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

(July, 2017)

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

Collaborative Project

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

D4.3: Final adaptive audiovisual communicative behaviour generation

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

(July, 2017)

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

STATUS: [DRAFT]

Deliverable Nature		
R	Report	
P	Prototype	
D	Demonstrator	**X**
O	Other	

Participant Number	**Participant Name**	**Acronym**	**Country**
1	University of Nottingham (coordinator)	UON	U.K.
2	Imperial College of Science, Technology and Medicine	ICL	U.K.
3	Centre National de la Recherche Scientifique (WP Leader)	CNRS	France
4	Universitat Augsburg	UA	Germany
5	Universiteit Twente	UT	The Netherlands
6	Cereproc LTD	CPRC	U.K.
7	La Cantoche Production SA	CNT	France

(July, 2017)

Table of Contents

(July, 2017)

1. Summary

This is the last demonstrator for Work Package 4: *"Final adaptive audio-visual communicative behaviour generation"*. The leader of this work package is CNRS, with involvement from the following partners: UON, ICL, UA, UT, CPRC and CNT.

The objective of this work package (WP 4) is to model expressive acoustic and visual behaviours for the ARIA agent. In particular, the output of this work package endows the ARIA agent with social interaction capabilities with emphasis on socio-emotional expression of stances towards its interlocutor and adaptation to unexpected events during the interaction such as interruptions. In this deliverable we describe the latest progress of tasks terminating at month 31 (T4.2, T4.4) and still ongoing terminating at month 36 (T4.3, T4.5 and T4.6).

In Task 4.2 (months 13-31, CNRS, CNT), Adaptive nonverbal communicative behaviour generation model, the goal is to create a computational model that generates the ARIA's agent nonverbal behaviour for expressing the agent's communicative intentions and social-emotional stances (e.g. interpersonal attitudes).

In Task 4.3 (months 18-36, CNRS, CNT), Emergence of synchrony during engagement phases between ECA and User, the aim is to put in synchrony the agent with the user through its nonverbal behaviour display reflecting its shared understanding and engagement level with the user.

In Task 4.4 (months 1-31, CPRC), Adaptive speech synthesis, the objective is to synthesise expressive and conversational speech with particular emphasis on reactivity and dealing with interruptions when the user starts speaking.

In Task 4.5 (months 25-36, CNRS, UON, TUM, CNT), Synthesis-Analysis feedback loops, the goal is to integrate feedback loops between the audio and video analysis and the speech and visual synthesis components of the ARIA-VALUSPA system.

In Task 4.6 (months 18-36, CNRS), Multimodal behaviour response model to unexpected situations, interruption of ongoing behaviour and related communicative intents are considered. At intentional level, the co-articulation of behaviour into another one (as the current intention is followed by the instantiation of a new intention) is considered; at behavioural level the merge of behaviour with another one should be produced.

This deliverable is structured as follows: Section 2 details the contributions of each partner and the progress made according to each task. Section 3 describes the outputs (e.g. publications, demos). Section 4 presents a conclusive summary of WP4.

2. Progress

2.1 Overall dynamic non-verbal communicative behaviour model (Task 4.1)

The overall nonverbal communicative behaviour model in ARIA-VALUSPA has been largely described in D4.1 and D4.2. However, we show the ARIA system architecture here in this deliverable in order to facilitate the reading throughout the remaining sections.

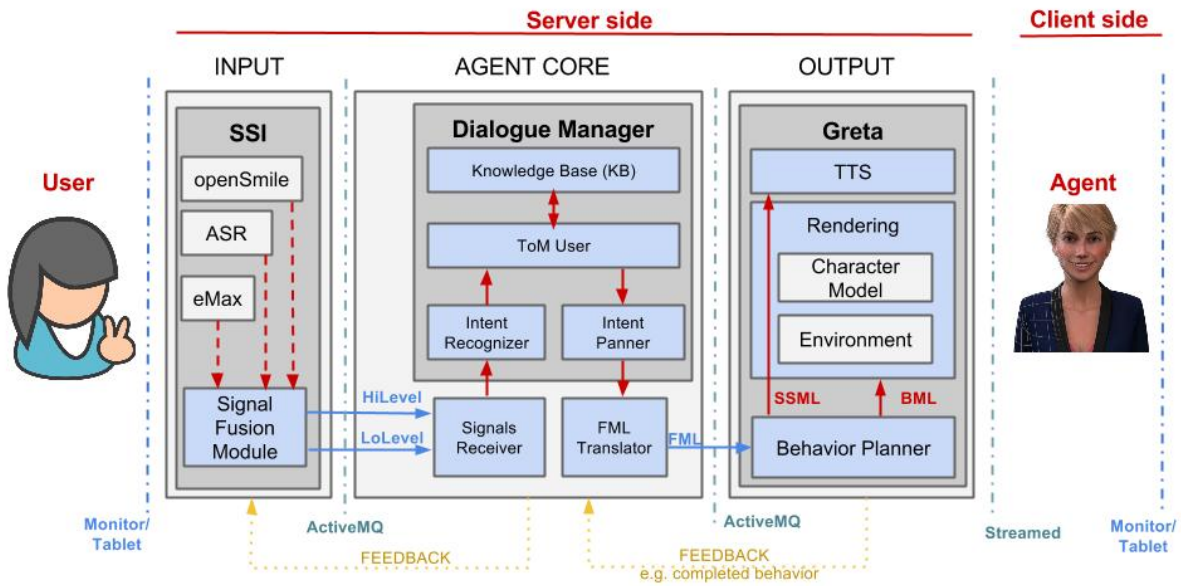


Figure 1: The latest ARIA system architecture.

2.2 Adaptive nonverbal communicative behaviour generation model (Task 4.2)

2.2.1 Updated FML Templates for the new Dialogue Manager

The new DM is currently able to handle Question-Answering and uses a template for sending this to the Behaviour Planner. Furthermore, we're adapting the dialogue templates to be compatible with the new features of FML templates described below. UTwente is currently developing the following features for more adaptive behaviour generation:

- Mapping user utterances independently of language to intents with the query-answer matcher.
- Dealing with interruptions by using goal markers (i.e. Important time markers) in the FML together with BML-feedback containing how much the agent has said (time mark of the interruption). We use these goal markers for marking the FML with information that is important and adapt our interruption strategy accordingly.

- Adapting the utterances of the agent by using words preferred by the user (found from the dialogue history) and by making them more positive or negative based on the valence and arousal of the user. More details are provided in Section

- 2.3 Emergence of synchrony between ECA and User (Task 4.3).
- Providing ARIA-Greta information about the agent's current interaction-state for displaying engaging behaviours. The interaction-state is updated by using information such as user's presence and voice activity. CNRS provided more details about the interaction states and the respective behaviours exhibited by the agent in each state are described in Section 2.5 Synthesis-Analysis feedback loops (Task 4.5).

Further details on these implementations are described in D3.3. The work package will be available in AVP 2.4 in September, with instructions and details described on the wiki page on GitHub.

CNRS worked on adapting the totality of FML templates described in D4.1 and D4.2 to the newest features offered by CereEngine (i.e. the TTS engine used in the ARIA system developed by Cereproc). Most importantly, as also described in 2.4.1 Speech Interruption API implementation, all FML Templates now support the <voice> tag and its emotion attribute that is used to set the emotional tone to the generated speech. In this way, the visual nonverbal behaviours can be coherently exhibited with the appropriate emotional tone when, for instance, the ARIA agent expresses specific emotions such as anger or joy. The full set of FML templates has been made available in the public release of ARIA as well[1].

