behaviour to an interruption. New videos of the agent are computed on the fly. Participants continue selecting videos of the agent until they are satisfied with the results. As such we were able to characterize precisely which multimodal behaviour an agent should display as a reaction to an interruption.

6.4.7 Advances made after July 2017

Since the last deliverable, the CNRS team worked on integrating the sequential attitude planner into the Aria platform. The dialogue manager sends the attitude change the agent should display. The attitude planner computes how to display the communicative intention with this attitude change. This integration of the attitude planner and the communicative intention planner was evaluated through perceptive study. The results of this study validated the computational model.

The CNRS team worked also on developing a perceptual study to characterize the reactive behaviour of the agent to an interruption. Through an interface, participants viewed four videos of the interruptee agent. Participants select which videos correspond best to a reactive behaviour. Participants can select videos as long as they wish. New videos are rendered on the fly where the behaviours of the agent are manipulated using a genetic algorithm. When they are satisfied with a video, they select it and precise their level of satisfaction.

6.5 WP5: Realisation of use-cases and portability

6.5.1 Task 5.1: Specification of use-cases

Specifications of the Book-ARIA and the industry-Associate ARIA have been delivered on time. Similar specifications were applicable in each case. These specifications determine what implementation and RnD efforts are necessary for the realisation of the Embodied Conversational Agent using the capabilities developed in WPs 1-4 and to ensure the cross-domain portability of the ARIA-VALUSPA technologies.

Uses cases utilise Living Actor technology and the dialogue management system integrated with Living Actor avatars and CereProc voice synthesis.

The POC is in English, French, and German languages and used for the realisation of Book-ARIA, using a HTML5 rendering system. The POC demonstrated several limitations regarding the rendering and the portability so the specifications where modified.

The behaviour of the avatar was fluid and able to react almost immediately to user events. Existing options to display real-time 3D animations in a web-based application (Unity, WebGL) may not have functioned properly on all mobile devices, but this was a limitation that was accepted as creating native applications for all mobile platforms was beyond the scope of this project. We decided to work on a streaming system to generate 3D animation on a remote server and the streaming of live videos, similar to a teleconference call.

Examples of specifications for the application include

- Vocal dialogue. In the solution provided by WP3, the user is able to interact with the ARIA directly by voice. This system is able to dynamically adapt the speech according to the user's reactions, take initiative if the user is passive, and deal gracefully with interruptions.
- This vocal solution could handle 3 languages: English, German and French.
- 3D real time avatar as streaming
- Empathic avatar: The animation of the avatar and its behaviour incorporated the developments from the WP4 to be more efficient and accurate.
- The ARIA analysed a live video capture of the user thanks to the developments of the WP2, in order to detect engagement, mood and specific behaviours that could enhance the Industry-ARIA experience. For example, if the user attention was

Figure 6.8: Alice model

caught by something else, or if the environment sound changed suddenly, the ARIA would be able to adapt the dialogue and wait for the user.

A detailed account of specifications can be found in the specification documentation.

6.5.2 Task 5.2 Realisation of Industry Associate-ARIA using affective technology

This section has been redacted from the public version of the deliverable.

6.5.3 Task 5.3 Realisation of Book-ARIA

The goal of the Book-ARIA was to provide an interactive book experience to users of all ages. The interaction was performed by an embodied agent simulating a character, the author, or another person related to the book. This demonstration application featured Alice from "Alice in Wonderland" (Lewis Carroll). In the demonstration application, the user stands in front of the computer and is able to interact with Alice in a natural way using voice to discuss the novel featuring her. Once the application detects the user's presence, the ARIA (Alice) initiates the dialogue. Alice introduces herself and asks about the user's knowledge of the Alice in Wonderland novel. If the user does not come up with a question to ask, the agent proposes some topics to discuss. The user is not required to have read the book.

This application was available in 3 languages: English, French and German. The agent adapts its behaviour to the user. From the onset of the project a virtual human representing the characterisation of a novel agent (called a Book-ARIA) was developed. Book-ARIAs are believed to have commercial value in their own right. More generally, the Book-ARIAs could function as a showcase of what rich personalities could be generated with ARIA-VALUSPA and how they could function as interfaces for information retrieval for more complex tasks, including, questions about the novel's content, characters, author, etc.

In the Book Personification scenario, a user could interact with a character representing the book, asking it questions related to the book and the character. This scenario was chosen because while Virtual Assistants are often very goal oriented, for reasons of optimizing public relations of companies, they do not allow for a diverse range of personalities. On the other hand, the book personifications that we developed were not similarly constrained, and allowed us to explore the interaction between users and exaggerated personalities within the well-defined context of the book they were based on. The selected book was determined and was intended to be different for each language taking in account that it was a classic novel of each language literature. The virtual character had the mission to personify the book.

Realising the Book-ARIA application discussed above required technical effort. For example, a new rendering mode was added to Living Actor™ 3D to send audio and video streams directly to a web page instead of rendering the avatar on the user's screen. The animation of the avatar was no longer rendered in the software window but in a memory buffer called offscreen, which can be stored in a file directly on the computer or sent as stream data. The real-time generated audio and video are sent through an http stream pipe as soon as they were generated by the 3D animation module. The data were then received by a server based on NodeJS which converts it to a websocket. The websocket is then received in a web page that plays it automatically.

As of December 2017, the streaming capability sending the video and audio data through an http stream pipe is technically working. The data is well sent and received, but there are some issues regarding the quality of the video displayed on the web page. It seems that the origin of the issue is located in the NodeJS server when converting the http flux to a websocket. Currently, developers have not succeeded in displaying an output that is non-pixilated, with no latency video output.

Finally, there is another issue directly linked to the 3D engine used by Living Actor™ 3D to render 3D avatars. When rendering to a memory buffer, an 'out of memory' error occurs regularly and most of the time the 3D rendering engine has to reboot in order to correctly render subsequent frames. Cantoche does not have control over the main memory leak inside the rendering engine. However, this was managed and reduced as best as possible, but the error still occurs and prevents developers from having the continuous flow of data streaming needed.

6.6 WP 6: Hypothesis testing, data collection and global evaluation

6.6.1 Task 6.1 Ethical Policies

For each partner of the project we obtained their ethical policies and documented them in deliverable 6.5. This mostly concerns the protection of data as well as the protocol for ethical clearance. All personal data collected in the project is treated according to EU law. In particular, in ARIA-VALUSPA personal is anonymised as best as possible (i.e. face and voice data will be kept intact). Anonymised demographic data is stored on user-level access controlled file servers.

6.6.2 Task 6.2 Experimental induction

As reported in deliverable 6.1 we initially set up two experiments. In this first stage of data collection, we aimed at collecting data that is necessary to develop the DM module with verbal inputs:

- Dataset 1. Set of recordings of a scripted dialogue between several users and the agent for testing the ASR module with contextualised vocabulary.
- Dataset 2. Set of recordings to obtain verbal inputs for the dialogue management system including unexpected situations, user engagement and general dialogue strategies in a Wizard of Oz (WoZ) scenario

For the first dataset, in order to collect realistic testing data for the ASR module a SSI pipeline of the user-agent interaction was set up. Interaction follows a static script in which the user and Alice talk in turns. The dialogue starts by presenting the user a sentence he or she should read out loud. After the end of a voice is detected, a picture of Alice is displayed and an appropriate answer is played back to the user (see Fig. 6.9. The results of the interpersonal stance ASR recording pipeline.). All speech input of the user is stored in separate audio files and will be used to refine the ASR models. Next, the scripted dialogue is replaced by the actual ASR output and the result is sent to the Dialogue manager.

We recorded 16 participants in Germany, France and the UK, all reading the English sentences. The training data was meant to improve the model by not only training on native speakers but also on persons with foreign dialects. The sentences were automatically separated into 16 chunks per participant and therefore automatically labelled.

As a second initial dataset University of Twente set up a wizard of of scenario with two goals in mind. Firstly, this data collection helped developing multi-modal interaction between users and agents. Secondly it supported the development and evaluation of a system that allows DM in the early development stages of virtual humans.

The participants for this study were split to run experiments under two conditions (8 per condition). In the first condition there were 7 male and 1 female participants (2 native English speakers) with an average age of 30.38 years. In the second condition were 4 male and 4 female participants (no native speakers) with an average age of 34.13 years. Most participants have seen an Alice in Wonderland movie, but none of them recently read the book. The data that was collected during the experiment allows to select the

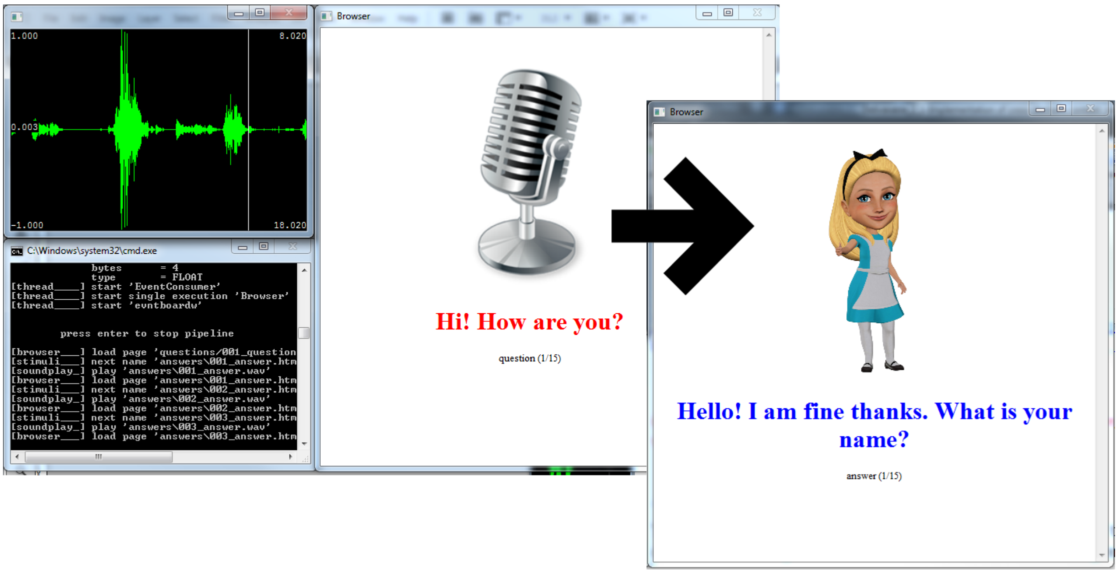


Figure 6.9: Example for the initial dataset of collected speech examples.

knowledge that is required to give appropriate answers in dialogues in later stages of the project.

6.6.3 Task 6.3 Recording of interactions with ARIA-VALUSPA platform

We opted to go for a mediated human-human interaction database already in early stages of the project that could be used for the development of the single components of ARIA-Valuspa. To this end we recorded the NoXi database, as described in 2.3. In NoXi an expert talks about a topic to a novice who is interested in the topic via a screen, which is close to the final interaction with an agent on the screen. We further added "unexpected" events to the recordings, such as phone calls and walk-ins. Overall we recorded 84 sessions with 2 participants per session, which results in an overall 25 hours of audio, video and depth information data.

6.6.4 Task 6.4 Annotation of emotion, social cues, etc., transcription of spoken content

In ARIA-Valuspa we went for a mixed approach for annotating the previously recorded NoXi database. Thereby we partly annotated the database in a manual effort, for other cues we used SSI's capabilities of automatically annotating specific social cues. Finally we used the NOVA tool for cooperatively annotating social cues, together with machine learning algorithms. The annotations are described in deliverable D6.2. In conclusion we manually labelled speech transcriptions for English and German sessions, Speech/Filler/Breath annotations that have been performed partly in a manual way, and partly with cooperative learning. We automatically performed annotations of interruptions by analysing the alignment of speech turns of both participants in the corpus, see Fig. 6.10.

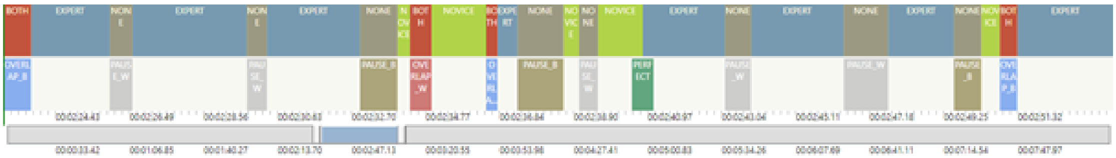


Figure 6.10: Example for automatically extracted interruption annotations in the NOVA tool.

Further we employed a couple of recognisers for gestures and expressiveness features:

Movement Energy/Power represents the dynamic properties of a movement (e.g. weak versus strong).

Overall Activation (represents the quantity of the movement (passive versus active). I

Spatial Extent is modelled as the space that is used for gesturing in front of the recorded person.

Fluidity differentiates smooth movements from jerky ones. This feature aims to capture the continuity between movements.

Gestures and Postures are recognised using an event-based mechanism that triggers an event each time the beginning or the end of a social signal, such as a particular arm configuration, is detected. The detected events are saved in an XML-based structure including a synchronised time stamp and the event's duration. For event-based behaviour analysis, our system makes use of the Full Body Interaction Framework (FUBI).

Further we automatically collected continuous measurements along the two affective dimensions valence and arousal. Each dimension is represented by a single value in the range [0..1]. For details see Chapter 2 'Nonverbal Social Signal Processing' in D2.1.

