GESTURAL CONTROL OF SINGING VOICE, A MUSICAL INSTRUMENT

Loic Kessous
Laboratoire de Mecanique et d'Acoustique - CNRS
Chemin Joseph Aiguier, 13008, Marseille, France

ABSTRACT

This paper describes research and experiments concerning digital musical instruments based on gestural control of singing voice synthesis. The purpose is to choose and use a two-handed control, a synthesis model and an adequate mapping strategy to allow an expressive pitch control and the articulation of the vowels. A visual feedback is provided to help the performer. By using different models, different aspects of synthetic singing voice as naturalness, expressiveness and vocal identity are explored in relation with gestural control. Several controllers have been tested to achieve a choice of controllers for the musical functionalities needed in this instrument.

1. INTRODUCTION

Today the composer can not only write notes, nuances and instrumental techniques, but also timbres and spectral evolutions. The performer can play, interpret, and possibly improvise the sound itself. One musical esthetic choice can be to not privilege pitch instead of spectral consideration or inversely but to find a way to have an optimized control for their cohabitation. A vowel singing voice synthesizer is a typical example of such a configuration: controlling the pitch precisely, modulating it expressively and modifying spectrum characteristics to articulate the vowels. Expressive gestural control of Singing voice synthesis is a challenge in itself but can also be a model for other digital musical instruments dealing with spectrum manipulation and expressive melodic phrasing.

2. SINGING VOICE SYNTHESIS

Since the beginning of voice synthesis some distinctions between singing voice synthesis and speech synthesis has been done. One of the more evident characterizations of the singing voice is the use of vibrato. Sound examples from the beginning of voice synthesis, like the Voder's demonstration, or Max Mathews's synthesis of "bicycle for two" [8] illustrate it evidently. A very popular synthesis method for voice is the source-filter method. The source models the vocal cords vibration and the air's flow turbulences. The filter models the vocal tract's resonances. The filter induces a formant structure to the spectrum, which is changing from a vowel to another. Sundberg has suggested the presence of a supplementary formant, "the singer formant". A lot of implementation structures are possible for voice synthesis. Choosing a particular one is a very discussed subject. One can use a high order filter, a bank filter (in cascade, parallel or lattice form), FOF or FM[3] and many others. When filters are used one can use also different implementations for the voiced source. One can use a sawtooth signal, a Dirac pulse train, a glottal signal model (LF, R++[9]) or a Physical model of glottis (mass string model). I will not try to define which of this implementation is better than another; I think it depends more on the musical application than in the intrinsic characteristics. Different implementations are presented here to explore different aspects of real-time control of singing voice. In particular, I am trying to link gesture to aspects like expressiveness, naturalness and "vocalness".

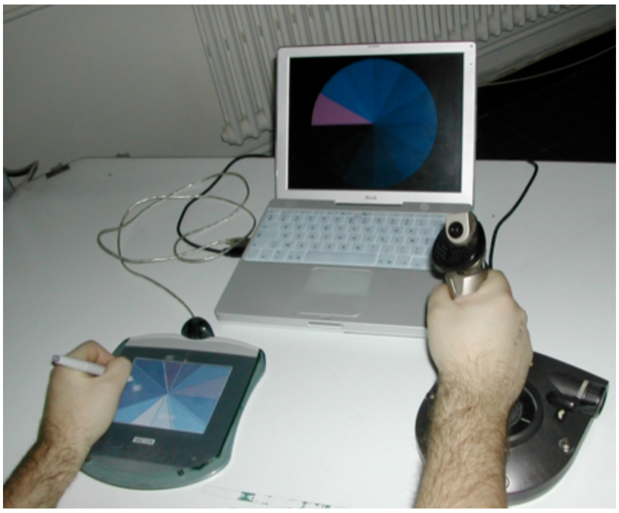

Figure 1. 'Joystick+Tablet' version of the Voicer

3. EXPRESSIVENESS, NATURALNESS AND "VOCALNESS"

3.1. Concept and definition

Expressiveness is the capacity to express something. It can be an emotion, a sentiment, a message and probably many others things. Design of digital instrument must allow enough flexibility and precision to allow performers to introduce nuances in their play with the most possible precision. We need the best time and data precision to have an expressive control, and particularly for continuous control. This is depending on the choice or design of controllers, the choice or implementation of communication tools and protocols between controllers and computer (even if there is other hardware possibilities) and also, and that is maybe the most important one, choice and implementation of a mapping strategy.