Moreover, the new FML Translator it now supports the GOAL markers, a new mechanism that the Dialogue Manager can use in order to emphasize a specific portion of an utterance in a given FML Template. More specifically, special temporal markers (TMs) can be placed within the utterance. Those TMs, when present and detected by the FML Translator support the production of multimodal behaviour when transforming the given FML to BML with specific communicative intentions (e.g. pitch accents, emphasis) as usual, but they restrict the focus on a particular portion of the utterance as illustrated in the FML example below (we name the special TM as DMImportantBegin/End for clarity but a more concise name can be used in the implementation).

```
<fml>
  ...[other FML tags]...
  <speech id="speech1">
  Hello <tm id="DMImportantBegin"/> my name is Alice<tm id="DMImportantEnd"/> and I am a virtual agent.
  </speech>
  ...[other FML tags]...
  <pitchaccent level="moderate" start=" DMImportantBegin" end=" DMImportantEnd" />
  ...[other FML tags]...
</fml>
```

This allows the DM to dynamically add those markers within utterance (or ask the users to manually annotate important parts of the utterances via authoring tools). The FML translator automatically falls back to a safe translation when, for instance, a

[1] https://github.com/ARIA-VALUSPA/ARIA-System/tree/master/Agent-Core-New/data/fmltemplates

communicative intention present in the template (e.g. pitch-accent) refers to the important TMs but those are not found in the utterance coming from the Dialogue Manager when dynamically filling in the FML Template with the required parameters (i.e. the utterance, etc.). For pitch-accents, the tag in the FML template is completely removed in case special GOAL time markers have not been included, whereas for other intentions (e.g. performatives) the eventual start and end attributes of those intentions referring to special GOAL markers are automatically fixed to refer to the whole utterance instead (i.e. start and end of the utterance). In the scenario of a user interrupting the ARIA agent, a goal marker, as shown in the example above, might be used by the Dialogue Manager to grab more information about the progress, in terms of semantic content, of what being generated by the ARIA agent. Therefore, the dialogue manager can be updated about whether the important part of the communicative intent (i.e. saying the agent's name) has been reached or not when the interruption occurs.

Finally, three new important adaptive features have been investigated and added in the ARIA system: interaction states and language (described in 2.5 Synthesis-Analysis feedback loops (Task 4.5)), and interpersonal attitudes (described below).

2.2.2 Expressing interpersonal attitudes

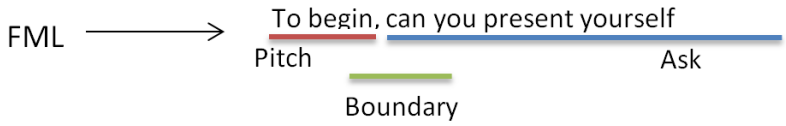
CNRS continued the work for making the ARIA agent capable of expressing different interpersonal attitudes, for example dominant or hostile, toward the user. We applied HCApriori , a temporal sequence mining technique, to extract from a multimodal corpus the most relevant temporal patterns (non-verbal sequences) representing four attitude variations: dominance increase, dominance decrease, friendliness increase and friendliness decrease.

In order to evaluate the extracted patterns, an empirical study relying on **Crowdflower**[2] was carried out to investigate whether non-verbal patterns extracted with our model for a given attitude are perceived as conveying the same attitude. The results indicated that patterns expressing dominance increase are evaluated as more dominant, more hostile and less friendly compared to a neutral behaviour. The patterns expressing dominance decrease are evaluated as more submissive and friendlier compared to the neutral behaviour. Patterns expressing friendliness decrease are evaluated as more hostile and more dominant compared to the neutral behaviour. However, patterns expressing friendliness increase are perceived as equivalent to the neutral expression.

Finally, 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 the attitudes variation. The Sequential Attitude Planner takes as input an FML file (as produced by the Dialogue Manager through the FML Translator depicted in Figure 1: The latest ARIA system architecture.Figure 1: The latest ARIA system architecture.) and the attitude variation that the agent will express toward the user. Figure 2: Example FML for the Sequential Attitude Planner describes a schematic view of an example FML

[2] https://www.crowdflower.com/

file where the utterance to be produced is: "*To begin, can you present yourself?".* The corresponding FML script includes three communicative intentions: a pitch-accent on the first part of the utterance, a boundary part around the punctuation marker (i.e. comma) and a set-question.



1. FML-to-BML sequence generation	head face gesture face gaze
2. BML Attitude sequence selection	posture head gesture face arms-crossed
3. BML sequence enrichment	posture head face gesture face gaze arms-crossed
4. Priority signals selection	posture head face gesture Eyebrows_Frown gaze arms-crossed

Figure 2: Example FML for the Sequential Attitude Planner

The proposed attitude planner works in four steps:

1. **From FML to BML Sequence generation**: the algorithm starts by generating a sequence of non-verbal signals that accomplish the communicative intentions in the FML. For example, the communicative intention "set-question" is expressed with a gesture represented in BML.
2. **BML Attitude sequence selection**: from a dataset of behaviours sequences expressing attitude variations (in the example, dominance increase), the algorithm selects the sequence (e.g. posture head gesture face arms-crossed) 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 (e.g. posture and arms-crossed) 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 each signal, for example "S1", in the final behaviour sequence with a mapped signal, say "S2", expressing a specific interpersonal attitude, if the probability of "S2" is higher than the probability of "S1" for expressing that attitude. In our example, the signal "face" is replaced by "eyebrows frown".

2.2.4 The two faces of ARIA with Greta and Living Actor technology

The Living Actor rendering solution has been merged to the ARIA-Framework with a native ActiveMQ interface. Some key algorithms of the Living Actor component are based on old libraries some of which are no longer maintained. Updating these libraries is a long-term task which has been initiated by this project. The ARIA-Framework uses a recent version of Active MQ. Despite the efforts of the developers of the consortium, it was not possible to compile this recent version in the Living Actor 3D component due to the technical debt of it. To bypass this issue, an intermediate tool – a connector - has been developed to manage the connection between the ARIA-Framework and the rendering component. The first version of this connector was linked to the Framework through Active MQ and wrote the BML query in a file on the hard drive of the system. Then, the 3D component read this file in order to play the BML animation. This quick and dirty solution was good enough for the first integration tests but not acceptable for the project due to poor response time due to the management of the filesystem. A new version of this connector has been included in the latest version of the platform. The connector is now tightly connected to the Living Actor 3D component by JMUI, this interface allows instant and bi-directional exchanges.

This integration considerably reduces the latency between the processing of the inputs and the restitution of the answer. The rendering of the animation generated by the dialog system is now running with almost no latency. It was a necessary step for the implementation of interruptions with the Living Actor avatars. It was not possible to generate smooth and relevant interruptions while the response time of the agent includes latency. This integration has also unlocked the feedback system; Living Actor is now able to notify the whole system of its current status. The 3D component now exchanges information in real time to inform when the BML has been played and that the agent is ready to play another animation, but also when an animation has been interrupted by another command. These feedbacks are used for example by the framework to manage the agent when he's listening to include backchannel animations. It still uses the specific lexicon created for D4.2 to match the available behaviours for the Alice avatar.

The Living Actor 3D player has now two output options: rendering in a viewer window on the computer where the software was launched or streaming the 3D render to a webpage.

2.3 Emergence of synchrony between ECA and User (Task 4.3)

2.3.1 Speech Synthesis

It has been observed that humans develop a synchrony during dialog interactions: they will adapt their behaviour and emotional response according to the conversation, but also adapt their speech rate, style and even accent to be more similar to their interlocutor; this behaviour is referred to as alignment. The question of whether it would be appropriate or desirable for an ECA to perform the same type of alignment is an open question, but as a research tool the ARIA framework should provide the ability to develop such an alignment. As the main driver for the agent behaviour generation, the speech synthesiser should therefore offer ways of controlling the speech rate and style.

CereProc's SDK[3] (and many other speech synthesizers) have been offering this type of control for many years, but thanks to the research and development conducted during this project their expressivity has been considerably improved.

Emotional adaptation

One of the unique offering of CereProc's technology is the rich expressivity of its voices, and in particular the ability to the emotions being expressed through simple xml tags:

```
<speech emotion='happy'>Hello world.</speech>
<speech emotion='sad'>Hello world.</speech>
```

Thanks to advanced vocoding and machine learning techniques, CereProc's SDK will soon be able to provide the ability to control the emotion at an even finer level, e.g.:

```
<speech level='0.5' emotion='happy'>I am a bit happy.</speech>
<speech level='2' emotion='sad'>I am very sad!</speech>
```

As a consequence, agent built with the ARIA-VALUSPA framework will have the ability to adapt the emotional state of the agent to unprecedented levels, allowing expressing gradual or very subtle changes in their emotional state.

Speech adaptation

Provided the agent is capable of measuring the speaking rate of the user reliably, there is also a way to adapt the speaking rate of the agent through the use of standard SSML tags, e.g.:

```
<prosody rate='1.1'>I am speaking a bit fast.</prosody>
<prosody rate='0.8'>I am now speaking very slowly.</prosody>
```

[3] M. P. Aylett and C. J. Pidcock. 2007. The CereVoice Characterful Speech Synthesiser SDK. In Proceedings of the 7th international conference on Intelligent Virtual Agents (IVA '07). Springer-Verlag, Berlin, Heidelberg, 413-414.

