RAMCESS: Realtime and Accurate Musical Control of Expression in Singing Synthesis

N. D’Alessandro¹, B. Doval², S. Le Beux², P. Woodruff³ and Y. Fabre¹
¹TCTS Lab, Faculté Polytechnique de Mons (Belgium), ²LIMSI-CNRS, Université Paris XI (France)

Abstract—The main purpose of this project is to develop a full computer-based musical instrument allowing realtime synthesis of expressive singing voice. The expression will result from the continuous action of an interpreter through a gestural control interface. That gestural parameters will influence the voice characteristics thanks to particular mapping strategies.

Index Terms—Singing voice, voice synthesis, voice quality, glottal flow models, gestural control, interfaces.

I. INTRODUCTION

XPRESSIVITY is nowadays one of the most challenging topics in view by the researchers in speech synthesis. Indeed, recent synthesizers provide acceptable speech in terms of intelligibility and naturalness but the need to improve human/computer intercations brought researchers to develop more “human”, more expressive systems. Some recent realizations have shown that an interesting option was to record multiple databases corresponding to a certain number of “labelled” expressions (e.g. happy, sad, angry, etc) [1]. At synthesis time, the expression of the virtual speaker is set by choosing the units in the corresponding database.

Last year, during eNTERFACE’05 [2], we decided to investigate the opposite option. Indeed, we postulated that “emotion” in speech was not the result of switches between labelled expressions but a continuous evolution of voice characteristics extremely correlated with context. Thus, we developed a set of flexible voice synthesizers “conducted” in realtime by an operator [3]. After some tests, it was clear that such a framework was particularly efficient for singing synthesis.

Remarkable achievements have been recently reached in singing voice synthesis. A review of state of the art can be found in [4]. Technology seems mature enough for replacing vocals by synthetic singing, at least for backing vocals [5] [6]. However, existing singing synthesis systems suffer from two restrictions: they are aiming at mimicking singers rather than creating new instruments, and are generally limited to MIDI controllers.

We think it worthwhile to extend vocal possibilities of voice synthesizers and design new interfaces that will open new musical possibilities. In a first attempt we decided to restrain our survey on voice quality control to the boundaries of natural voice production. As a matter of fact, it is always better trying to mimic one particular voice, as we are disposed to hear someone behind the synthesizer. This process enables to achieve analysis by synthesis : once we we are able to perceive more naturalness in the synthesized voice, this means that we understood something in voice production process. It is then easier to go astray from these limits when dealing with a musical application in a more creative way.

II. AIMS OF THE WORK

Our aims for this eNTERFACE’06 workshop can be summarized in three main axes. First, we target the implementation of intra- and inter-dimensional mappings driving low-level parameters of source models (e.g. complex interactions between vocal effort and tenseness, represented by the phonetogram). Then, we investigate the effects of the vocal tract in voice quality variations (e.g. the singer formant, lowering of the larynx). Finally, source/filter coupling effects (e.g. relations between harmonics and formants frequencies) are analysed, and several mechanisms are implemented (e.g. overtone, croatian, bulgarian, occidental singing).

III. BACKGROUND IN SINGING SYNTHESIS

Speech and singing both result from the same production system: the voice organ. However, the signal processing techniques developed for their synthesis evolved quite differently. One of the main reasons for this deviation is: the aim for producing voice is different for the two cases. The aim of speech production is to exchange messages. For singing, the main aim is to use the voice organ as a musical instrument. Therefore a singing synthesis system needs to include various tools to control (analyze/synthesize or modify) different dynamics of the acoustic sound produced: duration of the phonemes, vibrato, wide range modifications of the voice quality, the pitch and the intensity, etc. some of which are not needed in most of the speech synthesis systems. A pragmatic reason for that separation is that singing voice synthesizers target almost exclusively musical performances. In this case, “playability” (flexibility and real-time abilities) is much more important than intelligibility and naturalness. Discussions about various issues of singing synthesis can be found in [7] [8].

As described in [9], frequency-domain analysis/modifications methods are frequently preferred in singing synthesis research due to the need to modify some spectral characteristics of actual recorded signals. The most popular application of such a technique is the phase vocoder [10], which is a powerful tool used for many years for time compression/expansion, pitch shifting and cross-synthesis.

