Digitartic: bi-manual gestural control of articulation in performative singing synthesis

Lionel Feugère, Christophe d'Alessandro

Digitartic: bi-manual gestural control of articulation in performative singing synthesis

Lionel Feugère
LIMSI-CNRS, BP 133, F-91403 Orsay, France
UPMC Univ Paris 06, F-75005 Paris, France
lionel.feugere@limsi.fr

Christophe d'Alessandro
LIMSI-CNRS, BP 133, F-91403 Orsay, France
cda@limsi.fr

ABSTRACT

Digitartic, a system for bi-manual gestural control of Vowel-Consonant-Vowel performative singing synthesis is presented. This system is an extension of a real-time gesture-controlled vowel singing instrument developed in the Max MSP language. In addition to pitch, vowels and voice strength controls, *Digitartic* is designed for gestural control of articulation parameters, including various places and manners of articulation. The phases of articulation between two phonemes are continuously controlled and can be driven in real time without noticeable delay, at any stage of the synthetic phoneme production. Thus, as in natural singing, very accurate rhythmic patterns are produced and adapted while playing with other musicians. The instrument features two (augmented) pen tablets for controlling voice production: one is dealing with the glottal source and vowels, the second one is dealing with consonant/vowel articulation. The results show very natural consonant and vowel synthesis. Virtual choral practice confirms the effectiveness of *Digitartic* as an expressive musical instrument.

Keywords

singing voice synthesis, gestural control, syllabic synthesis, articulation, formants synthesis

1. INTRODUCTION

In speech, phoneme articulation is a task which involves coordination between different organs (such as vocal folds, tongue, lips or uvula) and adaptation to past and future phonemes.

In music, a typical additional constraint for the vocal apparatus is phoneme timing. Articulators have to constantly anticipate their target position to be on time with the metric. As any musical instrument, synthetic voice instruments should have a maximum latency of 10-20 ms (i.e. non perceptible) to be played accurately with tempo.

A second constraint is the expressiveness of articulation. In the same way that a continuous and subtle F_0 control is necessary to have an expressive and adapting intonation, we believe that expressiveness will be enhanced while articulating if we smoothly control it in real time. Then a large variety of articulations will be possible, such as different degrees of articulation (hypo- to hyper-articulation), or duration and intensity modification during the articulation stages.

Other constraints for an effective musical instrument are easiness of learning and a well-designed control model, in order to command a maximum of production parameters with as few as possible control parameters.

These constraints require a powerful and rapid voice production model and a high-resolution control interface. Existing systems based on physical models demand too much computing power for real-time applications. Concatenation synthesis or HMM basis speech synthesis [3] introduce at least one phoneme delay when produced in real time. Several signal voice models have been designed for real time applications. Some are intended for vowel synthesis and vowel articulation [6] [16] [10] [14], but they are unable to control consonant articulation. The only ones that allow consonant production are the ones that trigger consonants by selection gestures [7] [13] [4], without detailed control of the articulation process. Then expressiveness in articulation is lost. Moreover, a delay of at least 30-50 ms is created for the Consonant-Vowel (CV) syllables where the musical beat is on the final vowel, e.g. syllables with fricatives, semi-consonants or nasal occlusives. Although this is acceptable in the context of speech synthesis or for some particular musical applications, it is not accurate enough for performative singing synthesis.

Contrary to the systems cited above, our work enables a fine control of phoneme articulation owing to modification gestures (as defined in [5]), while allowing temporal precision as well as accurate control of pitch. The aim is not a full text synthesis, but rather a realistic syllable synthesis. This fits with musical styles such as slow onomatopoeia recitation of Indian percussion or slow sung syllables of scat music. In our work, a low-CPU-consuming method called formant synthesis is used. While other synthesis methods give a better quality (but with all the issues presented above), we believe that the quality of gestural control somehow compensates the relatively poor quality of this voice synthesis technique.

Besides the wish to design a musical voice instrument, synthesis gestured control of living process models allows to investigate the respective roles of voice production and control. With synthetic voice instruments, the voice production is separated from the gestural internal control while the borders between the two are not so prominent in natural voice. In addition, we propose to replace the internal gesture by manual gesture which is far to be trivial.