We manually annotated the conversational "Engagement" for all French and German sessions (English session are currently annotated) on a continuous scheme ranging from 0..1. These annotations represent the interest of a person in the interaction. Figure 6.11 shows the gold standard annotations of engagement that have been merged from annotations from multiple raters. The expert's (on the left) engagement is shown on the upper tier, the novice's (on the right) engagement is shown on the lower tier

6.6.5 Task 6.5 Recordings of voice talents for speech synthesis

A within domain 4,000 word script was created based on 100 example sentences and a list of 500 key within domain words resulting in 1.5 hours of within domain audio data. Acted conversational speech was recorded using two interviews with academics with the recording engineer acting as the dialog partner. The created a 14,000 word script

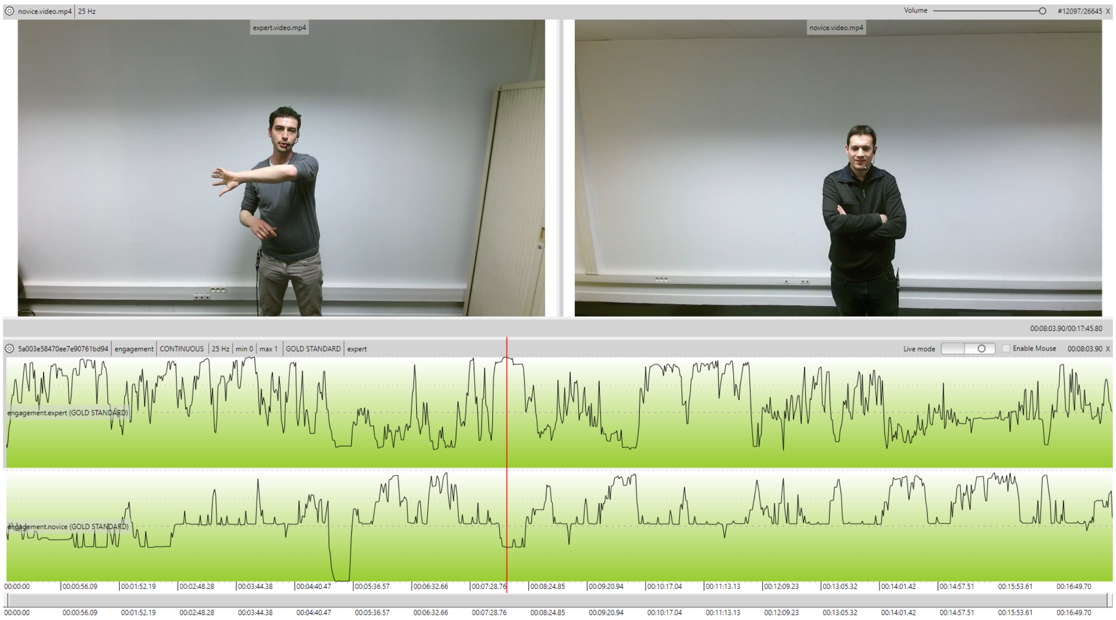


Figure 6.11: An example of the Gold Standard annotation of engagement in the NOVA tool.

resulting in 2.5 hours of acted conversational genre speech. Lombard speech was recorded by playing noise to the voice talent over headphones. A standard phonetically balanced script was used of 5,000 words resulting 2 hours of Lombard speech audio. Quality control was carried out on the recorded audio and a voice was released to the project.

Work on algorithmic extension of expressive speech was carried out on data already collected by CereProc which was used to build a freely available emotional voice (see section 3.3).

6.6.6 Task 6.6 Systematic evaluation of Industry Associate Demonstrator

In the following, we will describe the data recordings with the full interactive AVP system. Recordings took place at Nottingham University between 2017/10/26 and 2017/12/15. In total, we recorded 40 unique individuals during 226 sessions. 49.1% of people reported their gender as Female, and 6.64% as 'Other'. 49% claimed to be native English speakers, 44.24% to have high proficiency, and 6.63% intermediate proficiency in English. We started with the initial system, which we used in week 1 and 2 (wk12). Afterwards, we identified a couple of shortcomings. For instance, the agent sometimes repeated the same sentence within a session. Consequently, we enriched the knowledge base of the dialog manager with more alternatives. We prepared an improved system, which we used in week 3 and 4 (wk34). Afterwards we took a two weeks break to prepare the final system. Besides further improvements of the dialog manager we also added the possibility to

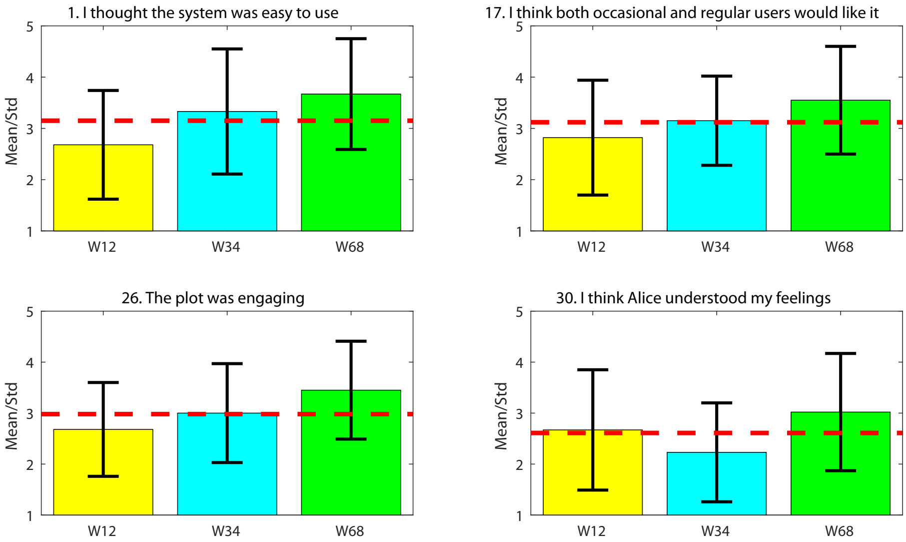


Figure 6.12: Some of the questions visualised as bar charts with respect to the three system stages (wk12=initial, wk34=intermediate, wk68=final). The height of the bar represents the average score (an error bar gives the standard deviation). The red line indicates the average over all sessions.

capture the agent screen along with the other signals. The total length of all recordings sums up to more than 20 h. Figure 6.13 shows a screenshot of the interaction visualised with the NOVA tool.

During the interaction the agent represented the little girl, who inspired Lewis Carroll to write his book, whose name is Alice Liddell. Users had to convince her to reveal some personal details of her life. In order, to receive the correct answer the user had to gain the trust of the agent first. The trust level took into account the verbal part of the conversation, as well as, the user's non-verbal expressions. After each quest, we asked participants to fill in a questionnaire. Each question had to be answered on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). We were particularly interested to measure the relative improvement between the initial (wk12), intermediate (wk34), and final system (wk68).

Results showed that the system usability generally improved over time and was less complex to follow. The final system was perceived user friendly and users believed that new subjects are able to quickly learn to interact with the system. We also noted that the users found the plot of the interaction more engaging in the final system. However, there seemed to be no improvement regarding the perception of the conversations. People interacting with the final version were generally more satisfied and had more fun using the system. They also noted that they would recommend the system to other users. The awareness of the user inputs increased with the improved system, too. This is true for the verbal aspects, as well as, the non-verbal aspects. The non-verbal responds of the agent was also judged more positively. However, there was no improvement regarding the verbal response. Some of the results are visualised in Figure 6.12. More details can be found in D6.4.

6.7 WP7: Impact Delivery

6.7.1 Task 7.1 Project website

We set up a project website immediately after starting the project describing the project's objective and motivation, as well as who is involved. The project can be found at `aria-agent.eu`. After the mid-term review, we have regularly added blob-posts to this, approximately one per month. The website also provides access to information about the current results of the project (reports, publications, software releases, and demo videos).

There are clear descriptions of and links to the three most successful outputs of ARIA-VALUSPA: the ARIA-VALUSPA Platform (AVP) on GitHub, our NoXi web database, and the NoVa annotation tool, again on GitHub.

6.7.2 Task 7.2 Data access

Access to almost all of the data collected in ARIA-VALUSPA is provided on-line through the NoXi database for both project internal and external use. See sections above as well as the dedicated deliverable D6.2. This include recordings of human-human and human-agent interactions, relevant annotations. As part of task 7.2, the NoXi web-database was set up, with an end-user license agreement for the corpus, which is in line with the project's and partner's IPR policy as well as providing and maintaining resources for hosting the data.

6.7.3 Task 7.3 Software releases

To release the AVP and NOVA software, we have used the popular service called GitHub. This service allows users to download data, suggest new edits to code, report issues and make suggestions for improvement. GitHub also provides a WiKi which serves as our detailed documentation of AVP and NOVA.

- The ARIA-VALUSPA Platform (AVP) can be found at `https://github.com/ARIA-VALUSPA/AVP`.
- NOVA can be found at `https://github.com/hcmlab/nova`.

6.7.4 Task 7.4 Contribution to standards

We have adhered to existing standards, but have not actively engaged in the creation of new or modification of existing standards.

6.7.5 Task 7.5 Workshops and tutorials

Two workshops were given to stakeholders, both industrial and academic, to allow them to get accustomed with AVP and NOVA. The workshops were held in 2017, the first one in London, the second one in Paris.

The two meetings followed the same structure:

1. a presentation on the overall motivation behind the ARIA project was given by the coordinator of the project, and two of the three main outputs of the project were presented (AVP and NoVa).

2. NoXi was only presented in passing as due to ethics and privacy constraints the NoXi database cannot be shared with non-academic entities.

3. NoVa was presented by Tobias Baur, representing the university of Augsburg and the 'Framework' aspect of the project.

4. After this, the stakeholders were matched in turn with a representative from the Behaviour Analysis, Dialogue Management, Behaviour Generation, and Framework modules of AVP. They could speak to them for 15 minutes at a time, after which the stakeholders moved to the next representative.

5. At the end of the meeting, the companies reported on how they thought they could use the AVP and/or NoVa.

The proceedings of these meetings have been redacted from this public version of the document.

6.7.6 Task 7.6 Writing of ARIA-VALUSPA Book

From our experience in the SEMAINE project, publishing a description of the entire final system in a traditional journal outlet is very difficult. Given the extent and tight integration of the project, it is very hard to fit the abundance of information describing it in a single paper. For ARIA-VALUSPA it will only be harder to write a similar paper, as the system is bound to be that much more complex. We will therefore write a book titled 'Building Virtual Humans', focused on teaching students and developers how to build their own Virtual Humans and/or interactive Artificial Intelligence agents. In addition it will cover some of the scientific and experimental elements, as well as describing the publicly available framework. The book will include hands-on examples complete with code snippets, as most of the source code will be publicly available. This way, the book will both be a description of the system, a guide to creating sensitive artificial listeners, and a manual to use the ARIA- VALUSPA framework.

6.7.7 Task 7.7 Development of business cases

Business cases have been developed by the Industry Partners Cantoche and CereProc, and are described in section 5. The Industry Associates wrote business cases for creation of Industry-ARIAs, which are described in D7.2 and D5.2.

7 Academic Outputs

In this section we enumerate our academic outputs.

7.1 Keynotes given

Keynotes and other talks given were previously listed in the Impact section 4.4.

7.2 Workshops and Tutorials organised

Workshop on Conversational Interruptions in Human-Agent Interactions. This workshop was held at IVA 2017 in Stockholm on August, 27th. The aim is to bring together researchers from a variety of fields interested in the study of conversational interruptions in multimodal human-human, human-agent (both virtual and robotic) or agent-agent interactions. Our aim is to address current challenges in this area (as well as identifying new ones) and to set a research agenda to make IVAs capable of believably react and adapt to unexpected situations such as conversational interruptions. Organisers: Angelo Cafaro, Eduardo Coutinho, Patrick Gebhard and Blaise Portard. Website: `http://workshopcihai2017.doc.ic.ac.uk`. Published proceedings: `http://ceur-ws.org/Vol-1943/`.

Audio/Visual Emotion Challenge and Workshop (AVEC 2017) @ACM Multimedia 2017 The Audio/Visual Emotion Challenge and Workshop (AVEC 2017) will be the seventh competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual, and audiovisual depression and emotion analysis, with all participants competing under strictly the same conditions. Organisers: Fabien Ringeval, Michel Valstar, Jonathan Gratch, Björn Schuller, Roddy Cowie, Maja Pantic. Website: `http://sspnet.eu/avec2017/`. Published proceedings: `http://www.sigmm.org/opentoc/AVEC2017-TOC`.

Audio/Visual Emotion Challenge and Workshop (AVEC 2016) @ACM Multimedia 2016 The Audio/Visual Emotion Challenge and Workshop (AVEC 2016) "Depression, Mood and Emotion" will be the sixth competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and physiological depression and emotion analysis, with all participants competing under strictly the same conditions. Organisers: Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Roddy Cowie, Maja Pantic. Website: `http://sspnet.eu/avec2016/`. Published proceedings: `http://www.sigmm.org/opentoc/AVEC2016-TOC`.

Multimodal Emotion Recognition Challenge (MEC 2017) @ 2018 Asian Conference on Affective Computing and Intelligent Interaction (AACII) The Multimodal Emotion Recognition Challenge (MEC 2017) will be the second competition event aimed at the comparison of multimedia processing and machine learning methods for automatic audio and visual emotion analysis, with all participants competing under strictly the same conditions. The goal of the Challenge is to provide a common benchmark data set and to bring together the audio and video emotion recognition communities, and to promote the research in multimodal emotion recognition. Organiser: Jianhua Tao, Björn Schuller Website: `http://www.chineseldc.org/htdocsEn/emotion.html`.

Workshop on Tools and Algorithms for Mental Health and Wellbeing, Pain, and Distress (MHWPD) @ACII 2017 This workshop is in the field of affective health computing, focusing on detection and intervention techniques for mental health and well-being, pain and distress. We invite contributions from researchers with multidisciplinary expertise (computer science, engineering, psychology and medicine), both in academia and industry, in the following domains: Distress - e.g. pain, panic, confusion, itching - in patients with restricted communicative verbal abilities such as neonates and children, somnolent patients and patients with dementia is difficult to diagnose. Organisers: Akane Sano, Steffen Walter, Ognjen (Oggi) Rudovic, Nadia Bianchi-Berthouze, Björn Schuller, Rosalind W. Picard. Website: `http://mhw.media.mit.edu/`.

Computational Paralinguistics Challenge (ComParE), Interspeech 2017 The Interspeech 2017 Computational Paralinguistics ChallengE (ComParE) is an open Challenge dealing with states and traits of speakers as manifested in their speech signal's acoustic properties. There have so far been eight consecutive Challenges at INTERSPEECH since 2009 (cf. the repository), but there still exists a multiplicity of not yet covered, but highly relevant paralinguistic phenomena. Thus, we introduce three new tasks by the Addressee Sub-Challenge, the Cold Sub-Challenge, and the Snoring Sub-Challenge. Organisers: Björn Schuller, Stefan Steidl, Anton Batliner, Elika Bergelson, Jarek Krajewski, Christoph Janott. Website: `http://emotion-research.net/sigs/speech-sig/is17-compare`.

Computational Paralinguistics Challenge (ComParE), Interspeech 2016 The Interspeech 2016 Computational Paralinguistics ChallengE (ComParE) is an open Challenge dealing with states and traits of speakers as manifested in their speech signal's acoustic properties. There have so far been seven consecutive Challenges at INTERSPEECH since 2009 (cf. the repository), but there still exists a multiplicity of not yet covered, but highly relevant paralinguistic phenomena. Thus, we introduce three new tasks by the Deception Sub-Challenge, the Sincerity Sub-Challenge, and the Native Language Sub-Challenge. Organisers: Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Eduardo Coutinho. Website: `http://emotion-research.net/sigs/speech-sig/is16-compare`.