One can also use the term ‘expressiveness’ to suggest the ability of an instrument to be used to play different styles of music. That can mean different tone scales, tempered or not, microtonal or just scales, but also different styles, different ways to articulate the whole phrase in term of time, energy or spectrum. In this case expressiveness is correlated with the capacity for an instrument to allow the performer to adapt his/her play to a context.

Naturalness is the attribute which quantifies the possibility of a voice heard to be a human real voice or a synthesized voice. One can also evaluate if the voice is near from a analyzed voice.

‘Vocalness’ can be defined as a signature of the voice, a vocal identification factor, which let us say if what we heard is a voice (natural or synthetic) or not. A Theremin has an expressiveness near to the expressiveness of the voice but we can perceive a difference related to the absence of formant and maybe more precisely formant articulation (a dynamic process). This signature is also present in the framework of audio effects. A wah-wah pedal, connected to a guitar, makes it sing by using a filter (depending of the implementation it can be just one filter, two or three filters modulated in one dimension of articulation). A simple model using only three filters can be perceived as a singing voice but will not if the articulation is not properly done, in this case one will perceive three filter changing of center frequency instead of a vowel articulation. It sometimes sounds like a diphonic singer and shows the importance of the interaction between sound perception and gestural control.

3.2. Experimenting different aspects with different synthesis models

The first model is made of a sawtooth signal filtered by three second-order filters in cascade. This implementation illustrates the robustness of the vocalness in term of vowel perception; three formants are enough to identify product of spectral response of filters as vowels. It also shows the importance of continuous control of pitch in singing voice and the importance to access by gesture to a precise and fine control, not obligatory by separating drastically the generic note (as a MIDINOTE message generated by a keyboard) and the pitch modulation (as the PITCHBEND message generated by a pitch wheel). A precise continuous pitch control is essential to reproduce singing voice expressiveness.

The second model is made of 5 FOFs. I chose to use FOFs principally because the Max/MSP CHANT Library provides externals to link by rules fundamental frequency and formant frequencies. The choice was also done to use easily the dictionary of 5 formants values for soprano singer. Another Max external allows the emulation of a vocal effort. As it reproduces and adds transformation to real voice spectrum analysis this implementation is, in a first step, the best candidate to have a natural singing voice.

The third model is made of a R++ glottal source model connected to a filter bank. With this model I can connect gesture to open quotient, asymmetry coefficient, and release time of the glottal pulse, which have spectral implications and are expressive control parameters. With this implementation we can focus separately on expressiveness factors related to the glottal source and to the vocal tract. Actually, an interaction between the vocal tract and the glottal source should be taken in consideration but can be taken apart in a first time.

4. DESIGN OF A MUSICAL INSTRUMENT

4.1. Mapping strategies

With acoustical instruments, designers try to find the best compromise between the abilities of human body and the physical constraints due to acoustic sound generation. The gestures used on acoustical instruments strongly depend on the physical properties of the instrument. Digital musical instruments don’t have physical constraints due to the sound generation: the designers of such instruments are free to choose the gestures they want and to choose the way they want to link these gestures to the sound.

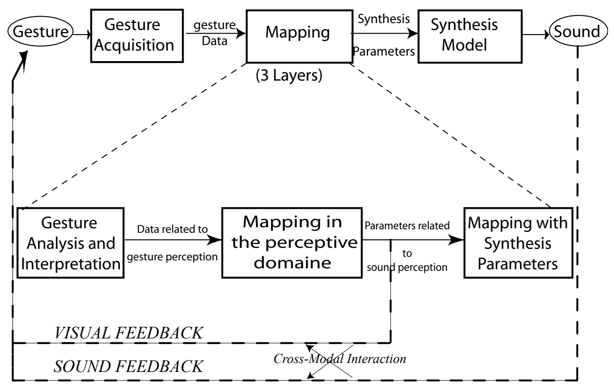
Gesture
Gesture Acquisition
gesture Data
Mapping
(3 Layers)
Synthesis Parameters
Synthesis Model
Sound
Gesture Analysis and Interpretation
Data related to gesture perception
Mapping in the perceptive domain
Parameters related to sound perception
Mapping with Synthesis Parameters
VISUAL FEEDBACK
SOUND FEEDBACK
Cross-Modal Interaction

Figure 2. A three layer mapping chain digital instrument design

This link, usually called the mapping, represents an important field of research in computer music [6]. Although a mapping strategy should be considered as a whole, I will first consider the articulation parameters, the pitch and the loudness as completely separable variables. The model of mapping used is a three layer based on considerations related to perception [1].