The production model of *Digitartic* is detailed in section 2. In section 3, the control model for continuous control of phoneme articulation is presented. Finally the interface is presented as well as the gestures necessary to produce syllables, and the synthesis sounds obtained are compared to natural syllables.

NIME'13, May 27 – 30, 2013, KAIST, Daejeon, Korea.

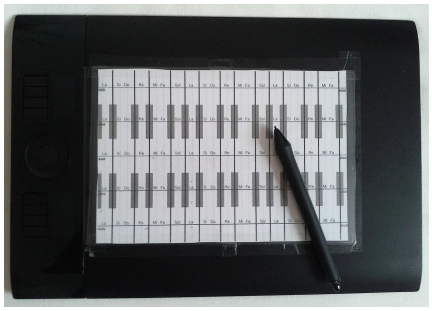


Figure 1: *Preferred hand Interface for controlling F_0, voice strength and vocalic color*

2. VOICE PRODUCTION MODEL

2.1 Vowels synthesis

Digitartic is implemented in Max/MSP [1]. It is a rule-based formant synthesizer, i.e. a source-filter model, including the following components :

- The RT-CALM [9] glottal source model, a real time version of the CALM model [11]. This model, working in the spectral domain, is based on an analysis of key temporal models proposed in the literature [12]. A white noise modulated by the shape of the glottal source model signal represents voice breathiness. Voice strength is modeled by an increase in the signal intensity and a decrease in the spectral slope of the glottal flow model derivative.
- Five resonant bandpass filters model the vocal tract resonances. They represent the formants of the vowels as well as the consonants. Thus, a database of formants defines the vowels and the consonantal targets by a set of amplitude / bandpass / frequency values. A formant frequency multiplier is added to modify the vocal tract size of the model in order to get different voice types.

For vowels synthesis, the control model is as follows: F_0 is mapped on the X-axis of a graphic tablet; continuous vocalic color are mapped on the Y-axis (based on vowels /i,e,a,o,u/ and their interpolations); vocal strength is mapped on the stylus pressure. The tablet is superposed by a printed layer, that allows for precisely targeting F_0 and vocalic color as shown in Fig. 1. This layer represents a piano keyboard with a continuous and linearised pitch control. Several keyboards are drawn in the Y-axis for each canonical vowel.

2.2 Temporal structure of VCV synthesis

French consonants can be categorized by the help of 3 features: place of articulation, manner of articulation, and voicing. Place of articulation is the location of the articulator constriction resulting in the partial or full obstruction of the airflow. For example, the place of articulation of /p/ (in ***p**apa*) is labial. The second feature is the manner of articulation, i.e. how the constriction is done. Plosives, fricatives and semi-consonants represent various degree of obstruction from full to slight, while french nasals are plosives and nasalized. For example, the manner of articulation of /j/ (in ***y**ellow*) is palatal. Lastly, voicing deals with the use or non-use of the vocal folds during the consonant production.

Fig. 2 gives a schematic representation of the *Digitartic* system. The instrument is able to articulate syllables of V_1CV_2 type, where: V_i is the vowel i among /i,e,a,o,u/ and their interpolations along the 1D axis /i,e,a,o,u/; C is a consonant among plosives /p,b,t,d,k,g/, fricatives /f,v,s,z,ʃ,ʒ/,

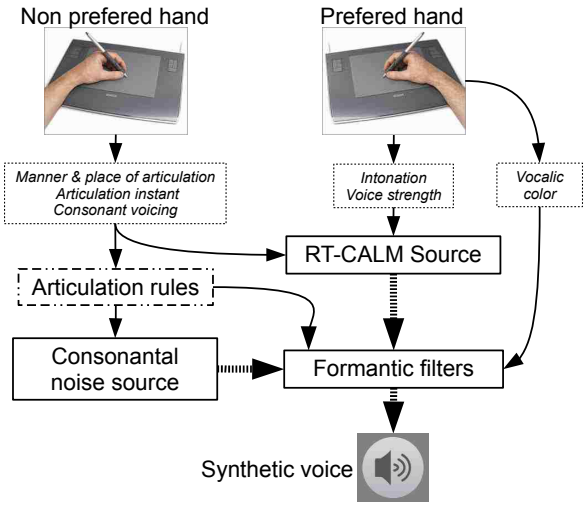


Figure 2: *Schematic representation of the Digitartic synthesizer*

Manner \ Place	**Labial**	**alveolar**	**palatal**
plosives	/p,b/	/t,d/	/k,g/
fricatives	/f,v/	/s,z/	/ʃ,ʒ/
semi-consonants	/w/	/ɥ/	/j/
nasals	/m/	/n/	/ŋ/