2.3.2 Verbal Alignment

Convergence of behaviour is an important feature of Human-Human (H-H) interaction that occurs both at low-level (e.g., body postures, accent and speech rate, word choice, repetitions) and at high-level (e.g., mental, emotional and cognitive level)[4]. In particular, dialogue participants automatically align their communicative behaviour at different linguistic levels including the lexical, syntactic and semantic ones[5]. A key ability in dialogue is to be able to align (or not) to show a convergent, engaged behaviour or at the opposite a divergent one. Such convergent behaviour may facilitate successful task-oriented dialogues[6] [7].

CNRS' goal is to provide a virtual agent with the ability to detect the alignment behaviour of its human interlocutor, as well as the ability to align with the user to enhance its believability, to increase interaction naturalness and to maintain user's engagement.

We aim at providing measures characterising verbal alignment processes based on repetitions between DPs. To this end, we have 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.

These measures make it possible to quantitatively characterise the strength and orientation of verbal alignment between DPs in a task-oriented dialogue. A promising perspective of this work lies in the exploitation of these measures to adapt and align the verbal communicative behaviour of a virtual agent.

2.3.3 Nonverbal Engagement

CNRS is currently manually annotating, using the NOVA tool developed for the ARIA project by UAugsburg, the non-verbal behaviour of both expert and novice for the French sessions of the NoXi database[8] (tot. 20 sessions). We have also added continuous annotations of engagement for both expert and novice. We have defined 5 options to annotate changes in the perception of engagement: strongly disengaged, partially disengaged, neutral, partially engaged, strongly engaged. For audio modality,

[4] Gallois C., Ogay T. and Giles H.H. 2005. *Communication accommodation theory: A look back and a look ahead.* W. Gudykunst (ed.): Theorizing about intercultural communication. Thousand Oaks, CA: Sage pages 121–148.
[5] Pickering M. J. and Garrod S. 2004. *Toward a mechanistic psychology of dialogue*. Behavioral and brain sciences 27(02):169–190.
[6] Nenkova A., Gravano A. and Hirschberg J. 2008. *High frequency word entrainment in spoken dialogue*. In Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies (ACL-HLT): Short papers. Association for Computational Linguistics, pages 169–172.
[7] Friedberg H., Litman D. and Paletz S. BF. 2012. *Lexical entrainment and success in student engineering groups*. In Spoken Language Technology Workshop (SLT). IEEE, pages 404–409.
[8] https://noxi.aria-agent.eu/

we extracted, using Praat[9], several prosodic and acoustic features such as pitch (i.e. the quality of a sound represented by the rate of vibrations), pause duration (i.e. the duration of silence time in an audio segment) and signal intensity. Table 1 summarizes the multimodal behaviors that we are annotating and the number of annotated sessions. The goal is to study the synchrony between the expert and the novice in terms of non-verbal behaviors and engagement variation.

Modality	**Labels**	**Number of annotated sessions**
Head movements	NOD — SHAKE	5
Head direction	FORWARD — BACK — UPWARDS — DOWNWARDS — SIDEWAYS — SIDE TILT	5
Eyebrow movements	RAISED — FROWN	5
Smiles	SMILE	5
Gaze direction	TOWARDS INTERLOCUTOR — UP — DOWN — SIDEWAYS — OTHER	5
Gestures	ICONIC — METAPHORIC — DEICTIC — BEAT — ADAPTOR	10
Hand rest positions	ARMS CROSSED — HANDS TOGETHER — HANDS IN POCKETS — HANDS BEHIND BACK — AKIMBO	20
Engagement	STRONGLY DISENGAGED … STRONGLY ENGAGED	20
Audio	PAUSE DURATION — PITCH — SIGNAL INTENSITY- TURN TAKING	20

Table 1: Annotation scheme for the multimodal behaviors and engagement annotations in NoXi.

[9] Paul Boersma. 2001. Praat, a System for Doing Phonetics by Computer. Glot International (2001), 341–345.

(July, 2017)

2.4 Adaptive speech synthesis (Task 4.4)

2.4.1 Speech Interruption API implementation

Being able to react in a realistic manner is a key requirement for an interactive speech synthesis system. There are situations when the output of a speech synthesis system (virtual agent) needs to be interrupted, for example, when a noise event occurs (banging door, passing train, or other environmental noises) that make it unlikely that the user will be able to understand the synthesis. Additionally, there will be times when the user wants to interrupt the system. While previous 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 at which point the system is expected to react, ideally in a convincingly authentic manner.

There is an implicit assumption that reactive synthesis has to be incremental. 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.

In contrast, there are clear examples where incremental systems are required for example synthesising typed text as the user is typing it[10] for performative synthesis where the interruption rate can be seen as being continuous and focused on spectral and duration properties rather than linguistic context[11]. In this work, we are focused on producing a system that can allow a conversational agent the ability to react rapidly and naturally to external interruptions and stimuli. Therefore, we present a system that incorporates re-planning as a strategy for dealing with user-initiated interruptions.

We created a demonstration of `Reactive Synthesis' for the British Science Museum Lates in London (March 2017) as an educational tool to provide a general introduction to speech technology. Prior to the presentation at the museum, we were interested in obtaining feedback from the general public as to their opinions and attitudes to speech synthesis in general and to the demo more specifically. The approach we took in this work was to evaluate by means of a focus group.

[10] Maël Pouget, Thomas Hueber, Gérard Bailly, Timo Baumann. HMM Training Strategy for Incremental Speech Synthesis. 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Dresden, Germany, pp.1201-1205.

[11] Astrinaki, M., d'Alessandro, N., Picart, B., Drugman, T., & Dutoit, T. (2012, December). Reactive and continuous control of HMM-based speech synthesis. In Spoken language technology workshop (SLT), 2012 IEEE (pp. 252-257). IEEE.

System Description

The re-planning and splicing approach is as follows: given a required latency, e.g. 200ms, the system must operate fast enough to re-synthesise 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 current chunk. The new audio can then be seamlessly re-splicing 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 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: 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 tr - t0 must be larger than the maximum system latency..

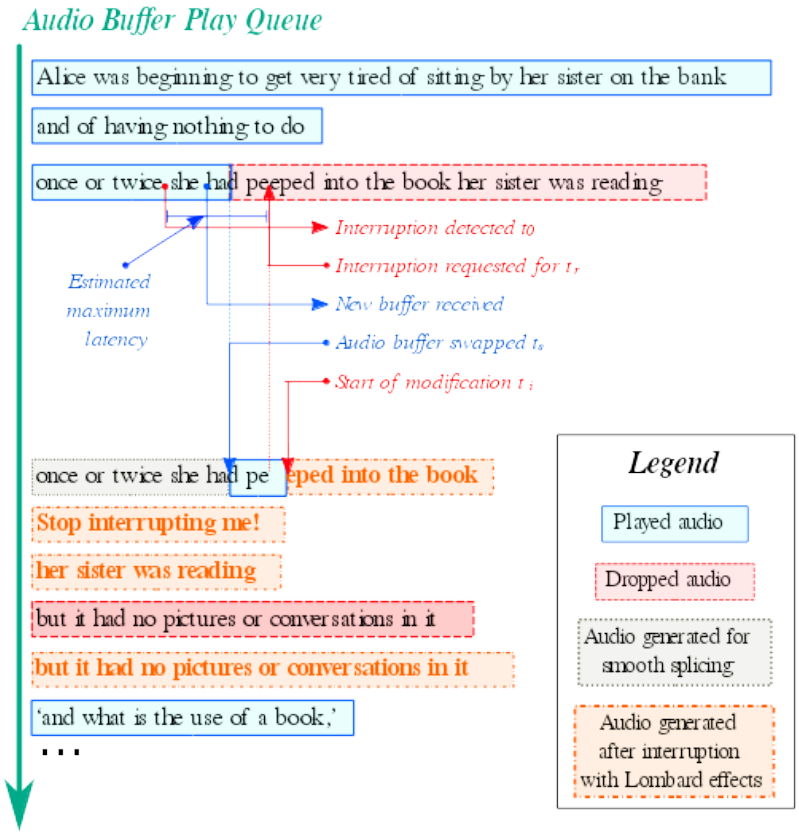


Figure 3: Example of the use of the interruption API, showing the changes in audio buffers. Final played audio is in blue and orange boxes, red and grey boxes are dropped. Note that tr - t0 must be larger than the maximum system latency.

