CoVoP – Collaborative Vocal Puppetry
Multi-User Performative Voice Synthesis
on Distributed Platforms

Principal Investigator — Nicolas d'Alessandro, Research Associate
Media and Graphics Interdisciplinary Centre (MAGIC)
University of British Columbia (Vancouver)

Candidate Participants — Sidney Fels, Johnty Wang, Cam Hassall (MAGIC), Bob Pritchard (Music), University of British Columbia (Vancouver, BC)

Thierry Dutoit, Maria Astrinaki, Onur Babacan (NUMEDIART), University of Mons (Belgium)

Abstract

Vocal interaction is primarily a gestural phenomenon. Our sense of vocal identity and intent is determined by subtle variations in the voice production mechanism. These variations are unconsciously interacting with each other in a conversational context, making the use of voice one of our most complex social experiences. From the bio-mechanical point of view, the understanding of voice production mechanism remains pretty elusive. There are fundamental barriers in being able to observe intraoral activity. In this project we aim at developing social aspects of voice production from a different perspective. We will develop a new interactive voice synthesizer and a social experience platform, in order to allow a group of users/performers to manipulate various aspects of *performative voice synthesis*, distributed on their own device (computer or mobile). We also want to conduct user studies and perceptive tests in order to evaluate the impact of such a collaborative and performative approach of voice synthesis on vocal intelligibility, naturalness and identity.

Objectives

In this project we aim at developing a new framework for performative voice synthesis, i.e. a realtime synthesis systems where voice is directly produced by gestural control, with no more reference to textual input. We want to address both phonetical and prosodical issues, with applications in speech and singing synthesis. The target of this new system is to extend the context in which performative voice synthesis is produced and develop a new concept : *voice synthesis as a framework for multi-user participation*. Indeed we have known for a long time that managing all the parameters of speech production is tough for one single performer. Consequently, we want to explore the idea of *collaborative vocal puppetry* and see how vocal intelligibility, naturalness and even identity can be addressed as a social game, involving the participation of multiple users on their own device.

Objective 1 – Multi-User Voice Synthesis Architectures

Most of current voice synthesis architectures are designed like a giant script which aims at writing down a waveform onto the hard drive. Eventually if we consider real-time and interactive voice synthesizers, we find out that their structure is quite monolithic and difficult to break down. We want to explore how to design voice synthesis architecture with the various components being detachable on heterogenous platforms – computers or mobile devices – while maintaing sound quality and low latency. The same voice synthesis process will be handled by a variable amount of performers.

Objective 2 – Interactive Control of Voice Production

Voice synthesis can be split in various types of typical issues to be solved: coarticulation of phonemes, speech timing management, intonation modelling, voice quality dimensions, etc. Most of these tasks refer to a significantly different representation of data and the development of appropriate human-computer interaction (HCI) models has not been widely studied for these tasks.. We want to develop new interaction paradigms for voice synthesis based on ubiquitous sensing technologies, such as multitouch screens, camera-based tracking, embodied sensors, etc.

Objective 3 – Software Platform for Distributed and Interactive Voice Synthesis

Objectives 1 and 2 require a software platform to be developed in order to support both the modularized and distributed voice synthesis and the new interaction models on a cloud of heterogenous devices like computers (with various kinds of sensors plugged in), cellphones or the newly successful tablets. We aim at extending and integrating several existing tools (such as openFrameworks) in order to provide a software platform in which new collective experiential models can be implemented, and deployed on a variable amount of devices, depending on the context.

Objective 4 – Assessing Collaborative Emergence of Vocal Intelligibility, Naturalness and Identity

There is a linguistic and sociological interest in questioning and validating several properties of voice (at various levels: intelligibility, naturalness and identity) when this voice is produced by a collaborative work. Indeed it questions the sense of social emergence of vocal characters and how they are perceived by the members of the collaborating group, but also an external audience.

Background

Voice production is primarily a gestural phenomenon. In the early years of artificial voice production, intraoral vocal gestures were being imitated by manual manipulation of physical objects, such as the von Kempelen's machine [1] or the Voder [2]. The simple models used at that time had quite poor intelligibility and naturalness but these systems provided a rich and embodied interaction paradigm, like a musical instrument. When voice synthesis was implemented on computers, Text-To-Speech (TTS) became the main trend, implicitly defining the keyboard input as the interaction model. From a human-computer interaction (HCI) perspective, the textual input is more of a historical heritage, based on similarities between first desktop computers and typewriters. The last twenty years brought many innovations to get rid of this textual input at the graphical level (e.g. mouse and icons) but voice synthesis has rarely been considered with more interactive input than text.

