Reactive Statistical Mapping: Towards the Sketching of Performative Control with Data

Project Proposal for the eNTERFACE 2013 Workshop

Project Leader Nicolas d'Alessandro, numediart, University of Mons

Team Candidates Maria Astrinaki, Onur Babacan, Alexis Moinet, Thierry Ravet, Joëlle Tilmanne, Hüseyin Cakmak, Thomas Hueber et al.

Abstract

MAGE is an open-source performative speech synthesis system built upon the HTS HMM-based speech synthesizer. The aim of this project is to broaden the scope of its potential applications, by generalizing the concepts of performative speech synthesis to other use cases and making MAGE the first tool for the performative HMM-based synthesis of potentially any phenomenon. HMM-based performative synthesis is an approach suited for many real-world phenomena which can be seen as time series, such as speech, singing, laughter or motion. However the complexity of its implementation can restrain most potential users from testing it, or prevent them to benefit from the whole control and flexibility that such an approach can give. In addition to generalizing MAGE to other domains than speech, we aim at making it a user-friendly system, lowering the difficulty to access HMM-based performative synthesis. Furthermore, we will extend the real-time properties of the system, and validate their consistency across the different aspects of the workflow: mapping, HMM solving, rendering and user interaction.

Objectives

The main objective of this eNTERFACE project is the sketching of the third major version of the MAGE software platform. To our current knowledge, MAGE is the first system to propose a performative approach to speech synthesis. For nearly a decade, the vision behind "performative speech synthesis" has been to bring a refined and reactive control of speech production properties to the user. This vision has been shaped iteratively through many past eNTERFACE projects, now leading to a collaborative and open-source platform. For this workshop, we decided to focus on three user-centered objectives:

Objective 1: MAGE should be independent from data sets and rendering engines

For the last decade, the use of Hidden Markov Modeling has emerged in various fields of signal processing research, for solving recognition and generation problem. Unlike several other statistical models, HMMs toolkits have developed in a rather idiosyncratic way. For instance, HTS [1], the most comprehensive HMM-based synthesis toolkit is designed around speech processing. As a result, the use of HMMs in any new research area, i.e. gait modeling [2], happens to be achieved by hacking these tools. We would like to propose a formalism to describe any kind of analyzed data for which we want to realize some HMM training and trajectory generation. We want to introduce the idea of "codec" in MAGE, and build the core HMM routines as totally blind to the actual data types and rendering techniques.

Objective 2: MAGE should be truly "real time": from duration to scheduling

MAGE 1 and 2 have been the opportunity to validate the possibility of creating HMM-based trajectories in a reactive way, i.e. with an ongoing impact of the user on the current output. However the current stage of development does not formally change the concept of time encapsulated in HTS duration models [3]. Nevertheless a truly reactive system should carry an explicit concept of what is the "real time", i.e. a clear and controllable mapping between the system and the user timelines. Practically, it means that we need to figure out ways for HMMs to be both solved (deducing the trajectory) and scheduled (deducing when the trajectory is actually applied on the renderer) from the models.

Objective 3: MAGE should be accessible to non-experts of the HMM processing

Many cases have shown that users who truly need innovative synthesis solutions (e.g. for speech or 3D images) are rarely the ones who create them in the first place. These days, the manipulation of HMMs is still highly restricted to experts. We would like MAGE to be the first platform to be user-friendly enough to actually disseminate in other research fields and be used by application designers. Adapting a HMM process to a given use case mainly consists in modifying the training and mapping routines. Our purpose is to lower the complexity of these two manipulations by using a specific API. This idea can be compared to the "creative coding" trend that has spread in fields like 3D rendering or sound synthesis.

Background

The challenge in producing artificial contents, such as synthetic speech, music, images or motion, has always been to "make it more real". Behind this idea of realism, there are some qualities that viewers and listeners have learnt to expect from the artificial medium. Disciplines of media synthesis have come across similar trends. For many decades, researchers have been targeting the ability for the artificial medium to "preserve the message", i.e. what is heard or seen is at least understood correctly. In speech synthesis, this is called *intelligibility* but we can easily transpose it to images or motion. Later, the target has become to "make it look more natural". The trend of *naturalness* is what has brought these research fields to use recording of real human performances. We can retrieve this concept in the emergence non-uniform unit selection for speech [4], giga-sampling for music [5], face scanning [6] or full-body motion capture [7].