Workshop on Affective Social Multimedia Computing ASMMC 2017 Affective analysis of social multimedia is attracting growing attention from industry and businesses that provide social networking sites, content-sharing services, distribute and host the media. This workshop focuses on the analysis of affective signals in interaction and social multimedia (e.g., twitter, wechat, weibo, youtube, facebook, etc). Organisers: Dong-Yan Huang, Björn Schuller, Jianhua Tao, Lei Xie, Jie Yang, Sven Bölte, Dongmei Jiang, Haizhou Li. Website: `http://www.nwpu-aslp.org/asmmc2017/content/committee.html`.

Agents in Practice - Designing for Dialogues. SIKS workshop/tutorial. 2017, March The Netherlands Research School for Information and Knowledge Systems (SIKS) organised a course on 'Trends and Topics in Multi Agent Systems'. As part of this course, we gave a tutorial session on the design of dialogues for agents. We showed the ARIA-demo and how you can use the components of our system to design an intelligent agent. In particular we focused on turn management in dialogues. Organisers: Merijn Bruijnes and Jelte van Waterschoot Website: http://www.siks.nl/Agent-2017.php

7.3 Papers published

At the mid-term review there was some concern that it wasn't easily visible how the papers we reported were associated with the project. Therefore, we have now organised all papers in the appendix, and for every paper we clarify which of the authors were (partly) employed by ARIA-VALUSPA, and we provide up to 100 words justifying why the paper relates to the project.

We have also conducted a citation impact analysis. Despite the project having finished only recently, the academic papers listed here have already attained 1,008 citations, according to Google Scholar on 13 February 2017. The project has an h-index of 14. Table 7.1 shows a distribution of paper citations.

Table 7.1: Citation statistics of ARIA paper outputs

Citation range	The number of papers
0	24
1-5	34
6-10	16
11-50	14
> 51	5
total	93
h-index	14

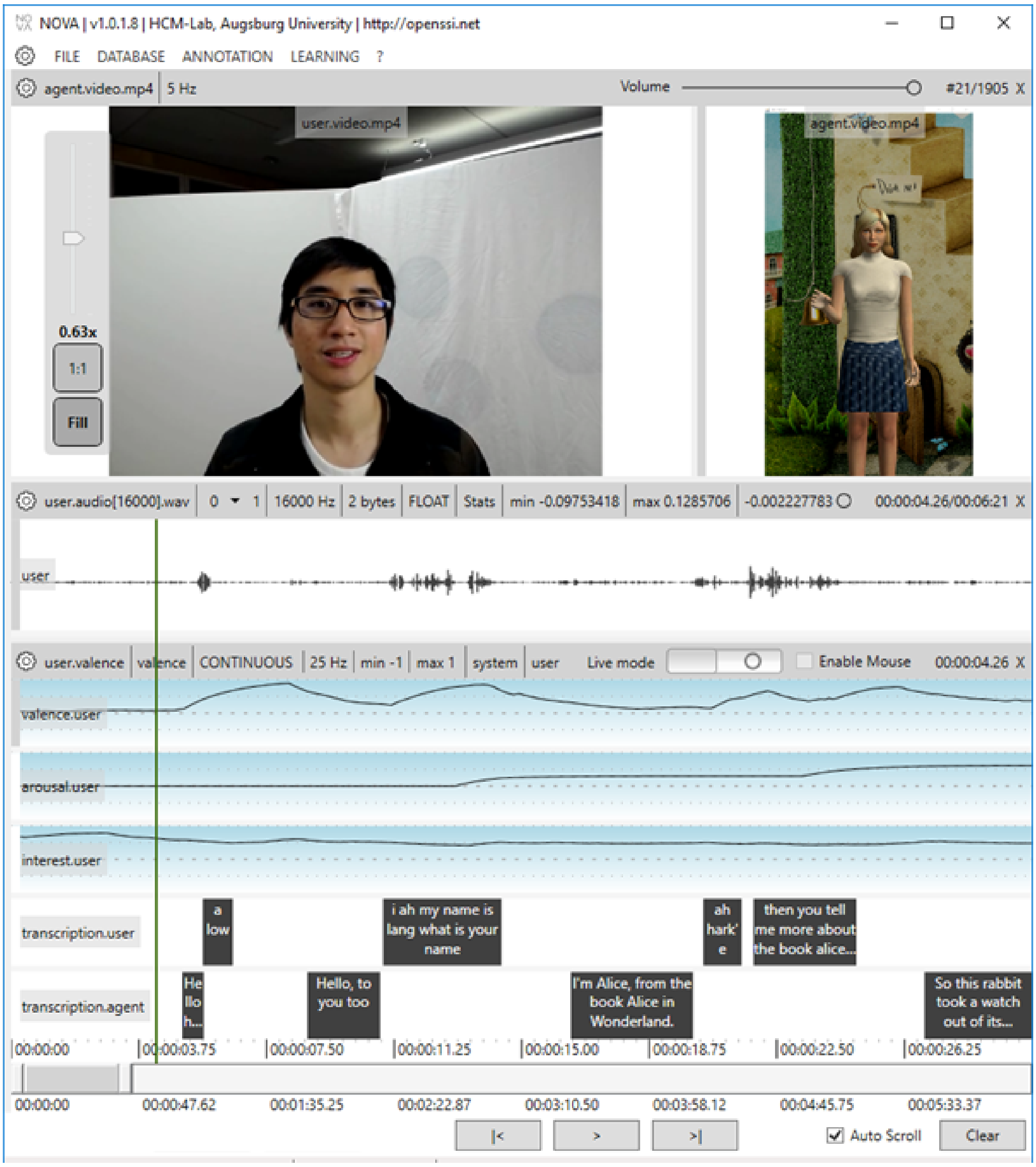


Figure 6.13: A user interaction in NOVA. Top: Videos of user and agent. Middle: Audio waveform and facial features. Bottom: Affect annotations and transcriptions.

References

[1] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.

[2] M. Argyle. *Bodily Communication.* Methuen and Co. Ltd, London, 1988.

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, 2014.

[4] Matthew P. Aylett, Rasmus Dall, Arnab Ghoshal, Gustav Eje Henter, and Thomas Merritt. A flexible front-end for HTS. In *Proc. Interspeech*, pages 1283–1287, 2014.

[5] Matthew P Aylett and Christopher J Pidcock. The Cerevoice characterful speech synthesiser SDK. In *Intelligent Virtual Agents (IVA)*, pages 413–414, 2007.

[6] Tadas Baltrusaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Facial Expression Recognition and Analysis Challenge, in conjunction with IEEE Int'l Conf. on Face and Gesture Recognition*, 2015.

[7] Timo Baumann and David Schlangen. INPRO_iSS: A component for just-in-time incremental speech synthesis. In *ACL 2012 System Demonstrations*, pages 103–108, 2012.

[8] Mike Brookes. VOICEBOX: Speech processing toolbox for MATLAB. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`. Accessed: 2017-10-13.

[9] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. Towards an iso standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010.

[10] Hendrik Buschmeier, Timo Baumann, Benjamin Dosch, Stefan Kopp, and David Schlangen. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 295–303, 2012.

[11] Angelo Cafaro, Merijn Bruijnes, Jelte van Waterschoot, Catherine Pelachaud, Mariët Theune, and Dirk Heylen. Selecting and expressing communicative functions in a saiba-compliant agent framework. In *International Conference on Intelligent Virtual Agents*, pages 73–82. Springer, 2017.

[12] Angelo Cafaro, HannesHÃűgni VilhjÃąlmsson, Timothy Bickmore, Dirk Heylen, and Catherine Pelachaud. Representing communicative functions in saiba with a unified function markup language. In Timothy Bickmore, Stacy Marsella, and Candace Sidner, editors, *Intelligent Virtual Agents*, volume 8637 of *Lecture Notes in Computer Science*, pages 81–94. Springer International Publishing, 2014.

[13] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel F. Valstar. The noxi database: multimodal recordings of mediated novice-expert interactions. In *19th ACM International Conference on Multimodal Interaction*, November 2017. Published in: Proceedings of 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, November 13–17, 2017 (ICMI'17) ISBN: 9781450355438 ; DOI:10.1145/3136755.3136780.

[14] Mathieu Chollet, Magalie Ochs, and Catherine Pelachaud. Mining a multimodal corpus for non-verbal signals sequences conveying attitudes. In *Language Resources and Evaluation Conference (LREC)*, 2014.

[15] Robert AJ Clark, Korin Richmond, and Simon King. Multisyn: Open-domain unit selection for the festival speech synthesis system. *Speech Communication*, 49(4):317–330, 2007.

[16] Soumia Dermouche and Catherine Pelachaud. Sequence-based multimodal behavior modeling for social agents. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, pages 29–36, Tokyo, Japan, 2016. ACM.

[17] Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. Automatic measures to characterise verbal alignment in human-agent interaction. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 71–81, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[18] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System (FACS): Manual*. A Human Face, Salt Lake City (USA), 2002.

[19] Evania Lina Fasya. Automatic question generation for virtual humans. Master's thesis, University of Twente, 2017.

[20] S. Ghosh, E. Laksana, S. Scherer, and L. P. Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 609–615, Sept 2015.

[21] Jonathan Gratch, Arno Hartholt, Morteza Dehghani, and Stacy Marsella. Virtual humans: a new toolkit for cognitive science research. *Applied Artificial Intelligence*, 19:215–233, 2013.

[22] Amogh Gudi, H. Emrah Tasli, Tim M. den Uyl, and Andreas Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *Facial Expression Recognition and Analysis Challenge, in conjunction with IEEE Int'l Conf. on Face and Gesture Recognition*, 2015.

[23] Shizhong Han, Zibo Meng, AHMED-SHEHAB KHAN, and Yan Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *Advances in Neural Information Processing Systems*, pages 109–117, 2016.

[24] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[25] Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 373–376. IEEE, 1996.

[26] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[27] B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic. Decision level fusion of domain specific regions for facial action recognition. In *International Conference on Pattern Recognition*, 2014.

[28] John Kominek, Christina L Bennett, Brian Langner, and Arthur R Toth. The blizzard challenge 2005 cmu entry-a method for improving speech synthesis systems. In *INTERSPEECH*, pages 85–88, 2005.

[29] Anton Leuski and David Traum. Npceditor: Creating virtual human dialogue using information retrieval techniques. *Ai Magazine*, 32(2):42–56, 2011.

[30] G. Fant & J. Liljencrants & Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4):001–013, 1985.

[31] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li Deng. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *Signal Processing Magazine, IEEE*, 32(3):35–52, 2015.

[32] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[33] Fabrizio Morbini, David DeVault, Kenji Sagae, Jillian Gerten, Angela Nazarian, and David Traum. Flores: a forward looking, reward seeking, dialogue manager. In *Natural interaction with robots, knowbots and smartphones*, pages 313–325. Springer, 2014.

[34] Kevin Patrick Murphy and Stuart Russell. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002.

[35] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[36] Daniel Povey, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow, Richard C. Rose, Petr Schwarz, and Samuel Thomas. The subspace Gaussian mixture model—a structured model for speech recognition. *Comput. Speech Lang.*, 25(2):404–439, 2011.

[37] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. The Kaldi speech recognition toolkit. *Proc. IEEE ASRU*, 2011.

[38] Charles Rich and Candace Sidner. Using collaborative discourse theory to partially automate dialogue tree authoring. In *Intelligent Virtual Agents*, pages 327–340. Springer, 2012.

[39] E. Sánchez-Lozano, G. Tzimiropoulos, B. Martinez, F. De l. Torre, and M. Valstar. A functional regression approach to facial landmark tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[40] Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, and Michel Valstar. Cascaded continuous regression for real-time incremental face tracking. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 645–661, Cham, 2016. Springer International Publishing.

[41] Alexander Sorin & Slava Shechtman. Semi parametric concatenative tts with instant voice modification capabilities. *Interspeech*, 2017.

[42] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 1003–1011, 2015.

[43] Mark ter Maat and Dirk Heylen. Flipper: An information state component for spoken dialogue systems. In *International Workshop on Intelligent Virtual Agents*, pages 470–472. Springer, 2011.

[44] Mark Ter Maat and Dirk Heylen. Flipper: An information state component for spoken dialogue systems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6895 LNAI:470–472, 2011.

[45] Timothy Leary. *Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation.* Ronald Press, New York, 1957.

[46] David R Traum and Staffan Larsson. The information state approach to dialogue management. In *Current and new directions in discourse and dialogue*, pages 325–353. Springer, 2003.

[47] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *Journal of selected topics in signal processing*, 2017.

[48] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–8. IEEE, 2015.

[49] Karel Veselỳ, Karel, Arnab Ghoshal, Lukáš Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *Proc. Interspeech*, pages 2345–2349, 2013.

[50] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM international conference on Multimedia*, MM '13, pages 831–834, New York, NY, USA, 2013. ACM.

[51] S. Xiao, S. Yan, and A. A. Kassim. Facial landmark detection via progressive initialization. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 986–993, 2015.

[52] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013.

[53] J. Yang, J. Deng, K. Zhang, and Q. Liu. Facial shape tracking via spatio-temporal cascade shape regression. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 994–1002, 2015.

[54] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Mixed excitation for HMM-based speech Synthesis. In *Proceedings of Eurospeech*, pages 2259–2262, 2001.

[55] Anil Yüce, Hua Gao, and Jean-Philippe Thiran. Discriminant multi-label manifold embedding for facial action unit detection. In *Facial Expression Recognition and Analysis Challenge, in conjunction with IEEE Int'l Conf. on Face and Gesture Recognition*, 2015.

[56] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan Black, and Keiichi Tokuda. The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. SSW6*, pages 294–299, 2007.

[57] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *icassp*, pages 7962–7966, 2013.

