KING'S SPEECH

Foreign language: pronounce with style!

Principal investigators: Georgios Athanasopoulos*, Céline Lucas* and Benoit Macq*

Candidates: Guillaume Gustin, Alessandro Cierro and Robin Guerit.

* ICTEAM-ELEN - Université Catholique de Louvain, Belgium

Project objectives

The principal investigators are developing the GRAAL[1] project which is concerned with developing a set of tools to facilitate self-training on foreign language pronunciation, with the first target being learning French.

This set of tools includes a method to align a sentence uttered by a native speaker and the sentence repeated by a learner. Following the alignment, the sentence is decomposed in parts and different features are extracted. After analysis of the identified vowels/consonants, prosody, intonation, rhythm, etc., an evaluation feedback is given to the learner. The GRAAL project also aims to provide a multimodal interface for rendering the native speaker's discourse and for returning feedback to the learner. This interface makes use of different approaches such as 3D avatars, sagittal sections, as well as speech visualization tools.

The goal of KING'S SPEECH is to develop new interaction modalities and evaluate them in combination with existing functionality aiming to better personalize GRAAL to the taste and specificities of each learner. This personalization will rely on a machine learning approach and an experimental set-up to be developed during eNTERFACE'17.

The eNTERFACE'17 developments could be based on a karaoke scenario where the song is replaced by some authentic sentences (extracts of news, films, publicities, etc.). Applications like SingStar (Sony) or JustSing (Ubisoft) could also serve as a source of inspiration, e.g., using a smartphone as a microphone while interacting with avatars.

The main tasks to be developed will be:

- Audio processing modules for real environments (e.g., echo cancelation, noise suppression, source identification, phrases alignment, etc.)
- New modes of interactions between an animated avatar (tutor) and the learner (including rhythm, prosody, phonetics, …)
- Visual speech synchronization (lips motion tracking and alignment)
- Explore the link between speech rhythm and wearable tactile stimulation
- Engagement through interface gamification
- Empowerment by machine learning (best combination of modalities/optimal fusion-fission)

The expected deliverables are:

- Design, implementation and integration of new interfaces and interaction modalities
- Building training scenarios and test corpus based on multiple modalities
- Test and optimization of the developed application, including machine learning

[1] GRAAL: Guidage en Réalité Augmentée pour l'Apprentissage des Langues

Background Information and Proposed Approach

Signal acquisition

The goal of a multimodal human-machine interface is to analyze, understand and synthesize (i.e., perception and production) multiple communication means in real-time. In the scope of KING'S SPEECH, different modalities of communication could be considered for serving the scenario to be developed, with each mode of communication requiring different types of sensors (microphones, cameras, depth sensors, etc.).

Depending on the scenario specificities, various types of sensors could be selected. An essential element in KING'S SPEECH is the acquisition of the user's speech. In this scenario, the audio interface could make use of a single or multiple microphones. The microphones could be located in different positions, such as embedded in webcams (i.e., static placement) or smartphones (whose position can change over time). Moreover, several low cost microphones could be used for creating a smart-room environment.

The signals recorded by the various microphones should be further processed in order to allow a smooth interaction in a real-word environment, where disturbances such as ambient noise and reverberation are often present. Besides, the system's own audio responses will be also captured by the microphones and should be therefore separated from the user's speech input. Typical techniques that can be considered for addressing the above acoustic interferences are noise suppression, beamforming, echo cancellation, source separation, etc. [1].

Real-Time Spoken Language Visualization

Visual patterns generated from a speech signal can deliver a stimulating evaluation feedback to the learner. Hence, one of the objectives of KING'S SPEECH is the exploration of existing techniques, as well as the development of new tools for the real-time visualization of the user's speech. These tools could rely on low level signal's characteristics such as the intensity, spectral contain (e.g., spectrograms), etc. Speech characteristics such as the pitch and formants could also provide input to higher level visualizations such as the position of the tongue, lips rounding, the degree of nasalization, etc.

Rendering and Grading of the Spoken Language

For further analyzing the learner's speech, KING'S SPEECH project will get access to a set of GRAAL project tools, which allow the alignment of a sentence uttered by a native speaker and the sentence repeated by a learner [2]. Following this alignment, different features are extracted facilitating the offline comparison of the two utterances. The analysis is leading to the detection of pronunciation, intonation and rhythmic errors, which can be used as an input for the avatar based rendering and trigger events in the game scheduler. Furthermore, the speech analysis results can feed the empowerment and personalization modules, as well as facilitate the adaptation of the system to the modality that is most appropriate for the user.