Nowadays our target is to create expressive outputs. There are basically two main trends in how researchers deal with *expressivity*. The first one is "more data": they try to cover even more contents in their recordings, hoping to have enough utterances of various expressions to achieve good synthesis [4]. The second one is "back to modeling": data are now used to train statistical models and these models are then used for controlling production models [8]. This come-back to parametric models leads to more flexibility in how expressivity is approached. It has also been shown in various studies [9,10] that the trend of expressivity is not only a matter of passively experiencing the contents. It rather involves a significant *enactive* component and bring human factors in the actual design space [11].

Hidden Markov Modeling

In this project, we use HMMs as the statistical model for its unique properties in capturing time-series data. Classical HMMs are not designed for synthesis applications when considering most real world phenomenon since they assume constant statistical modeling of the observations within a given HMM state and a state duration probability modeled by a decreasing exponential. Over the last 15 years, the HTS working group [12] has developed an extension which alleviates the limitations of these classical HMMs. By taking into account the dynamics of the data in addition to static parameters, and by explicitly modeling state duration, their trajectory HMMs are now a state-of-the-art generative model for speech. This HMM-based synthesis approach opens a broad range of possibilities to the user: style control, speaker adaptation, voice modification, voice reconstruction, etc. More recently, we have seen the great application of this approach to gait modeling [13].

Performative Speech Synthesis

Even if TTS became the main trend, the speech community has always continued to develop several speech synthesis engines that were partially or entirely aimed at being user-controlled. This trend goes back to early speaking machines such as the von Kempelen's [14] or Bell Lab's Voder [15] and has been preserved in the digital age through HCI-related communities, such as NIME [16] or p3s [17]. In 2009, we have started the MAGE project as an attempt to develop such performative machines on the top of the state-of-the-art HMM-based speech synthesis technologies [18]. MAGE 2.0 is the first software to enable high-quality speech synthesis with on-the-fly control of both phonetic labels and prosodic parameters. As revealed by several studies on articulatory-to-speech synthesis [19], the introduction of gestural input in the realm of HMM-based synthesis shifts the role of HMMs towards being more of a *mapping* tool vs. purely a generative one. We think this is a promising turn in how HMMs could be used in interactive multimodal applications.

Technical Description

In this project, we aim at bringing together a new platform for *reactive statistical mapping*, capitalizing on several years of MAGE development and the expertise of various other HTS contributors. The development of this new software will consist in consolidating various core HMM-based functionalities in order to be independent from the use case. Such a consolidation will be benchmarked by *live-testing* the platform among at least three categories of applicative cases for which we already have contributing numediart researchers and/or solid collaborations: voice synthesis (speech, singing, laughter), gait animation and facial animation. All our use cases will be respecting a MVC design:

1. View: a front-end where the contents with rendered in real-time: sound, images
2. Model: a middleware component instantiating MAGE and its codecs and extensions
3. Controller: a back-end with a user interface to input gestures: GUI, TUI or NUI

This part of the proposal gives more details on the different technologies that are envisioned and gives the main research and development axes that will be followed in order to build the new system. We also give greater insights about the devices, environments and prototyping strategies that will be aligned in this project. Finally we also describe the project management strategies that will be deployed in the team.

In this section, most of the following text refers to module names that are depicted in Figure 1. We also highlight the workpackages of the project and introduce a naming convention (WP_N) that will be reused in the section where the schedule is described.

