Real-time CALM Synthesizer
New Approaches in Hands-Controlled Voice Synthesis

N. D'Alessandro
TCTS Lab (FRIA Researcher)
Faculté Polytechnique de Mons
B-7000 Mons, Belgium
nicolas.dalessandro@fpms.ac.be

C. d'Alessandro, S. Le Beux, B. Doval
LIMSI - CNRS
Université Paris Sud XI
F-91403 Orsay, France
{cda, slebeux, boris.doval}@limsi.fr

ABSTRACT

In this paper, a new voice source model for real-time gesture–controlled voice synthesis is described. The synthesizer is based on a causal-anticausal model of the voice source, a new approach giving accurate control of voice source dimensions like tenseness and effort. Aperiodic components are also considered, resulting in an elaborate model suitable not only for lyrical singing but also for various musical styles playing with voice qualities. The model is also tested using different gestural control interfaces : data glove, keyboard, graphic tablet, pedal board. Depending on parameter-to-interface mappings, several instruments with different musical abilities are designed, taking advantage of the highly expressive possibilities of the synthesis model.

Keywords

Singing synthesis, voice source, voice quality, spectral model, formant synthesis, instrument, gestural control.

1. INTRODUCTION

Remarkable achievements have been recently reached in singing voice synthesis. A review of state of the art can be found in [1]. Technology seems mature enough for replacing vocals by synthetic singing, at least for backing vocals [2] [3]. 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. On the one hand, a voice synthesizer should be able to reproduce several voice quality dimensions, resulting in a wide variety of sounds (e.g. quasi-sinusoidal voice, mixed periodic aperiodic voice, pressed voice, various degrees of vocal effort, etc.). On the other hand, vocal instrument being embodied in the singer, multidimensional control strategies should be devised for externalizing gestural controls of the instrument.

In this paper, a new elaborate voice source model able to produce various voice qualities is proposed. It is based on spectral modelling of voice source [4]. Links between spectral parameters and auditory effects are relatively straightforward. Then playing instruments based on spectral modelling seems very intuitive. Another key point is gesture-to-parameter mapping. Following the pioneering work by Fels [5], we found data glove particularly well suited to vocal expression. Recent work on hand-controlled vocal synthesis include series of instruments presented by Cook [6] and the Voicer by Kessous [7]. It must be pointed out that musical possibilities offered by an instrument strongly depend on mapping and interfaces. Then, depending on intended musical aims, different instruments are proposed. This paper is organized as follows. In section 2, the voice synthesis model is reviewed. In section 3, control devices and mapping of voice quality dimensions onto control parameters are discussed. Section 4 presents two musical instruments built on basis of synthesis model and vocal dimensions. Section 5 presents a discussion of results obtained and proposes directions for future works.

NIME 06, June 4-8, 2006, Paris, France

2. VOICE SYNTHESIS MODEL

In this section, we first give an overview of mechanisms involved in voice production. Then, we focus on the glottal source and present the causal-anticausal linear model developed by d'Alessandro/Doval/Henrich in [4]. We also explain the nature of non-periodical components we introduced in the model. Finally, we describe structure and possibilities of the real-time glottal flow synthesizer based on CALM (RT-CALM) we developed and integrated in following singing instruments.

2.1 Voice 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 [8]. Sounds produced by the larynx are then propagated in oral and nasal cavities which can be seen as time-varying filtering. Finally, the volume velocity flow is converted into radiated pressure waves through lips and nose openings (cf. Figure 1).

In the context of signal processing applications, and particulary in singing synthesis, some simplifications are usually accepted. First, lips and nose openings effect can be seen as derivative of the volume velocity flow. It is generally processed by a time-invariant high-pass first order linear filter [9]. Vocal tract effect can be modelized by filtering of glottal signal with multiple (4 or 5) second order resonant linear filters (formants).