To increase flexibility, short-time signal frames can be modeled as sums of sinusoids (controlled in frequency, amplitude and phase) plus noise (controlled by the parameters of a filter which is excited by a white noise). HNM (Harmonic plus
Noise Model) [11] provides a flexible representation of the signal, which is particularly interesting in the context of unit concatenation. That representation of signals is thus used as a basis in many singing synthesis systems [12] [13] [14] [15].

Another approach is to use the source/filter model. Several models of glottal pulse has been proposed with different quality and flexibility. A complete study and normalisation of the main models can be found in [16]. For example, the R++ model has been used in the famous Voicer [17]. LF [18] and CALM [19] models have been used during eNTERFACE’05 [3]. Other differences appear in the method used to compute the vocal tract transfer function. Some systems [20] compute the formants from the magnitude spectrum: a series of resonant filters (controlled by formants frequencies, amplitudes and bandwidths). Some other systems compute an acoustic representation of the vocal tract, as a cascade of acoustic (variant-shape) tubes. For example, the SPASM synthesizer [21] uses digital waveguides [22] to model acoustic features of oral, nasal cavities and throat radiation (driven by a frequency-domain excitation model). The model was extended to variable length conical sections by Välimäki and Karjalainen [23].

There exist also some particular approaches like FOF (Formes d’Ondes Formantiques) synthesis [24], used in CHANT [25], which performs synthesis by convolving a pulse train with parallel formant wave functions (time-domain functions corresponding to individual formants resonance).

IV. V.OICE PRODUCTION

Voice organ is usually described as a "source/filter" system. Glottal source is a non-linear volume velocity generator where sound is produced by complex movements of vocal folds (larynx) under lungs pressure. A complete study of glottal source can be found in [26]. Sounds produced by the larynx are then propagated in oral and nasal cavities which can be seen as time-varying filtering. Finally, the flow is converted into radiated pressure waves through lips and nose openings (cf. Figure 1).

A. The Causal/Anticausal Linear Model (CALM) [19]

Modelling vocal tract in spectral domain (with resonant filters central frequency, amplitude and bandwidth) is very powerful in term of manipulation because spectral description of sounds is close to auditory perception. Traditionally, glottal flow has been modeled in time domain. A spectral approach can be seen as equivalent only if both amplitude and phase spectra are considered in the model.

For amplitude spectrum, two different effects can be isolated (cf. Figure 2). On the one hand, an amount of energy is concentrated in low frequencies (i.e. below 3 kHz). This peak is usually called "glottal formant". We can see that bandwidth, amplitude and position of the glottal formant change with voice quality variations. On the other hand, a variation of spectrum slope in higher frequencies (called "spectral tilt") is also related to voice quality modifications.

Considering both "glottal formant" and "spectral tilt" effects, two cascading filters can be used. A second order resonant low-pass filter ($H_1(z)$) for glottal formant, and a first order as derivative of the volume velocity signal. It is generally processed by a time-invariant high-pass first order linear filter [27]. Vocal tract effect can be modelized by filtering of glottal signal with multiple (usually 4 or 5) second order resonant linear filters.

V. THE GLOTTAL SOURCE

In this section, we describe the work related to the realtime generation of the glottal source signal. We first explain our theoretical basics: the modelization of the glottal flow as the response of a causal/anticausal linear system (CALM). Then, we will describe two different implementations achieved during this workshop: a buffered computation of a causal stable filter (v1.x) and a sample-by-sample computation of a causal unstable filter (v2.x).
low-pass filter \(H_2(z)\) for spectral tilt. But phase information indicates us that this system is not completely causal. Indeed, as it is illustrated on Figure 3, glottal pulse is a combination of a “increasing” (or active) part and a “decreasing” (or passive) part. The decreasing part, called the return phase, mainly influences the spectral tilt and hence is causal. And we can also show that the second order low-pass filter has to be anticausal in order to provide a good phase representation. This information is sometimes refered as the mixed-phase representation of voice production [32].