Table 1: Produced consonants by the *Digitartic* depending on place and manner of articulation

semi-consonants /w,ɥ,j/, nasals /m,n,ŋ/ and their interpolations along the place of articulation for a same articulation manner and voicing (/p,t,k/, /b,d,g/, /f,s,ʃ/, /v,z,ʒ/, /w,ɥ,j/, /m,n,ŋ/). Any combination of V_1CV_2, like $V_1C_1V_2$-C_2V_3, or any sub-part of V_1CV_2 like CV can be produced, as well as adding combinations of silence. Tab. 1 summarizes the possible consonants. Note that although the reference language is french, it is possible to produce consonants with various places of articulation to synthesize also consonants in other languages.

Temporal structures of VCV disyllables can be classified into two categories, depending on their symmetry of their temporal structure. The first category is made of symmetric disyllables: fricatives, semi-vowels and nasals. Among this 3 manners of articulation, only fricatives may be unvoiced, with some consonantal noise. This friction noise occurs because articulators partly obstruct the vocal tract, creating a turbulent air stream. The second category is made of plosives which present an asymmetrical time structure. The three phases of a plosive are closure, then silence (or a murmur for voiced plosive), then noise burst and coarticulation with the following vowel. The burst due to the sudden release of air from the complete closure of the articulators only appears during the CV transition stage. Therefore VCV disyllable synthesis adopt different strategies. Although symmetric transitions are used for fricatives, semi-vowels or nasals, in the case of plosives, the voiced part of the synthesis is treated symmetrically, while the noise portion of plosive bursts is dealt with asymmetrically between syllables CV and VC.

2.3 Formant targets

A vocal tract resonance gives rise to a formant, and the set of formants at a given time or its evolution will help for the

phoneme identification.

Different formant databases [15] [17] [2] were used and have been manually tuned and adapted to our synthesizer. All consonantal values have been adjusted in /a/-*C*-/a/ context and we extrapolate this values to other vocalic contexts. This is a quite rough approximation as the vocal tract shape on a consonant depends of the following and preceding vowel. Consequently, the instrument performs in principle better for a /a/-*C*-/a/ disyllable. A band-stop filter is added to model the main nasal anti-resonance for the nasal consonants (no nasal vowel is available).

2.4 Consonantal noises

Two types of consonantal noise are considered: frication and burst noise. Friction noise comes from the air passage between two articulators close enough to make a noise. It lasts as long as the passage of air through the vocal tract remains. Burst noise occurs when releasing the total constriction of the articulators in the early CV stage of plosive consonants. Unlike friction noise, it is brief and the duration of 10-30 ms is hardly controllable. Its spectrum depends on the place of articulation and adjacent phonemes.

We make the assumption that the consonantal noise spectrum, independently from its time evolution, will only depend on 3 parameters: the place of articulation which influences the source noise spectrum by dividing the vocal tract in several cavities from both sides of the constriction; the vocalic configuration for a same place of articulation, i.e. the preceding and following phonemes which modify the position of the articulators while keeping the same place of articulation; voicing, i.e noise modulation by the amplitude of the glottal source signal.

Therefore it is assumed that for a same place of articulation, friction and burst noise spectra are similar. In our model, the difference mostly comes from its duration and intensity evolution. The procedure for consonantal noise synthesis is as follows:

1. Each place of articulation is associated to a specific noise source, made of Gaussian white noise filtered by a 2^{nd} order bandpass butterworth filter.
2. Noise amplitude is time-modulated according to several parameters: *articulatory phase* (i.e. articulation position between the preceding and following phonemes); manner of articulation; consonant voicing.
3. Noise is modulated by the glottal flow wave in case of voiced consonant.
4. For the sake of coarticulation, noise is filtered according to the spectrum of the preceding vowel in end of the VC phase and according to the spectrum of the following vowel in beginning of the CV stage.