Contrary to this "standard" vocal tract implementation, plenty of models have been developed for represen-

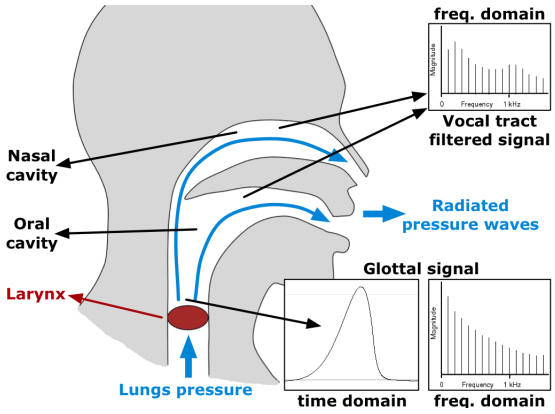

Figure 1: Voice production mechanisms : vocal folds vibrations, vocal tract filtering and lips/nose openings radiation.

tation of glottal flow, with differences in accuracy and flexibility. Usual models are KLGLOTT88 [10], R++ [11], Rosenberg-C [12] and LF [13] [14]. We present now the causal-anticausal linear model (CALM) [4], explain why we worked with this spectral approach and propose adaptations of the existing algorithm to ease real-time manipulation.

2.2 CALM : causal-anticausal linear model

We have seen that 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 modelized in time domain. A spectral approach can be seen as equivalent only if both amplitude and phase spectra are considered in the model.

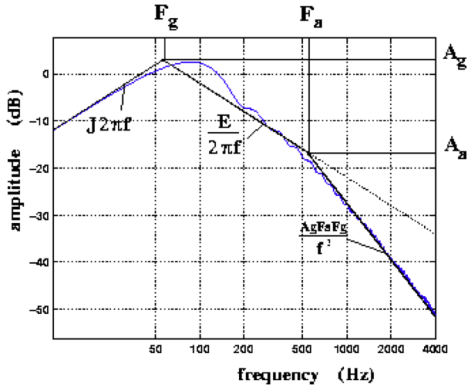

Figure 2: Amplitude spectrum of the glottal flow derivative : illustration of glottal formant (F_g, A_g) and spectral tilt (F_a, A_a).

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.

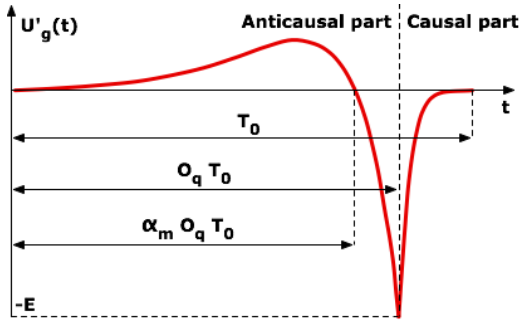

Figure 3: Time-domain representation of derivated glottal pulse : anticausal part and causal part.

Considering both "glottal formant" and "spectral tilt" effects, two cascading filters are implemented. A second order resonant low-pass filter (H_1) for glottal formant, and a first order low-pass filter (H_2) 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.

A complete study of spectral features of glottal flow, detailed in [4], gives us equations linking relevant parameters of glottal pulse (F_0 : fundamental frequency, O_q : open quotient, α_m : asymetry coefficient and T_l : spectral tilt, in dB at 3000Hz) to H_1 and H_2 coefficients. Note that expression of b_1 has been corrected. [4] also contains equations linking this time-domain parameters with spectral-domain parameters.

Anticausal second order resonant filter :

$$H_1(z) = \frac{b_1 z}{1 + a_1 z + a_2 z^2}$$

where :

$$a_1 = -2e^{-a_p T_e} cos(b_p T_e),\ a_2 = e^{-2a_p T_e}$$

$$b_1 = \frac{E}{b_p} e^{-a_p T_e} sin(b_p T_e)$$

$$a_p = -\frac{\pi}{O_q T_0 tan(\pi \alpha_m)},\ b_p = \frac{\pi}{O_q T_0}$$

Causal first order filter :

$$H_2(z) = \frac{b_{T_L}}{1 - a_{T_L} z^{-1}}$$

where :

$$a_{T_L} = \nu - \sqrt{\nu^2 - 1},\ b_{T_L} = 1 - a_{T_L}$$

$$\nu = 1 - \frac{1}{\eta},\ \eta = \frac{\frac{1}{e^{-T_L/10ln(10)}} - 1}{cos(2\pi \frac{3000}{F_e}) - 1}$$

2.3 Non-periodical components

As described theoretically in [4], the glottal flow is a deterministic signal, completely driven by a set of parameters. Adding naturalness involves the use of some random components we propose to describe.