A complete study of spectral features of glottal flow, detailed in [19], gives us equations linking relevant parameters of glottal pulse \(F_0\): fundamental frequency, \(O_q\): open quotient, \(\alpha_m\): asymetry coefficient and \(T_i\): spectral tilt, in dB at 3000Hz) to \(H_1(z)\) and \(H_2(z)\) coefficients. Expression of \(b_1\) as been corrected, compared to [19] and [33].

\[
\begin{align*}
\text{Anticausal second order resonant filter:} \\
H_1(z) &= \frac{b_1z}{1 + a_1z + a_2z^2} \\
a_1 &= -2e^{-a_p T_e} \cos(b_p T_e) \\
a_2 &= e^{-2a_p T_e}, b_1 = ET_e \\
a_p &= -\frac{\pi}{O_q T_0 \tan(\pi \alpha_m)}, b_p = \frac{\pi}{O_q T_0} \\
\end{align*}
\]

\[
\begin{align*}
\text{Causal first order filter:} \\
H_2(z) &= \frac{b_T_L}{1 - a_T_L z^{-1}} \\
a_T_L &= \nu - \sqrt{\nu^2 - 1}, b_T_L = 1 - a_T_L \\
\nu &= 1 - \frac{1}{\eta}, \eta = \frac{1 - e^{-4T_L/\max(\nu)}}{\cos(2\pi 3000/P_e)} - 1 \\
\end{align*}
\]

Full anticausal processing is only possible offline, by running algorithms backwards on data buffers. In a realtime context, anticausal response can be processed with two differents methods. On the one hand, the response of a causal version of \(H_1(z)\) is stored backwards \((v1.x)\). On the other hand, \(H_1(z)\) is replaced by an unstable causal filter and the "divergent" impulse response is truncated \((v2.x)\). We can also note that in order to be usefull our implementations have to be able to produce correct glottal flow (GF) and glottal flow derivative (GFD). Indeed, the GFD is the acoustical signal used to synthesize the voiced sounds, but the GF is important in the synthesis of turbulence sounds, involved in unvoiced and breathy sounds.

\[\text{B. RealtimeCALM v1.x Implementation}\]

This implementation is the continuation of the development tasks of eNTERFACE’05 [3] and work presented to NIME’06 [33]. In this algorithm, we generate the impulse response by period-synchronous anticausal processing. It means that in order to achieve the requested waveform, the impulse response of a causal version of \(H_1\) (glottal formant) is computed, but stored backwards in a buffer. This waveform is truncated at a length corresponding to instantaneous fundamental frequency \((F_0+Jitter)\). This algorithm is now integrated in both Max/MSP [34] [35] and Pure Data [36] external objects (for Mac OS X, Windows and Linux): almPulse\(_{\sim}\) v1.x. Then the resulting period is filtered by \(H_2\) (spectral tilt). This algorithm is also integrated in both Max/MSP and Pure Data external objects: stFilter\(_{\sim}\) v1.x. Coefficients of \(H_1\) and \(H_2\) are calculated from equations described in subsection The Causal/Anticausal Linear Model (CALM) and [19]. Thus, both time-domain and spectral-domain parameters can be sent.

Actually, we take advantage of physical properties of glottis to propose this real-time algorithm. Indeed, glottal pulse corresponds to opening/closing movements of vocal folds. It means that impulse responses generated by \(H_1\) and \(H_2\) filters can’t overlap. Thus, impulse responses can be stored backwards and truncated period-synchronously without changing too much their spectral properties.

Truncation of the CALM waveform at each period gives quite good synthesis results. Nevertheless, several configurations of parameters (e.g. high value of \(\alpha_m\) plus low value of \(O_q\)) make the impulse response oscillating inside the period, which gives signals that are no more related to glottal source phenomena and changes voice quality perception. Thus, earlier truncation points and windowing options have been tested (e.g. first zero crossing of the GF, first zero crossing of the GFD). This study has shown us that it is not possible to set a truncation point inside the period which gives simultaneously correct synthesis results on the GF and the GFD (even with a synchronized half-Hanning window\(^1\)). This modelization problem and limitations due to the use of period buffer drove us to change the architecture of this synthesis module \((v2.x)\). Discontinuity in GFD due to GF truncation is illustrated at the Figure 4.

\[\text{C. RealtimeCALM v2.x Implementation}\]

This part explains another version of the anticausal filter response computation. It avoids the use of period buffer. Main

\(^1\)This windowing method multiplies the increasing part of the glottal pulse (flow or derivative) – meaning the part between the zero crossing and the positive maximum – by the left part of a Hanning window.
idea behind this solution was to decrease memory allocations, in order to be able to generate simultaneously the glottal flow and the glottal flow derivative, with their own truncation points and windowings.