2.5 Consonant and vowel transition rules

The basic rules of *Digitartic* only include targets for parameter evolution. The rules are independent of the transition duration, because it is assumed that timing is being managed by the player's gestures.

For the production of syllables, a serie of parameters varies continuously between the vowel and the consonant targets: consonant formant values (central frequency, amplitude, and bandpass); filter coefficients of the consonantal noise; voicing evolution; noise and the aspiration amplitudes.

As stated earlier, the parameters evolution is symmetric for the CV stage and the VC stage, plosive bursts excepted (they are involved only during the CV stage). Fig. 3 displays an example of parameters variation for the VCV articulation with C=/p/. Note that the consonantal noise burst appears in the CV transition stage.

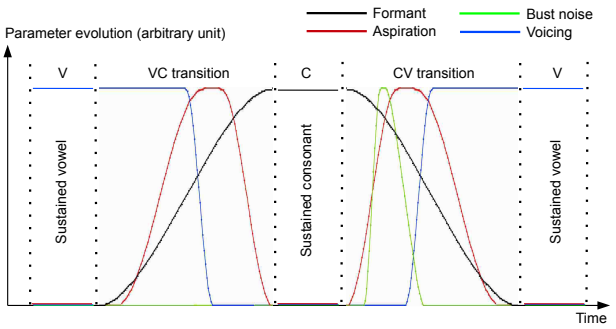


Figure 3: *Example of evolution of articulatory transition parameters for V-/p/-V sequence.*

3. ARTICULATORY PHASE AND PLACE OF ARTICULATION AS HIGH LEVEL CONTROL PARAMETERS

In our control model, 3 main high level parameters of articulation are available: the first one, denoted by $y1_{voyCible}$, represents the targeted vowel; the second one, $x2_{conCible}$, corresponds to the targeted consonant; the third one, $y2_{articu}$, deals with the articulation phase, i.e. the evolution of the transient parameters of CV or VC articulation.

3.1 Controlling the place of articulation

For a given manner of articulation and depending on voicing, $x2_{conCible}$ determines the targeted consonant and controls the interpolation degree of the parameters which characterize the two nearest consonants on an axis corresponding to the place of articulation (for a same manner of articulation). This parameter enables to continuously modify the targeted consonant in real time, even during a CV or VC transition.

With a linear interpolation between reference consonants, targeting the reference consonants is quite difficult, as a slight shift of the parameter will change the consonant. In order to be able to target the reference articulation place in an easier way, a non-linear factor is applied to $x2_{conCible}$ with respect to the place of articulation, as shown in the Fig. 4.

The system allows for production of intermediate consonants, located between two reference consonants of a same manner of articulation. From the formants values of frequency / amplitude / bandwidth and the filter coefficients of

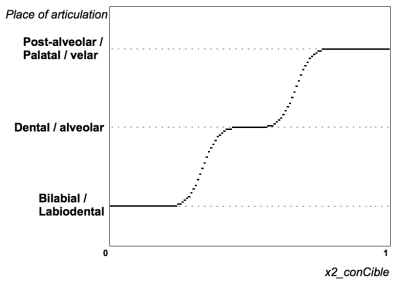
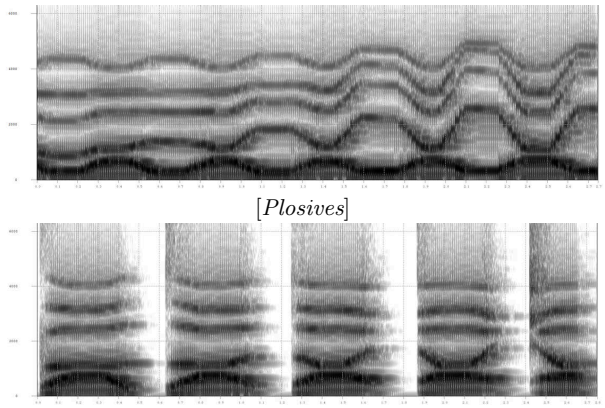