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 re-synthesised differently, or empty the queue entirely, depending on the application.

Demo set-up

As the demonstrator was intended for a science exhibit (the British Science Museum Lates), and more specifically to provide a general introduction to speech technology to the general public, the voice of the system was chosen for its clarity and apparent patience. The text to be read is a general introduction about speech synthesis, which is available online[12].

[12] http://derstandard.at/1325485448039/Talking-Technology

(July, 2017)

It is well known that combining speech synthesis with a strong visual modality has the side effect of reducing the overall impact of the audio modality. To maximise the impact of the speech synthesis, the visual aspects of the system were thus kept to a minimum, and only displays the text and the current position within it, and a basic interface reflecting its internal state (see Figure 4).

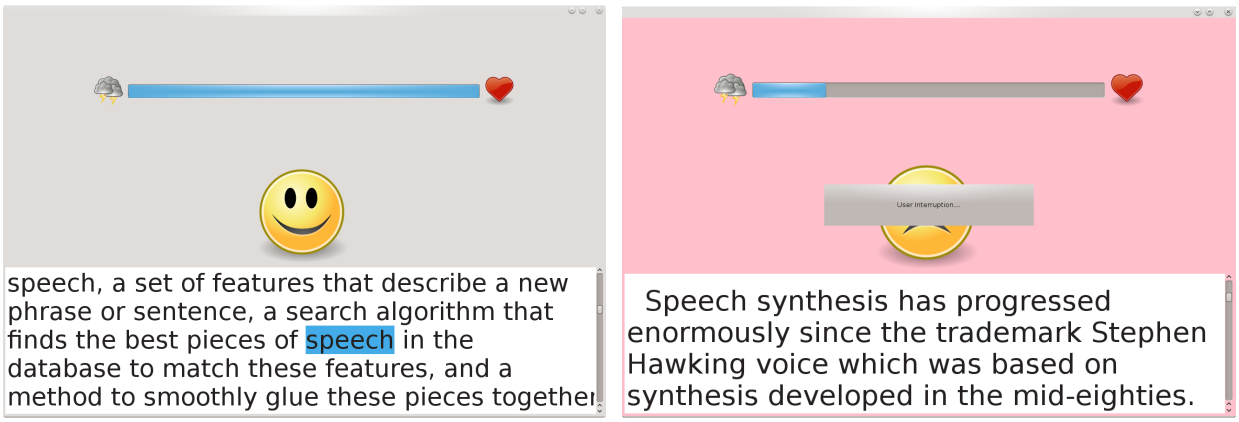


Figure 4: Visualisation of the system in action. On the left, the bar and the emoticon indicate the system's mood. The text being read is displayed at the bottom, with the current word being read highlighted. On the right, when an interruption occurs, the background colour is modified and an explicit dialogue window is displayed.

(July, 2017)

Rather than an exhaustive study of the whole range of possible reactions to interruptions, we selected a small subset in order to simulate three distinct system reaction styles.

- First, a baseline system that simply did not react whatsoever when the user attempted to interrupt, it simply continued reading.
- The second system demonstrated a basic reactive system: it detected user interruptions and stopped its speech gracefully, almost instantly. It then waited for the user to stop speaking before resuming the reading task.
- The third system was intended to demonstrate a more complex behaviour of the system: it initially reacted gracefully to users' interruptions, but as the interactions continued, the type of reactions evolved to demonstrate an increased level of irritation of the system when interrupted. After about 10 interruptions, the system simply decided it had had enough and it would leave the interaction.

For the demo, Voice Activation only was used to detect interruptions. No linguistic or recognition processing was carried out. More specifically, the system used a Voice Activity Detection (VAD) module[13] to detect speech-based interruptions from the user. The system did not distinguish between speech feedback from the user or actual voluntary interruptions, considering any event that triggered an active state of the VAD module as an interruption.

When an interruption was detected, a reaction was generated, comprising 5 stages:

1. Modify the currently playing phrase to surrender the floor to the user as described above; normally stopping at a natural point or tailing off politely (maximum latency is 50ms).
2. Pause to let the user speak; the system would stay silent until a silence of at least 200ms long was detected.
3. Generate a reaction from the system about the interruption that had just happened. At the beginning of the interaction, reactions were rather polite and invited the user to carry on speaking, but as the interaction progressed the reactions turned into vocal gestures and in the final phase clearly irritated remarks.
4. Optionally pause to let the user express him/herself; the duration of the pause becomes shorter as the system becomes more irritated.
5. Resume the reading task.

In practice, all three systems were built the same way, simply disabling some of the functionalities of the "full"' system:

- The first system had the VAD module disabled.
- The second system had its mood change disabled, and always performed the most polite of reactions.

[13] https://github.com/dpirch/libfvad

Evaluation

There is no established approach to evaluating the effectiveness of reactive synthesis. Many studies look at the quality or naturalness of the output synthetic speech after applying incremental processing compared to a standard non-incremental version of their system HMM training strategy for incremental speech synthesis[14]. However, in such cases the responsiveness of the synthetic speech is not considered in the evaluation.

Rather than try and obtain judgements on naturalness and quality of the synthetic speech we were interested in obtaining feedback from the general public on a higher level regarding their opinions and attitudes to speech synthesis in general and to the `reactive synthesis' demo created for educational purposes more specifically. The approach we took to evaluation was by means of a focus group.

Focus Group

In setting up the focus group we followed the Guidelines for Conducting a Focus Group[15]. In addition to the `Reactive Synthesis' demo there was another demo `Bot or Not' which had also been designed for the Science Museum Lates. All the various stages in the focus group meeting are described below. However in terms of feedback we only present responses relevant to the `Reactive Synthesis' demo. Six people (3M/3F) were selected to take part in the focus group. They were recruited through the Edinburgh University Careers Service MyCareerHub. The only requirements were that the participants were able to speak English to a native level and were able to attend the meeting.

Two facilitators ran the focus group meeting. The meeting included the following stages:

- Welcome to participants – filling in of consent forms.
- Playing with interactive (speech technology related) toys.
- Filling in of a general speech synthesis questionnaire.
- The first demo `Bot or Not' was carried out in pairs .
- Group conversation discussing the interactive toys and `Bot or Not'.
- `Reactive Synthesis' demo was run in three different ways:
  1. System talked -- no reaction to interruptions at all.
  2. System stopped talking when interrupted -- then continued.
  3. System stopped talking and reacted every time it was interrupted -- reactions started polite and ramped up a step each interruption until very irritated.
- Group conversation discussing the reactive synthesis demo.

[14] Maël Pouget, Thomas Hueber, Gérard Bailly, Timo Baumann. HMM Training Strategy for Incremental Speech Synthesis. 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Dresden, Germany, pp.1201-1205.

[15] Wolff B., John E. Knodel and Werasit Sittitrai. "Focus Groups and Surveys as Complementary Research Methods: Examples from a Study of the Consequences of Family Size in Thailand." PSC Research Report No. 91-213. 5, 1991.

- End of focus group meeting - subjects thanked and remunerated for their time and effort.

Feedback on reactive synthesis demo

The goal of the focus group meeting was to get a more general, higher level of input from people who may have previously not considered speech synthesis. One of the first striking elements that became clear during the discussions and when considering the responses of the participants (P) to the speech synthesis questionnaire was that speech recognition and speech understanding were seen as part and parcel of speech synthesis, for example answers to the questions ``Do you use synthetic speech? If no, why not?'' included:

P3:``*I'd be worried I'd find because I talk fast and with a broad Scottish accent, I'd not get understood."*

P4: ``*I find it's too laborious and time-consuming. You have to repeat yourself over and over – irritating.''*

The gist of the feedback that was given after the reactive synthesis had been demonstrated was that overall participants preferred the reactive mode over the non-reactive. And of the two reactive modes they preferred the more simple approach, in which the system stopped when interrupted and when it resumed it went back and repeated the spurt[16] it was interrupted in. Furthermore, there was agreement that emotion is not the point of synthetic speech and that strong emotions get in the way of efficiency. A few of the participants' quotes that illustrate the above summary:

P3: ``*The second seemed the most appropriate for most situations, given that the third one is – it's great that it's maybe more emotion, but it's often not going to react in exactly the right way, and getting that is so difficult''*

P1: ``*I agree that the second version is best, cause it's reactive but it doesn't get in the way.''*

P4: ``*I quite like the other one actually (the third one), the fact that it kind of spoke back with you. Maybe, like, it being a bit sassy was a bit – not appropriate at the time, but I think I'd quite like that to have on my phone. ''*

P6: ``*If it's personable it's fine, but if it's angry then probably not.''*

16 A spurt is a section of speech with silence before and after it. The silence can be very short (as little as 20ms).

Some of the participants also mentioned that they had been concentrating more on how the synthesis sounded in the demo rather than paying attention to the interface and the interruptions.