4.1.1. Mapping Strategy for pitch control

To control the pitch in a large register, a circular and incremental strategy is used. To allow control within one octave and from one to the other, we divide the tablet's active space into 12 sectors (12 equal angular parts where each part corresponds to a semitone of the chromatic scale). One can find here a metaphor also used in the literature to describe pitch perception. The transition between two sectors is continuous but not mandatory linear. The user can specify the power of transition (figure 1). This facilitates gestures such as portamento or vibrato because tuning control is more powerful on limits of the angular sectors, according to the transition power set by the user. Using a non-linear transition helps for vibrato and others pitch modulation gestures. The pitch control is continuous and circular, turning clockwise changes pitch from low to high. Crossing the origin in the positive direction allows controlling the pitch in the higher octave. Inversely, crossing the origin in the negative direction allows controlling the pitch in the lower octave. This possibility is only active when the middle button of the joystick is pressed; this prevents to have a non-desired octave jump.

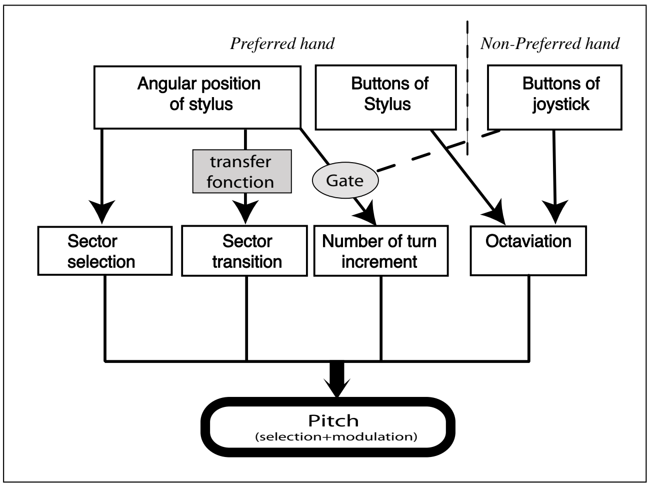

Figure 3. Mapping used for pitch control

The user can also go from a note to its lower or higher octave by pressing the stylus lateral button up or down. One can also use the left and right joystick buttons located under the thumb. The first mapping step is to define, in which of the twelve angular parts is the pen. Then we look if it is more or less centered on it. Finally we need to know how many turns have been made around the center of the tablet. I have tried to respect the natural relation between distal-to-proximal consideration and precision needed. This pitch control strategy allows, for example, to make a glissando with a large range and to finish his gesture with a vibrato. This is not so easy, actually quasi-impossible, with a keyboard and a pitch modulation wheel but that's a musical gesture currently used by singers (and also other instruments). Jitter of fundamental frequency can also be controlled, for example by using the rotation axis of the joystick. One can also map jittering amplitude with the inverse of the force pressure applied on the stylus, related to the catch intensity of the stylus. This can be measured with a FSR pressure sensor fixed on the stylus.

4.1.2. *Mapping Strategy for articulation control*

When one wants to articulate between two vowels, the most evident way to do is to change the formant frequencies from a vowel to another. In some case one can observe that the "road" used can sound non-natural. A natural way is to articulate by acting on two dimensions related to tongue hump position and degree of constriction of the vocal tract [4]. This is done by using a 2 dimensional navigation by the way of an interpolator where lists of parameters associated to vowels are adequately positioned in the interpolation plan.