Figure 4: *Non-linearity of the consonantal interpolation parameter*

consonant noises, intermediate pseudo-consonants are built on bilabial - alveo-dental - palatal axis. The parameter $x2_{conCible}$ is continuous, thus we can produce an infinite number of pseudo-consonants (consonants that are not in the phonological system of the target language) by interpolation. Fig. 5 presents a series of voiceless plosive and semi-consonant spectrograms from the Digitartic instrument (each followed by the vowel /a/) for which the place of articulation gradually moves from bilabial to palatal. Therefore, the configuration goes through the syllables /pa/, /ta/ and /ka/ for plosives, and /wa/, /ɥa/ and /ja/ for semi-vowels. While interpolation has a physical meaning for semi-vowels (i.e. the degree of articulation), it is less obvious for plosives, whose perception is categorical. In a given language no phoneme exists in between e.g. bilabial and alveolar plosives. However, from a musical point of view, virtual hybrid consonants can be useful, giving new human-like sounds, even if they do not exists as a linguistic elements.



[*Plosives*]

[*semi-consonants*]

Figure 5: *Spectrogram* $(0 - 6000$ *Hz) of successive /a/-C-/a/ sequences produced by Digitartic, with the articulation place of the consonant moving along the bilabial - alveo-dental - palatal axis*

3.2 Controlling articulation timing

An another high level control parameter, called *articulatory phase* and denoted by $y2_{articu}$, deals with the time evolution of phoneme parameters between a vowel and a consonant. Physically, it corresponds to the position of the articulators between two articulatory reference positions: the targeted vowel ($y1_{voyCible}$) and the targeted consonant ($x2_{conCible}$).

Fig. 6 schematizes the transition from an intermediate targeted vowel to an intermediate targeted consonant. Here, *intermediate* means that the vowel and consonant are an interpolation of canonical vowels and consonants respectively.

This parameter is continuously controlled by the player. There is no delay between the gesture and sound production, contrary to other approaches, like real-time voice synthesis using HTS [3] where the sound corresponding to whole phoneme or diphone is computed before being played. This point is very important for musical performance, because of the rhythmic precision requirement.

4. GESTURE DYNAMICS, INTERFACE AND MAPPING

4.1 Interface & Mapping

The instrument is bi-manual, and requires the use of two pen tablets. With the preferred hand, a first WACOM pen tablet enables control of vowels and the glottal source

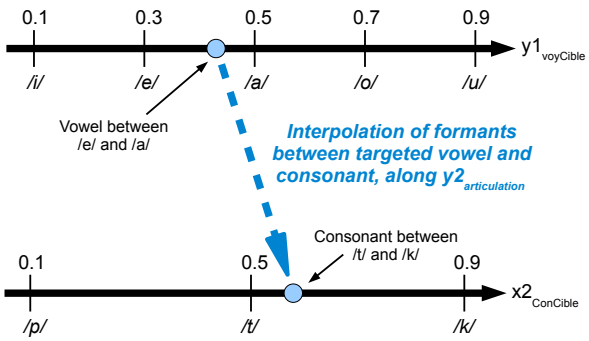


Figure 6: *Example of formant interpolation between two targeted vowel and consonant*

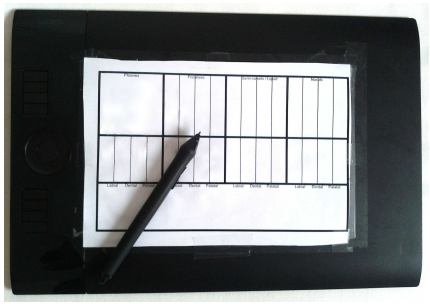


Figure 7: *Pen tablet equipped with its mark layer for controlling* VCV *articulation*

(F_0 and voice strength) as explained and illustrated in section 2.1. A second WACOM pen tablet is used to control consonant production with the help of the non-preferred hand. The stylus coordinates and pressure values on the tablet are sent through USB and UDP protocol to the Max/MSP synthesis application using the *s2m.wacom* Max external [1].

Pen tablet allows to capture gestures fast enough to reproduce natural articulatory gestures, thanks to its high temporal resolution (5 ms). The rapid movement of the pen on the tablet over a short distance can be easily done in a few tens of milliseconds and therefore correctly reproduce the articulation dynamics by a smooth interpolation.