Instead of computing a causal version of the impulse response offline and then copying it backwards into a fixed buffer, the computation is here straightforward. The iterative equation corresponds indeed to the unstable anticausal filter. Anyway, the explosion of the filter is avoided by stopping the computation exactly at the Glottal Closure Instant (GCI). We can also note that glottal flow and glottal flow derivative can both be achieved with the same iterative equation, only changing the values of two first samples used as initial conditions in the iteration process.

One other main implementation problem is that the straightforward waveform generation has to be synchronized with the standard Pure Data’s performing buffer size. This standard size is 64 samples which, at an audio rate of 44100Hz, corresponds to a frequency of approximately 690 Hz. Most of the time, the fundamental frequency of the glottal flow is less than 690 Hz, which means that several buffers are necessary to achieve the complete computation of one period. But whenever a buffer reaches the end, the main performing routine is called and thus the iterative equation is computed (signal processing, acoustics, phonetics, singing), completed, and described in a formalized set.

D. Dimensional Issues

The next step in the realization of our singing tool was to define perceptual dimensions underlying the control of voice quality, and implement analytic mapping functions with low-level synthesis parameters. Dimensional features of voice were first collected from various research fields (signal processing, acoustics, phonetics, singing), completed, and described in a formalized set.

- *Melody (F₀):* short-term and long-term elements involved in the organization of temporal structure of fundamental frequency;
- *Vocal Effort (V):* representation of the amount of "energy" involved in the creation of the vocal sound. It makes the difference between a spoken and a screamed voice;
- *Tenseness (T):* representation of the constriction of the voice source. It makes the difference between a lax and a tensed voice;
- *Breathiness (B):* representation of the amount of air turbulences passing through the vocal tract, compared to the amount of voiced signal;
- *Hoarseness (H):* representation of the stability of sound production parameters (especially for fundamental frequency and amplitude of the voice);
- *Mecanism (Mᵢ):* voice quality modifications due to phonation type involved in the sound production.

E. Description of Mapping Functions

Once dimensions are defined, two main tasks can be investigated. First, the implementation of mapping functions between these dimensions and low-level parameters. Then, identification and implementation of inter-dimensional phenomena. In this area, many different theories have been proposed relating several intra- or inter-dimensional aspects of voice production.

- [28] [41] [43] [44] [45] [46]. We decided to focus on some of them -- like direct implementation of tenseness and vocal effort, realization of a phonetogram, etc. – and design our synthesis platform in order to be able to extend it easily (e.g. correct existing relations, add new mapping functions, etc.). All current parameters are defined for a male voice.

Relations between Dimensions and Synthesis Parameters

During this workshop, we focused on several aspects of the dimensional process. First, we consider relations between a limited number of dimensions (F₀, V, T and Mᵢ) and synthesis parameters (O_q, α_m and T_i). Then, we decided to achieve our data fusion scheme by considering two different "orthogonal" processes in the dimensional control. On the one hand, vocal effort (V) (also related to F₀ variations, cf. next paragraph: *Inter-Dimensional Relations*) and mecanisms (Mᵢ) are controlling "offset" values of parameters (O_q, α_m, T_i). On the other hand, tenseness (T) controls "delta" values of O_q and α_m around their offsets (ΔO_q, Δα_m). Considering this approach, effective values of synthesis parameters can be described as:

\[ O_q = O_{q₀} + ΔO_q \]

2We can observe that our method will change the link between those two waveforms. Indeed, if two separated truncation points and windowings are applied, what we call "glottal flow derivative" is no more the derivative of the glottal flow.
\[ \alpha_m = \alpha_{m0} + \Delta \alpha_m \]
\[ T_i = T_{i0} \]

Following equations consider \( V \) and \( T \) parameters normalized between 0 and 1 and \( M_i \) representing the \( i^{th} \) phonation mecanism.