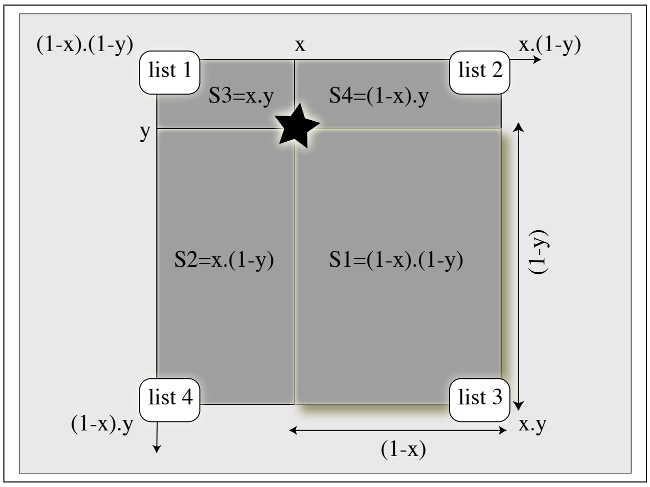

Figure 4. Vowel articulation by using 'surfacic' interpolation

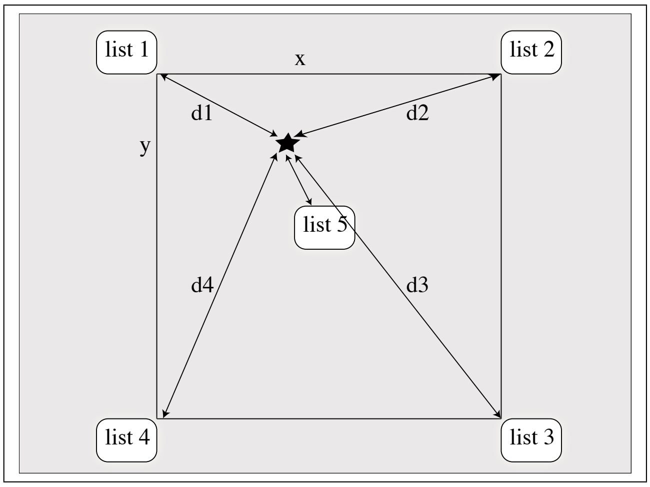

Figure 5. Vowel articulation by using 'weighted' interpolation

The first step is to choose key vowels and to position them in the interpolation plan, the second step is to define the rule of interpolation, which can change radically the expressiveness when gestural control is used, the last step is to connect sensors or peripherals to the dimensions of the interpolator. General formulae for interpolation between two lists of parameters must express a resulting list in function of the others and the weights associated to each

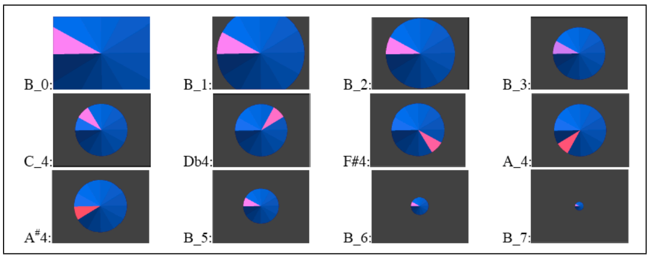

Figure 6. Pitch control Visual feedback for the Voicer

of them. Using a lot of vowels don't seem to be very important but a neutral vowel at the center of the space seems to be useful.

5. VISUAL FEEDBACK

In digital musical instrument the Sensory Feedback channels are generally broken. Sensory feedbacks, haptic or visual, enhance the capacity of regulation and the emotional immersion of the performer. Haptic feedback is probably fundamental but requires a complete study and an adequate technology. I chose to focus here on the relation between sound and visual feedback and their implication in information affluence considerations. I hope Visual feedback can help, first for beginners and later for more expert users at the level of the cognitive processes.

5.1. Pitch control visual feedback

A visual feedback is provided for the pitch control part of the Voicer. 12 angular sectors appear in different blue intensities to be seen individually. The sector pointed by the stylus is lighted by a red component; the red intensity depends on the pressure; so, it is linked with loudness. The radius of the disc (made by all the sectors) depends on the number of turns done; so, it is linked to the octave played. The goal is to transmit to the user the most of pertinent information with the less of perception 'bandwidth' usage.

5.2. Vowel articulation visual feedback

In order to not introduce an important asymmetry in the visual feedback compared to the asymmetry in the gestural control, the articulation of vowels should also represented. One could suggest to use two screens or to separate the screen in two parts. In this case we could have one visual feedback for the pitch control and loudness and another for vowel articulation.