[58] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

8 Appendix A - Academic Papers Published

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Afunctional regression approach to facial landmark tracking	E. Sánchez-Lozano, G. Tzimiropoulos, B. Martinez, F. De la Torre, M. Valstar	Enrique Sánchez-Lozano and Brais Martinez and Michel Valstar	IEEE Transactions on Pattern Analysis and Machine Intelligence (in press)	2017	Behaviour Analysis	Computer Vision, Face Tracking, Face Analysis	WP2, T2.1	This paper presents the theoretical development of Continuous Regression, the key component of iCCR, the state of the art face tracking system, capable of performing incremental learning in real time. The paper presents a novel approach to solve the least squares problem, the main learning problem in Cascaded Regression, and derives a close-form solution that includes the infinite set of facial landmarks configuration. The new solution results in a much faster training algorithm, that also benefits from a real-time incremental learning approach (i.e. incorporating the user's information into the model). The facial tracker system of eMax is currently the implementation of the prototype developed for the paper.
The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions.	Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, Michel Valstar.	Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Catherine Pelachaud, Elisabeth André, Michel Valstar.	In Proceedings of the 19th ACM International Conference on Multimodal Interaction	2017	Corpus	Multimodal Corpora, Database	WP6	This work describes a multi-lingual database of natural dyadic novice-expert interactions, named NoXi, featuring screen-mediated dyadic human interactions in the context of information exchange and retrieval. NoXi has been designed to provide spontaneous interactions with emphasis on adaptive behaviors and unexpected situations (e.g. conversational interruptions). A rich set of audio-visual data, as well as continuous and discrete annotations is publicly available through a web interface. Audio-visual data and available annotations are used transversally used within the ARIA-VALUSPA project for designing, building and testing modules of the ARIA-System.
Selecting and Expressing Communicative Functions in a SAIBA-Compliant Agent Framework.	Angelo Cafaro, Merijn Bruijnes, Jelte van Waterschoot, Catherine Pelachaud, Mariët Theune, Dirk Heylen.	Angelo Cafaro, Merijn Bruijnes, Jelte van Waterschoot, Catherine Pelachaud, Mariët Theune, Dirk Heylen.	In Proceedings of the 17th International Conference on Intelligent Virtual Agents (IVA'17).	2017	Behaviour Generation	Dialogue management, communicative function, FML, multimodal behaviour, SAIBA	WP3, WP4	In SAIBA-compliant agent systems, the Function Markup Language (FML) describes the agent's communicative functions that are transformed into utterances with appropriate non-verbal behaviours. This work defined an improvement of the FML standard in order to create FML Templates that the Dialogue Manager in the ARIA framework can dynamically select and fill with a variety of parameters (e.g. emotional expression) that are used by the behavior generation system (ARIA-Greta) to generate behaviors.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions.	Angelo Cafaro, Nadine Glas, Catherine Pelachaud.	Angelo Cafaro, Catherine Pelachaud.	In Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'16).	2016	Behavior Generation	Turn-taking, interruptions, interpersonal attitude, engagement, empirical evaluation	WP4	This user study investigated the effects of interruption strategies (disruptive vs. cooperative) and reactions (long, medium, short overlap) between two agents on a third person human observer. Results have been used to define a taxonomy of interruption strategies (for the interrupter) and reactions (for the interruptee) in the context of unexpected situations in ARIA.
Speech Synthesis for the Generation of Artificial Personality	Aylett, M.P., Vinciarelli, A. & Wester, M.	Aylett, M.P., & Wester, M.	IEEE Transactions on Affective Computing	2017		Speech synthesis, Unit selection, parametric synthesis, Emotion, Personification.	WP4 T4.4 WP5 T5.3	Speech synthesis can be used to personify the interface. In this paper we investigate the direct relationship between expressive synthesis and the perception of character in order to support the development of personification in the book ARIA.
Real-time reactive speech synthesis: incorporating interruptions.	Wester, M., Braude, D.A., Potard, B., Aylett, M.P., & Shaw, F.	Wester, M., Braude, D.A., Potard, B., Aylett, M.P., & Shaw, F.	In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017)	2017		Reactive speech synthesis. Adaptive speech synthesis.	WP4 T4.4	This work directly addresses the challenge of improving the reactivity of and embodied conversational system. It relates closely with work carried out by CNRS on implementing reactivity in Greta and is the subject of a current patent submission by CereProc.
Bot or Not? Exploring the Fine Line between Cyber and Human Identity.	Wester, M., Aylett, M.P. & Braude, D. A.	Wester, M., Aylett, M.P. & Braude, D. A.	In Proceeding of the 19th ACM International Conference on Multimodal interaction(ICMI 2017)	2017		Expressive speech synthesis. Personification. Emotion.	WP4 T4.4	This paper describes the Science Lates demo Bot or Not. This work explored the extent expressive speech synthesis and modified speech affected the perception of authenticity. The results form this work support the personification work carried out in WP5 T5.3
Don't Say Yes, Say Yes: Interacting with Synthetic Speech Using Tonetable	Aylett, M.P., Pullin, G. Braude, D.A., Potard, B., Henning, S. & Antunes Ferreira, M.	Aylett, M.P., Braude, D.A. & Potard, B.	In Proceedings of the 2016 HI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)	2016		AAC, interactive media, speech synthesis	WP4 T4.4	Developing expressive speech synthesis is a key requirement for WP4 T4.4. This paper describes a technical probe created in collaboration with Dundee college of art which explored subjects experience and needs for expressivity in limited conversational environments.
Cross Modal Evaluation of High Quality Emotional Speech Synthesis with the Virtual Human Toolkit	Potard, B., Aylett, M.P. & Braude, D.A.	Potard, B., Aylett, M.P. & Braude, D.A.	In Proceedings of the 16th International Conference on Intelligent Virtual Agents (IVA'16).	2016		Speech synthesis, Unit selection, Expressive speech synthesis, Emotion, Prosody, Facial animation	WP4 T4.4	Creating and assessing the effect of emotional speech synthesis is central to creating expressive speech synthesis. How this interacts with a rendered graphical head supports ARIA integration with Greta.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Idlak Tangle: An Open Source Kaldi Based Parametric Speech Synthesiser based on DNN	Potard, B., Aylett, M.P., Braude, D.A. & Motlicek, P.	Potard, B., Aylett, M.P., & Braude, D.A.	In Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016)	2016		Speech synthesis, Kaldi, Idlak, HTS, DNN	WP4 T4.4	As stated in the proposal we remain agnostic on whether parametric or unit selection speech synthesis techniques for embodied conversational agents. DNN approaches to speech synthesis have demonstrated increased expressivity compared to other parametric techniques. This work describes a baseline freely available DNN system developed within a Kaldi framework and support expressive speech synthesis work in the parametric domain.
Demo of Idlak Tangle, An Open Source DNN-Based Parametric Speech Synthesiser	Potard, B., Aylett, M.P. & Braude, D.A.	Potard, B., Aylett, M.P. & Braude, D.A.	In Proceesings of the 9th ISCA Speech Synthesis Workshop (SSW2016)	2016		Speech synthesis, Kaldi, Idlak, HTS, DNN	WP4 T4.4	This paper describes the live demo of the Idlak DNN parametric system 'Tangle' which forms a baseline for expressive synthesis work using parametric approaches to speech synthesis (such as Wavenet etc).
Automatic Measures to Characterise Verbal Alignment in Human-Agent Interaction	Dubuisson Duplessis, G.; Clavel, C.; Landragin, F.	Dubuisson Duplessis, G.; Clavel, C.	18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)	2017	Emergence of synchrony between ECA and User	Adaptation, NLG, verbal alignment	WP3, WP4	This work provides verbal alignment measures based on repetition between dialogue participants at the lexical level. These measures can be leveraged in a NLG process to adapt system utterance to the user.
Shared acoustic codes underlie emotional communication in music and speech - evidence from deep transfer learning.	Eduardo Coutinho, Björn Schuller	Eduardo Coutinho, Björn Schuller	PLOS ONE, 12 (e0179289): 1-24, June 2017.	2017	Behaviour Analysis	speech, music, emotion, deep tranfer learning	2.1	Music and speech exhibit striking similarities in the communication of emotions in the acoustic domain. From a Machine learning perspective, the overlap between acoustic codes for emotional expression in music and speech opens new possibilities to enlarge the amount of data available to develop speech (and music) emotion recognition systems. In this research we investigated the Transfer Learning between these domains and demonstrated an excellent cross-domain generalisation performance in both directions. This directly feeds into the improvement of speech emotion recognition.
Automatically estimating emotion in music with deep longshort term memory recurrent neural networks.	Eduardo Coutinho, George Trigeorgis, Stefanos Zafeiriou, and Björn Schuller.	Eduardo Coutinho, Björn Schuller	Proceedings of MediaEval Multimedia Benchmark Workshop, satellite of INTERSPEECH, volume 1436, Wurzen, Germany, September 2015. CEUR.	2015	Behaviour Analysis	speech, music, emotion, deep tranfer learning	2.1	In this work we applied our speech emotion recongnition techniques to music emotion recognition. Given the striking similarities in the communication of emotions in music and speech, this work directly contributed to the improvement os speech emotion recognition models.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Exploring the importance of individual differences to the automatic estimation of emotions induced by music.	Hesam Sagha, Eduardo Coutinho, and Björn Schuller.	Eduardo Coutinho, Björn Schuller	Proceedings of International Workshop on Audio/Visual Emotion Challenge (AVEC)	2015	Behaviour Analysis	perceived emotion, affect induction, personality, emotional intelligence, mood states, physiological signals	2.3, 4.1	The goal of this study was to evaluate the impact of the inclusion of listener-related factors (individual differences) on the prediction of music induced affect (the context in which the work was developed). We identified individual trists that have a significant explanatory power over the affective states induced in the listeners. Our results show that incorporating information related to individual differences permits to identify more accurately the affective states induced in the listeners, which differ from those expressed by the music. This work has direct implications to user-adaptation in the context of emotional speech synthesis.
The icl-tum-passau approach for the mediaeval 2015 'affective impact of movies' task	George Trigeorgis, Eduardo Coutinho, Fabien Ringeval, Erik Marchi, Stefanos Zafeiriou, and Björn Schuller.	Eduardo Coutinho, Björn Schuller	Proceedings of MediaEval Multimedia Benchmark Workshop (satellite of INTERSPEECH), vol. 1436, Wurzen, Germany, September 2015. CEUR	2015	Behaviour Analysis	face, speech, emotion, deep learning	2.1, 2.4	In this paper we describe our participation in the MediaEval's 'Affective Impact of Movies' challenge, which consists in the automatic detection of affective (arousal and valence) and violent content in movie excerpts. This effort had a direct impact on the audio-visual emotion recognition work developed in this project.
Building autonomous sensitive artificial listeners (extended abstract).	Marc Schroeder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Björn Schuller, Etienne de Sevin, Michel Valstar, and Martin Woellmer.	Björn Schuller, Michel Valstar	Proceedings of Conference on Affective Computing and Intelligent Interaction (ACII), pages 456-462, Xi'an, P. R. China, September 2015. IEEE.	2015	Behaviour Analysis	real-time interactive multimodal dialogue system, non-verbal interaction, sensitive artificial listener	All WPs	This paper describes a substantial effort to build a real-time interactive multimodal dialogue system with a focus on emotional and non-verbal interaction capabilities a fully autonomous integrated real-time system created in previous work. The systems combines incremental analysis of user behaviour, dialogue management, and synthesis of speaker and listener behaviour of an artificial character displayed as a virtual agent. Principles that should underlie the evaluation of these systems are also discussed. This paper described the groundwork that established a departure point for ARIA.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Sentiment analysis and opinion mining: On optimal parameters and performances.	Björn Schuller, Amr El-Desoky Mousa, Vryniotis Vasileios.	Björn Schuller	WIREs Data Mining and Knowledge Discovery, 5:255-263, September 2015.	2015	Behaviour Analysis	Sentiment analysis, opinion mining, parameters, performance, machine learning	2.1, 2.3	Sentiment analysis is the task of identifying the polarity and subjectivity of documents using a combination of machine learning, information retrieval, and natural language processing techniques. This paper focuses on practical issues in statistical machine learning, and specifically to tedtermine the best feature selection methods, dimensionality reduction algorithms and classification techniques. This work has a direct impact to the recognition of affective information from linguistic context of the users' speech.
Face reading from speech - predicting facial action units from audio cues.	Fabien Ringeval, Erik Marchi, Marc Méhu, Klaus Scherer, and Björn Schuller.	Björn Schuller	Proceedings of INTERSPEECH, pages 1977-1981, Dresden, Germany, September 2015. ISCA.	2015	Behaviour Analysis	face, speech, facial action units, audio cues, machine learning	2.1, 2.4	We present in this paper the very first attempt in using acoustic cues for the automatic detection of FACS AU, as an alternative way to obtain information from the face when such data are not available. Results show that features extracted from the voice can be effectively used to predict different types of FACS AU. This work provides a new solution to improve the robusteness of emotion recognition from facial cues, even when the image information is not available.
AVEC 2015: The 5th international audio/visual emotion challenge and workshop.	Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic.	Björn Schuller, Michel Valstar	Proceedings of ACM International Conference on Multimedia, pages 1335-1336, Brisbane, Australia, October 2015. ACM.	2015	Behaviour Analysis	audio, video, emotion, challenge	2.4	The fifth Audio-Visual Emotion Challenge and workshop AVEC 2015 was held in conjunction ACM Multimedia'15. The workshop/challenge addresses the detection of affective signals represented in audio-visual data in terms of high-level continuous dimensions. The goal of the Challenge is to provide a common benchmark test set for multimodal information processing and to bring together the audio, video and physiological emotion recognition communities, to compare the relative merits of the three approaches to emotion recognition under well-defined and strictly comparable conditions and establish to what extent fusion of the approaches is possible and beneficial.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
AV+EC 2015 - the first affect recognition challenge bridging across audio, video, and physiological data.	Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic.	Björn Schuller, Michel Valstar	Proceedings of International Workshop on Audio/Visual Emotion Challenge, pages 3-8, Brisbane, Australia, October 2015. ACM.	2015	Behaviour Analysis	audio, video, physiology, emotion, challenge	2.4	This paper described the first Audio-Visual+ Emotion recognition Challenge and workshop (AV+EC 2015) aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and physiological emotion analysis. The goal of the Challenge is to provide a common benchmark test set for multimodal information processing and to bring together the audio, video and physiological emotion recognition communities, to compare the relative merits of the three approaches to emotion recognition under well-defined and strictly comparable conditions and establish to what extent fusion of the approaches is possible and beneficial.
Speech analysis in the big data era.	Björn Schuller	Björn Schuller	Proceedings of International Conference on Text, Speech and Dialogue, volume 9302 of Lecture Notes in Computer Science (LNCS), pages 3-11. Springer, September 2015.	2015	Behaviour Analysis	Speech analysis, paralinguistics, big data, self-learning	WP2	This is a position paper on issues related to data scarcity in spoken language analysis tasks. This contribution shows the de-facto standard in terms of data-availability in a broad range of speaker analysis tasks that can be explored to improve the state-of-the-art in spoken language analysis, including the work developed in the context of ARIA for achieving these goals (e.g., 'cooperative' learning, dynamic aactive learning, multitask learning). New directions are also discussed.
Automatic estimation of biosignals from the human voice.	Eduardo Coutinho and Björn Schuller	Eduardo Coutinho, Björn Schuller	Science, Special Supplement on Advances in Computational Psychophysiology, 350(6256):114:48-50, October 2015.	2015	Behaviour Analysis	computational paralinguistics, applications, biosignals estimation	2.1, 2.3	This is an invited contribution in which we provide an overview of how Computational Paralinguistics can offer new solutions for health care?the recognition of physiological parameters (biosignals) from the voice alone. This work has a direct impact to the recognition of user states (affective and cognitive).
Semi-supervised active learning for sound classification in hybrid learning environments.	Wenjing Han, Eduardo Coutinho, Huabin Ruan, Haifeng Li, and Björn Schuller	Eduardo Coutinho, Björn Schuller	PLoS ONE, 11(9):e0162075, 2016.	2016	Behaviour Analysis	active learning, sound events classification, annotation effort reduction	2.3	This work described the application of the algorithms developed in ARIA for the reduction of annotated effort and increasing the ammount annotated data for an improvement of modelling tasks to a new domain - environmental sounds classification. This work has a direct impact to the recognition of contextual cues in the content human-machine communication.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Deep recurrent music writer: Memory-enhanced variational autoencoder-based musical score composition and an objective measure	Romain Sabathe, Eduardo Coutinho, and Björn Schuller	Eduardo Coutinho, Björn Schuller	Proceedings Of International Joint Conference on Neural Networks (IJCNN), pages 3467-3474, May 2017.	2017		automatic music composition, evaluation metric, variational autoencoders, generative models	None	This work advances state-of-the-art in automatic music composition through the use of truly generative models based on Variational Autoencoders. We also introduce and evaluate a new metric for an objective assessment of the quality of the generated pieces. We demonstrate that our model can generate music pieces that follow general stylistic characteristics of a given composer or musical genre, and that the newly proposed measure permits investigating the impact of various parameters and model architectures on the compositional process and output.
Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network.	George Trigeorgis, Fabien Ringeval, Raymond Bruckner, Erik Marchi, Mihalis Nicolaou, Björn Schuller, and Stefanos Zafeiriou	Björn Schuller	Proceedings of ICASSP, pages 5200-5204, Shanghai, P. R. China, March 2016. IEEE.	2016	Behaviour Analysis	end-to-end learning, raw waveform, emotion recognition, deep learning, CNN, LSTM	2.1, 2.4	The automatic recognition of spontaneous emotions from speech is a challenging task. On the one hand, acoustic features need to be robust enough to capture the emotional content for various styles of speaking, and while on the other, machine learning algorithms need to be insensitive to outliers while being able to model the context. In this paper, we propose a novel solution to the problem of 'context-aware' emotional relevant feature extraction in order to automatically learn the best representation of the speech signal directly from the raw time representation (end-to-end speech emotion recognition). This research established the ground work for ARIA's audio-visual emotion recognition system.
Does my speech rock? automatic assessment of public speaking skills.	Lucas Azais, Adrien Payan, Tianjiao Sun, Guillaume Vidal, Tina Zhang, Eduardo Coutinho, Florian Eyben, and Björn Schuller	Eduardo Coutinho, Björn Schuller	Proceedings of INTERSPEECH, pages 2519-2523, Dresden, Germany, September 2015. ISCA.	2015	Behaviour Analysis	Automatic Public Speech Assessment, database, classification, regression, prosody	2.2, 2.3	This paper describes an investigation of which suprasegmental speech features can be used for evaluating oratory speaking skills. It also provides a new annotated database for the development of Automatic Public Speech Assessment (APSA) models. In the context of ARIA, this work is particularly relevant for the development of realistic speech synthesisers by providing a framework for evaluation of the oratory quality of the synthesized speech.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Assessing the prosody of non-native speakers of english: Measures and feature sets.	Eduardo Coutinho, Florian Hoenig, Yue Zhang, Simone Hantke, Anton Batliner, Elmar Noeth, and Björn Schuller	Eduardo Coutinho, Yue Zhang, Björn Schuller	Proceedings of Language Resources and Evaluation Conference (LREC), pages 1328-1332, Portoroz, Slovenia, May 2016. ELRA.	2016	Behaviour Analysis	non-native speech, prosody, feature evaluation	2.2, 2.3	In this paper, we describe a new annotated database with audio recordings of non-native (L2) speakers of English, and the perceptual evaluation experiment conducted with native English speakers for assessing the prosody of each recording. We also compared the relevance of different feature groups modelling prosody in general (without speech tempo), speech rate and pauses modelling speech tempo (fluency), voice quality, and a variety of spectral features. We also discuss the impact of various fusion strategies on performance. This work is directly relevant user-adaptation in terms of individual differences in linguistic fluency that are relevant from Automatic Speech Recongnition.
Enhanced semi-supervised learning for multimodal emotion recognition	Zixing Zhang, Fabien Ringeval, Bin Dong, Eduardo Coutinho, Erik Marchi, Björn Schuller	Zixing Zhang, Eduardo Coutinho, Björn Schuller	Proceedings of ICASSP, pages 5200-5204, Shanghai, P. R. China, March 2016. IEEE.	2016	Behaviour Analysis	Multimodal emotion recognition, enhanced semi-supervised learning	2.1, 2.4	In this paper, we propose an enhanced semi-supervised learning (SSL) approach to address two issues of SSL in the context of emotion recognition: I) performance degradation; 2) noise accumulation problem. Initially, we exploit the complementarity between audio-visual features to improve the performance of the classifier during the supervised phase. Then, we iteratively re-evaluate the automatically labeled instances to correct possibly mislabeled data and this enhances the overall confidence of the system's predictions. This work directly contributed to the improvement of multi-modal emotion recognition models by leveraging largely-available unlabelled data.
Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks	Zixing Zhang, Fabien Ringeval, Jing Han, Jun Deng, Erik Marchi, Björn Schuller	Zixing Zhang, Björn Schuller	Proceedings of INTERSPEECH, pages 3593-3597, San Francisco, CA, September 2016. ISCA	2016	Behaviour Analysis	emotion recognition, spontaneous speech, additive and convolutional noises, feature enhancement, autoencoder, LSTM Neural Networks	2.3	This work address the performance degradation problem when putting the speech emotion recognition systems in real-life conditions, where environmental additive and convolutional noises severely impact the system performance. We proposed to evaluates the impact of a front-end feature enhancement method based on an autoencoder with long short-term memory neural networks, for robust emotion recognition from speech. Support Vector Regression is then used as a back-end for time- and value-continuous emotion prediction from enhanced features.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Towards intoxicated speech recognition	Zixing Zhang, Felix Weninger, Martin Woellmer, Jing Han, Björn Schuller	Zixing Zhang, Björn Schuller	Proceedings of International Joint Conference on Neural Networks (IJCNN), pages 1555-1559, 2017. IEEE.	2016	Behaviour Analysis	speech recognition, speaker intoxication	2.1, 3.5	In a real-life scenario, the acoustic characteristics of speech often suffer from the variations induced by diverse environmental noises and different speakers. Almost all previous studies only considered the speakers' long-term traits, such as age, gender, and dialect. Speakers' short-term states, for example, affect and intoxication, are largely ignored. This paper address one particular speaker state, alcohol intoxication, which has rarely been studied. This work has direct implications to user-adaptation in the context of speech recognition.
Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices	Maximilian Schmitt, Erik Marchi, Fabien Ringeval, Björn Schuller	Björn Schuller	Proceedings of ITG Symposium on Speech Communication (ITG SC), pages 264-268, Paderborn, Germany, 2016. VDE, IEEE	2016	Behaviour Analysis	autism detection, cross-linguistic	2.1, 2.2	The work focusses on automatic diagnosis of Autism Spectrum Conditions (ASC) from the voice in cross-lingual situation, which is rarely studies previously. We conducted extensive cross-lingual evaluations based on four databases collected in English, French, Hebrew, and Swedish. The datasets contain speech of children with ASC and typically developing (TD) children matched in both age and gender. We demonstrate automatic ASC vs TD classification to be feasible despite such variation with a remaining error. In the context of ARIA, this work is particularly relevant to the analysis of user profiling.
Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis	Kun Qian, Christoph Janott, Vedhas Pandit, Zixing Zhang, Clemens Heiser, Winfried Hohenhorst, Michael Herzog, Werner Hemmert, Björn Schuller	Zixing Zhang, Björn Schuller	IEEE Transactions on Biomedical Engineering, 64(8):1731-1741, 2017. IEEE	2017	Behaviour Analysis	Snore Sound Classification, Multi-Feature Analysis	2.1, 2.2	This work systematically compares different acoustic features, and classifiers for their performance in the classification of the excitation location of snore sounds. Snore sounds from 40 male patients have been recorded during Drug-Induced Sleep Endoscopy, and categorized by ENT experts. Crest Factor, Fundamental Frequency, and the others have been extracted and fed into several classifiers. Using the ReliefF algorithm, features have been ranked and the selected feature subsets have been tested with the same classifiers. In the context of ARIA, this work is particularly relevant to the analysis of user profiling.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Bird sound classification by large scale acoustic features and extreme learning machine	Kun Qian, Zixing Zhang, Fabien Ringeval, Björn Schuller	Zixing Zhang, Björn Schuller	Proceedings of IEEE Global Conference on Signal and Information Processing (GlobalSIP), pages 1317-1321, Orlando, FL, December 2015. IEEE.	2015	Environment Anaylysis	Bird Sounds, p-centre, openS-MILE, ReliefF, Extreme Learning Machine	2.3	In this paper, we present a novel framework for bird sounds classification from audio recordings. Firstly, the p-centre is used to detect the 'syllables' of bird songs, which are the units for the recognition task; then, we use our openSMILE toolkit to extract large scales of acoustic features from chunked units of analysis (the 'syllables'). ReliefF helps to reduce the dimension of the feature space. Lastly, an Extreme Learning Machine (ELM) serves for decision making.This work has a direct impact to the recognition of environmental/contextual cues in the content human-machine communication.
Non-linear prediction with LSTM recurrent neural networks for acoustic novelty detection.	Erik Marchi, Fabio Vesperini, Felix Weninger, Florian Eyben, Stefano Squartini, and Björn Schuller	Björn Schuller	Proc. of International Joint Conference on Neural Networks (IJCNN), pages 1-7, Killarney, Ireland, July 2015. IEEE.	2015	Behaviour Analysis	Acoustic novelty detection, realistic and unex-pected situations	2.2, 3.5, 4.6	Novelty detection is a challenging task, and it aims at recognising situations in which unusual events occur. In this paper, we present a novel approach based on non-linear predictive denoising autoencoders. The autoencoder is trained on a public database which contains recordings of typical in-home situations such as talking, watching television, playing and eating. The evaluation was performed on more than 260 different abnormal events. In the result, our novel approach significantly outperforms existing methods. This work directly contributes to the ARIAs' abilities for handling unexpected situations and environmental conditions.
Adeep matrix factorization method for learning attribute representations.	George Trigeorgis, Konstanti-nos Bousmalis, Stefanos Zafeiriou, and Björn Schuller	Björn Schuller	arxiv.org	2015	Behaviour Analysis	face recog-nition, blind audio source separation	2.2	Non-negative matrix factorization (NMF) can be a successful dimensionality reduction technique over a variety of areas including, but not limited to, environmetrics, microarray data analysis, document clustering, face recognition, blind audio source separation and more. In this work, we propose a novel model, Deep Semi-NMF, that is able to learn such hidden representations that allow themselves to an interpretation of clustering according to different, unknown attributes of a given dataset. Within the context of ARIA, the novel deep framework for matrix factorization is suitable for clustering of multimodally distributed objects such as faces.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Deep canonical time warping.	George Trigeorgis, Mihalis A. Nicolaou, Stefanos Zafeiriou, and Björn Schuller	Björn Schuller	Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5110–5118, Las Vegas, NV, June 2016. IEEE.	2016	Behaviour Analysis	Temporal Alignment of multiple data sequences	2.2	In this work, we present the Deep Canonical Time Warping (DCTW), a method which automatically learns complex non-linear representations of multiple time-series. On four real datasets, we show that the representations learnt via the proposed DCTW significantly outperform state-of-the-art methods in temporal alignment, elegantly handling scenarios with highly heterogeneous features, such as the temporal alignment of acoustic and visual features. In ARIA, potential applications range from the temporal alignment of facial expressions and motion capture data, to the alignment for human action recognition, and speech.
Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions.	Yue Zhang, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn Schuller.	Yue Zhang, Eduardo Coutinho, Björn Schuller	Proc. of International Conference on Multimodal Interaction (ICMI), pages 275-278, Seattle, WA, November 2015. ACM.	2015	Behaviour Analysis	Active learning, annota- tion, NOXI	2.2, 6.3	Active Learning (AL) is a technique to reduce human effort and thus costs and time for the annotation of emotion, social cues, etc. In this work, we propose a novel AL algorithm, termed Dynamic Active Learning (DAL), which?makes the decision on a per instance level how many human annotators are required to determine the gold standard label.?To this end, an early stopping criterion based on inter-rater agreement is proposed. Due to its high efficiency, this novel approach has a direct positive?impact on the project by considerably accelerating the annotation process of NOXI.