- \( O_{q0} = f(V|M_i) \)
  
  \[ O_{q0} = 1 - 0.5 \times V|M_1 \]
  \[ O_{q0} = 0.8 - 0.4 \times V|M_2 \]

- \( \alpha_{m0} = f(M_i) \)
  
  \[ \alpha_{m0} = 0,6|M_1 \]
  \[ \alpha_{m0} = 0,8|M_2 \]

- \( T_{i0} = f(V) \)
  
  \[ T_{i0}(dB) = 55 - 49 \times V \]

- \( \Delta O_q = f(T) \)
  
  \[ \Delta O_q = (1 - 2T) \times O_q + 0.8T - 0.4|T \leq 0,5 \]
  \[ \Delta O_q = (2T - 1) \times O_q + 2T + 1|T > 0,5 \]

- \( \Delta \alpha_m = f(T) \)
  
  \[ \Delta \alpha_m = (0,5T - 1) \times \alpha_m - 1,2T + 0,6|T \geq 0,5 \]
  \[ \Delta \alpha_m = (0,25 - 0,5T) \times \alpha_m + 0,4T - 0,2|T > 0,5 \]

Last adaptation on parameters concerns a perceptual distortion of \( O_q \) (square distorsion) and \( \alpha_m \) (square root distorsion) between their ranges of variation (\( O_q \): 0,4 to 1; \( \alpha_m \): 0,6 to 0,8).

**Inter-Dimensionnal Relations: the Phonetogram**

One important characteristic of human voice production is that we are not able to produce any fundamental frequency (\( F_0 \)) at any vocal effort (\( V \)). A strong relationship exists between these two perceptual features. For example, one could not produce a very low pitch (around 80Hz) at a sound pressure level higher than 80dB (for a male speaker) or conversely to produce a high pitch at low intensity. Trying to do so results in a sudden stop of vocal production. This relationship is called phonetogram, and the evolution of this dependency is varying very much from one speaker to another, considering for example that the subject is a trained singer or not, male or female, has pathological voice or not, etc. As a first approach, we decided to implement an average phonetogram, relying on the work of N. Henrich [46]. Figure 5 and Figure 6 represent two average phonetograms for male and female.

Moreover, this phenomenon involves different types of laryngeal configurations. We here dealt with mainly two configurations, first and second mechanisms of the vocal folds (\( M_1 \) and \( M_2 \)). This two laryngeal mechanisms are, in the common singing typology, referred as chest and falsetto registers. Hence, as shown on Figure 5 and Figure 6, it is also not possible to produce any frequency in both mechanisms, but the two configurations have an overlapping region in the middle of the phonetogram. This region enables the passing between the two mechanisms. Following the work presented in [47], the frequency range where this passing can occur is about one octave (or 12 semi-tones). The main characteristic of this passing is to provoke a break in the fundamental frequency (\( F_0 \)). Thus, when producing an increasing glissando from \( M_1 \) to \( M_2 \), there is an average 8 semi-tones break, whereas it is approximately 12 semi-tones when performing a decreasing glissando. Breaking intervals probabilities are depicted on Figure 7 and Figure 8. On the first one we can actually note that the frequency breaks also depends on the fundamental frequency where it occurs.

So as to say that this phenomenon introduces an hysteresis. For most of untrained speakers or singers this break is uncontrollable whereas trained singers are able to hide more or less smoothly this break, although they cannot avoid mechanism switch.

**VI. THE VOCAL TRACT**

In this section, we describe the implementation of a vocal tract model. This module is based on a physical "tubes-based" representation of vocal tract filter, which is simultaneously
controllable with geometrical (areas) and spectral (formants) parameters.

A. The lattice filter

A geometrical approach of vocal tract representation

Linear Predictive Coding [48] is a method for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. The order of the filter is related to the complexity of the envelope, and also the number of control parameters. Thus, for representing a five-formant singing vowel, a filter containing five pairs of conjugated poles (for the resonances), and two simple poles (for the glottic wave) is needed, adding up to a total of fourteen parameters.