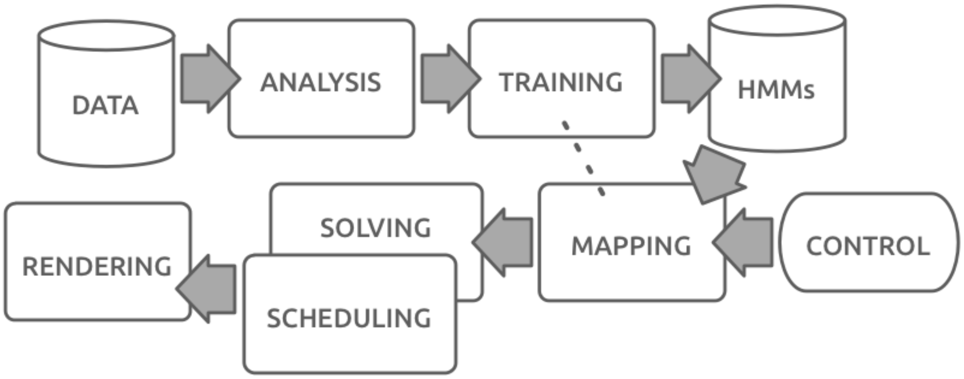

Figure 1 - Architecture of MAGE 3 as a reactive statistical mapping system.

Research and Development Axes

The MAGE 3.0 system will be composed of five main components, identifying the groups of researchers working on these problematics: the HMM training algorithm, the solving/scheduling algorithm, the mapping space, the rendering front-ends (primarily voice synthesis, gait and face animation) and the user interaction back-end:

1. **HMM Training Algorithm (WP_1)**

   It has been clear over the last few years that complex context-based HMM training techniques have exciting applications far beyond the field of speech processing. In our research group, we are exploring at least for four other major fields: singing, laughter [20], gait [2] and face animation, as well as multimodal cases such as voice and face motion synthesis [21]. However these examples still rely on the HTS training process, which is fairly hacked in order to accommodate the new areas of exploration.

   In this workpackage, we would like to tear down the HTS training process to its fundamental needs and routines and propose a more generalized approach to it. The first aspect to consider is the broad range of analyzed data that a HMM training process could potentially have to handle (cf. Figure 2). We need to understand a fair range of the possible variability that can be encountered in these data dimensionality, rate, pattern structure. Moreover, we need to develop an abstract representation of how to possibly "construct" a HMM training process, including well-known processes like *adaptation* [8]. We foresee that such a descriptive task can be achieved with a markup language for which MAGE would have data-independent parsing routines.

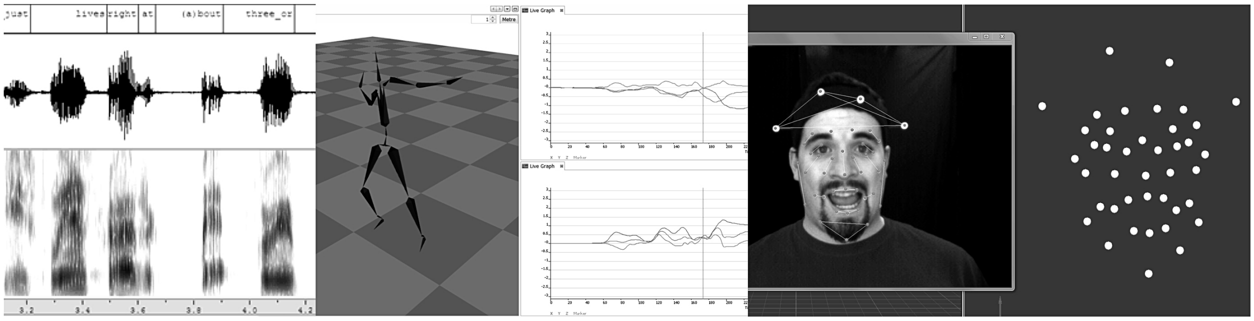

Figure 2 - Various data types on which to apply HMM training, from left to right: voice sound, full-body motion capture with IGS, face tracking with OptiTrack.