Existing Systems

For the last ten years, computers are fast enough to run heavy sound synthesis algorithms. As a consequence, the idea of *performative voice synthesis* [3] has been reinvestigated within the digital instrument making context. Performative voice synthesis is the production of digital voice directly from realtime gestures, without any reference to textual input, or musical score in the case of singing. We can cite the Voicer [4] as a pioneer investigation. Our teams are investigating this research axis for the last ten years and are have been developing systems such as :

- the RAMCESS synthesizer : a concatenative voice synthesizer with realtime control of the phonetic stream (several vowels and consonants), the intonation and the voice quality [5];
- the HandSketch digital instrument : a tablet-based singing instrument giving a refined control on the voice quality parameters, allowing to play melodic lines with expressive quality [6];
- the DiVA system : a fully embodied digital instrument that captures 3D positions and postures of hands and remap them to vocal tract shapes (all consonants and vowels) in realtime [7].

Voice Modelling Trade-off

Most of the actual vocal behaviour is still really difficult to observe both on the vocal apparatus [8] and recorded audio signals [9]. Consequently, voice modelling has been addressed in two very different ways. On the one hand, spectral models and concatenative synthesis produce high quality sounds but are usually detached from any refined laryngeal description, making the integration of expressive control difficult. On the other hand, articulatory synthesis gives full access to physiological properties but tuning complexity has a significant impact on the overall synthesis quality.

Human-Computer Interaction

The real-time control of voice properties starts to be more and more studied from the human-computer interaction point of view. We find several research exploring the control and imitation of voice intonation with hand gestures both for speech [10] and singing [11]. The idea of replacing the text keyboard by a phoneme-based typing interface has been investigated, but initiators of this project are the firsts to approach phonetic control (modifying vocal tract) on multitouch screens [3].

Technical Description

In this project we aim at bringing several performers together and make them collaborating on heterogenous devices – computers, cellphones, tablets – in order to arouse the voice of a single virtual character. This part of the proposal describes 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 provide a more technical description of the environments and devices that will be used. Finally we also describe the project management strategies that will be deployed in the team.

Research & Development Axes

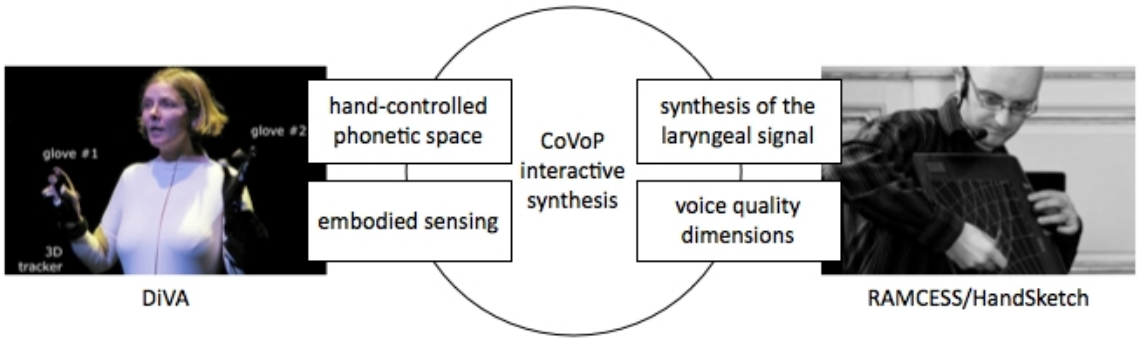
The CoVoP 1.0 system will be composed of three main components: an interactive voice synthesis engine, a distributed social experience platform and various kinds of human-computer interaction models distributed on various kinds of devices (computers, cellphones and tablets).

1. Interactive Voice Synthesis (**WP 1**)

   The first aspect of this research is to develop a new synthesis engine. We currently have two well advanced systems – RAMCESS [5] and DiVA [7] – which address complementary issues.

   RAMCESS (*Realtime and Accurate Musical Control of Expression in Singing Synthesis*) is a pitch-synchronous concatenative synthesis engine, using advanced spectral analysis techniques [9] for modifying some inner properties of the voice quality. The access to these voice quality dimensions is structured as a perceptual space: *pitch*, *vocal effort* and *tenseness*. As a result, RAMCESS is a high quality vocoder with expressive control of the whole laryngeal behaviour. The main musical instrument using RAMCESS is the HandSketch [6], a digital tablet where the control dimensions are displayed on a “playing fan” and several extra FSRs.