I have experiment to replace the circle by an ellipse and to map the 2 radius of the ellipse to vowel articulation but it is not conclusive because of the confusion with the variation of the radius due to octave representation. However, an adequate common representation could be interesting, particularly if it could reflect linkage in the sound domain.

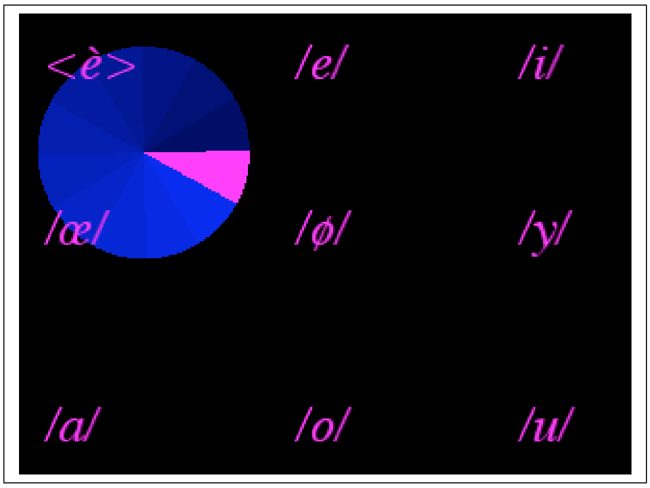

Figure 7. Pitch and articulation control 'combined' Visual feedback

6. CHOICE OF CONTROLLERS

Currently, we are studying the use of different types of controllers to achieve control functionalities needed in the design of this instrument. These controllers are a touch screen, a tactile surface, a graphic tablet, and a pen-based touch screen. The touch screen is used with a glove equipped with FSR pressure sensors at the extremities of fingers to allow, as the others peripherals used for the preferred hand, the use of the pressure applied. The first involved consideration is the catch of an object or not. This consideration implies two sub-considerations: to catch an object (here, a stylus) for pointing and to use the fingers in cooperation to manipulate it precisely (in particular, using the thumb-index pliers).

A tactile screen or a tactile surface allows pointing with whichever finger, in a monophonic implementation, or a polyphonic one, if the surface is a multi-point sensing device.

6.1. Direct or indirect interaction

Another consideration is the direct or indirect relation between manipulation and visualization. The visual feedback can be projected on the screen front of the user, or one can use a pen-based touch screen. Making a choice between these two controllers is a part of more general problem, which is the realization of a specific task, induced by the mapping strategy, based on efficiency, functionality and sensory considerations.

With a touch screen, or a pen-based touch screen, the visual feedback and the gesture have the same localization. For the others controllers the user can't look simultaneously at the visual feedback and at his own gestures.

6.2. Other considerations

Although these peripherals are different, they are used for the same functionalities. I have experimented here the

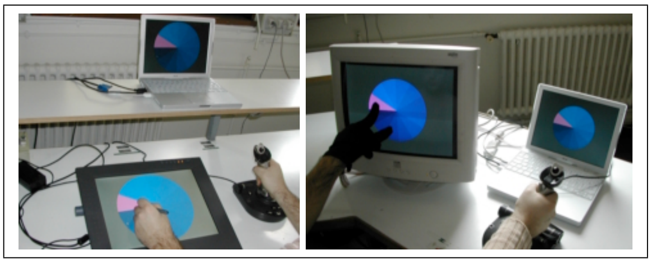

Figure 8. Visual feedback with direct interaction using a stylus, and a tactile screen with a data glove with finger pressure sensors

way to do the adequate choice of a peripheral for a specific function and more precisely for a specific musical function. A tactile screen doesn't seem to provide a precise enough position sensing but has the advantage to allow the use of all the fingers and to provide a common localization for gestures and visual feedback. A common localization tablet is stimulant but leads to a problem due to an asymmetry of visual feedback, which tends to focus the attention of the user to pitch control instead of providing a good compromise between pitch control and vowel articulation. The use of a tablet for the non-preferred hand seems to be inadequate because of the lack of proprioceptive feedback and the absence of spring-loaded return to a neutral position. Using symmetric functionalities is also in contradiction of models of asymmetric bimanual motor behavior [5].