2. **Solving/Scheduling Algorithm (WP_2)**

   In HTS, the trajectory generation process ultimately goes down to a giant matrix operation [8]. This matrix is constructed from the concatenation of selected HMMs and the output trajectory literally results from the *solving* of this matrix system. In this process, the temporal structure of the trajectory is implicitly determined through the same process, as each HMM state duration is based on state-related *duration models*. The optimal sequence of state durations is actually solved offline.

However the temporal nature of HMM-based parameter trajectories should be based on quite different principles when the synthesis system is supposed to work in “real time”, i.e. with an inherent concept of “time passing” inside the system. In a real-time situation, there is a tighter binding between a) how much of the future statistical context is known (e.g. upcoming models), b) how the models get solved considering unpredictable future and user inputs, and c) how/when the decision gets causally propagated through the system and reach the renderer. These aspects actually narrow down the need for the solving algorithm to be completed with a *scheduling* one. The solver/scheduler toolbox will need to address typical aspects of HMM-based trajectory generation such as *context-checking* (find the most appropriate models), *decoding* (find the most likely state sequence), *inferring* (find the most likely observed sequence), but also operations on models such as *interpolation*, *extrapolation* and *inversion* [22].

3. **Mapping Space (WP$_3$)**

In this project, our third objective envisions that most of the above-mentioned tasks should be accessible to non-experts of HMM programming. However we acknowledge that there is an inherent complexity in how these statistical models should get manipulated in a reactive way and how decisions get taken and propagated to the rendering phase according to user interaction. This is actually a quite systematic issue when users have to deal with real-time systems, as it appears that non-experts really have difficulties with time-passing, generative, evolutive processes [23].

In various fields, *creative coding* communities has seemed to emerge in response to these issues. Applying the user-centered mindset that we have defined as fundamental for this project, these communities have tried to change the affordance of such technological toolboxes by working on appropriate spaces, languages or APIs. A coding space/language/API is the formalism which draws the clear line between describing *what the application does* and *how it actually does it*. Max or Pd might be the greatest examples of such a mindset for audio processing, as it presents a patching metaphor to the user but completely hides the complex audio scheduling process. We draw the same conclusion with how Processing has changed computer graphics.

We think that the discussion on how to make complex statistical modeling more affordable for non-experts still has to be initiated. Some great environments like EyesWeb or Wekinator [24] have started to deal with these issues but no one is giving a satisfying answer when it comes to Hidden Markov Modeling. This project will be our chance to propose one way to simplify the access to HMM-based mapping.

4. **Rendering Front-Ends (WP$_4$)**

There is an inherent bound between the types of analyzed data and the rendering engines to which we will have to tie the MAGE middleware. Indeed, a voice database will lead to control a MFCC-based vocoder, full-body motion capture data will ultimately animate the joints of a 3D character, face tracking data will enable to animate a 3D face model through key points. Although this project is really about leaping forward with the MAGE middleware, there is nothing that we can really validate without observing practical synthesis results. In numediart, we already have developed several rendering machines such as speech [18], singing and laughter vocoders [20], skeletons and bindings to character building software [13] for both body and face animation (cf. Figure 3). We also have maintained solid collaborations with researchers working e.g. on face articulatory models [19,25,26]. Some of these tools are already real-time, others are not yet. The purpose of this workpackage is to have access to several real-time renderers during the project, as for testing the MAGE trajectory generation and discuss what is the best interface format to hook the middleware up with real-world rendering engines. Indeed network protocols like OSC have been used a lot to glue various pieces of software together but we have come across the fact that it is not always appropriate.