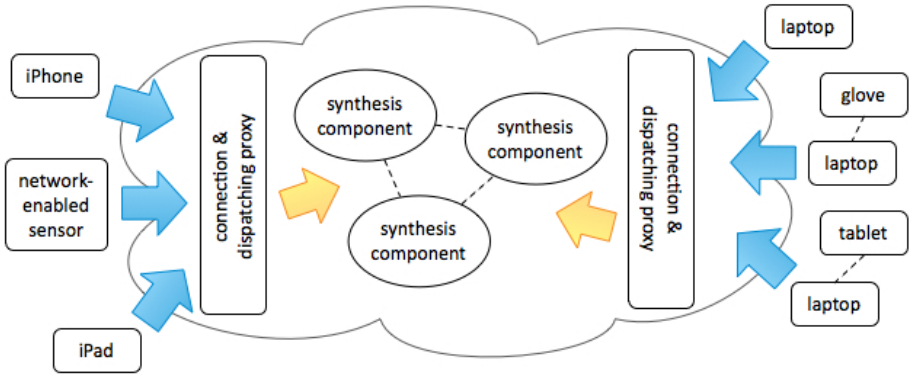
   DiVA (*Digitally Ventriloquized Actor*) is an embodied musical instrument based on two virtual reality gloves and one embodied 3D hand tracking device. The right hand glove is a CyberGlove and contains eighteen bend sensors, covering most of the hand joint angles. The left hand glove is custom-made and contains eight contact sensors, distributed on four fingers. The right wrist position is captured in 3D by a Polhemus magnetic field sensor. This gestural configuration is used to train and trigger a neural network with a given gestural language that associates one posture for each phoneme of English, and allows for intonation control. The sound synthesis is achieved with the Holmes synthesizer [12], a parallel-formant algorithm.



The overall trends of both DiVA 3.0 and RAMCESS 3.0 new software development is modularity, extendability and portability of some foundation components, following the idea of migrating from monolithic softwares to flexible toolboxes and APIs. This context is particularly appropriate for blending the two systems together. From the DiVA side, we will keep the hand-controlled phonetic space, i.e. the aspect of continuously controlling the vocal tract shape with some gestures. We also think that the embodied property of the sensing (hand postures, wrist position, accelerations) fits the overall objectives of the project. From the RAMCESS side, the laryngeal model and the voice quality dimensional mapping will be used.

2. Distributed Social Experience Platform (**WP 2**)

One other significant aspect of the CoVoP project is the distribution of various parts of the synthesis process on several heterogenous devices. More than defining a convenient modularization of the synthesis engine, we need to architecture a distributed peer-to-peer service over the network. Each introduced device – a laptop with sensors, a cellphone or a tablet – will require for some resources (a part of the voice synthesis process) and display an appropriate control space to the user. As a result this platform considers the voice synthesis process as a collection of interconnected resources to be handled by the cloud of devices, with priority and dispatching strategies. The devices themselves bring information to the whole network, such as screen size, CPU capacity, sound properties or embedded sensors. The whole process of connecting and participating to the collaborative experience has to handle these compatibility issues and particularize the experience for each performer.



3. Human-Computer Interaction Models (**WP 3**)

We already investigated several human-computer interaction models for controlling various aspects of voice synthesis. For example the 2D space for browsing the vowels and the posture-based browsing of consonants in DiVA [7], several one-handed or two-handed touchless mappings for the pitch or the (pitch, tension, effort) “fan” mapping of the HandSketch [6]. However the introduction of multitouch and portable devices is new in the project. We would like to investigate the extension of several of these above-mentioned control paradigms for multitouch screens (e.g. 2D vowel space or HandSketch fan, browsed by touch gestures). We also would like to prototype new HCI models, especially for these devices. More generally, there is a discussion about the scalability and form factor diversity of voice synthesis control.

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 study (**WP 4**) and validating in a real performance context (**WP 5**). Updating the case study will consist in constantly refining the vocal puppetry scenarios that are developed. The current starting points are: 1- the blending of HandSketch and DiVA performance paradigms, keeping the most valuable components of each; 2- replacing several controllers of these two systems, e.g. with multitouch versions. Along with the iterative development of the platform, these starting points might evolve. We want to maintain a relevant amount of work in discussing these user experience scenarios within the team. The other aspect is evaluation of the system in a real performance context. As we are talking about a performing system, we will involve other participants of the workshop in creating little performing groups and/or attending the performance and we will collect various kinds of user feedbacks.