Agreement-based dynamic active learning with least and medium certainty query strategy.	Yue Zhang, Eduardo Coutinho, Zixing Zhang, Caijiao Quan,?and Björn Schuller.	Yue Zhang, Eduardo Coutinho, Björn Schuller	Proc. of Advances in Active Learning: Bridging Theory and Practice Workshop held in conjunction with the International Conference on Machine Learning (ICML), Lille, France, July 2015. IMLS.	2015	Behaviour Analysis	Confidence measure, annota- tion, NOXI, SSI	2.2, 6.3	Cooperative Learning (CL) is a recently introduced methode to efficiently share the labellling work between human and machine, when facing label scarcity as a ever-present issue in data-driven fields. The idea is to adopt the machine labels for the instances predicted with high confidence, and only query human oracles in case of medium / low confidence. Based on our previous works on CL and DAL, we scrutinise the confidence levels for maximum efficiency. For speech emotion recognition, our results show that the DAL approach yields the same accuracy, but requires up to 67% less human annotations for the medium certainty and 79% for the least certainty query strategy. This approach has been intergrated into the ARIA core engine SSI.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
On rater reliability and agreement based dynamic active learning.	Yue Zhang, Eduardo Coutinho, Zixing Zhang, Michael Adam, and Björn Schuller.	Yue Zhang, Eduardo Coutinho, Björn Schuller	Proc. of Affective Computing and Intelligent Interaction (ACII), pages 70?76, Xi?an, P. R. China, September 2015. IEEE.	2015	Behaviour Analysis	Rater reliability, inter-rater agreement, NOVA	2.2, 6.3	In this work, we propose several variations of the DAL method, taking into account inter-rater reliability and inter-rater agreement. By query the most reliable rater first, we could achieve further cost reduction for large-scale data annotation.?This approach opens up new realisation possibilities for the NOn-Verbal Annotator (NOVA) tool developed within the ARIA project.
Multitask deep neural network with shared hidden layers: Breaking down the wall between emotion representations.	Yue Zhang, Yifan Liu, Felix Weninger, and Björn Schuller	Yue Zhang, Björn Schuller	Proc. of ICASSP, New Orleans, LA, March 2017. IEEE.	2017	Behaviour Analysis	Affect recognition,?emotion representations	2.1, 2.4	For speech emotion recognition, label scarcity presents a particular challenge as the limited available databases are usually associated to diversified emotion conceptions, derived from categorical, dimensional, and appraisal-based approaches. In this work, we advocate the usage of multi-task deep neural networks with shared hidden layers.?In this way,?an utterance can be interpreted in manifold ways according to various emotion representations, which is of particular importance for the affect recognition system (WP 2).
Language proficiency assessment of English L2 speakers based on joint analysis of prosody and native language.	Yue Zhang, Felix Weninger, Anton Batliner, Florian H?nig, and Björn Schuller	Yue Zhang, Björn Schuller	Proc. of ACM International Conference on Multimodal Interaction (ICMI), Tokyo, Japan, November 2016. ACM. 274?278.	2016	Behaviour Analysis	Native language identification, language proficiency assessment, user adaptation	2.4	The first Language (L1) influences the non-native prosody of users speaking English as a second language (L2). ?Language proficiency assessment is highly important for the user adaptivity of the ARIA system. ?For example, context-aware spoken dialogue systems can exploit accent-specific acoustic models, adapt the tempo of speech synthesis to the language proficiency of individual speakers, or even switch to a user's native language in case of difficulties with the interaction in the default language. Realising these capabilities in automatic systems conceivably leads to more natural and human-like interaction.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Semiautonomous data enrichment based on cross-task labelling of missing targets for holistic speech analysis.	Yue Zhang, Yuxiang Zhou, Jie Shen, and Björn Schuller	Yue Zhang, Björn Schuller	Proc. of ICASSP, pages 6090?6094, Shanghai, P. R. China, March 2016. IEEE.	2016	Behaviour Analysis	Label enrichment, data aggregation	2.2	In this work, we propose a novel approach for large-scale data enrichment, addressing the scarcity of multi-label databases. The idea of our work is to join existing data resources into one universal database with a multi-dimensional label space by using semi-supervised learning techniques to predict missing labels. We evaluated the proposed method for likability, personality, and emotion recognition as exemplary tasks from the Computational Paralinguistic Challenge (ComParE). This work enables the model training on aggregated data for the holistic speaker analysis, which is a key component for the ARIA system.
Sincerity and deception in speech: Two sides of the same coin? a transfer- and multi-task learning perspective.	Yue Zhang, Felix Weninger, Zhao Ren, and Björn Schuller	Yue Zhang, Björn Schuller	Proc. of INTERSPEECH, pages 2041?2045, San Francisco, CA, September 2016. ISCA.	2016	Behaviour Analysis	Non-verbal social cues, context understanding	2.2, 2.3	Identifying contexts and capturing non-verbal behavioural cues present a central aspect of social intelligence, and thus is highly important for the ARIAs to be able to naturally interact with human users. In this work, we propose a novel multi-task learning method for recognising speech deception and sincerity.?To this end, we employ our previously introduced method for data aggregation by semi-supervised cross-task label completion. In the result, our approach achieves significant error rate reductions compared to state-of-the-art systems.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring.	Björn Schuller, Stefan Steidl, Anton Batliner, Elika Bergelson, Jarek Krajewski, Christoph Janott, Andrei Amatuni, Marisa Casillas, Amanda Seidl, Melanie Soderstrom, Anne Warlaumont, Guillermo Hidalgo, Sebastian Schnieder, Clemens Heiser, Winfried Hohenhorst, Michael Herzog, Maximilian Schmitt, Kun Qian, Yue Zhang, George Trigeorgis, Panagiotis Tzirakis, and Stefanos Zafeiriou	Yue Zhang, Björn Schuller	Proc. of INTERSPEECH, pages 3442?3446, Stockholm, Sweden, August 2017. ISCA.	2017	Behaviour Analysis	Speaker analysis	2.1, 2.2, 2.3	The INTERSPEECH 2017 Computational Paralinguistics Challenge addresses three different problems for the first time in research competition under well-defined conditions: In the Addressee sub-challenge, it has to be determined whether speech produced by an adult is directed towards another adult or towards a child; in the Cold sub-challenge, speech under cold has to be told apart from ?healthy? speech; and in the Snoring sub-challenge, four different types of snoring have to be classified. Thus, this work addes new aspects for the speaker analysis of the ARIA system.
The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception & Sincerity.	Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini	Yue Zhang, Eduardo Coutinho, Björn Schuller	Proc. of INTERSPEECH, pages 2001?2005, San Francisco, CA, September 2016. ISCA.	2016	Behaviour Analysis	Speaker analysis	2.1, 2.2, 2.3	The INTERSPEECH 2016 Computational Paralinguistics Challenge addresses three different problems for the first time in research competition under well-defined conditions: classification of deceptive vs. non-deceptive speech, the estimation of the degree of sincerity, and the identification of the native language out of eleven L1 classes of English L2 speakers. Given the three different languages used in the ARIA system, native language identification plays a particularly important role for user adaptation.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Towards human-like holisitc machine perception of speaker states and traits.	Yue Zhang, Yifan Liu, and Björn Schuller	Yue Zhang, Björn Schuller	Proc. of the Human-Like Computing Machine Intelligence Workshop (MI20-HLC), pages 1?3, Windsor, U.K., October 2016. Springer.	2016	Behaviour Analysis	Speaker analysis	2.1, 2.2, 2.3	In this work, we advocate the usage of multi-task deep neural networks with shared hidden layers for various paralinguistic tasks. To this end, the feature transformations are shared across dierent tasks, while the softmax layers are separately associated with each target label. As a new milestone in holistic speech processing, we constructed a multilabel database, thus enabling large-scale data aggregation for better recognition performance. In ARIA, this work allows SSI to emcompass all paralinguistic speech phenomena featured in the ComParE challenges.
A paralinguistic approach to speaker diarisation.	Yue Zhang, Felix Weninger, Boqing Liu, Maximilian Schmitt, Florian Eyben, and Björn Schuller	Yue Zhang, Björn Schuller	Proc. of ACM International Conference on Multimedia, pages 387?392, Mountain View, CA, October 2017. ACM.	2017	Behaviour Analysis	Speaker diarisation	2.2, 3.5, 4.6	Speaker diarisation is the task of determining ?who speaks when? in an audio stream.?In this work, we present a new view on automatic speaker diarisation, based on the recognition of speaker traits such as age, gender, voice likability, and personality. Since in real-life, ARIAs would encounter situations when they deal with multiple users at once, speaker diarisation is highly relevant for ARIAs to handle challenging and unexpected situations.
Cross-domain classification of drowsiness in speech: The case of alcohol intoxication and sleep deprivation	Yue Zhang, Felix Weninger, and Björn Schuller	Yue Zhang, Björn Schuller	Proc. of INTERSPEECH, Stockholm, Sweden, August 2017. ISCA.	2017	Behaviour Analysis	Speaker analysis	2.1, 2.2, 2.3	In this work, we study the drowsy state of a speaker, induced by alcohol intoxication or sleep deprivation. In particular, we show that an effective, general drowsiness classifier can be obtained by aggregating the training data from both domains.?Since ARIAs can be integrated in self-driving cars and other safty and security sensitive environments, these recognition models can be highly useful in practical use.
Infected Phonemes: How a Cold Impairs Speech on a Phonetic Level	Johannes Wagner, Thiago Fraga-Silva, Yvan Josse, Dominik Schiller, Andreas Seiderer, and Elisabeth Andr'e	Johannes Wagner, Elisabeth Andr'e	Proc. of INTERSPEECH, Stockholm, Sweden, August 2017. ISCA.	2017	Behaviour Analysis	Speaker analysis	2.1, 2.2, 2.3	In this work we investigate the audible effects of a cold on a phonetic level. Results on a German corpus show that the articulation of consonants is more impaired than that of vowels. With such knoweldge we can improve the robustness of the paralinguistic analysis integrated in the ARIA system.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
MobileSSI: Asynchronous Fusion for Social Signal Interpretation in the Wild	Simon Flutura, Johannes Wagner, Florian Lingenfelser, Andreas Seiderer, and Elisabeth André	Johannes Wagner, Elisabeth André	Proceedings of the 18th ACM International Conference on Multimodal Interaction	2016	Behaviour Analysis	Speaker analysis	2.1, 2.2, 2.3	In this paper MobileSSI, a port of the Social Signal Interpretation (SSI) framework to Android and embedded Linux platforms is introduced. It is tested to what extent it is possible to run sophisticated synchronization and fusion mechanisms in an everyday mobile setting and compare the results with similar tasks in a laboratory environment. This can be helpful to run parts of the ARIA detection system natively on a mobile device.
Asynchronous and Event-based Fusion Systems for Affect Recognition on Naturalistic Data in Comparison to Conventional Approaches	F. Lingenfelser and J. Wagner and J. Deng and R. Bruckner and B. Schuller and E. André	Johannes Wagner, Elisabeth André	IEEE Transactions on Affective Computing	2017	Behaviour Analysis	Speaker analysis	2.1, 2.2, 2.3	Recognition results gained on a naturalistic conversational corpus show a drop in recognition accuracy when moving from unimodal classification to synchronous multimodal fusion. In this article, we taggle this problem and present a novel real-time system for affect recognition in a naturalistic setting. Since ARIA makes use of multiple modalities, the tested techniques may be applied in the future to improve the robustness of the uni-modal recognizers.
Combining Hierarchical Classification with Frequency Weighting for the Recognition of Eating Conditions	Wagner, Johannes and Seiderer, Andreas and Lingenfelser, Florian and Andre, Elisabeth	Johannes Wagner, Elisabeth André	INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015	2015	Behaviour Analysis	Speaker analysis	2.1, 2.2, 2.3	In this paper we classify whether a speaker is eating or not, and if so, which type of food the speaker is currently tasting. To allow for a fine-grained adaption to the characteristic spectrum of single food types we adopt a hierarchical tree structure and decompose the classification task into a sequence of binary decisions. This may help to improve the robustness of the paralinguistic analysis integrated in the ARIA system.
Cumulative attributes for pain intensity estimation	Edege, J. and Valstar, M.	Michel Valstar	Proceedings of the 19th ACM International Conference on Multimodal Interaction	2017	Behaviour Analysis	Pain estimation; Attribute learning; Multi-output regression; Relevance Vector Machines	WP2, T2.1	This paper presents a novel approach to automatic pain estimation, in which different set of features (appearance and shape) are used to predict an output vector lying in what is referred to as a Cumulative Attribute (CA) space. The CA encodes all the ordinal levels of pain up to the one that corresponds to the target frame. The CA outputs are finally regressed to give a final pain estimate. The paper is relevant to ARIA as it offers an efficient method for pain estimation, using a technique that can be extended to predicting the intensity of Action Units, which are used in the visual part.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Automatic Analysis of Facial Actions: A Survey	Brais Martinez, Michel F Valstar, Bihan Jiang, Maja Pantic	Brais Martinez and Michel Valstar	IEEE Transactions on Affective Computing (in press)	2017	Behaviour Analysis	*facial expression recognition, action units	WP2, T2.1	This paper presents a thorough review of the state of the art techniques and databases in Facial Action Units detection and intensity estimation. The paper includes a detailed review of the main components that face analysis systems need. It also summarises the existing and available databases, and includes an overview of the challenges that remain to be solved. The paper is relevant to ARIA to illustrate what are the challenges and opportunities in the task of Action Unit detection and intensity estimation. These tasks are key in the visual part of eMax, as it is responsible for returning the values for the Action Units intensities.
Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation	Joy Egede, Michel Valstar, Brais Martinez	Michel Valstar and Brais Martinez	12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 689-696	2017	Behaviour Analysis	*feature fusion, pain estimation	WP2, T2.1	The paper proposes an efficient method for feature fusion to automatically estimate the level of pain from the face. Considering the lack of annotated data for the task of pain estimation, a deep neural network tasked with detecting Action Units, is used, given that these correlate with the level of pain. The deep learned features are thus those extracted in the last level of the network. These features are combined with hand-crafted features, such as those given by the appearance (HOG) and the facial landmarks (geometric). Each of the three set of features is the input of a corresponding Relevance Vector Regressor (RVR), and the output of each RVR feeds a second-level RVR, which returns the final score. The results of this paper are of interest for future issues of an ARIA framework, desired to understand whether a user is suffering pain, for instance in a task-assisted scenario.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Automatic Detection of ADHD and ASD from Expressive Behaviour in RGBD Data	Shashank Jaiswal, Michel F Valstar, Alinda Gillott, David Daley	Michel Valstar	12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 762-769	2017	Behaviour Analysis	*ADHD, ASD, Action units	WP2, T2.1	The paper presents a novel method to aid the diagnostic of Attention Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD). The questionnaires that are generally employed by experts to evaluate the patient behaviour are shown and analysed automatically by the proposed system. This system employs RGBD data from a Kinect 2 camera, and extracts a set of high-level descriptors, including Action Units, Kinect Animation Units, Head Pose, Speed of Head movement, Cumulative distance, and response times. These features are then fed to a classifier that returns the expected ADHD/ASD score. The results of this paper are of interest for future issues of an ARIA framework capable of automatically report these specific disorders.
ACNN cascade for landmark guided semantic part segmentation	Aaron S Jackson, Michel Valstar, Georgios Tzimiropoulos	Michel Valstar	Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III, pp. 143-155	2016	Behaviour Analysis	*CNN, face segmentation, face analysis	WP2, T2.1	The paper presents a deep Convolutional Neural Network (CNN) cascade for facial parts segmentation. These segments correspond to meaningful parts of the face, such as the mouth or the eyes. The CNN performs first facial localisation, and it is followed by the part segmentation. The results of the proposed approach show that guiding the segmentation from the landmark localisations improves the performance drastically. The paper is of relevance to ARIA in the sense that it provides with an accurate understanding of facial parts, which can thereafter guide the detection of Action Units.
Cascaded Continuous Regression for Real-time Incremental Face Tracking	Enrique S?nchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, Michel Valstar	E. S?nchez-Lozano, M. Valstar, B. Martinez	Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII, pp. 646-661	2016	Behaviour Analysis	*face tracking, incremental learning, continuous regression, functional regression	WP2, T2.1	The paper presents a novel approach to facial landmark tracking using Cascaded Regression, by extending the linear regression problem to the continuous domain. Instead of generating samples, the paper proposes to use a first-order Taylor approximation of the feature space, yielding a close-form solution for the linear regressor. Its inclusion into the Cascaded Regression approach, and the development of the incremental learning rules allows the method to perform incremental learning in real-time, being the first (and so far the only) tracker that incorporates this capacity. The tracker resulting out of this paper was the preliminary one used in the visual part of ARIA.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Cascaded regression with sparsified feature covariance matrix for facial landmark detection	Enrique S?nchez-Lozano and Brais Martinez and Michel Valstar	E. S?nchez-Lozano, M. Valstar, B. Martinez	Pattern Recognition Letters, 73, pp 19-26	2016	Behaviour Analysis	Supervised descent method, Cascaded regression, Facial point localisation	WP2, T2.1	The paper presents a Cascaded Regression approach for facial landmark localisation in which the linear regression problem considers the correlation that exists between different landmarks. The solution to the least-squares problem considers the covariance of the features extracted at the landmark localisations. This paper proposes a sparsification of the covariance matrix towards removing the influence of the features extracted at positions that might be highly uncorrelated with the target landmark. The paper is relevant to the visual part of ARIA, given that the tracker?s initialisation uses this method to detect the facial landmarks for the first frame.
Deep Learning the Dynamic Appearance and Shape of Facial Action Units	Shashank Jaiswal and Michel Valstar	M. Valstar	IEEE Winter Conference on Applications of Computer Vision (WACV)	2016	Behaviour Analysis	*deep learning, action units, lstm, optical flow	WP2, T2.1	The paper presents a Deep Learning approach to Facial Action Unit detection. The proposed approach first slides the image into regions of interest, and then computes a combination from the appearance of the first frame with the optical flow computed for the rest of the frames, as well as a set of binary masks for each of the frames, responsible of encoding the shape variations. Each of the regions for each of the images of a given sequence feeds a Convolutional Neural Network that computes a set of features that are subsequently used with a Bi-directional LSTM, responsible for encoding the temporal consistency. The proposed approach is key to the ARIA visual part, as it has proved to be the best framework on the FERA 2015 dataset. This system is being currently used to extract the Action Units, which are part of the features returned by eMax.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
L2;1-based regression and prediction accumulation across views for robust facial landmark detection	Brais Martinez and Michel Valstar	B. Martinez and M. Valstar	Image and Vision Computing, vol. 47, pp. 36-44. 2016	2016	Behaviour Analysis	*facial landmark localisation, multiview, face alignment, cascaded regression	WP2, T2.1	The paper presents a Cascaded Regression approach to facial landmark tracking, in which the regression problem, typically learnt through Least-Squares, is replaced by the L2,1 norm, making it more robust to poor initialisations or partial occlusions. The L2,1 norm provides a sparse representation of the per-landmark localisation error, thus enforcing the error of occluded landmarks not to contribute to the total cost in a dramatic way. Besides, the paper proposes to use multiple initialisations that are efficiently combined to improve the accuracy of the final landmark localisations. The paper is relevant to the ARIA visual system, and was used at first as an estimator, before the development of the accurate and fast face tracker described in this report.
IProbe, Therefore I Am: Designing a Virtual Journalist with Human Emotions.	Bowden, K., Nilsson, T., Spencer, C., Cengiz, K., Ghitulescu, A. and van Waterschoot, J.	Alexandru Ghitulescu and Jelte van Waterschoot	Proceedings of the 12th Summer Workshop on Multimodal Interfaces (eNTERFACE ?16), pp. 47-53, July 18 ? August 12, Enschede, The Netherlands	2017	Dialogue Management	user adaptation, emotions	WP 3.2, 3.3	In this paper we discuss a Virtual Human Journalist, a project employing a number of novel solutions from these disciplines with the goal to demonstrate their viability by producing a humanoid conversational agent capable of naturally eliciting and reacting to information from a human user. We argue that naturalness should not always be seen as a desirable goal and suggest that deliberately suppressing the naturalness of virtual human interactions, such as by altering its personality cues, might in some cases yield more desirable results.
HAI Alice - An Information-Providing Closed-Domain Dialog Corpus	Jelte van Waterschoot, Merijn Bruijnes, Guillaume Dubuisson Duplessis and Dirk Heylen	Jelte van Waterschoot, Merijn Bruijnes, Guillaume Dubuisson Duplessis and Dirk Heylen	In Press	2018	Behaviour Synthesis	user adaptation, verbal alignment	WP 3.3	The contribution of this paper is twofold 1) we provide a public corpus for Human-Agent interaction (where the agent is controlled by a Wizard of Oz) and 2) we show a study on verbal alignment in Human-Agent interaction to exemplify the corpus' use. The goal of the data collection was to create a corpus with unexpected situations that can occur during a conversation between a virtual agent and a user, such as misunderstandings, (accidental) false information, and interruptions by another person. The HAI Alice-corpus consists of 15 conversations and more than 900 utterances. We transcribed the corpus and as a use-case example we measured the verbal alignment between the user and the agent. The paper contains information about the set-up of the data collection, the unexpected situations and a description of our verbal alignment study.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition.	Brais Martinez and Michel Valstar	B. Martinez and M. Valstar	Book Chapter in ?Advances in Face Detection and Facial Image Analysis?, pp 63-100.	2016	Behaviour Analysis	*facial expression recognition, face analysis, action units	WP2, T2.1	This book chapter summarises the state of the art techniques in Facial Expression Recognition (FER), including a review of all the blocks of which these techniques build on. It first starts describing the possible use cases of a FER system, by means of the model target. These are the categorical emotions (anger, happiness?), the FACS code system (Action Units), or the dimensional emotions (valence and arousal). Then, it describes the standard pipeline: face and landmark localisation and tracking, feature extraction, and machine learning techniques. Finally, it reviews the remaining challenges. This work is relevant to ARIA to illustrate the drawbacks of existing FER systems, given that these are needed to analyse users? interaction with the agent.
Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection	Timur Almaev, Brais Martinez, Michel Valstar	B. Martinez and M. Valstar	Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3774-3782	2015	Behaviour Analysis	*action units, multi-task learning	WP2, T2.1	The paper proposes a novel approach to Multi-Task learning, in which the latent structure between different Action Units is also considered. The method approaches targeting specific Action Units when these are only annotated for a subset of the training images. Thus, it is possible to train a classifier for all the target Action Units using different datasets even when these barely share the target annotations. The proposed approach is relevant to the ARIA visual part, as it was an efficient and relatively cheap method for Action Unit detection, needed to generate the visual features of the visual part.
TRIC-track: Tracking by Regression with Incrementally Learned Cascades	Xiaomeng Wang, Michel Valstar, Brais Martinez, Muhammad Haris Khan, Tony Pridmore	B. Martinez and M. Valstar	Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4337-4345	2015	Environment Anaylysis	*object tracking, supervised descent method, cascaded regression	WP2, T2.1	The paper presents a method for part-based object tracking, in which the only given information is the bounding box of the target object in the first frame of a video sequence. The proposed approach trains a set of cascaded regressors on the go, whilst encoding certain shape constraints. These regressors are subsequently updated as the tracking is ongoing, updating the appearance around the current location of the target object, making it more accurate. The proposed framework was used at an early stage of the ARIA visual part to perform the needed face tracking.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
AVEC 2017 ? Real-life Depression, and Affect Recognition Workshop and Challenge	Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmi, Maja Pantic	B. Schuller and M. Valstar	Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge. New York : ACM, 2017, p. 3-9, held in conjunction with ACM Multimedia	2017	Behaviour Analysis and Corpus	Affective Computing; Social Signal Processing; Automatic Emotion/Depression Recognition	WP2, T2.1, T2.4	The paper presents the seventh competition and workshop aimed at comparing methods for automatic audiovisual depression and emotion analysis. A new dataset, SEWA, is used in this edition for the Affect Sub-Challenge. The paper presents a baseline system and encourages participants to submit their methods for a strict comparison under the same benchmark. The paper is relevant to ARIA as it brings a gathering of state of the art methods in emotion and depression recognition, and helps understand the current limitations in the field. This is important for the visual input of ARIA, eMax, as it is responsible of outputting the affectional dimensions of users in unconstrained conditions.
FERA 2017-Addressing Head Pose in the Third Facial Expression Recognition and Analysis Challenge	Michel F Valstar, Enrique S?nchez-Lozano, Jeffrey F Cohn, L?szl? A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, Maja Pantic	E. S?nchez-Lozano, M. Valstar	12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)	2017	Behaviour Analysis and Corpus	*facial expression recognition, face analysis, action units, multiview	WP2, T2.1	The paper presents the third FERA challenge. The novelty with respect to previous challenges resides in that the database is now prepared to cover a wide range of head poses, toward developing the state of the art in facial expression recognition to unconstrained scenarios, in which the camera view is not predefined. A new database, synthesised using 3D models of an extended version of the FERA 2015 data (BP4 dataset), is released, and results are given both overall and per-view, illustrating the challenges yet to be covered in the field. The paper points out the conditions on which facial expression recognition systems are prone to fail. These are of importance to ARIA as the framework is expected to work in unconstrained scenarios
AVEC 2016 ? Depression, Mood, and Emotion Recognition Workshop and Challenge	Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, Maja Pantic	M. Valstar, Bjorn Schuller	Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 3-10. Held in conjunction with ACM Multimedia	2016	Behaviour Analysis and Corpus	*Affective Computing, Social Signal Processing, Automatic Emotion/Depression Recognition	WP2, T2.1, T2.4	The paper describes the challenge and baseline for the sixth series of competitions on multimedia processing and machine learning for automatic video, visual, and physiological depression and emotion analysis. AVEC 2016 basically re-runs AVEC 2015 in the emotion recognition sub-challenge, but introduces a novel dataset for the depression severity estimation sub-challenge. The paper aims at proposing a common benchmark to evaluate the state of the art on emotion recognition, which is of relevance for ARIA both in the audio and the visual systems, which are responsible of analysing the users? emotions.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
ChaLearn Looking at People and faces of the world: face analysis workshop and challenge 2016	Sergio Escalera, Mercedes Torres Torres, Brais Martinez, Xavier Bar?, Hugo Jair Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian Corneou, Marc Oliu, Mohammad Ali Bagheri, Michel Valstar	B. Martinez and M. Valstar	Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop "Looking at People"	2016	Behaviour Analysis and Corpus	*complements, face analysis	WP2, T2.1	This paper presents a three-competition challenge, addressing age estimation, accessory classification and smile and gender classification, respectively. The dataset has been collected and labelled following a crowd-sourcing approach, using a custom-built application. The paper includes the results attained by participants. The work aims at providing a unified framework for facial analysis in unconstrained conditions, and presents an analysis of state of the art methods, gathering under the proposed benchmark. The paper sheds light on the state of the art in three important challenges that are necessary in facial analysis, and thus are of relevance to the visual part of ARIA (eMax).
Ask Alice: an artificial retrieval of information agent	Michel Valstar, Tobias Baur, Angelo Cafaro, Alexandru Ghitulescu, Blaise Potard, Johannes Wagner, Elisabeth André, Laurent Durieu, Matthew Aylett, Soumia Dermouche, Catherine Pelachaud, Eduardo Coutinho, Björn Schuller, Yue Zhang, Dirk Heylen, Mari?t Theune, Jelte van Waterschoot	Michel Valstar, Tobias Baur, Angelo Cafaro, Alexandru Ghitulescu, Blaise Potard, Johannes Wagner, Elisabeth André, Laurent Durieu, Matthew Aylett, Soumia Dermouche, Catherine Pelachaud, Eduardo Coutinho, Björn Schuller, Yue Zhang, Dirk Heylen, Mari?t Theune, Jelte van Waterschoot	Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 419-420	2016	Dialogue Management	Virtual Humans, Technology Demonstrator, Affective Computing,	*WP5	The paper presents a demonstration of the ARIA framework, in which Alice, the virtual human placed on top of the framework, acts as the expert on the book ?Alice in Wonderland?. The framework incorporates the recent advances in the project in facial and speech analysis, and can deal with interruptions in a gracefully way. This work incorporates all the building blocks of ARIA. The Core Agent block keeps the agent?s information state, and is responsible for making queries to its domain-knowledge database to answer questions. It is also responsible for deciding which information, and which intents, the agent will express. This work is key to ARIA as it was a proof of concept at an intermediate state of the project