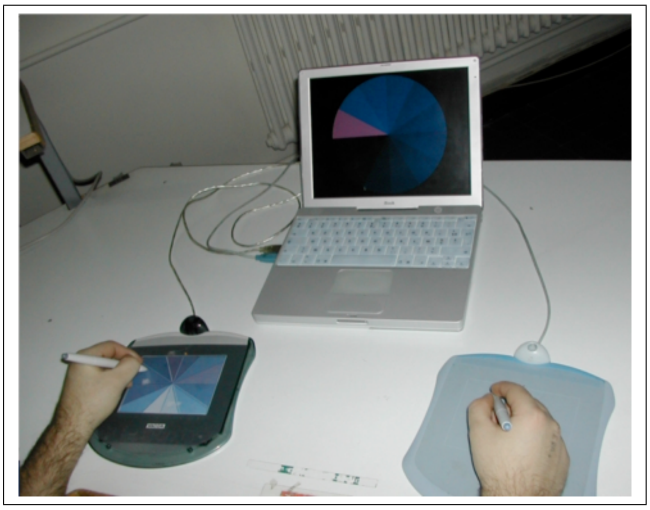

Figure 9. 'tablet+tablet' version of the Voicer

My first conclusion is to note the fact that as some of this controllers can be used as expressive musical controllers. They can fit to the functionalities needs for pitch control strategy used in this singing voice instrument. My second conclusion is to note that none of them is really perfectly adequate for this task but experimenting with them has provided good indication to build a specific controller including eventually an haptic feedback.

7. COMPOSITION, INTERPRETATION AND IMPROVISATION: TOOLS FOR EVALUATION

The better evaluation of a new musical instrument is to make music with it. I would like this kind of new instrument could preserve the possibility to improvise in many musical contexts as the voice or a violin, two very expressive musical instrument, can play indifferently different musical scales or music styles. In "phrasing" music: there is usually a level where one can hear entities named phrases, and in this sense interpretation would be inside a phrase, while improvisation would be at the level of the construction of phrases and their connections. Moreover one could say that interpretation is an "In-time" process, while improvisation needs an "Out-of-time" (figure 10).

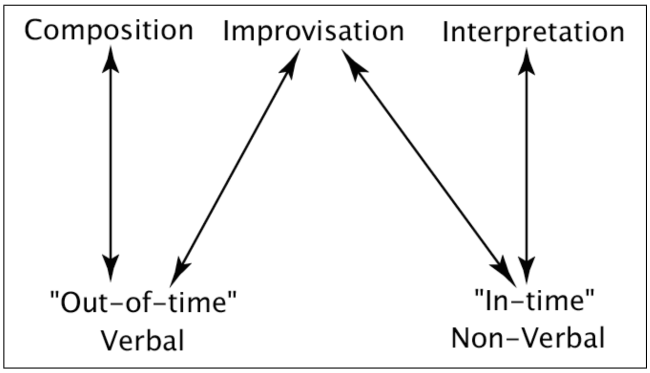

Figure 10. Cognitive implications of composition, interpretation and improvisation in music

At the beginning of the learning process of a musical instrument, a tonality can sometimes be privileged. This is particularly true for autodidact musicians. For example, E major is intensively used in guitar blues musical works from pioneers of the Delta blues. Composition contributes evidently to the evolution of a musical instrument but can also have some implications in the evaluation and the modification of its design. I have written a musical work named "d'ici et d'ailleurs" which version evolve continuously in parallel to the evolution of this singing voice instrument (named the Voicer).

The three movements of this musical work are in the same tonality but each of them is in a particular mode (see figure 11). Other secondary modes are also used inside each movement. Playing different modes based on the same tonality is a easiest way to evaluate the pitch control than playing different tonality. However, I have also used the Voicer in another tonality in a musical works of Daniel Arfib named 'route 729'. Different parts of this musical work offer a diversity of styles from rhythmic to arrhythmic, and from elementary sound object manipulation (glottal impulse at sub-audio frequencies with delay lines and others digital audio effects). As this a composition by itself this musical work is also a tool for improving the design of the instrument and specially the mapping used.