Jitter

Jitter is a natural unstability in the value of fundamental frequency. It can be modelized by a random value (gaussian distribution, around 0 with variance depending on the

amount of jitter introduced), refreshed every period, added to the stable value of fundamental frequency.

Shimmer
Shimmer is a natural unstability in the value of the amplitude. It can be modelized by a random value (gaussian distribution, around 0 with variance depending on the amount of shimmer introduced), refreshed every period, added to the stable value of amplitude.

Turbulences
Turbulences are caused by additive air passing through vocal folds when glottal closure is not complete. It can be modelized by pink noise filtered by a large band-pass (tube noise), modulated in amplitude by glottal pulses.

We can note here that we kept a direct control on irregularities (based on *Jitter*, *Shimmer* and *Turbulences* rates). Other models were developed, involving granular synthesis coupled with self-organizing dynamic systems [15], and could be considered in further works.

2.4 RT-CALM framework

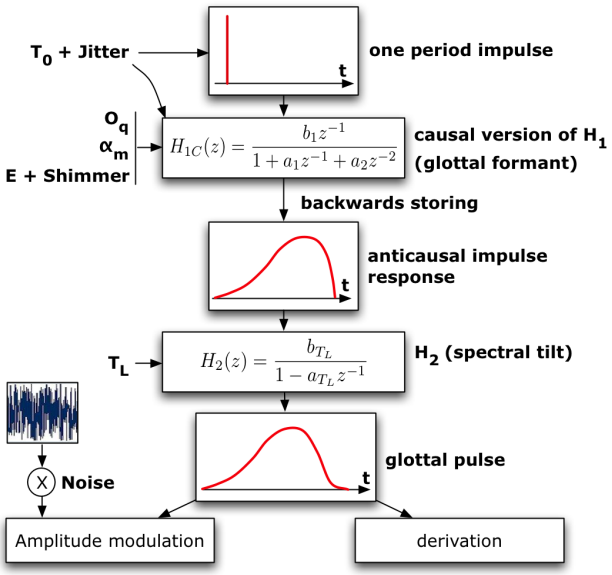

Figure 4: Framework of RT-CALM algorithm, allowing real-time synthesis of glottal pulses based on causal-anticausal linear model.

Full anticausal processing is only possible offline, by running algorithms backwards on data buffers. Anyway, in this context, we can take advantage of physical properties of glottis to propose a 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, if ranges of parameters are correctly limited, impulse responses can be stored backwards and truncated period-synchronously without changing too much their spectral properties.

To achieve the requested waveform, impulse response of causal version of H_1 (glottal formant) is computed, but stored backwards in the buffer. This waveform is truncated at a length corresponding to instantaneous fundamental frequency (F_0+*Jitter*). Then the resulting period is filtered by H_2 (spectral tilt). Coefficients of H_1 and H_2 are calculated from equations described in subsection 2.2 and [4]. Thus, both time-domain and spectral-domain parameters can be sent. On the one hand, glottal pulses are derivated to produce pressure signal (cf. Figure 3). On the other hand, it is used to modulate the amount of additive noise. Complete RT-CALM algorithm is illustrated at Figure 4.

3. VOICE QUALITY DIMENSIONS

Voice synthesis model is driven by a set of low-level parameters. In order to use these parameters in singing, they must be organized according to musical dimensions. Mappings between parameters and dimensions, and between dimensions and controllers are essential parts of instrument design. In this section, we describe main musical dimensions for voice source (cf. Figure 5) and vocal tract (cf. Figure 6).