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Playing with Social and Emotional Game Companions	Andry Chowanda, Martin Flintham, Peter Blanchfield, Michel Valstar	M. Valstar	International Conference on Intelligent Virtual Agents, pp. 85-95	2016	Dialogue Management and Behaviour Analysis	Social Signal Processing	*WP3	The paper presents a Game Companion, conceived as a Non-player Character (an agent) the users interact with when playing a game. The game companion analyses the user?s facial expressions, and introduces a dialogue management system, that helps the user develop an affective and social relation with the agent. The paper also studies the effect of this agent in users? engagement, by studying the interaction of users with two agents with opposite personality. The paper is relevant to ARIA in the sense that it introduces a social behaviour to the agent, which helps the user to engage with the system.
Topic Switch Models for Dialogue Management in Virtual Humans	Wenjue Zhu, Andry Chowanda, Michel Valstar	M. Valstar	International Conference on Intelligent Virtual Agents, pp. 407-411	2016	Dialogue Management	Social relationship, Framework, Game-agents, Interactions	*WP3	The paper presents a novel Topic Switch Model for a Dialogue Management, which learns connections between topics and between topics and utterances, enabling the selection of sentences that match the considered topic. The Dialogue Management works in a text-based manner. The agent responds to the user?s topic whilst maintaining and overlapping set of topics, and switching between them according to certain statistics. The paper is relevant to the ARIA project given that Alice needs to switch between subtopics in a natural way to the user, should they want to do so
Computational Models of Emotion, Personality, and Social Relationships for Interaction in Games (Extended Abstract).	Andry Chowanda, Peter Blanchfield, Martin Flintham, Michel Valstar	M. Valstar	Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, pp. 1343-1344	2016	Dialogue Management	Game Environment, Computational Models, Social Relationships,	*WP3	The paper presents a novel model for Non Player Character, endowed with emotions, personality, and social relationships. The NPC is introduced into a commercial game, to help users improve their experience. The study shows that players reported significant changes in their social relationship with the two different types of agents. It is shown that players appear to display an enhanced emotional attachment to the NPCs, and appear to forge relationships with them. The paper is relevant to ARIA because it demonstrates how users perceive a better experience when interacting with agents in a social manner.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Play SMILE Game with ERiSA: a user study on game companions	Andry Chowanda, Peter Blanchfield, Martin D Flintham, Michel F Valstar	M. Valstar	Workshop on Engagement in Social Intelligent Virtual Agents in Fifteenth International Conference on Intelligent Virtual Agents	2015	Dialogue Management	Social Interactions	*WP3	The paper presents a study conducted to evaluate the interaction between two different virtual agents and a set of participants. The virtual agents were differing in their personality, by means of extraversion and neuroticism. Users were interacting with the agent whilst playing the smile game, meant to enforce different facial expressions. A different set of annotations were collected, including the topic of the conversation with the agent, the facial expressions, and the turns. The paper is relevant to ARIA in the sense it helps understanding how users interact with agents and how these are being perceived.
Learning to combine local models for facial action unit detection	S. Jaiswal and B. Martinez and M. Valstar	B. Martinez and M. Valstar	IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015), FERA 2015 Challenge and Workshop	2015	Behaviour Analysis	*action untis, facial expression recognition	WP2, T2.1	This paper presents a simple neural network approach to facial action unit detection. Each of the network?s input neurons is assigned with the bin of the histogram of pixels computed at a specific predefined image region. Instead of using a fully connected network, that would need an exponential number of parameters, the network reduces the dimensionality locally and gets to local predictions, which are ultimately fused in a low-dimensional network. This makes the network sparse. The approach meant the first steps towards the ARIA?s visual system, which computes the user?s emotion in real time, thus requiring such a sparse representation
FERA 2015 ? Second Facial Expression Recognition and Analysis Challenge	M. F. Valstar and T. Almaev and J. M. Girard and G. McKeown and M. Mehu and L. Yin and M. Pantic and J. F. Cohn	M. Valstar	IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015), FERA 2015 Challenge and Workshop	2015	Behaviour Analysis	*action untis, facial expression recognition, FACS system	WP2, T2.1, T2.4	This paper presents the second challenge on facial expression recognition and analysis, in which participants are encouraged to submit their systems to be evaluated following a pre-defined protocol, using the new dataset BP4D, collected at Binghamton University. The goal of the challenge is to provide the community with a unified framework and benchmark for the evaluation of facial expression analysis systems. Such a framework helps defining the challenges and opportunities in the field. The broad acknowledge of this challenge indicates its successfulness. The paper is relevant to ARIA?s visual system, given that eMax is responsible of analysing users? emotions, and therefore it is important to have a deep understanding of the state of the art in facial expression recognition