Avatar Rendering and Game Scheduler

The use of avatars can be, in many ways, a wonderful tool. This is why we believe it is important in the scope of KING'S SPEECH to create both 3D avatars and 2D sagittal planes for

many uses such as comprehensive and learning visual tool. To do so several steps need to be developed:

1. Analysis of all the possible postures of the human vocal organs
2. Lips motion and vocal organs alignment with 2D sketches and 3D avatars
3. Designing, modelling and drawing the avatars and the sagittal planes
4. 3D and 2D animation
5. Optimization and rendering

For the scenario to be selected, some steps need to be identified and implemented, such as:

1. Analysis of the background and the environment of the game (scenario)
2. Understanding the interest of some assistance from 3D models
3. 3D modelling
4. Optimization and rendering
5. Analysis of the extra benefits of Virtual Reality or other devices

In addition, for supporting the scenario of choice, a game scheduler shall be incorporated acting as the orchestrator of the different actions and tasks to be performed. The following actions are suggested as a starting point of this activity:

1. Experimenting using Unity 3D [3] or an equivalent cross-platform game engine
2. Integration of the GRAAL technologies and all the visual artefacts and 3D animated models into a new standalone (game) software
3. Consideration of empowerment and gamification principles, as further detailed in the following section

Empowerment through Gamification

In recent years, gamification, which can be defined as the application of game-design elements and game principles in non-game contexts, has become a trending topic in many fields. Gamification is particularly suitable in the case of education, where it can be integrated effectively to motivate students and enhance learning [4,5,6]. There are some obvious overlaps between games and the classroom that makes gamification of curriculum a logical approach: game players work to achieve specific goals and win while students in the classroom work to achieve specific learning objectives and succeed in exams; game players progress from level to level based on performance while in the classroom students must pass prerequisite courses and show some level of understanding before adressing an upper academic level. Based on a review of popular gamification taxonomies, 6 inseparable gamification persuasive strategies that can be enumerated as follows [7]:

1. Clear goals and challenges setting
2. Constant feedback on performance
3. Reinforcement through rewards (not punishments)
4. Progress monitoring and comparison with self and others
5. Social connectivity
6. Fun and playfulness

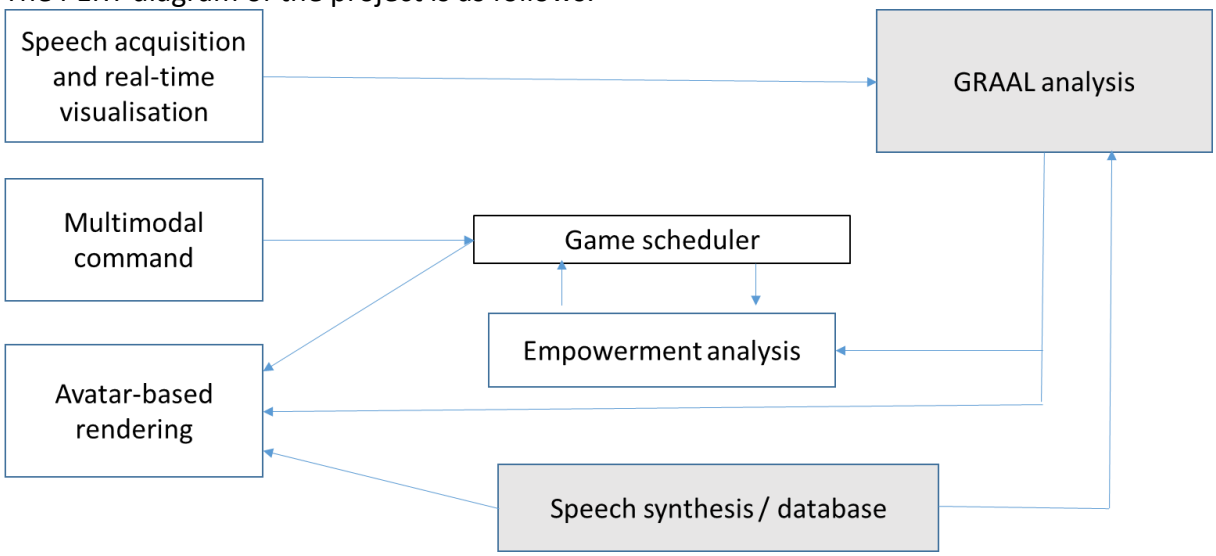
KING'S SPEECH intention is to employ throughout the chosen scenario these strategies, which are the broad principles that make gamification addictive, along with other popular tactics such as [8]:

1. Points and - increasing difficulty- levels
2. Badges

3. Leaderboards
4. Challenges and quests
5. Social engagement loops and onboarding
6. Avatar

Technical description

The PERT diagram of the project is as follows:



The project will also get access to a speech synthesizer (granted by the Acapela group company [9]) and to a corpus of speech sentences. This corpus will be previously validated by expert professors in French as a Foreign Language (FFL), among others Mrs Emmanuelle Rassart and Geneviève Briet who wrote a book about pronunciation in the classroom [10]. The selection of the sentences difficulty will follow a specific pedagogical progression for learning pronunciation in connection with the European indicator of languages objectives (Common European Framework of Reference for Languages CEFR [11]).. Subtitled movies could also be used to complement the corpus.

The progression of the work of pronunciation of the French language is a spiral progression. This spiral consists of a dozen notions and skills to acquire and to work in an iterative way by exercises whose items become more complex as the learning progresses.

These notions and skills are grouped around 5 themes which are 1 / syllabation, 2 / accentuation, 3 / intonation 4 / the correct production of vowels and consonants of French.

The exercises will include both a phase of perception to identify prosody (music of the language) and phonetics (consonants and vowels) in the form of quizzes, associations, etc. and a production phase with feedback via the GRAAL tools.

The global goal of the project is, as explained in the introduction, the design and implementation of a language training system based on the “karaoke” concept. To reach this goal we have decomposed it into subsystems to be designed and interconnected.

- Workpackage 1: Speech acquisition and real-time visualization: this workpackage will include the following subtasks:
    - Hardware setups for speech signal acquisition
    - Real-time speech patterns (spectrograms, pitch, intonation, …) analysis
    - Real-time feedback of speech production
- Workpackage 2: Multimodal commands: Multimodal interface development including touch, gesture and vocal commands.Workpackage 3: Avatar based rendering:
    - Analysis of all the possible postures of the human vocal organs
    - Lips motion and vocal organs alignment with 2D sketches and 3D avatars
    - Designing, modelling and drawing the avatars and the sagittal planes
    - 3D and 2D animation
    - Optimization and rendering

For the scenario to be selected, some steps need to be identified and implemented, such as:

- Analysis of the background and the environment of the game (scenario)
- Understanding the interest of some assistance from 3D models
- 3D modelling
- Optimization and rendering
- Analysis of the extra benefits of Virtual Reality or other devices

- Workpackage 4: Empowerment analysis: the goal of this module is to develop a finite-state machine which constructs profiles of each learner and provide a personalized focus of attention and progress plan. We aim at developing a structured learner electronic record which can be used intra-learner but also inter-learner to derive learning profile patterns.
- Workpackage 5: Game Scheduler: The initial version of the game scheduler will be based on Hierarchical Tree Networks, which will transform the spiral structure of the learning into specific goals to reach according to a linear tree structure. The implementation will make use of the open-source version of the Unity 3D game engine.

Prototype 1 after 10 days containing speech acquisition V1, Avatar rendering V1 and Multimodal command V1

Prototype 2 after 20 days= prototype V2 + empowerment analysis and game scheduler (including the definition of some game scenario).

Required resources:

- computers with Unity license installed
- Speech acquisition devices (microphone, mic arrays, …)
- Multimodal interactions (Kinect cameras, accelerometers, smart-phones)
- VR equipement (Oculus drift hardware and software)

Project management:

The project will be managed by a team of 3 persons:

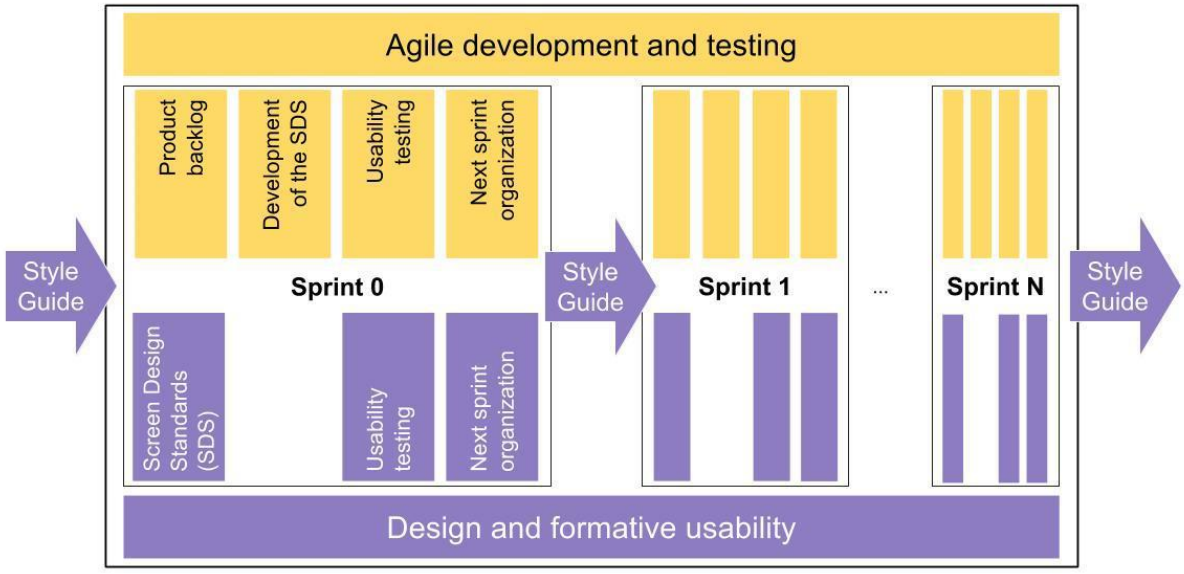
- Prof Benoit Macq as a general coordinator: Benoit has a long experience in managing European projects and spin-offs launching. He was coordinator of the SIMILAR project which gave rise to eNTERFACE experiences.
- Dr. Georgos Athanassopoulos will be the signal processing and software architecture coordinator.
- Mrs Céline Lucas, a language specialist, will coordinate the User Experience aspect of the project.

Workplan

The workplan will include 2 tracks. One “component” track working on tools development in the workpackage and one “integration” track working on the development of a prototype as a proof-of-concept.

The time-sharing between the two tracks will be 2/3-1/3.

The workplan of the second track will follow the Agile-UX methodology that was used for the egle project (www.egle.be) and which is described in [13]. The workplan of this part will follow the figure below with only 2 short sprints:



Benefits of the project

The project will be published as an example of “serious game” for language learning.

The project will also publish in “open-source” some of the developed components like its multimodal command, the voice acquisition tools, the Unity game-scheduler, empowerment tool based on FSM. Those components should be generic enough to be re-used in other applications (bots for m-health for example …)

Profile Team

The core of the team will be based on Prof. B. Macq, Dr. G. Athanassopoulos, C. Lucas, R. Guerit and A. Cierro from the Image and Signal Processing Group of UCL. The competence of this team covers project management, speech processing, language learning, Unity-based game development, VR, avatars ….

Who do we need from outside in the project:

We would like 2 to 4 additional young researchers/developers from outside active either in speech processing or in the field of Artificial Intelligence for gaming.

References

[1] W. Kellermann, "Acoustic Signal Processing for Next-Generation Human/Machine Interfaces", Proc. of the 8th Int. Conference in Digital Audio Effects, 2005

[2] S. Brognaux. Expressive speech synthesis: Research and system design with Hidden Markov Models. PhD thesis, Université catholique de Louvain (UCL) – Université de Mons (UMONS), 2015

[3] Unity website, https://unity3d.com/

[4] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work? A literature review of empirical studies on gamification," in System Sciences (HICSS), 2014 47th Hawaii International Conference on, pp. 3025–3034, IEEE, 2014.

[5] Jackson, M. (2016). Gamification in Education: A Literature Review

[6] Danowska-Florczyk, E., & Mostowski, P. (2012). Gamification as a new direction in teaching Polish as a foreign language. ICT for Language Learning

[7] B. Cugelman, "Gamification: what it is and why it matters to digital health behavior change developers," JMIR Serious Games, vol. 1, no. 1, 2013

[8] A. S. Miller, J. A. Cafazzo, and E. Seto, "A game plan: Gamification design principles in mhealth applications for chronic disease management," Health informatics journal, 2014

[9] Acapela Group website, http://www.acapela-group.com/

[10] G. Briet, V. Collige, E. Rassart. La prononciation en classe. Presses universitaires de Grenoble, 2014

[11] The Common European Framework of Reference for Languages, http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

[12] Martin, O. (2007). *Mixed reality interactive storytelling: acting with gestures and facial expressions* (Doctoral dissertation, UCL.).

[13] KIEFFER, Suzanne, GHOUTI, Aissa, et MACQ, Benoit. The Agile UX Development Lifecycle: Combining Formative Usability and Agile Methods. In : *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2017.