Software Environments

The development of the software will be achieved with openFrameworks (www.openframeworks.cc). The openFrameworks project aims at providing an open-source creative environment for aggregating various software components, initially for computer graphics (OpenGL, image management, Quicktime, etc). One significant advantage of openFrameworks is to be cross-platform (Mac, Linux, Windows, iOS). Moreover the open and free add-on format attracted a lot of third party developers to wrap a large amount of new functionalities: OpenCV [13], XML, Open Sound Control [14], network protocols, etc. We aim at taking advantage of these existing modules for building real-case prototypes during the workshop. We also aim at contributing to this international open-source community with the various advances of the project (**WP 6**): voice synthesis, the social distributed platform, GUIs, etc.

Facilities and Equipment

The team will essentially work with available devices brought by the participating labs: laptops, the CyberGlove, the Polhemus tracker, several Wacom tablets, several iPods, iPhones and iPads. As performing the devices is an important part of the work, we might need a separate room for running the audio performances without disturbing the other groups. We would also require usual team work facilities, such as a projector, a screen, a whiteboard, meeting space, etc.

Project Management

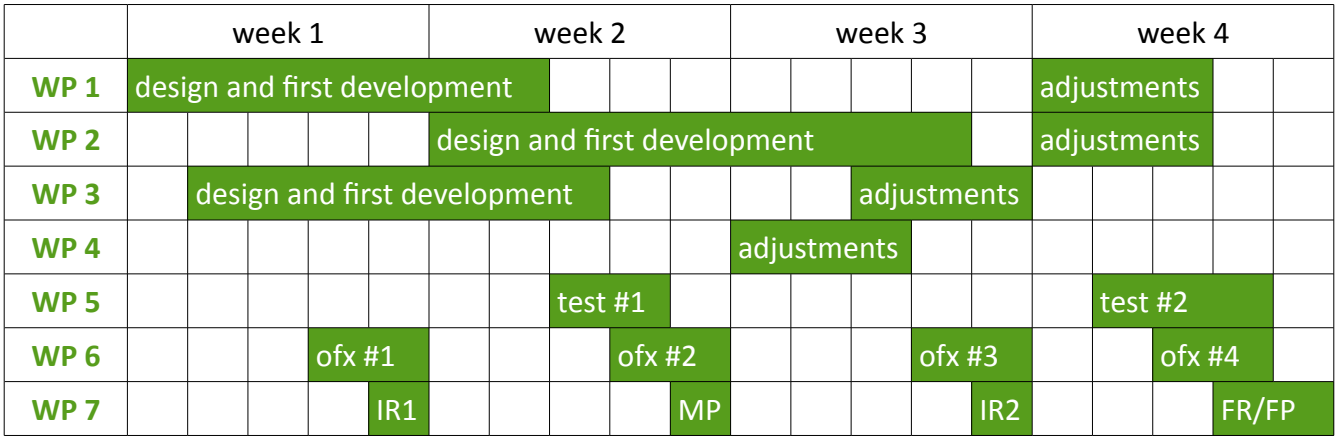
The whole project will be supervised by Nicolas d'Alessandro, from the University of British Columbia (but with five years of work and currently two PhD students in the University of Mons). He should stay on the site of the workshop for the whole period. Based on the subscribed participants, sub-teams will be gathered around the specific work packages of the project. Several experts in these various fields have already mentioned their interest in attending the workshop: Prof. Thierry Dutoit (speech synthesis), Prof. Bob Pritchard (performance with artificial voice), Prof. Sidney Fels (HCI).

As it has been done during previous eNTERFACE workshops (2005-2008), we consider important that all participants can go for interdisciplinary interest and discussions, especially regarding the performing side of the project. Thus we will foster that everybody can perform the prototypes.

Schedule

In this part we gather the various work packages that have been highlighted in the technical description (**WP N**) and set them down on a one-month schedule. We also added one typical work package which concerns the reporting task (preparing slides, writing intermediate and final reports).

