Rule-based performative synthesis of sung syllables

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

In this demonstration, the mapping and the gestural control strategy developed in the *Digitartic* are presented. *Digitartic* is a musical instrument able to control sung syllables. Performative rule-based synthesis allows for controlling semi-consonants, plosive, fricative and nasal consonants with a same gesture, despite the structural differences in natural production of such vocal segments. A graphic pen tablet is used for capturing the gesture with a high sampling rate and resolution. This system allows for both performing various manners of articulation and having continuous control over the articulation.

Keywords

Singing voice, gestural control, phoneme articulation, performative synthesis

1. CONTEXT

Phoneme articulation in speech and singing involves synchronization of many vocal tract and glottal organs. When simulating voice synthesis and attempting to control it in real time, it becomes very difficult to control so many parameters, and above all with the temporal constraints of a music band performance.

In some works, articulation is controlled by only triggering the transition between vowel (V) and consonant (C) or between C and V. Simple triggering is used together with formant synthesis by rules in projects such as Glove-Talk [6], or with corpus-based concatenation synthesis such as in the Luna Park performance [2].

Other works address the more complex question of continuous transition control. Using fixed syllable sequences, one can apply time-streching techniques like PSOLA-based voice modification [8] or pre-synthesized phonemes lists, as in the voice physical model SPASM [3]. Continuous phoneme control is possible as well with HMM-based synthesis, but this results in at least one phoneme delay, like in the MAGE system [1].

In the present demonstration, all the temporal phases of Vowel-Consonant-Vowel sequences are accurately controlled by hand gestures. In phonetics, Vowel-Consonant-Vowel dissyllables (VCV dissyllables) can be classified according to the manner of articulation of the consonant. Consonants which obstruct completely and temporally the vocal tract are called oral plosives (e.g. /p,b,t,d,k,g/) while the ones which obstruct completely and temporally only the oral tract are called nasal plosives (e.g. /m,n,ŋ/). Consonants which partially obstruct the vocal tract, resulting in a turbulent noise at the constriction are called fricatives (e.g. /f,v,s,z,ʃ,ʒ), while the ones which let the air flow free without friction (or very little) are called semi-consonants (e.g. /w,ɥ,j/).

NIME'14, June 30 – July 03, 2014, Goldsmiths, University of London, UK.

2. STYLUS CONTROLLED ARTICULATION

2.1 Digitartic

Digitartic (DIGIT(al) ARTIC(ulation synthesizer)) is a rule-based formant singing synthesizer [7]. With the help of two graphic tablets (e.g. Wacom Intuos5), the user can control pitch, voicing, vocal effort, vocalic color, articulation phasing, manners and places of articulation, through a source-filter model using the RT-CALM glottal wave model [5] (Figure 1 gives a picture of the Digitartic control). In this demonstration, the focus is put on the synthesizer control, particularly on the detailed and accurate control of all articulation phases between vowel and consonant.

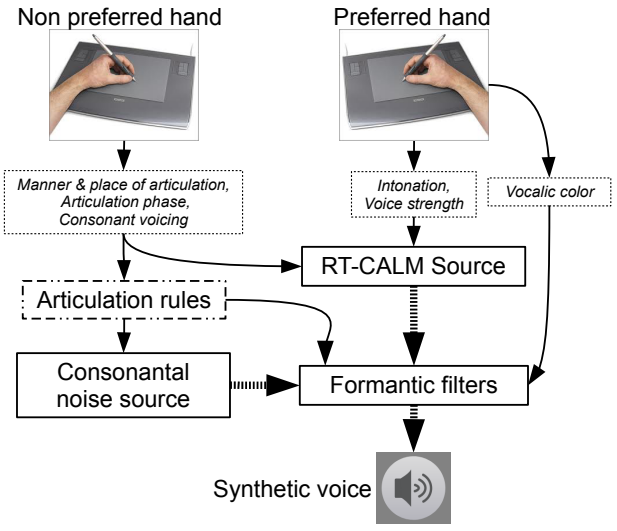

Figure 1: *Principles of Digitartic control.*

2.2 Continuous control of articulation phasing

Articulation is continuously controlled by a stylus in the tablet plane. The tablet is divided in several areas, corresponding to manner of articulation. After the selection of a given consonant (and then a specific area) the user can

control acoustic phases in phonemic transition with a continuous motion. As soon as the gesture begins, the articulation is modified until the end of the movement. The main advantage is that timing is fully under control: the player can synchronize accurately with other musical events. Note that this is not possible in triggered systems (e.g. the Luna Park system in the case of semi-consonants and fricatives) for which the musical attack is located at the end of the articulatory transition.