P3: ``*cause I was listening to the voice and him getting more annoyed, I didn't notice that there was a bar of increasingly, like, one end was thunderstorm and one was happy, right?*''

P1: ``*I didn't notice the face or anything because I was trying to work out, well what is good and what's bad about this voice.*''

Furthermore, they mentioned they had never given reactive speech synthesis any thought nor ever dealt with it.

P1: ``*I've probably never dealt with reactionary stuff before so perhaps the fact that it's stopping, I don't know quite how to judge what is good and what is bad*''

Finally, the focus group liked the idea of synthesis as an aid in human-human conversation rather than an interference.

P5: ``*It could be used in a lecture type setting.*''

P2: ``*like it could work with a lecturer or somebody as opposed to in place of them. Yeah I think that's a good idea.*''

Summary

The focus group gave some useful feedback regarding reactive synthesis, and a renewed perspective of what a group of people from the general public think about speech synthesis. In our focus group, participants had not previously come across reactive speech synthesis. Their first reaction –objecting to the system being annoyed– makes sense as they assumed the system should be cooperative. However, having annoyed or obstructive agents has a clear use case in a virtual environment for games or training purposes. These are types of use cases that people have yet to really come across.

We have discussed the relationship between incremental processing, re-planning and splicing. It should be clear that we consider reactiveness to depend on being able to re-plan the output of a synthesis system and splice it in at the right time. Incremental speech synthesis can be seen as a way of speeding up processing, but is not sufficient to achieve a reactive system. This leads to the following design guidelines for a reactive speech synthesis system:

1. **Fast enough.** Whatever the chunk is (utterance, phrase, etc.) the system must be able to synthesise replacement chunk within the required latency (we would suggest 200ms as a minimum). For example, in our system where the chunk is a spurt (or intonational phrase) and 95% are less than 2s long, then the system has to be at least 10x real-time to achieve a 200ms latency.

2. **Splice audio in.** You need to be able to tightly control audio output, i.e. be able to alter queued audio while it is waiting to be played and to know almost exactly what audio has been played. For multi-modal systems this would extend to the video output as well.
3. **Know how to respond.** The appropriate response to an interruption varies considerably by application, as we found in our focus group, helpful systems should pause politely and rephrase, however for virtual characters a whole set of human responses to interruptions including rudely continuing may need to be implemented.

2.4.2 Adaptive prosody modelling for emotional speech synthesis

One of the main contributions in terms of adaptive speech synthesis is the on-going development of a "hybrid" speech synthesis system, which is based on unit selection but uses prosody targets models that are generated by a neural network.

This approach has many advantages: the neural network, trained on a large variety of speech and in various conditions, and helped with a large amount of contextual information, can generate very realistic prosodic models for the sentences that need to be spoken. These prosodic models are then used during the unit selection process to guide the selection towards units close to these prosodic targets, from a pool of pre-recorded units from a human actor, leading to high overall acoustic quality and naturalness.

The prosodic model neural network are trained taking into account the context in which the speech was recorded, and thanks to a large database of emotional speech recorded by CereProc, can also adapt the prosodic model targets to express a specific emotion.

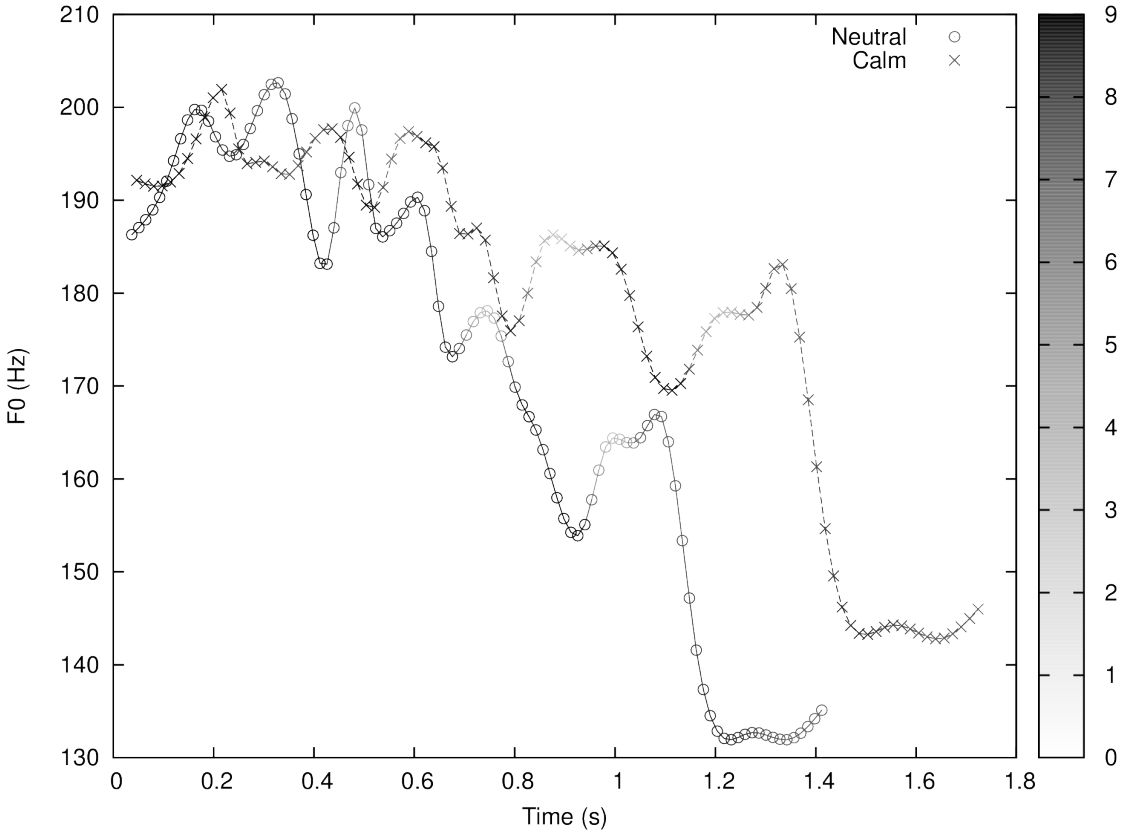


Figure 5: Synthetic prosody contours for the sentence *"I am having a great time."* with the "calm" and "neutral" emotional modalities. The shade of the curve indicates the energy of the signal (in a logarithmic scale). The "calm" variant has a smaller fundamental frequency range, a lower energy amplitude, and is noticeably slower.

These parametric models of prosody considerably reinforce the effect of the emotional speech synthesis, as they will offer the following benefits:

- Possibility to make a better use of "neutral" units in emotional speech synthesis.
- Possibility to perform gradual emotional response by changing the weight of the emotional factor provided to the neural network.

Neural network input and output features

The input features are rich contextual features about the utterance being spoken: phonetic context (current phoneme, two previous phoneme, two subsequent phonemes), features related to the phoneme identity (voicing, front / back position, etc.), as well as relative position within the syllable / word / phrase / sentence, Part Of Speech information, etc.

The prosodic features that are being modelled are the fundamental frequency of speech (F0), the energy of the signal, and the duration of phonemes in the speech utterance.

2.5 Synthesis-Analysis feedback loops (Task 4.5)

Currently, the dialogue manager has reactive behaviour templates that mirror the behaviour of the user, for example, a high valence value (received from SSI in the INPUT module of the ARIA system) of the user translates to an agent that is also positive (agent emotion set to joy in the FML Template for the FML Translator). However, it is possible to deviate from this default mirroring if necessary.