The LPC parameters (commonly named $a_i$) are non-linearly interpolable. This implies that, for two configurations $[a_1 a_2 ... a_n]$ and $[b_1 b_2 ... b_n]$ corresponding to two vowels, a linear interpolation between both of these vectors will not correspond to a linear interpolation between the two spectra, and could even lead to unstable combinations. For these reasons, we will use another implementation of the LPC filter: the lattice filter. The control parameters of such a filter are called reflection coefficients (commonly named $k_i$). Such a filter is represented in Figure 9. It is composed of different sections, each characterized by a $k_i$ parameter.

The reflection coefficients correspond to physical characteristics of the vocal tract, which may be represented by a concatenation of cylindrical acoustic resonators, forming a lossless tube. This physical model of the lattice filter is represented in Figure 10. Each filter section represents one section of the tube; the forward wave entering the tube is partially reflected backwards, and the backward wave is partially reflected forwards. The reflection parameter $k_i$ can then be interpreted as the ratio of acoustic reflections in the $i^{th}$ cylindrical cavity, caused by the junction impedance with the adjacent cavity. This value varies from 1 (total reflection) to -1 (total reflection with phase inversion), and is equal to 0 when there is no reflection.

B. Coefficients Conversion Framework

In order to use the area parameters of the lattice filter ($A_i$), a Max/MSP object was created to convert them to $k_i$ values which are used in the lattice filter. Several sets of $A_i$ parameters corresponding to different vowels were calculated. After selecting one of these presets, certain sections of the
vocal tract can be modified by a percentage $\Delta A_i$, which has the effect of opening or closing that section of the oral cavity.

A second approach to controlling the lattice filter was considered: a formant-based scheme was used to represent the spectral envelope, and the formant features, $F_i$, were converted to $k_i$ parameters (after conversion to the LPC $a_i$ coefficients), and then to $A_i$ areas to control the lattice filter. This allowed us to easily model certain phenomena that are well known in speech processing, like overtone singing or the singer formant [49] [50], by acting on analytical parameters (the formants) rather than geometrical parameters (the areas). Similarly to the control of the areas, the formants have presets for different vowels and can be modified by a percentage $\Delta F_i$.

The parameters conversion framework described above is represented in Figure 11.

![Fig. 11. Coefficients conversion and presets/modifications framework.](image)

**VII. ABOUT THE REAL-TIME CONTROL OF VOICE SYNTHESIS**

In this section, we comment some experimentations we realized in order to evaluate expressive and performing abilities of systems we developed. Modules were integrated together inside Max/MSP and various control devices (and various combinations) were dynamically connected with mapping matrix. This set of tests allowed us to reach efficient configurations, considering several performing styles (classical singing, overtone singing, etc) which were demonstrated at the end of the workshop.

**A. Concerning Voice Source**

In order to be able to compare expressive skills of this system with the one developed before [3] [33] [37], we decided to keep the same control scheme: a graphic tablet. In that way, we were able to evaluate really clearly ameliorations achieved with this new mapping functions. Early experimentations demonstrated us that independant control of tenseness and vocal effort is really increasing performing possibilities. Anyway, current mapping equations still provide some unlikely parameters combinations, resulting e.g. in "ultra-tensed" perception or unwilling dynamics variations.

The implementation of the phonetogram is also a major improvement in term of naturalness. It also gives better results in terms of expressivity than without monitored control of loudness (more linear). Although we deeply investigated this phenomenon, we did not yet integrated this frequency break in the system, as we did not find a satisfying solution for controlling it. It is not straightforward to translate this frequency break in the control domain, as our hand gestures are mainly continuous and as basic switch from one configuration to another is not really satisfying from a musical point of view, as it results in a break in frequency range and thus "wrong" notes.

**B. Concerning Vocal Tract**

The vocal tract was controlled using a data glove (P5glove [51]) as shown in Figure 12. The glove was mapped to the area parameters of the lattice filter in four different ways:

- The folding of the fingers control the opening angle of the mouth (represented in Figure 13) (see Figure 14)
- The hand movement along the $z$-axis controls the position of the "tongue" in the vocal tract (towards the back or the front of the mouth)
- The hand movement along the $y$-axis controls the vertical position of the tongue (near or far from the palate) (see Figure 14)
- The hand movement along the $x$-axis changes the vowels (configurable from one preset to another, for example from an /a/ to an /o/)

![Fig. 12. Vocal tract control with a data glove: 5 finger flexion sensors and 3 dimensions (x,y,z) tracking.](image)
achieve refined tweaking techniques (e.g. lowering the vocal tract, changing tongue position, etc.) and that way increasing expressivity.