3.1 Glottal source

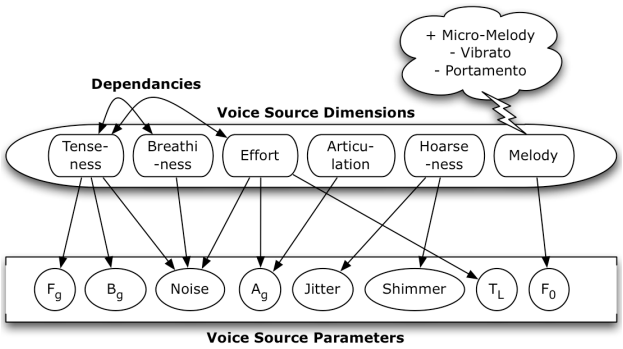

Figure 5: Mapping of the vocal source

Melodic dimension
For singing, this dimension can be decomposed into two parts. On the one hand, it seems important to sing in tune i.e. to make use of notes with well-defined pitches. On the other hand, micro-melodic variations are essential for expressive and natural singing (portamento, vibrato, etc.). Two different controls seem necessary for melodic dimension. This dimension mainly depends on parameter F_0. Anyway, a more precise vibrato synthesis should also involve amplitude variations.

Hoarseness dimension
This dimension is linked to structural aperiodicities in voice source, like *Jitter* and *Shimmer*.

Breathiness dimension
This dimension is linked to aspiration noise in voice source. It controls the relative amount of voicing vs. whispering, using the *Noise* parameter.

Pressed/lax dimension
This dimension is mainly linked to the position of the glottal formant F_g and its bandwidth B_g. It is often linked to breathiness and vocal effort. The pressed/lax dimension is used in some styles of singing e.g. Japanese noh theater or belt singing.

Vocal effort dimension
This dimension is linked to spectral tilt T_l and of course to gain parameter A_g.

3.2 Vocal tract

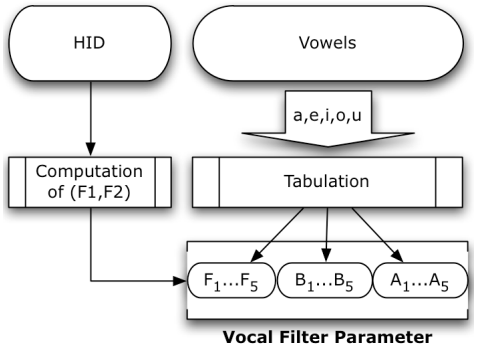

Figure 6: Mapping of the vocal tract

Vocalic space
This space is defining vocal genre (male/female/child), phonemes, and other expressive features (lips rounding, lips spreading, tongue position). This space can also be used for harmonic singing. The vocalic space is defined by formant parameters F_1, B_1, A_1, F_2, B_2, A_2,... F_N, B_N, A_N.

Articulation dimension
Finally, notes attacks and decays are controlled by an articulation dimension. "Articulation" is taken here in its musical meaning i.e. transitions between notes. It is essentially controlled by gain parameter A_g.

3.3 Musical control of vocal dimensions

Playing with melody
Melodic playing usually requires precise pitches. Then "selection" gestures are needed using e.g. a keyboard. However, natural vocal note transitions are generally slow, with more or less portamento and vibrato. Small and controlled pitch variations are therefore needed, and the "selection" gesture must be accompanied by a "modification" gesture, using e.g. hand position in one dimension of space. Another elegant solution offering accurate pitch control and smooth micro-melodic variation is using a graphic tablet. A virtual guitar board can be emulated this way. Well tuned pitches are not required in some singing styles imitating speech, like Sprechgesang (parlar cantando). Then only one control gesture is needed, that can be achieved by position of hand in one spatial dimension.

Playing with timbre : vocalic space
Playing with vocalic timbre is often used on slow moving melodies e.g. harmonic singing. The basic vocalic space needs two dimensions for contrasting vowels e.g. a joystick or a graphic tablet. One dimension is sufficient for harmonic singing (moving only second formant frequency), using a slider or position of hand in one spatial dimension. But a third dimension would be needed for signaling facial movements like lips spreading or rounding, using e.g. a data glove.