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

2.5.1 Interaction States

The ARIA agent can be in 4 different states with respect to a dyadic interaction with a user as depicted in **Error! Reference source not found.**. These states are:

- IDLE
- ENGAGING
- ENGAGED
- DISENGAGING

The analysis-synthesis loop is started by the INPUT module that informs the ARIA system with a variety of signals coming from hardware sensors such as camera and microphone. The Dialogue Manager is responsible for keeping track of this state, update it when input audio-visual signals change, or the analysis of those reveals new bits of information about the current user in interaction with the agent. Finally, the Dialogue Manager informs all other components, including ARIA-Greta that synthesizes appropriate multimodal behavior according to the current interaction state of the ARIA agent.

The agent is IDLE when no one is visually detected (i.e. the user is not detected by the camera) and there is absence of voice activity (i.e. the microphone does not capture any voice). In this state the ARIA agent displays idle behaviors such as gaze wondering around and various postures with arms crossed, hands on hips, etc.

The user's presence triggers a transition to an ENGAGING state in which the agent shifts the attention to the user with a gaze and a posture change. In this state the interaction with a user has started but not the user and the agent haven't yet exchanged any verbal behavior (i.e. the conversation hasn't started yet). The transition from this state to ENGAGED happens either when the user starts speaking to the agent or when a timeout triggers the opposite, which means that the agent proactively initiates the conversation with the user. In the ENGAGED state all the mechanisms described in D3.3 are activated, therefore the Dialogue Manager chooses the appropriate communicative functions to accomplish and via the FML Templates sent to the FML Translator module the multimodal behavior of the agent is generated.

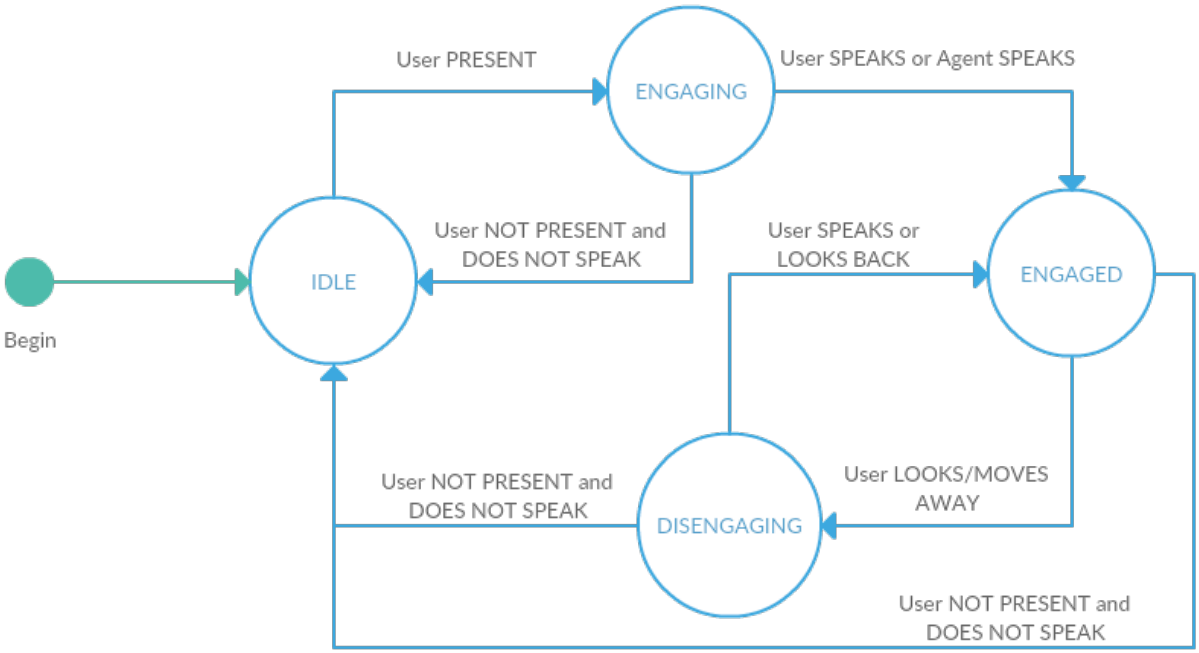


Figure 6: The four interaction states of an ARIA agent in a dyadic interaction with a user.

While in the ENGAGED state, it can happen that the user looks away, or simply moves out of the visual field of view of the camera sensor. This triggers a transition to an intermediate state named DISENGAGING where the agent is still in interaction with the user but the possible outcomes can be either a return to IDLE if the user completely disappears, or a return to ENGAGED if the user addresses the agent again.

2.5.2 Language switch

ARIA agents are multilingual (English, French and German) and should offer linguistic adaptation 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 are now working on the real-time detection of user's language changes (among the three system languages). This would support 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.

2.6 Multimodal behaviour responses to unexpected situations (Task 4.6)

At UTwente, we have developed two different scenarios for unexpected situations. The handling of interruptions in both scenarios is done by keeping track of the interaction state, the talking state, how talkative the agent is, previous interruptions and the goal markers. The scenarios are described as follows:

1. **Question-Answering (User interrupts agent).**
   In this scenario, the user can ask the agent questions. At any given point, the user can interrupt the agent. Based on the internal emotional state, the agent responds differently to user-initiated interruptions. This can be linked to ARIA-Greta and Cereproc's implementation of the graceful interruption handling of nonverbal behaviour (in Greta) and speech synthesis (in CereVoice).
2. **Read along (Agent interrupts user).**
   In this scenario, the user is reading a part of Alice in Wonderland. Whenever the agent notices particular words (keyword spotting), it will interrupt the user, taking also the user's and the agent's emotions into account. The agent will mention related (or unrelated) topics and see how the user responds. The agent's interruption will contain an opinion or an informative statement about a topic.

Both scenarios but in particular the Question-Answering one require appropriate multimodal responses when the user interrupts the agent. At CNRS we first made use of the graceful interruption handling of ongoing agent synthesised speech as described in D4.2, we then looked in more detail at the visual reaction in terms of nonverbal behaviours displayed by the agent as a reaction when an interruption occurs.

2.6.1 Visual nonverbal reactions to unexpected situations

Existing literature in human-human interaction and social psychology did not provide us many insights about the behaviours that people exhibits in case of unexpected conversational interruptions by the interlocutor. Researchers mainly focused on the impact that an interruption may have on the interruptee in terms of social attitudes[17], gender effects[18] and semantic meaning[19] (e.g. overlaps in conversation). Researchers in the field of embodied conversational agents looked at generating agent's verbal content when handling barge-in user's interruptions [20] or adaptive interruptible speech synthesis[21]. However none of them looked in detail the exact nonverbal behaviours that

[17] Farley, S. D. (December 01, 2008). Attaining Status at the Expense of Likeability: Pilfering Power Through Conversational Interruption. Journal of Nonverbal Behavior, 32, 4, 241-260.
[18] Beattie, G. W. (1981). Interruption in conversational interaction, and its relation to the sex and status of the interactants. Linguistics, 19(1-2), 15-36.
[19] Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. Language in society, 29(1), 1-63.
[20] Crook, N., Field, D., Smith, C., Harding, S., Pulman, S., Cavazza, M., ... & Boye, J. (2012). Generating context-sensitive ECA responses to user barge-in interruptions. Journal on Multimodal User Interfaces, 6(1-2), 13-25.
[21] Kopp, S., van Welbergen, H., Yaghoubzadeh, R., & Buschmeier, H. (2014). An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. Journal on Multimodal User Interfaces, 8(1), 97-108.

humans/agents exhibit when a conversational interruption occurs. This is crucial for us because our aim is to display such behaviours in synthetic agents in a realistic manner, but we miss references/examples of inspiration for models that could be taken from human-human interaction.

For this reasons, we first looked at the data available in NoXi, our database of expert-novice interactions created for this project. We examined 10 sessions recorded in French and we looked at occurrences of interruptions. This job was facilitated by our previous work consisting of automatically annotate conversation states as described in D4.2. We had automatic annotations describing who is speaking (i.e. expert, novice, both and none) and more detailed annotations automatically computed about overlaps and pauses within the turn of the same speaker or between turns of two different speakers. We could rapidly skim through videos and jump directly to parts where an overlap between turns was detected and annotated. An overlap between turns means that an interlocutor, for instance the Expert, was speaking and an overlap (which is likely to entail an interruption) occurred, followed by a turn change in favour of the other interlocutor, the Novice in this example. This meant that the Novice interrupted the Expert and successfully grabbed the turn.