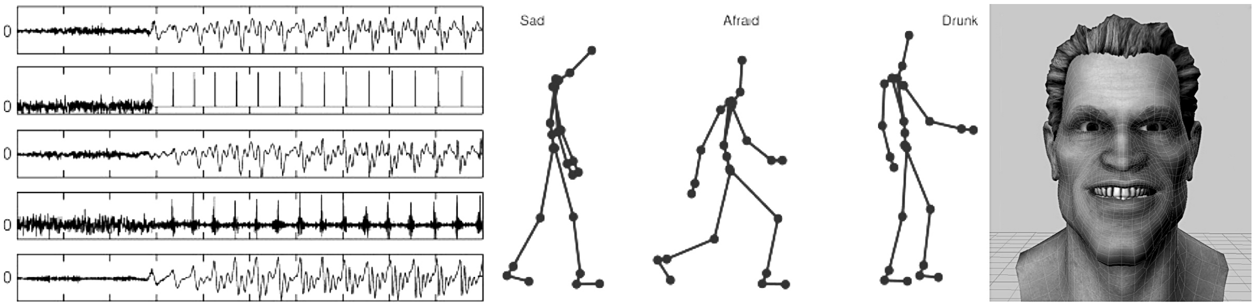

Figure 3 - Various rendering engines on which to apply MAGE HMM-generated trajectories, from left to right: speech synthesizers, 3D walking characters [13], 3D face models.

5. **User Interaction Back-End (WP$_5$)**

The ultimate nature of the demonstrators that we want to build is that they will actually react to user interaction. It means that the whole statistical process of generating parameter trajectories (for the renderers) will actually react to various continuous and discrete user gestural inputs. These gestural inputs can take various forms. Obviously it can be graphical user interfaces (GUIs) with several controls (knobs, sliders, x-y maps, etc.) but the numediart lab has developed, over the last years, a greater expertise in controlling signal processing processes with rather tangible and natural user interfaces (TUIs and NUIs). It has been shown that the manipulation of digital media through either physical or embodied controls triggers this more fundamental kind of intelligence called *enaction*. We are exploring enactive control of speech synthesis for at least a decade and we came up with various controllers that we will use in this project. Particularly, we will use several HandSketch prototypes [27] that have the particularity of proposing many degrees of control in an intuitive set of bimanual compound gestures (cf. Figure 4a). We have also used NUIs to control speech properties with facial expression (cf. Figure 4b).

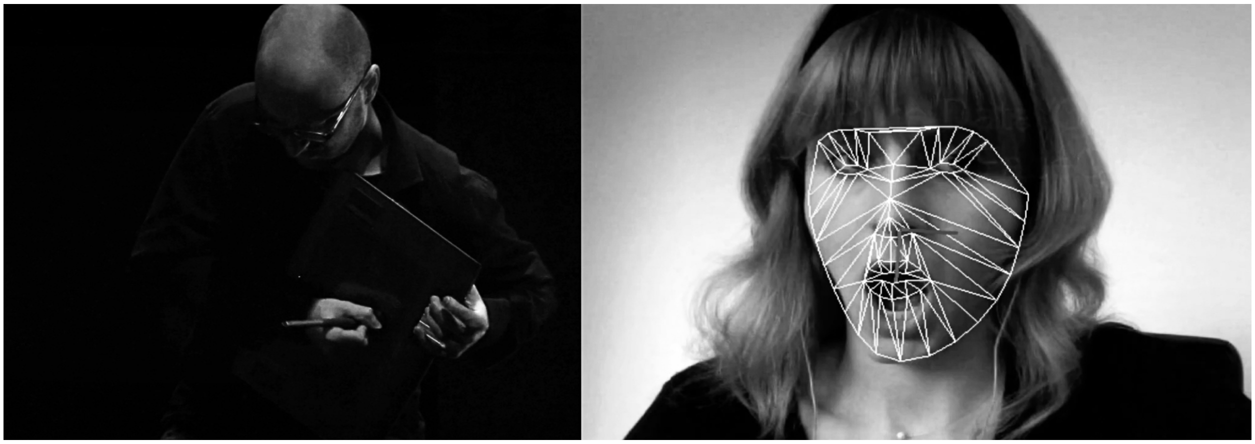

Figure 4 - HandSketch as a tangible user interface (left) [27] and real-time face tracking as a natural user interface (right) for the control of speech synthesis.