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Topic-Based Personalization of Dialogues with a Virtual Coach	J. van Waterschoot and M. Theune	J. van Waterschoot and M. Theune	Proceedings of the workshop on Persuasive Embodied Agents for Behavior Change, at Intelligent Virtual Agents 2017	2017	Dialogue Management	dialogue management, personalization, topic management	Task 3.3 User-adaptive dialogue strategies	We present an approach for establishing a personal relationship between a human and a virtual agent by employing topic management in human-agent dialogues. We describe a data-driven method for determining topic recognition and topic transition strategies, and discuss how the personalized agent can be evaluated. Although originating from the ARIA VALUSPA project, in this paper these ideas are discussed in the context of a virtual coach application.
Topic recognition and management in conversational agents	J. van Waterschoot	J. van Waterschoot	Young Researchers? Roundtable on Spoken Dialog Systems 2017 (poster abstract)	2017	Dialogue Management	dialogue management, personalization, topic management	Task 3.3 User-adaptive dialogue strategies	This poster briefly discusses the turn-taking and interruption management strategies employed in ARIA-VALUSPA, as well as the ideas about topic management in ARIA agents. It focuses on the plans for designing data-driven transition strategies, describing how a Wizard-of-Oz corpus study could be used to find out how the agent can gracefully introduce new topics or steer towards topics that it wants to discuss, and how it can recognize the user's topics of interest.
Interacting with Virtual Agents in Shared Space: Single and Joint Effects of Gaze and Proxemics	J. Kolkmeier, J. Vroon and D.K.J. Heylen	D.K.J. Heylen	International Conference on Intelligent Virtual Agents (IVA) 2016	2016	Behaviour Generation	virtual humans, gaze, proxemics	Task 4.1 Overall dynamic non-verbal communicative behaviour model	In human-human interactions, gaze and proxemic behaviours work together in establishing and maintaining intimacy. In this study we examine how these behaviours affect the perceived personality of virtual agents in immersive Virtual Reality. Agents that exhibited more directed gaze and reduced interpersonal distance were attributed higher scores on intimacy related items than agents that exhibited averted gaze and increased interpersonal distance. These findings could inform the behaviour model of the ARIA agents.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Sequence-based Multimodal Behavior Modeling for Social Agents	Soumia Dermouche, Catherine Pelachaud	Soumia Dermouche, Catherine Pelachaud	ACM International Conference on Multimodal Interaction ICMI 2016	2016	Behaviour Generation	Virtual agent, interpersonal attitudes; non-verbal behavior, Temporal Sequence Mining	Task 4.2 Adaptive non-verbal communicative behaviour generation model	The goal of this work is to model a virtual character able to converse with different interpersonal attitudes. To build our model, we rely on the analysis of multimodal corpora of non-verbal behaviors. The interpretation of these behaviors depends on how they are sequenced (order) and distributed over time. To encompass the dynamics of non-verbal signals across both modalities and time, we make use of temporal sequence mining. Specifically, we propose a new algorithm for temporal sequence extraction. We apply our algorithm to extract temporal patterns of non-verbal behaviors expressing interpersonal attitudes from a corpus of job interviews. We demonstrate the efficiency of our algorithm in terms of significant accuracy improvement over the state-of-the-art algorithms.
Computational Model for Interpersonal Attitude Expression	Soumia Dermouche	Soumia Dermouche	ACM International Conference on Multimodal Interaction ICMI 2016, Doctoral Consortium	2016	Behaviour Generation	Virtual agent, interpersonal attitudes; non-verbal behavior, Temporal Sequence Mining	Task 4.2 Adaptive non-verbal communicative behaviour generation model	This paper presents a plan towards a computational model of interpersonal attitudes and its integration in an embodiedconversational agent (ECA). The goal is to endow an ECA with the capacity to express different interpersonal attitudes depending on the interaction context. Interpersonal attitudes can be represented by sequences of non-verbal behaviors. In our work, we rely on temporal sequence mining algorithms to extract, from a multimodal corpus, a set of temporal patterns representing interpersonal attitudes. Specifically, we propose a new temporal sequence mining algorithm called HCApriori and we evaluate it against four state-of-the-art algorithms. Results show a significant improvement of HCApriori over the other algorithms in terms of both pattern extraction accuracy and running time. The next step is to implement the temporal patterns extracted with HCApriori on an ECA.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Beat Gesture Prediction using Prosodic Features	Varun Jain, Chlo? Clavel and Catherine Pelachaud	Varun Jain, Catherine Pelachaud	3rd Workshop on Virtual Social Interaction VSI'17	2017	Behaviour Generation	virtual agent, gesture, prosody	Task 4.1 Overall dynamic non-verbal communicative behaviour model	In this work we present a machine learning approach to gesture prediction using prosodic features.We use conditional random fields to predict the presence of beat gestures using the following prosodic features: pitch, pitch-derivatives, intensity and absence or presence of syllable nuclei. These features are calculated over overlapping sliding windows big enough to average out the high frequency variations associated with pitch and intensity at the syllable level. We found that the results improve remarkably when the classification is treated as a multi-class problem as opposed to a binary problem with the two classes: presence and absence of gesture.
Automatic Measures to Characterise Verbal Alignment in Human-Agent Interaction	Dubuisson Duplessis, G.; Clavel, C.; Landragin, F.,	Dubuisson Duplessis, G.	18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)	2017	dialogue	verbal alignement	Task 3.3 User-adaptive dialogue strategies	This work aims at characterising verbal alignment processes for improving virtual agent communicative capabilities. We propose computationally inexpensive measures of verbal alignment based on expression repetition in dyadic textual dialogues. Using these measures, we present a contrastive study between Human-Human and Human-Agent dialogues on a negotiation task. We exhibit quantitative differences in the strength and orientation of verbal alignment showing the ability of our approach to characterise important aspects of verbal alignment.
AWeb-Based Platform for Annotating Sentiment-Related Phenomena in Human-Agent Conversations,	Langlet, C.; Dubuisson Duplessis, G.; Clavel, C.	Dubuisson Duplessis, G.	17th International Conference on Intelligent Virtual Agents (IVA 2017)	2017	dialogue	verbal content annotation, virtual agent, sentiment analysis	Task 3.3 User-adaptive dialogue strategies	This paper introduces a web-based platform dedicated to the annotation of sentiment-related phenomena in human-agent conversations. The platform focuses on verbal content and deliberately sets aside non-verbal features. It is designed for managing two dialogue features: adjacency pair and conversation progression. Two annotation tasks are considered: (i) the detection of sentiment expressions, (ii) the ranking of user?s preferences. These two tasks focus on a set of specific targets. With this demonstration, we aim to introduce this platform to a large scientific audience and to get feedback for future improvements. Our long-term goal is to make the platform available as open-source tool.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
From analysis to modeling of engagement as sequences of multimodal behaviors	Soumia Dermouche, Catherine Pelachaud	Soumia Dermouche, Catherine Pelachaud	The 11th International Conference on Language Resources and Evaluation: LREC	2018		Engagement, Non-verbal behavior; ECA, Human-agent interaction	WP6	In this work, two types of manual annotation of NoXi corpus were conducted: non-verbal signals such as gestures, head movements and smiles; engagement level of both expert and novice during the interaction. Then, we used a temporal sequence mining algorithm to extract non-verbal sequences eliciting variation of engagement perception. Our aim is to apply these findings in human-agent interaction to analyze user's engagement level and to control agent's behavior. The novelty of this study is to consider explicitly engagement as sequence of multimodal behaviors.
Expert-Novice Interaction: Annotation and Analysis	Soumia Dermouche, Catherine Pelachaud	Soumia Dermouche, Catherine Pelachaud	MULTIMODAL CORPORA 2018: Multimodal Data in the Online World, MMC	2018			WP6	In this demonstration, we present the NoXi corpus of expert-novice interactions, our annotations and analysis. To analyze the data we apply HCApriori, a Temporal Sequence Mining algorithm to extract relevant behavior sequences for both expert and novice. NoXi provides over 25 hours of dyadic interactions recorded in different languages, mainly English, French, and German. The annotation tool NOVA allows annotating data using discrete and continuous schema. We use NOVA to manually annotate behaviors (discrete annotation) and engagement levels (continuous annotation).
Social Context Disambiguates the Interpretation of Laughter.	Curran, William; McKeown, Gary; Rychlowska, Magdalena; André, Elisabeth; Wagner, Johannes; Lingenfelser, Florian.	Elisabeth André, Johannes Wagner	Frontiers in Psychology, Vol. 8, 2342	2018	Behaviour Analysis	Laughter, Multimodal Corpus	WP6	Despite being a pan-cultural phenomenon, laughter is arguably the least understood behaviour deployed in social interaction. As well as being a response to humour, it has other important functions including promoting social affiliation, developing cooperation and regulating competitive behaviours. Understanding these functions can lead to a better contextual interpretation of laughter during a conversation, e.g. during the interaction with agent (like in ARIA).
Real-time Sensing of Affect and Social Signals in a Multimodal Framework: a Practical Approach,	Johannes Wagner, Elisabeth André	Johannes Wagner, Elisabeth André	The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition	2018	Behaviour Analysis	Social Signal Processing, Real-time recognition, Multimodal fusion	WP6	The most promising way to encourage developers to put more effort into building online systems is by providing adequate tools that take as much work off their hands as possible. In this book chapter we present the open-source framework SSI, which has been called to life for this very purpose. The ARIA recognition system is implemented with SSI.