We carefully observed all those segments for both expert and novice in 5 sessions for a total of 96 minutes (for each interlocutor) of observation. We took note of all nonverbal reactions that the interruptee displayed at the moment the interruption occurred. While doing this we considered the interruption classification schema proposed in D4.2 and implemented in ARIA-Greta for speech synthesis interruptions with CereVoice. In this schema we distinguished 3 types of reactions to an interruption:

- **HALT**: the current speaker halts and yields the turn to the interrupter.
- **OVERLAP**: the current speaker continues speaking and grabs the turn.
- **REPLAN:** the current speaker halts but immediately responds with a new utterance thus keeping the turn.

We noted for each type of reaction a variety of nonverbal behaviours that was common to all three types. Interruptees throughout the various sessions reacted with their eyes, gaze, head, smile, gesture and posture behaviours. More specifically, while reacting to an interruption they closed the eyes, raised/lowered the eyebrows, gazed at the interrupter, tilted, nodded or tossed their head, smiled, held, retracted or expanded their current gesture for a variable amount of time, leaned forward their torso.

Those behaviours were sometimes exhibited in concert with others, other times in isolation. Table 2 summarizes the frequencies of those occurrences for each behaviour and for the three different types of reaction every time a speaker (interruptee) was interrupted by the interlocutor. The first column on the left, after the types, shows the total number of occurrences of a specific nonverbal of reaction.

Type	Total	Eye lid down	Eyebrown down	Eyebrown raise	Gaze at other	Head tilt	Head nod	Head toss	Smile	Gesture hold	Gesture retract	Gesture expand	Lean forward
HALT	60	21	10	7	8	10	11	2	10	22	19	0	2
OVERLAP	9	4	0	4	0	2	1	2	3	0	0	3	1
REPLAN	3	1	0	2	0	0	1	1	1	0	0	0	2

Table 2: The frequencies of observed nonverbal reactions to interruptions exhibited by the interruptees in NoXi.

In the HALT case, the interruptee had two distinct gesture behaviours that we named GESTURE HOLD and RETRACT. Those correspond to two phases during the reaction where they held their current gesture and, sometimes, they retracted towards a rest pose with hands down by performing another pause (i.e. gesture freeze) of a certain duration that we called GESTURE RETRACT. We did not report exact timings for all these observations because our task was to observe at a first glance the behaviours that emerged during such reactions and not a precise annotation of all aspects of behaviours.

In the OVERLAP case, some participants reacted by expanding the current gesture, meaning that while overlapping their speech they enlarged the movement of their hands thus resulting in an expansion of their current gesture.

As we can see from the chart in Figure 7, most of the reactions (in blue) were common to the HALT type, i.e. when the current speaker halted and gave the turn to the interrupter, with a significant amount of eye, head and gesture behaviours among the others.

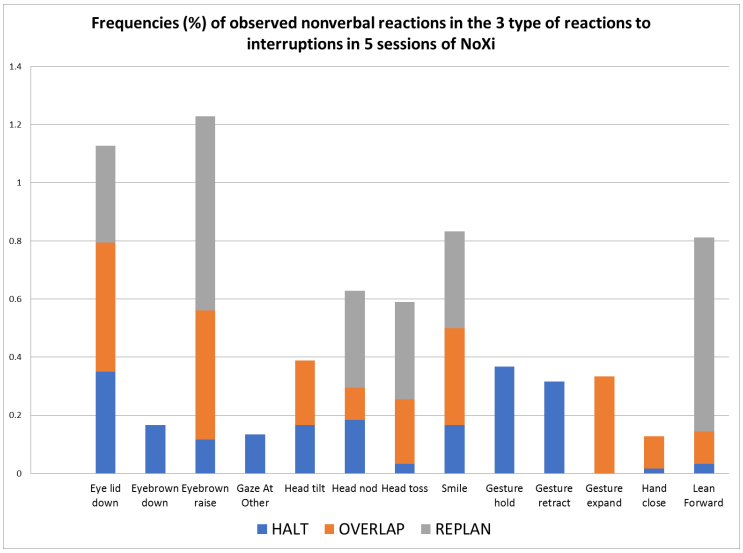


Figure 7: The frequencies of nonverbal reactions in %

2.6.2 Implementation of a toolbox in ARIA-Greta for nonverbal reactions

The observed nonverbal cues in NoXi in human-human conversational interruptions provided us a rich set of information that we transferred into the ARIA system via a toolbox implemented in the ARIA-Greta component. The objective of this toolbox, depicted in Figure 8, is to have a testing module that allows us to generate and observe synthetic nonverbal reactions with our ARIA agent. We took this intermediate step prior to creating a model for the automatic generation because our goal is not only to generate those reactions, but also to discover the parameters that are needed in order to control this generation and obtain realistic and believable animations when displayed in real-time by the ARIA agents.

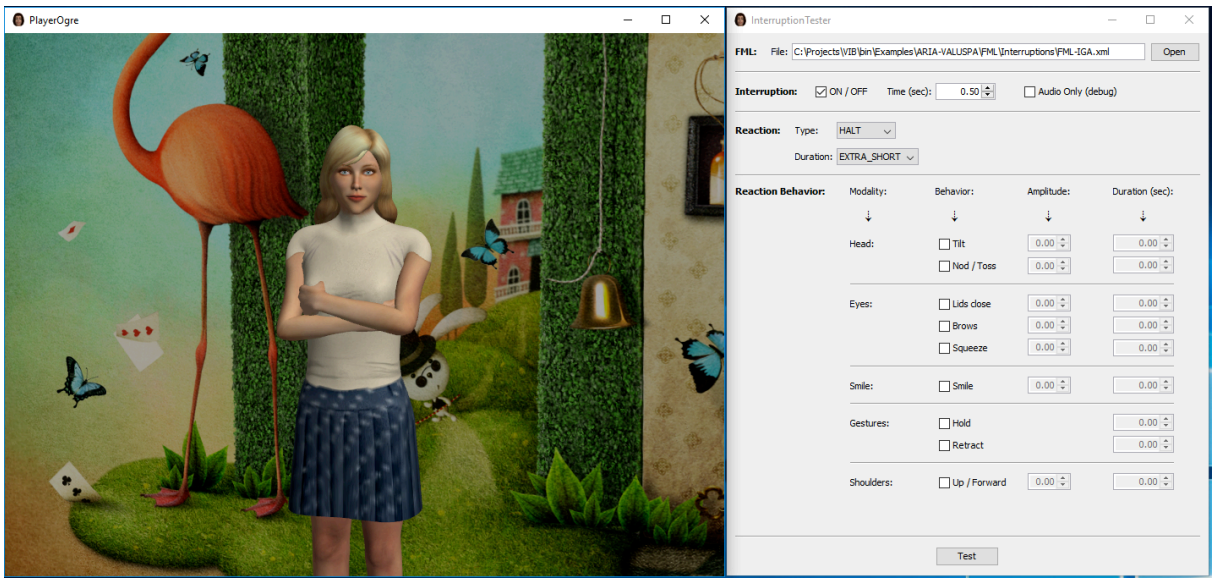


Figure 8: The tool developed in ARIA-Greta for testing nonverbal reactions controlled via their Amplitude and Duration.

The tool depicted in the figure on the right, supports the automatic simulation of a user's interruption at a given time. A dropdown menu offers the selection of the reaction type (i.e. HALT, OVERLAP, REPLAN). The full set of observed behaviours has been implemented and can be chosen for testing, i.e. displayed by the agent when the simulated interruption occurs, via the Modality selector. We identified two common continuous parameters for all nonverbal reactions: amplitude and duration. The amplitude (0..1) sets the amount of visual effect for the nonverbal reactions and duration specifies the duration of the animation. For instance, the eyelids can be closed with a continuous amount ranging from 0 (open) to 1 (fully closed), and the duration sets the total time of the animation from the initial position (e.g. eyelids fully open) to the desired amount (e.g. fully closed) and back to the original position.

This tool enabled us to manually generate and observe a great number of nonverbal reactions, by choosing behaviours in isolation or in concert with others, and by

specifying different amplitudes and durations. This is needed for the next and last step towards the automatic generation of those reactions.

2.6.3 Automatic interruption reaction generation

This last step is ongoing work towards an automatic generation of nonverbal reactions by the ARIA agent when a user's interruption occurs. We aimed at exploring the (almost) totality of the parameters' space given the multitude of possibilities when combining reaction type, nonverbal signals (i.e. modality), amplitude and duration.