Prototyping Cycle

As in any HCI application development, the team workflow is made of iterations between various phases, including research & development (described above) but also updating the case studies on which we are working (WP_6) and validating our output results in front of a panel of external observers (WP_7). Updating the case studies will consist in constantly revisiting the scenarios on which we are working. These scenarios are attempts to control several rendering tools with user interfaces through the MAGE middleware. There are many ways to balance between these different components (choices in the modalities that are used, mapping strategies implemented in MAGE, human-computer interaction models that are tested) and we want to stay open to test various ideas. We hope to broadly open the discussion for other eNTERFACE participants, in order to come and exchange their ideas on the use of HMMs in multimodal interactive applications. The other aspect is the evaluation of the system outputs (sound and images) by external listeners/observers. We wish to demonstrate our synthesis results to as many eNTERFACE researchers as possible and establish a first informal benchmarking of successful strategies.

Facilities and Equipment

The team will essentially work with available devices brought by the participating labs. Obviously we will bring our own laptops. Moreover, we will try to bring several acquisition systems, probably the OptiTrack and some voice recording hardware with electroglottographic sensors (shared with another eNTERFACE project submitted by numediart). It is also possible that the GIPSA lab will bring ultrasonic equipment for capturing intraoral gestures. On the user interface side, we will also bring HandSketchs.

The only pieces of equipment that we would eventually require would be several extra speakers and secondary monitors that we might not be able to transport on site. We would also require a quiet room (because we work with sound synthesis) and some space for setting up booths for motion capture (which require free space and controlled light).

Project Management

The whole project will be supervised by Nicolas d'Alessandro, project leader in the numediart institute, University of Mons. He should stay on the site of the workshop for the whole period. Based on the subscribed participants, sub-teams will be gathered around the specific workpackages of the project. The methodology which is promoted in this project aims at staying flexible and adapt to our successive prototyping cycles. We will work with guidelines inspired by various Agile techniques, such as organizing scrum meetings or collectively defining the development backlog. MAGE is already an open-source project, so we will respect this philosophy for the development of core MAGE functionalities. However some more specific aspects, like attempts on the mapping modules, could be integrated as pre-compiled libraries in the project in order to protect some participants' source code.

Project Schedule

In this part we gather the various workpackages that have been highlighted in the technical description (WP_N) and set them down on a one-month schedule, plus some extra tasks:

1. **WP_1 - HMM Training Algorithm**: generalization of the training routines for a data-independent HMM training framework, proposition of a description format
2. **WP_2 - Solving/Scheduling Algorithm**: real-time and interactive sequence of state solving, trajectory generation enabling real-time operation on model parameters
3. **WP_3 - Mapping Space**: user-friendly access to HMM-based mapping and synthesis
4. **WP_4 - Rendering Front-Ends**: have MAGE interact with real-time renderers that correspond to our targeted use cases: vocal, full-body and facial animation
5. **WP_5 - User Interaction Back-End**: develop human-computer interaction models that use gestural inputs to real-time control our studied rendering processes
6. **WP_6 - Iteration on Case Studies**: inline reassessment of the gesture-to-rendering scenarios we use in our case studies for vocal, gait and facial animation
7. **WP_7 - Assessment of Synthesis Results**: organization of external observation regarding our synthesis results in the various media we have addressed
8. **WP_8 - Reporting and Publishing**: general dissemination tasks