Playing with timbre : noise and tension
Some musical styles are also playing with noise and tension. These parameters are moving relatively slowly, on a limited scale, and gestures must not be extremely precise. They can be naturally associated to flexion of fingers in a data glove.

Playing with articulation and phrasing The data glove proved also useful for articulation (in the musical meaning of note attack and release) and phrasing. Hand movements in space are well suited to phrasing and finger flexions are well suited to articulation.

4. CALM-BASED INSTRUMENTS

This section describes two setups we realised. Main purpose of this work was to realize extensive real-time tests of our CALM synthesis model and voice quality dimensions mappings. No dedicated controllers were designed for this purpose. Only usual devices such as tablets, joyticks or keyboards were used.

4.1 Instrument 1

In this first instrument implementation, we use a keyboard to play MIDI notes in order to trigger the vowels at different tuned pitches. Thus, by using keyboard, we are able to set glove free for fine tuning of F_0 so as to achieve vibrato, portamento of other types of melodic ornaments. Accurate control of F_0 by glove position alone proved difficult because well tuned notes references were missing, due to approximative nature of hand gestures.

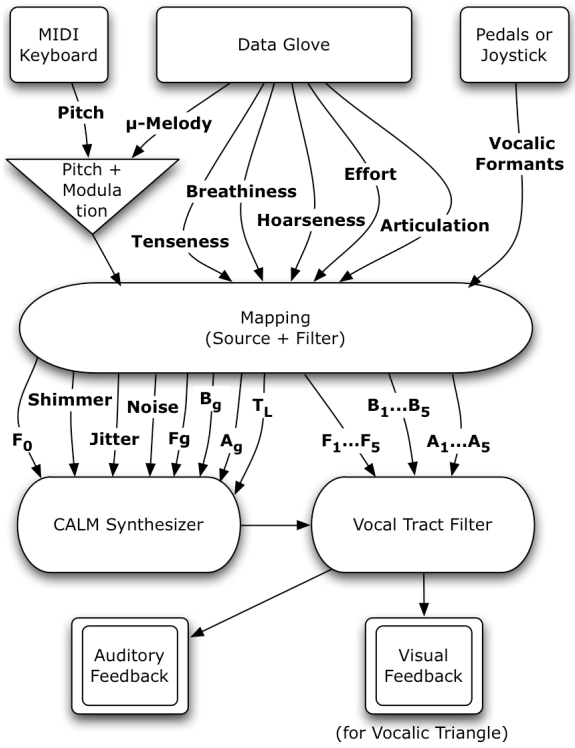

Figure 7: Structure of Instrument 1

Tempered notes (or other conventions) delivered by keyboard can be modified to a certain extent, thanks to tracking of glove position along a certain axis (transversal axis gives better ergonomics as one don't have to fold the elbow to achieve vibrato). General gain is mapped onto longitudinal axis of the glove. Then both vibrato and amplitude envelopes of sound can be produced by circular hand movements. Other vocal dimensions are controlled by flexion of

data glove fingers. First finger controls vocal effort (spectral tilt), second finger controls breathiness (linked to additive noise), third finger control the pressed/lax dimension (linked to the glottal formant), fourth finger controls hoarseness (linked to jitter and shimmer). Voice quality modifications are achieved by closing/opening movements of whole hand or selected fingers. Preset vowels are associated to keys of computer keyboard. Vowel formants can also be modified by additional devices, like pedal board or joysticks.

In summary, for this first instrument :

1. left-hand controls the keyboard (tempered notes)
2. right-hand movements control both fine pitch modulation, and note phrasing.
3. right-hand fingers control tension, effort, hoarseness, and breathiness.

In this implementation, note phrasing results of relatively large hand movements. An alternative solution is to couple effort and note phrasing in fingers movements, and to keep one dimension of hand movement for controlling another vocal dimension (e.g. breathiness). Then, phrasing is controlled by smaller and quicker finger movements. Overall description of this instrument and its various components is illustrated on Figure 7.