This is unpractical in a controlled empirical study; we therefore opted for a crowdsourcing method that will allow us to explore the set of generated reactions for many given combination of all parameters. In order to limit those combinations to a feasible number, we first defined the range of durations that remains visually observable and within an acceptable length (i.e. a reaction lasting over 2 seconds is not believable in the context of reactions to interruptions). We also discretized the amplitudes with a step of 0.25 from 0 to 1 because smaller steps were not perceived when observing the generated behaviours with our tool. We then decided to begin by focusing on the HALT type given that the majority of observed reactions in NoXi were found within this type.

The final step is to generate videos corresponding to a specific combination of those parameters and ask human users to perform, in a crowdsourcing context, a selection of the videos in relation to a specific social label (e.g. stance), such as for instance a friendly reaction to an interruption or a hostile reaction. The results of this job will be labelled combinations of parameters that can be instantiated by ARIA-Greta as reactions to interruptions.

3. Outputs

In this section we indicate the outputs with pertinence to WP 4 for the Months 23-31.

Peer-reviewed conferences papers and journals:

- **ICMI 2017 (submitted):** Cafaro A., Wagner J., Baur T., Dermouche S., Torres Torres M, Pelachaud C., André E. and Valstar M.F. *"The NoXi Database: Multimodal Recordings of Mediated Novice-ExpertInteractions"*, In Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI'17). November 13-17, 2017, Glasgow, Scotland, UK.
- **IVA 2017**: Cafaro A., Bruijnes M., van Waterschoot J., Pelachaud C., Theune M. and Heylen D. *"Selecting and Expressing Communicative Functions in a SAIBA-Compliant Agent Framework"*. In Proceedings of the 17th International Conference on Intelligent Virtual Agents (IVA'17). Stockholm, Sweden.
- **SIGDial 2017**: Dubuisson Duplessis G., Clavel C. and Landragin F. *"Automatic Measures to Characterise Verbal Alignment in Human-Agent Interaction"*, 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial 2017), pp. 11.
- **VSI 2017:** Varun J., Clavel C. and Pelachaud C. *"Beat Gesture Prediction Using Prosodic Features"*, 3rd International Workshop on Virtual Social Interaction (VSI 2017), Bielefeld, Germany.

Dissemination activities (meetings, talks, demos, workshops):

- **January 2017**: NoXi web interface is online.
- **January 2017**: Joint meeting between EU H2020 projects ARIA-VALUSPA and KRISTINA at the Dagstuhl seminar in Germany.
- **March 2017**: ARIA-VALUSPA demo at France IA, national day about artificial intelligence in cooperation with the French Ministries of Economy and Education.
- **March 2017**: Demo of the ARIA System at the AI Working group meeting in Paris at the University Pierre and Marie Curie.
- **May 2017**: Video demo of the ARIA system at the RGLab during the Roland Garros tournament.
- **June 2017:** ARIA-VALUSPA demo during the 10th anniversary of the ISIR with an international delegation of guests.
- **June 2017:** Demo of the ARIA-VALUSPA system during the working group day of the "GT ACAI" in Paris, France.
- **July 2017:** Invited talk of Catherine Pelachaud (WP4 leader) discussing the ARIA-VALUSPA project at the Virtual Social Interaction workshop in Bielefeld, Germany.
- **August 2017:** Invited talk of Catherine Pelachaud at INTERSPEECH2017 in Stockholm, Sweden.
- **August 2017:** Organized the Workshop on Conversational Interruptions win Human-Agent Interaction at the International Conference on Intelligent Virtual Agents (IVA 2017) in Stockholm, Sweden.

4. Conclusions and Last Period Plan

4.1 Conclusions

This work package enabled the generation of adaptive audio-visual communicative behaviour for our ARIA agents. While accomplishing the tasks of WP4 we, as consortium, collaboratively advanced research and techniques in the domain of artificial embodied conversational agents.

Our adaptive nonverbal communicative behaviour generation model makes use of common standards for communicative intentions and behaviour description (FML, FML Templates and BML). Therefore, it supports interoperability among the ARIA system components and other SAIBA-compliant systems. We advanced state-of-the-art generation of audio-visual behaviour taking into account synchrony with users (e.g. verbal alignment, nonverbal engagement), adaptive features (e.g. language, speech prosody) and unexpected situations (multimodal responses to interruptions).

4.2 Last Period Plan

In the last period of the project CNRS, with respect to task 4.2, more specifically concerning the expression of interpersonal attitudes, aims at evaluating the *Sequential Attitude Planner* in order to assess whether the output of the proposed algorithm convey attitudes that are recognizable by users. For this we will perform an empirical study and ask participants to compare an agent playing FML scripts without attitude expression and an agent exhibiting behaviours including those in output from the *Sequential Attitude Planner.* Furthermore, a last step is to finalize the annotations (as described in Section 2.2.2 Expressing interpersonal attitudes) of the French sessions in NoXi. Then, we will study the synchrony between the expert and the novice in terms of non-verbal behaviors and engagement variation.

For the last period of the project concerning Task 4.3, 4.5 and 4.6, Cantoche has planned to create two characters, Alice and a special character for Unilever. Alice has been done for the POC and the first implementation of the ARIA Valuspa platform. These steps have revealed the limitation of blending animations of the Living Actor avatars. Actually, the emotion is focused on the face of the agent and have no impact on the body of it. The emotions of the agent (e.g. sadness, joy, etc.) should impact the shoulders and the posture of the agent to have a natural behaviour.

This is why the Living Actor component has been updated and will be enhanced to manage several channels of animations, each of them are animated manually by experts in 3D animation, and will be merged by the software to produce smart animations:

- Head
- Torso

- Shoulders
- Arms

This split allows random combination of animation to extends the range of the animation capabilities of the avatar, and add naturalness and variation in the triggered behaviours. In the same time the intensity and the speed of the animation can be adjusted to match with the behaviour.

Living Actor is waiting for the finalisation of the UNILEVER-ARIA casting to provide a new avatar built with this set of animations. If the casting failed, we will update Alice with a new set of animations.

As for Task 4.3, future work at CNRS aims at using the developed framework along with the verbal alignment metrics to give the ability to the ARIA agent to verbally align with the user. This involves:

1. The development of a context-aware verbal alignment score of an utterance at the surface text level that takes into account dialogue history,
2. The over-generation of a pool of candidate system utterances, and
3. The building of a verbal alignment strategy to dynamically choose the best candidate in the pool of available system utterances, based on the score obtained in step 1.

Step 1 is currently being developed at CNRS. Steps 2 and 3 are developed in collaboration between CNRS and UTwente.

Regarding the text-to-speech development of CEREPROC and Task 4.4, we will focus our efforts in the last period on the two following topics:

- **Voice quality transformations**. One of the ways to enrich the system's expressiveness is to perform some post-processing that will modify the voice quality of the TTS output for various effect. For example, adding creak to a voice will make it "warmer". CereProc has been working on several voice quality transformations in the last few months (such as creak and jitter) and will make them accessible to the ARIA framework through custom xml tags.
- **Emotional corpora augmentation**. The emotional speech in our unit selection voices does not sound as good as the neutral speech due to sparsity of data: we cannot afford to record as much emotional speech as neutral due to the strain it imposes on the voice talents. By using advanced machine learning techniques, we work on creating mappings from neutral speech to emotional speech, thus allowing us to augment the emotional sub-corpora.

Concerning Task 4.5, UTwente's plan is to create a module that modifies the stance of the agent dynamically, such that ARIA-Greta can display these stances as interpersonal attitude changes, taking advantage of the model described in Section 2.2.2 Expressing interpersonal attitudes.

Finally, for Task 4.6, UTwente is planning to evaluate two scenarios (described in D3.3) in September. The idea is to plan a setup for both scenarios and come up with metrics for measuring the interaction on for example 'adaptiveness' and 'naturalness'. Currently, the DM does not make full advantage of the feedback sent by ARIA-Greta. The goal is to have the DM better adapted to the FML generation, such that the agent does not produce the same utterance twice for instance. Furthermore, we need to automate the creation of agent dialogue moves that translate to valid FML with parameters for (non)-verbal behaviour.

CNRS will finalise the work on the generation of nonverbal reactions in response to conversational interruptions.