- **WP 1** – interactive voice synthesis : integrating RAMCESS and DiVA into a new synthesizer;
- **WP 2** – social experience platform : developing a cloud application for connected devices;
- **WP 3** – human-computer interaction models : prototyping new gestural control paradigms;
- **WP 4** – vocal puppetry scenarios : proposing new multi-user experiences with voice synthesis;
- **WP 5** – performance / assessment : testing the prototypes, as a performer, as an audience;
- **WP 6** – openFrameworks integration : wrapping and sharing technologies as oF add-ons;
- **WP 7** – reporting: preparing the intermediate and final documents (slides, reports).



ofx #n : nth session of wrapping the code as openFrameworks add-ons;
IR1 : internal report #1; MP: mid-term presentation;
IR2: internal report #2; FR/FP: final report/presentation.

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. A new interactive synthesizer: when various modules of RAMCESS/HandSketch and DiVA will be integrated together, it will result in one of the most advanced and flexible voice synthesizers with built-in interactive properties, for both phonetic and prosodic control.
2. A validated implementation of the first social experience model around the voice synthesis of a single character, with various integrations in heterogenous devices (laptops and iOS devices).
3. A large amount of openFrameworks add-ons – including the interactive synthesizer and the social experience platform – will be distributed open-source for the community. OpenFrameworks has a really powerful distribution model that will eventually bring contributors and users to these new tools.
4. We will provide two sessions of performing tests open to all the people of the workshop and we will communicate the results of the user study achieved around these performances.
5. A scientific report will be provided at the end of the workshop, in the requested format.

Team Profile

Leader: Nicolas d'Alessandro, PhD (University of British Columbia)

Staff proposed by the leader: Prof. Thierry Dutoit, Onur Babacan, Maria Astrinaki (University of Mons), Prof. Bob Pritchard, Prof. Sidney Fels, Johnty Wang, Cameron Hassall (University of British Columbia).

Researcher profiles needed: voice analysis/synthesis, real-time audio software architecture, distributed architecture, human-computer interaction with some major interests in digital instrument making, linguistics, digital art performers, C++ development, iOS device development.

References

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

[2] H. Dudley. *The Carrier Nature of Speech*. Bell System Tech., 19:495–515, 1940.

[3] N. d'Alessandro, B. Prichard, J. Wang and S. Fels, "*Interactive Manipulation of Speech and Singing on Mobile Distributed Platforms,"* submitted to CHI 2011, Vancouver, May 2011.

[4] L. Kessous and D. Arfib, *"Bi-Manuality in Alternate Musical Instruments,"* in Proc. New Interfaces for Musical Expression, pp. 140–145, 2003.

[5] N. d'Alessandro, O. Babacan, B. Bozkurt, T. Dubuisson, A. Holzapfel, L. Kessous, A. Moinet, and M. Vlieghe, *"RAMCESS 2.x Framework - Expressive Voice Analysis for Realtime and Accurate Synthesis of Singing,"* Journal of Multimodal User Interfaces, vol. 2, no. 2, pp. 133–144, 2008.

[6] N. d'Alessandro and T. Dutoit, *"Handsketch Bi-Manual Controller: Investigation on Expressive Control Issues of an Augmented Tablet,"* In New Interfaces for Musical Expression, pages 78–81, 2007.

[7] S. Fels, B. Pritchard, and A. Lenters, *"Fortouch: A Wearable Digital Ventriloquized Actor,"* In New Interfaces for Musical Expression, pages 274–275, 2009.

[8] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, *"On the Use of the Derivative of Electroglottographic Signals for Characterization of Non-Pathological Phonation,"* Journal of Acoustical Society of America, vol. 115, pp. 1321–1332, 2004.

[9] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, *"Zeros of the Z-Transform Representation with Application to Source-Filter Separation in Speech,"* IEEE Signal Processing Letters, vol. 12, no. 4, pp. 344–347, 2005.

[10] C. d'Alessandro, N. d'Alessandro, J. Simko, S. Le Beux, F. Cetin, H. Pirker, *"The Speech Conductor: Gestural Control of Speech Synthesis,"* eNTERFACE'05 Summer Workshop on Multimodal Interfaces, Mons, Belgium, 2005.

[11] N. d'Alessandro, C. Ooge, T. Dutoit, S. Fels, *"Analysis-by-Performance: Gesturally-Controlled Voice Synthesis as an Input for Modelling of Vibrato in Singing,"* submitted at the International Computer Music Conference, Huddersfield, 2011.

[12] J. M. Rye and J. N. Holmes, *"A Versatile Software Parallel-Formant Speech Synthesizer,"* Joint Speech Research Unit Report, no. 1016, 1982.

[13] http://opencv.willowgarage.com

[14] http://opensoundcontrol.org