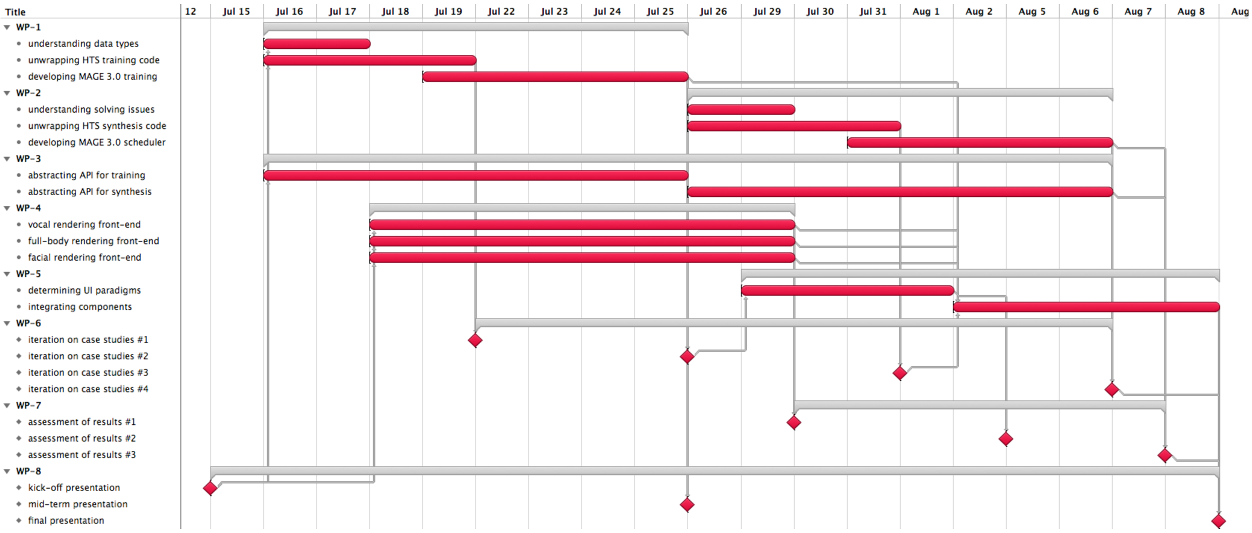

Figure 5 - Scheduling of the project (workpackages, sub-tasks and milestones).

Deliverables and Benefits

In this part we describe what are the main deliverables and benefits that the team will provide at the end of the workshop:

1. the third version of MAGE released as an open-source software
2. collection of new data corresponding to the considered use cases
3. new interactive setups corresponding to the considered use cases
4. synthesis examples corresponding to the considered use cases
5. open sessions during the working for welcoming observers
6. a scientific report, distributed in the required format

Team Profile

Project leader: Nicolas d'Alessandro (software design, singing, HCI | University of Mons)

Team currently proposed: Joëlle Tilmanne (project management, motion capture | University of Mons), Maria Astrinaki (software design, speech | University of Mons), Alexis Moinet (software design, speech | University of Mons), Thierry Ravet (software design, computer graphics | University of Mons), Thomas Hueber (statistics, speech | Grenoble Institute of Technology), Hüseyin Cakmak (laughter, face tracking | University of Mons), Onur Babacan (singing, University of Mons). We are also currently in touch with researchers from the Center for Speech Technology Research (Edinburgh, UK), Trinity College (Dublin, Ireland) and the Media and Graphics Interdisciplinary Centre (Vancouver, BC).

Collaborators that we are looking for: As described in the project schedule, this workshop will need pretty advanced software developers for most of the time. The first half of the month will be more oriented towards data analysis and statistical modeling, as the second half will require expertise in real-time applications and system scheduling. The second half of the project will also involve more testing of the synthesis results and human-computer interaction properties of our software. Therefore for this second half, we are also looking for HCI / COG profiles, i.e. people who can run user studies and observation sessions.

References

[1] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black and K. Tokuda, " *The HMM-Based Speech Synthesis System (HTS) version 2.0,"* in Proc. of the 6th ISCA Workshop on Speech Synthesis, August 2007.

[2] J. Tilmanne, A. Moinet and T. Dutoit, " *Stylistic Gait Synthesis Based on Hidden Markov Models,"* in the EURASIP Journal on Advances in Signal Processing, 2012:72.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, *"Duration Modeling for HMM-Based Speech Synthesis,"* in Proc. of the Fifth International Conference on Spoken Language Processing, pp. 29-32, 1998.

[4] A. Raux and A. W. Black, *"A Unit Selection Approach to F0 Modeling and its Applications to Emphasis,"* in Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 700-705, December 2003.

[5] E. Lindemann, *"Music Synthesis with Reconstructive Phrase Modeling," in the* IEEE Signal Processing Magazine, pp. 80-91, vol. 24, no. 2, March 2007.