4.2 Instrument 2

The key point of this second instrument is simplicity of learning and using. Different choices have been made to achieve that result. First, we decided to focus on voice quality. Vocal tract control would be limited to vowel switching. Then, we took advantage of our natural writing abilities to map all glottal flow features only on tree dimensions of a graphic tablet (x axis, y axis and pressure).

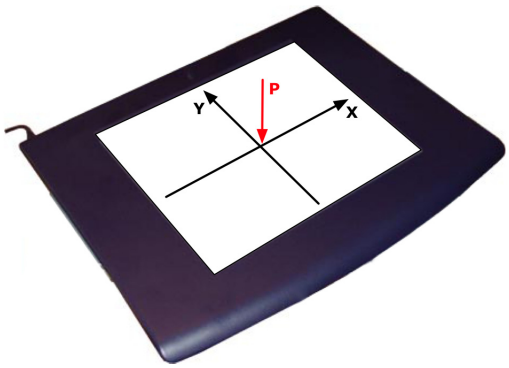

Figure 8: Mapping on the graphic tablet. X axis : fundamental frequency, Y axis : pressed/lax and vocal effort dimensions, Pressure (P) : general volume.

As described on Figure 8, horizontal axis is mapped to fundamental frequency. Tests have been made showing that, after a few training, 2 or 3 (even 4) octaves can be managed on a *Wacom Graphire* tablet. Anyway, transposition and surface scaling features have been implemented. Vertical axis control both pressed/lax and vocal effort dimensions. Mapping is made by using Y value as an interpolation factor between two different configurations of parameters O_q, α_m and T_L, from a "quiet" voice to a "tensed" voice (cf. Figure 9). Finally, pressure parameter is mapped to the gain (E).

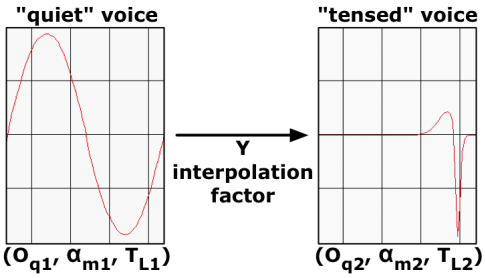

Figure 9: Interpolation between "quiet" voice and "tensed" voice made by Y axis of the graphic tablet.

Regression of voice quality control on an overall expressive axis makes main manipulations of voice source possible with simple "drawings" (i.e. bidimensional + pressure shapes). This compromise makes this instrument really intuitive. Indeed, as it can be done e.g. with a guitar, interpreter only needs graphic tablet to play. MIDI controller (e.g. pedal board) is just used for changing presets (cf. Figure 10).

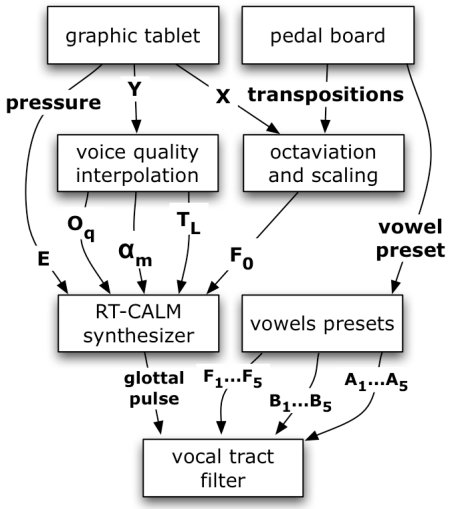

Figure 10: Structure of Instrument 2.

5. CONCLUSIONS AND FUTURE WORK

The two instruments implemented so far are suitable for musical use. Instruments have a truly human sound, and new possibilities offered by gestures to sound mapping enable intuitive playing. Compared to other voice synthesis systems, more emphasis is put on voice quality controls. It is then possible to play with expressive musical dimensions inherent to wind instruments, like effort, pressure and noise. These dimensions are exploited in acoustic instruments like saxophones and brass, and of course voice, but are generally ignored in singing synthesis. Hand movements in space and hand/fingers closures/openings are intuitively associated to such dimensions as effort or voice pressure.