A drawing layer is superposed to the pen tablet (Fig. 7) in order to give visual indications for accurate control of the synthesis process. These marks are detailed in Fig. 8, which displays a schematic top view of the different control zone of the tablet. The following mapping is realized with the interface :

- The place of articulation $x2_{conCible}$ corresponds to the X-axis position of the stylus for each zone of articulation manner. The vertical purple dotted lines corresponds to the canonical place of articulation of french.
- The articulation phase $y2_{articu}$ corresponds to the Y-axis position of the stylus (upward for VC articulation and downward for CV articulation)
- A voice strength factor corresponds to the stylus pressure over the tablet
- Consonantal voicing (on/off) is controlled with one of the stylus button

[1] `http://metason.cnrs-mrs.fr/Resultats/MaxMSP/index.html`

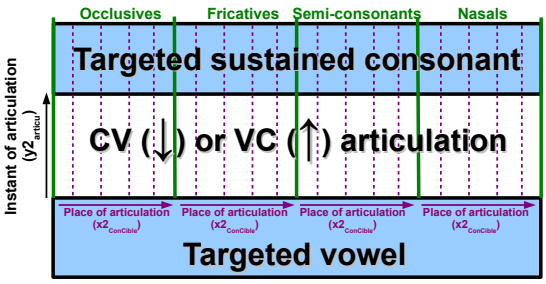


Figure 8: *Schematic top view of the different control areas of the tablet intended for articulation control.*

- The manner of articulation corresponds to different tablet areas, divided along the X axis (vertical green lines)

Note that other types of controllers have been considered and tried (trackpad, wiimote, accelerometer, optic captors) but they never equalled the resolution, latency or ease of use of the WACOM graphic tablet. Also other mappings have been tried, using selection or modification gestures, but no one was considered better than this one. Indeed, with only one pen, almost all consonants are available, and fast, continuous and expressive articulation is possible.

Originally, the idea was to get inspired from percussive stroke movement to control the syllables as it occurs very fast and the resulting sound changes depending on the location and manner the percussion is stroked, in the same way that consonants are dependent of place and manner of articulation. The capture of this 3-D movement not being possible in such small latency, we decided to "project" the percussive movement onto the Y-axis of the tablet plane.

4.2 Controlling the transition dynamics

Singing is made of long vocalic segments and generally shorter consonantal transitions. This transition is controlled continuously, reversibly and without perceptible latency through $y2_{articu}$. Any sequence or subset of type $C_1V_1...V_NC_2V_{N+1}...V_MC_3...C_K$ can be produced.

The distance between the tablet area of the sustained consonant to the vowel is the same regardless of the mode and place of articulation. However, the transition time between a sustained consonant and a vowel (and vice versa) depends on the consonant and above all on the manner of articulation. Thus, this duration will be controlled by the user's gesture. The type of gesture, through its speed and the short distance, added to the tablet resolution, enables to continuously control the transition in real time. On Fig. 9 is represented the position of the hand for a sustained consonant (left side) and the one for a vowel (right side). Gesture (indicated by a red arrow) is a fast movement of the index (VC syllable on left side, CV syllable on right side).

We compare some VCV sequences produced by *Digitartic* with the same VCV sequences produced by natural voice: three semi-vowels and three plosives are plotted in Fig. 10 and 11. The natural voice was recorded in an anechoic chamber, while the synthesis voice was produced by mimicking the natural voice. All the spectrograms are 400 ms long and are plotted from 0 to 6000 Hz.

The general formant dynamics is well rendered by *Digitartic*. The slightly different formant values between synthetic and natural voices are due to the fact that the voice used for comparison is not the same as the analyzed voice used for building the synthesis rules.

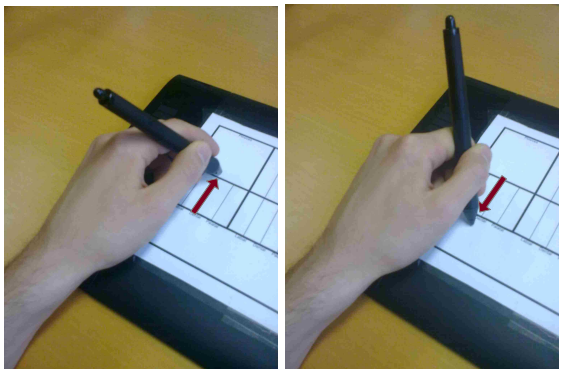


Figure 9: *Movements for VC and CV production, in case of a syllable with hypoarticulated consonant. Left: position for a sustained consonant. Right: position for an hypoarticulated vowel*