[6] D. Schabus, M. Pucher and G. Hofer, *"Building a Synchronous Corpus of Acoustic and 3D Facial Marker Data for Adaptive Audio-Visual Speech Synthesis,"* in Proc. of the Eighth International Conference on Language Resources and Evaluation, May 2012.

[7] A. Menache, *Understanding Motion Capture for Computer Animation and Video Games*, by Morgan Kauffman Publishers Incorporated, 2000.

[8] H. Zen, K. Tokuda and A. W. Black, *"Statistical Parametric Speech Synthesis,"* in Speech Communication, vol. 51, no. 11, pp. 1039-1064, 2009.

[9] J. Piaget, *Play, Dreams and Imitation in Childhood*, London:Routledge, 1951.

[10] F. J. Varela, E. Thompson and E. Rosch, *The Embodied Mind*, MIT Press, 1991.

[11] N. d'Alessandro, B. Pritchard and S. Fels, *"A Design Space for Gesture to Performative Speech and Singing Synthesis,"* in Proc. of the First International Workshop on Performative Speech and Singing Synthesis, pp. 13-23, March 2011.

[12] *HTS Working Group*. Online: http://hts.sp.nitech.ac.jp

[13] J. Tilmanne, *Data-Driven Stylistic Humanlike Walk Synthesis*, PhD thesis, 2013.

[14] W. von Kempelen, *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*, 1791.

[15] H. Dudley, *The Carrier Nature of Speech*, Bell System Tech, vol. 19, pp. 495-515, 1940.

[16] *New Interfaces for Musical Expression*. Online: http://www.nime.org

[17] *Performative Speech and Singing Synthesis*. Online: http://www.magic.ubc.ca/p3s

[18] M. Astrinaki, N. d'Alessandro and T. Dutoit, 2012, *"MAGE: A Platform for Tangible Speech Synthesis," in* Proc. of the Twelfth International Conference on New Interfaces for Musical Expression, pp. 353-356, 2012.

[19] T. Hueber, G. Bailly and B. Denby, *"Continuous Articulatory-to-Acoustic Mapping using Phone-Based Trajectory HMM for a Silent Speech Interface,"* in Proc. of Interspeech, 2012.

[20] J. Urbain , H. Cakmak and T. Dutoit, *"Development of HMM-based Acoustic Laughter Synthesis"*, in Proc. of the Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalizations in Speech, Ireland, October 2012.

[21] J. Urbain, R. Niewiadomski, E. Bevacqua, T. Dutoit, A. Moinet, C. Pelachaud, B. Picart, J. Tilmanne and J. Wagner, *"AVLaughterCycle: Enabling a Virtual Agent to Join in Laughing with a Conversational Partner Using a Similarity-Driven Audiovisual Laughter Animation,"* in the Journal on Multimodal User Interfaces, vol. 4, no. 1 , pp. 47-58, 2010.

[22] J. Tilmanne and T. Dutoit, *"Continuous Control of Style through Linear Interpolation in Hidden Markov Model Based Stylistic Walk Synthesis"*, in Proc. of the International Conference on Cyberworlds 2011, pp. 232-236, 2011.

[23] D. Shiffman, *The Nature of Code*, 2012.

[24] R. Fiebrink, P. R. Cook, and D. Trueman. *"Human Model Evaluation in Interactive Supervised Learning,"* in Proc. of the SIGCHI Conference on Human Factors in Computing, Vancouver, May 2011.

[25] Z. Ling, K. Richmond, J. Yamagishi and R. Wang, *"Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis,"* IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1171-1185, August 2009.

[26] N. d'Alessandro, J. Wang, B. Pritchard and S. Fels, *"Bringing Bio-Mechanical Modeling of the OPAL Complex as a Mapping Layer for Performative Speech Synthesis,"* in Proc. of the International Seminar on Speech Production, June 2011.

[27] N. D'Alessandro and T. Dutoit, *"Advanced Techniques for Vertical Tablet Playing A Overview of Two Years of Practicing the HandSketch 1.x,"* in Proc. of the International Conference on New Interfaces for Musical Expression, pp. 173–174, 2009.