Another challenging point for singing synthesis is accurate yet flexible F_0 control, like in fretless string instruments. This has been implemented in two ways in our

instruments (graphic tablet and glove controlled F_0). This flexible F_0 control enables the player all possible types of intonation, from singing to speech. Melodic ornaments like e.g. vibrato or portamento are easily controlled.

Spectral processing of voice quality proved also useful for "spectral" singing styles. Overtone singing, formant melodies, various types of throat singing are easily produced and controlled in this framework.

Instruments can also be considered as tools for studying singing, because they produce very natural sounding and controlled signals. Then they can be used for investigating musical gestures involved in singing.

Apart from the two instruments presented here, we are also investigating other types of data gloves and elaborated 3D joysticks for refining control of the synthesizer. However, this will not change the nature and number of useful vocal dimensions, but improve precision and ergonomics.

Of course, singing is an instrument that mixes together music and language. Thus, our next challenge is to control the "speech" part of singing. This point has been only marginally considered in the present research and will be the object of future work. Addition of speech articulations would drive us to more accurate modelization of vocal tract, eventually based on existing databases. Considering interfaces, syntaxic abilities of controllers have to be determined in order to achieve syllables, words or sentences synthesis.

6. ACKNOWLEDGMENTS

This work originated from the Speech Conductor project, a part of the eNTERFACE'05 workshop organized by Prof. Thierry Dutoit (Faculté Polytechnique de Mons) within the SIMILAR Network of Excellence (European Union - FP6). We would like to thank all these institutions that provided excellent working conditions.

7. REFERENCES

[1] M. Kob, "Singing Voice Modelling As We Know It Today," *Acta Acustica United with Acustica*, Vol. 90, pp. 649–661, 2004.

[2] Virsyn Corporation, "The Cantor Singing Synthesis Software," 2005-present, url : http ://www.virsyn.de/

[3] Yamaha Corporation, "The Vocaloid Singing Synthesis Software," 2003-present, url : http ://www.vocaloid.com/

[4] B. Doval, C. d'Alessandro and N. Henrich, "The Voice Source as an Causal-Anticausal Linear Filter," *Proc. ISCA ITRW VOQUAL'03*, Geneva, Switzerland, August 2003, pp. 15–19.

[5] S. Fels and G. Hinton, "Glove-TalkII : A Neural Network Interface which Maps Gestures to Parallel Formant Speech Synthesizer Controls," *IEEE Transactions on Neural Networks*, Vol 9, No. 1, pp. 205–212, 1998.

[6] P. Cook, "Real-Time Performance Controllers for Synthesized Singing," *Proc. NIME 2005*, Vancouver, Canada, May 2005, pp. 236–237

[7] L. Kessous, "Gestural Control of Singing Voice, a Musical Instrument," *Proc. of Sound and Music computing 2004*, Paris, October 20-22, 2004.

[8] N. Henrich. *"Etude de la source glottique en voix parlée et chantée."* Thèse de doctorat à l'Université Paris VI, 2001.

[9] G. Fant. *"Acoustic Theory of Speech Production"*, Gravenhage, 1960.

[10] D. Klatt and L. Klatt, "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers," *J. Acoust. Soc. Am.*, 87(2) :820–857, 1990.

[11] R. Veldhuis, "A Computationally Efficient Alternative for the Liljencrants-Fant Model and its Perceptual Evaluation," *J. Acoust. Soc. Am.*, 103 :566–571, 1998.

[12] A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," *J. Acoust. Soc. Am.*, 49 :583–590, 1971.

[13] G. Fant, J. Liljencrants, and Q. Lin, "A Four-Parameter Model of Glottal Flow," *STL–QPSR*, 85(2) :1–13, 1985.

[14] G. Fant, "The LF-Model Revisited. Transformations and Frequency Domain Analysis," *STL–QPSR*, 2–3, 119–56, 1995.

[15] E. R. Miranda, "Generating Source-Streams for Extralinguistic Utterances". *Journal of the Audio Engineering Society (AES)*, Vol. 50 N°3, March 2002.