The Voicer has been used in several concerts and situations. First interpreted with a acoustic band composed

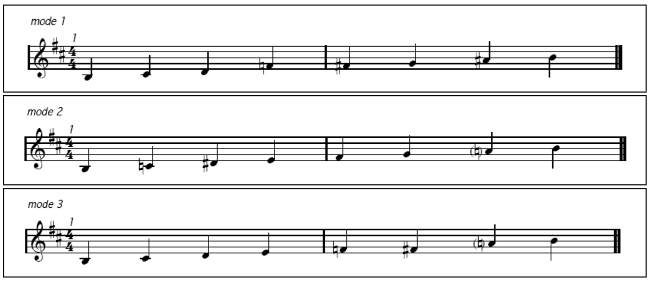

Figure 11. The three modes used in the musical musical work ‘D’ici et d’ailleurs’

of a guitar, a darbouka and a flute, it has been played in a second configuration with a electric band composed of a guitar, a bass, an electronic (or acoustic) saxophone and a drummer, in a third configuration with a an electronic percussion with a gesture controlled processing and a tape, and finally with a fourth configuration composed of a digital piano, an bass, an electric saxophone and a tape. Along this performance the musical work ‘D’ici et d’ailleurs’ has been rearranged and modified according to the evolution of the instrument, the context, and of course, my time of practice.

8. FUTURE AND RELATED WORKS

The work presented here is a part of a more general work on bimanual digital musical instruments. Three others bimanual instruments have been created including the Scangloves [7], a granular scrubber-sampler controller and a guitar inspired controller. One of the goals of these realizations is to explore manual and bimanual mapping strategies for specific musical purpose. Elements of this work can also have implications to vocal-related digital audio effects and to their gestural control, a first example of this has been presented in [2]. Another interest of digital instrument is the interaction with the computer and the possibility to use algorithmic but gesture controlled accompaniment as it has been used sometimes used in performance with a special version Voicer. Most of expressive features for secondary voices should then be specified by rules instead of being driven directly by gesture and suggest the necessity of a more extensive work on this point. A more powerful version of this instrument can probably be made by extending and improving concepts used in the work presented here and by extracting the better of the controller configurations experimented. I hope this instrument to be a tool for analysis of expressiveness and other aspects of voice as one can use ‘Analysis by synthesis’ of sound.

9. CONCLUSION

This work has provided a musically usable expressive instrument based on singing voice synthesis. It has also pointed on the importance of the relation between perception, mapping strategy and design or choice of controllers to make an effective musical instrument. Using alternate controllers and not controllers which imitates acoustic instruments can fit better to functionalities needed but, as they can not benefit from previous experiences of design, one should take care of sensory perception and motor behavior.

10. REFERENCES

[1] Arfib D., Couturier J.-M., Kessous L. and Verfaille V., ”Strategies of mapping between gesture parameters and synthesis model parameters using perceptual spaces”, *Organised Sound*, Cambridge University press, pages 127-144, volume 7, number 2, august 2002.

[2] Arfib D., Couturier J.-M., Kessous L., “Gestural Strategies for Specific Filtering Processes”, *Proceedings of the 5th International Conference on Digital Audio Effects (DAFX-02)*, Hamburg, Germany, 2002.

[3] Mathews, Max V. and Pierce, John R. , *Current Directions in Computer Music Research*, 1989.

[4] Flanagan, J. L., *Speech Analysis, Synthesis, and Perception*, Springer-Verlag ISBN 0-387-05561-4.

[5] Guiard Y., “Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a model”. *Journal of Motor Behavior* 19 (1987), 486- 517, 1987.

[6] Hunt, A. and Wanderley, M. M., ”Mapping performance parameters to synthesis engine”, *Organised Sound*, Cambridge University press, pages 103-114, volume 7, number 2, august 2002.

[7] Kessous L., “The Scangloves: a video-music instrument based on Scanned Synthesis”, *Proceedings of the Symposium on Gesture Interfaces for Multimedia Systems*, COST-286, University of Leeds, UK, 2004.

[8] Mathews M. V., “Bicycle built for two (1961)”, *Computer Music Currents n°13, The Historical CD of Digital Sound Synthesis*, Vergo.

[9] Veldhuis R., “A computationnally efficient alternative for the liejencrants-Fant model and its perceptual evaluation”, *Journal of acoustical Society of America* 103, 566-571, 1998.