C. Transversal Remarks

In overall, at this stage of development, the synthesizer allows to control 17 parameters which are namely: pitch, vocal effort, tenseness, mechanisms, the first two formants, the singer’s formant, vocal tract length, gain, transition between vowels, width of the vocal tract, position of the tongue and mouth opening (5 parameters). Considering all these parameters, only the actions on mechanisms is not a continuous parameter, so as to say that 16 parameters have to be monitored thanks to continuous parameters. From the controllers side, we have all in all 17 continuous parameters (out of 33), meaning that we are actually theoretically able to control all needed parameters. However, the problem is that from user’s side, it is impossible to manipulate three interfaces at the same time. There are actually two solutions: one is to have multiple users (2 or 3) being in control of the interfaces, the other one is to use one-to-many mappings, allowing the performer to control several parameters with the same controller.

VIII. Conclusions

In this workshop, our main aim was to build a performant singing musical instrument allowing a wide range of expressive possibilities. Our actual work results in the implementation of new models for voice source and vocal tract, working in real-time, which are strategic tools in order to be able to work further. Improvement of expressivity in this new system really encourage us to go forward with this approach. Moreover, our modular architecture drives us to go to a widely extensible synthesis platform which will be really useful in order to continue to integrate other results (existing and coming) from voice production sciences.

Acknowledgment

The authors would like to thank SIMILAR Network of Excellence (and through it the European Union) which provides resources allowing researchers from all Europe to meet, share and work together, and then achieving really exciting results. We also would like to thank croatian organization team (responsible: Prof. Igor Pandzic) which maintain local structures in order to manage the work and the life of more than 50 people. Finally, we would like to thank our respective laboratories (in our case: TCTS Lab, Mons, Belgium and LIMSI-CNRS, Paris, France) which adapt their research agendas in order to allow us to participate to those annual summer events.
Nicolas D’Alessandro holds an Electrical Engineering degree from the Faculté Polytechnique de Mons (FPMs) since 2004. I did the master’s thesis in the Faculty of Music of the University de Montréal (UdeM) (supervisor: Prof. Caroline Traube). That work gathered the development of applications based on perceptual analogies between guitar sounds and voice sounds, and a study of mapping strategies between (hand-based) gestures and speech production models (source/filters and concatenative approaches). He started a PhD thesis in September 2004 in the TCTS Lab of the FPMs (supervisor: Prof. Thierry Dutoit) related to the real-time control of unit-based synthesizers.

Boris Doval holds an Engineering degree from the "Ecole Centrale de Paris" since 1987. He did his master thesis at IRCAM and his PhD thesis at "Université Paris VI" on fundamental frequency estimation of sound signals. He then joined LIMSI-CNRS as an associate professor where he concentrated until now on voice source analysis, synthesis and modelisation. In particular, he coorganized the ISCA workshop VOQUAL’03 on voice quality in 2003.

Sylvain Le Beux graduated from Master in Electronics, Telecommunications and Informatics from CPE Lyon engineer school in 2004. During his training he did an internship at Infineon Technology A.G. in Munich for one year, and achieved his master’s thesis at IRCAM which topic was about speech recognition. So he actually helped IRCAM wreck on a nice beach using calm insense. He then graduated from a Master thesis on embedded systems and data processing from Orsay University in 2005, and then integrated LIMSI Laboratory where he is currently achieving his PhD focused on the gestural control of speech synthesis and relationship between intention and expressivity.

Pascale Woodruff holds an Electrical Engineering degree from FPMs since June 2004. She is currently working on a project which aims to improve the workflow in industrial maintenance by equipping technicians with a multimodal wearable system allowing them to access maintenance information using speech and/or other modalities.

Yohann Fabre is ending a master thesis in audiovisual technologies at the ISIS (Ingénierie des Systèmes, Image et Son) of Valenciennes. He is currently working as a trainee on voice synthesis at the TCTS Lab of FPMs.