A printed layer attached to the tablet gives visual references of the transition phases. The pen must slide along the transition path without losing contact. The printed layer is divided in 3 areas as shown in Figure 2:

- A vowel area (at the bottom of the tablet) which corresponds to the vowel target (in terms of formants). No sound modification occurs when moving the pen inside this area;
- A consonant zone (at the top of the tablet) which corresponds to the sustained targeted consonant (in terms of formants, and friction noise if any). No sound modification occurs when moving the pen inside this area, except the vocal effort depending on the pressure;
- A transition area (in the middle of the tablet, between the vowel and consonant areas). If one moves from the bottom to the top areas, the articulation goes from the vowel to the consonant. If one moves from the top to the bottom areas, the articulation goes from the sustained consonant to the vowel. If a forward and downward movement is performed, then a VCV disyllable is produced. Formants evolve depending on the stylus position in this zone, as well as burst noise for plosives for the CV transition (no burst for the VC transition).

The passage from one to another adjacent area is continuous in terms of articulation. Duration of the consonantal transition is determined by the stylus velocity. In this way, it is possible to simulate any degree of hypo- or hyper-articulation.

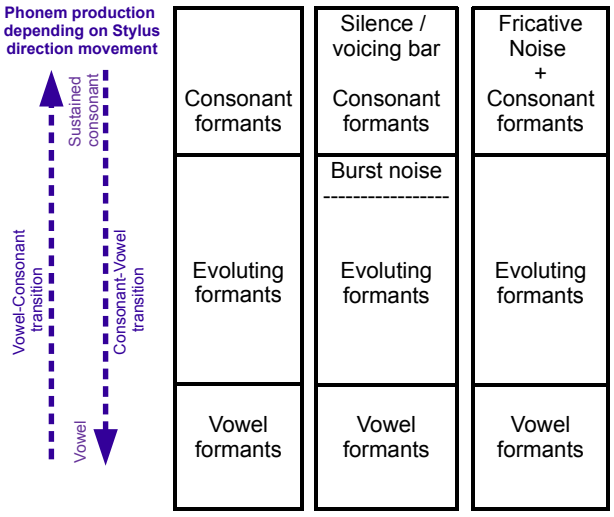

Figure 2: *Articulation zones on the tablet*

The same type of movement is used for each manner and place of articulation. The only difference lies in the required stylus velocity:

- A fast movement is required for plosives. More specifically, velocity is important for producing an effective burst noise. If the gesture is too slow, the burst noise will be perceived rather as a friction noise.
- A medium velocity is required for fricatives, nasals and semi-consonants

Digitartic has been effectively used for performing music at several occasions e.g. NIME 2013 in Korea and JDEV 2013 in France. A video extract of this last performance is available online [1] as well as some VCV sounds recorded all at once[2] (/apa/, /apa/, /ata/, /aka/, /ava/, /aza/, /aʒa/, /awa/, /aɥa/, /aja/, /ama/, /ana/).

3. PERSPECTIVES

Digitartic proved successful at performing intelligible syllables with accurate timing control. However, it is more difficult to articulate multi-syllabic words. As a perspective, we plan to implement a playing mode where the user can chain pre-defined syllables (with about 3 different consonants) using the same gestures as above.

Finally, comparing the voice apparatus gestures and hand gestures would be interesting to study further how the hand manages to imitate the voice. Gesture controlled vocal instruments could be compared to voice articulation in the same way as vocal prosody has been compared to chironomic prosody [4].

4. REFERENCES

[1] 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.

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

[3] 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.

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

[5] N. d'Alessandro, C. d'Alessandro, S. Le Beux, and B. Doval. Real-time calm synthesizer : new approaches in hands-controlled voice synthesis. In *Proceedings of the 6th International Conference on New Interfaces for Musical Expression (NIME'06)*, pages 266–271, Paris, France, June 2006.

[6] S. S. Fels and G. E. Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE Transactions on neural networks*, 3(6):1–7, November 1992.

[7] L. Feugère and C. d'Alessandro. Digitartic: bi-manual gestural control of articulation in performative singing synthesis. In *Proceedings of the 13th Conference on New Interfaces for Musical Expression (NIME'13)*, pages 331–336, Daejeon, Korea Republic, May 2013.

[8] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.

[1] `http://youtu.be/d4TV-IcK8c8?t=6m40s`

[2] `http://groupeaa.limsi.fr/_media/membres:feugere:these_apatakavazajawauayamana.wav`