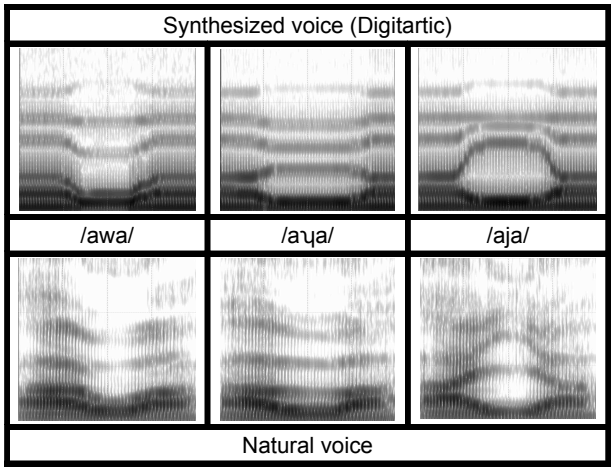


Figure 10: *Spectrograms of semi-consonants from Digitartic and natural voice ($0 - 6000$ Hz, 400 ms long for each)*

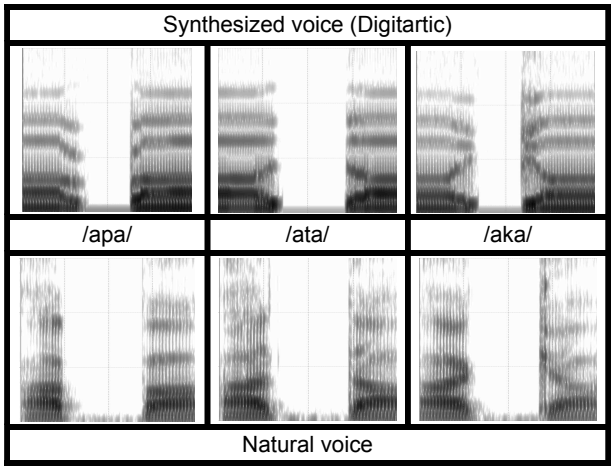


Figure 11: *Spectrograms of plosives from Digitartic and natural voice ($0 - 6000$ Hz, 400 ms long for each)*

4.3 On the analogy between articulator and hand gestures

In *Digitartic*, nasals, semi-consonants and fricatives are symmetrical with respect to the sustained consonant in a VCV sequence as described in section 2.2. All features of the consonants (formants, nasality, friction noise) appear gradually along the transition zone. Then hypo-articulation is possible by targeting a point inside the transition zone, instead of targeting the sustained consonant zone as in Fig. 9. Indeed, the articulatory phase parameter $y2_{articu}$ doesn't reach its maximal value. The consequence is a consonant with less amplitude in formants motion. Similarly, hypo-articulated vowel is also easily achieved.

To produce the burst of a plosive, the pen must reach first the sustained consonant zone of the tablet (corresponding to the sustained closure of the vocal tract), as the burst doesn't occur in the VC part of the transition, but only as a short potion in the CV part of the transition. In this case it is not possible to reduce the pen trajectory for hypo-articulation, because the burst would not occur. Nevertheless, hypo-articulation can be achieved in another way (like for other consonants), i.e. by a voice strength decrease once the sustained consonant is reached. The effect produced is a decrease of the sound level during the consonant attack.

The analogy between articulatory gesture of the vocal apparatus and the manual articulatory gestures is reinforced, as a manual gesture hypo-articulation naturally results in synthetic syllables hypo-articulation (except for plosives). Indeed, the manual trajectory is very much analogous to voice articulators trajectories meaning. Then too rapid manual gestures will not reach the consonantal targets on the tablet and then produce hypo-articulated singing, and conversely slower gestures will be allow to produce hyper-articulated singing. This analogy is musically very relevant.

Controlling articulation phase is very useful to improve musical expressiveness in syllables. In *Digitartic*, accents can be produced at once by changing the voice strength and by changing the degree of articulation.

5. CONCLUSION AND PERSPECTIVES

The *Digitartic* instrument demonstrates that it is possible to precisely and accurately control convincing consonant transitions using manual gestures. The analogy between gesture of articulation in the vocal apparatus and hand gesture seems promising.