Title	All authors	Authors funded by ARIA	Journal / proceedings	Year	Category	Keywords	Related Task	100-word justification of relevance
Modeling User?s Social Attitude in a Conversational System	Tobias Baur, Dominik Schiller, Elisabeth André	Tobias Baur, Elisabeth André	Emotions and Personality in Personalized Services pp 181-199	2017	Behaviour Analysis	User Modelling, Engagement	2.3	In this article we describe an approach for modelling Social Attitudes, such as the Engagement of a User in an Interaction with a Conversational System, such as the ARIA-Valuspa virtual humans
Applying Cooperative Machine Learning to Speed Up the Annotation of Social Signals in Large Multi-modal Corpora	Johannes Wagner, Tobias Baur, Yue Zhang, Michel F. Valstar, Björn Schuller, Elisabeth André	Johannes Wagner, Tobias Baur, Yue Zhang, Michel F. Valstar, Björn Schuller, Elisabeth André	https://arxiv.org/abs/1802.02565	2018	Behaviour Analysis	Annotation, Cooperative Learning, Interactive Learning, Data Collection	6.2, 6.3	In this paper we describe the novel cooperative machine learning approach, developed in the ARIA-Valuspa Project. We shortly introduce the NoXI Database and the NOVA tool, and describe an evaluation of the approach on the NOXI corpus
Context-Aware Automated Analysis and Annotation of Social Human–Agent Interactions	Tobias Baur, Gregor Mehlmann, Ionut Damian, Florian Lingenfelser, Johannes Wagner, Birgit Lugrin, Elisabeth André, Patrick Gebhard	Tobias Baur, Johannes Wagner, Elisabeth André	ACM Transactions on Interactive Intelligent Systems (TiiS) - Special Issue on Behavior Understanding for Arts and Entertainment (Part 1 of 2) Volume 5 Issue 2	2015	Behaviour Analysis	Annotation, User Modelling, Engagement	6.2, 2.3	In this journal article we introduce of the NOVA tool in earlier stages and of the user model to detect enaggement. We examplified this with a virtual agent