Place of articulation and phase of articulation, i.e. position between two phoneme targets, can be controlled in real time without delay. This is interesting for playing expressive and reactive articulated syllables when singing or scatting.

The interface enables to produce any VCV sequences, among all the french consonants and vowels, except nasal vowels and liquid consonants.

The synthesis of articulation is gradually introduced into digital choir practice, extending the vocal possibilities of virtual singers.

Besides qualitative spectrogram comparisons with natural voice we have presented, formal tests must evaluated the perceptual quality of this synthetic syllables, as well as the expressive articulatory possibilities (articulation degree and dynamics). We hope to demonstrate an analogy between articulator and hand gestures as it was done between intonation and manual gesture [8].

Another type of validation is musical performance, that is the ultimate goal of this research. We are confident in the fact that our virtual choral practice will confirm *Digitartic* effectiveness as an expressive musical instrument.

6. REFERENCES

[1] http://www.cycling74.com/products/maxmsp.html, website lastcheck on 22/01/2013.

[2] http://www.phonetique.ulaval.ca/illust.html, Laboratoire de Phonétique et Phonologie, Université Laval, Québec, website lastcheck on 22/01/2013.

[3] M. Astrinaki, N. d'Alessandro, B. Picart, T. Drugman, and t. Dutoit. Reactive and continuous control of hmm-based speech synthesis. In *IEEE Workshop on Spoken Language Technology (SLT 2012)*, Miami, Florida, USA, December, 2-5 2012.

[4] G. Beller. Gestural control of real-time concatenative synthesis in luna park. In *P3S (Performative Speech and Singing Synthesis)*, 2011.

[5] C. Cadoz. Instrumental gesture and musical composition. In *Proceedings of the 1988 International Computer Music Conference*, pages 1–12, San Francisco, 1988.

[6] P. R. Cook. Spasm, a real-time vocal tract physical model controller; and singer, the companion software synthesis system. *Computer Music Journal*, 17(1):30–44, 1993.

[7] P. R. Cook. Real-time performance controllers for synthesized singing. In *Proceedings of the 5th Conference on New Interfaces for Musical Expression (NIME'05)*, Vancouver, BC, Canada, May 26-28 2005.

[8] C. d'Alessandro, A. Rilliard, and S. Le Beux. Chironomic stylization of intonation. *J. Acoust. Soc. Am.*, 129(3):1594–1604, March 2011.

[9] N. d'Alessandro, C. d'Alessandro, S. Le Beux, and B. Doval. Real- time calm synthesizer : new approaches in hands-controlled voice synthesis. In *Proc. of New Interfaces for Musical Expression 2006*, pages 266–271, Paris, France., 2006.

[10] N. d'Alessandro and T. Dutoit. Handsketch bi-manual controller, investigation on expressive control issues of an augmented tablet. In *Proceedings of the 7th Conference on New Interfaces for Musical Expression (NIME'07)*, New York, USA, 2007.

[11] B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In ISCA, editor, *Proceedings of Voqual'03 : Voice Quality : Functions, analysis and synthesis*, Geneva, Switzerland, 2003.

[12] B. Doval, C. d'Alessandro, and N. Henrich. The spectrum of glottal flow models. *Acta Acoustica*, 92:1026–1046, 2006.

[13] S. S. Fels and G. E. Hinton. Glove-talk ii : a neural network interface which maps gesture to parallel formants. *IEEE*, 9(1):205, 1998.

[14] L. Feugère, S. Le Beux, and C. d'Alessandro. Chorus digitalis : polyphonic gestural singing. In *1st International Workshop on Performative Speech and Singing Synthesis (P3S 2011)*, Vancouver (Canada), 14/03 au 15/03 2011.

[15] M. Garnier-Rizet. *Elaboration d'un module de règles phonético-acoustiques pour un système de synthèse à partir du texte pour le français.* PhD thesis, Université de la Sorbonne nouvelle, 1994.

[16] L. Kessous. Bi-manual mapping experimentation, with angular fundamental frequency control and sound color navigation. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME'02)*, pages 113–114, 2002.

[17] K. N. Stevens. *Acoustic Phonetics.* The MIT Press, 1998.