Chorus Digitalis: experiments in chironomic choir singing

Sylvain Le Beux, Lionel Feugère, Christophe d'Alessandro

Audio and Acoustics group, LIMSI-CNRS, 91403 Orsay, France

{slebeux,lionel.feugere,cda}@limsi.fr

Abstract

This paper reports on experiments in real-time gestural control of voice synthesis. The ability of hand writing gestures for controlling singing intonation (chironomic singing synthesis) is studied. In a first part, the singing synthesizer and controller are described. The system is developed in an environment for multi-users music synthesis, allowing for synthetic choir singing. In a second part, performances of subjects playing with the system are analyzed. The results show that chironomic singers are able to control melody with accuracy, to perform vibrato, portamento and other types of fine-grained intonation variations, and to give convincing musical performances.
Index Terms: singing synthesis, intonation, gesture control, performative synthesis.

1. Introduction

Research in gestural control of speech and singing synthesis can be traced back to the famous Kempelen speaking machine. The next step arose with the venue of electrical circuitry for speech synthesis, and reached maturity with the Voder [1]. This electro-mechanical approach needed a large amount of training. Then, along with the development of computers, automatic text-to-speech synthesis emerged, on the one hand, with intelligibility but few expressivity, and computer music on the other hand, with musical expressivity, but of course no linguistic intelligibility. The first encounter between speech synthesis and gestural control seems to be "glove talk" [2] a pioneering work using two data gloves and a foot pedal. Experiments in singing synthesis using a graphic tablet were reported in [3, 4, 5].

Intonation stylization using "chironomy", i.e. the analogy between hand gestures and prosodic movements, has recently been studied in some details, in the context of speech intonation [6]. The experiments used an intonation mimicking paradigm: the task of the subjects was to copy the intonation patterns of sentences with the help of a stylus on a graphic tablet, using a system for real-time manual intonation modification. Distance measures between gestural copies, vocal imitations, and original sentences, together with perceptual testing showed that vocal imitation and chironomic imitation are comparable. Moreover, the best stylized contours using chironomy seem perceptually indistinguishable from natural contours. This indicates that chironomic stylization is effective, at least for synthesis of speech intonation.

The present research addresses chironomic control of intonation in singing. The precision needed for intonation in singing is about 4 hundredth of semi-tone, although distances between chironomic stylized contours and natural intonation contours in speech were around one semi-tone. Other differences between speech and singing are that the melody is explicitly specified in singing, using musical notation, that tones are generally very sable (i.e. no melodic glissando), and that melodic ornaments like vibrato are often used. Choral singing is a source of motivation for electronic musicians as well as for singers: it involves synchronization, musical nuances, etc. It is therefore a very interesting framework for performance studies.

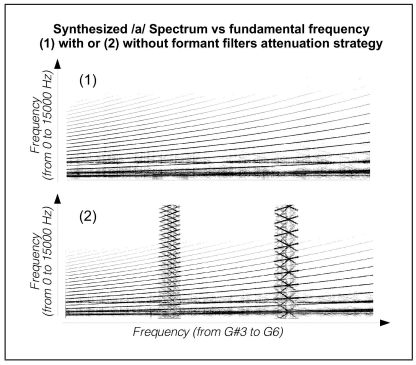

Figure 1: Source-Formant interaction for avoiding"harmonic whistling".

The paper is organized in 2 parts. In the next Section, the system for chironomic singing synthesis called "Cantor Digitalis" and the environment for chironomic choir singing (i.e. multi-user chironomic singing synthesis), called "Chorus Digitalis", are presented. Experiments with Chorus Digitalis are reported in Section 3.

2. Chironomic choir

2.1. Cantor Digitalis: an improved real-time formant synthesizer

Cantor digitalis is an improved parallel real-time formant synthesizer, including the following features:

1. the voice source is the CALM model, allowing precise spectral control of source characteristics.
2. voice categories (barytone, tenor, alto, soprano) are specified, with specific timbre, voice range profile, vowels and voice registers.
3. special attention is paid to pitch continuity, even for very high voices (pitch steps are never audible).
4. source-filter interaction is computed for smoothing timbre variations according to pitch variations (formants/pitch tuning).

The CALM model [7] allows for spectral control of voice source parameters. Source parameters, like voice open quotient, spectral tilt, amplitude of voicing, are controlled using higher-level vocal dimensions like voice tension, breathiness, roughness, or vocal effort. For singing high pitched notes, the source component is sampled initially at $8 \times 44\,100$ Hz , then filtered

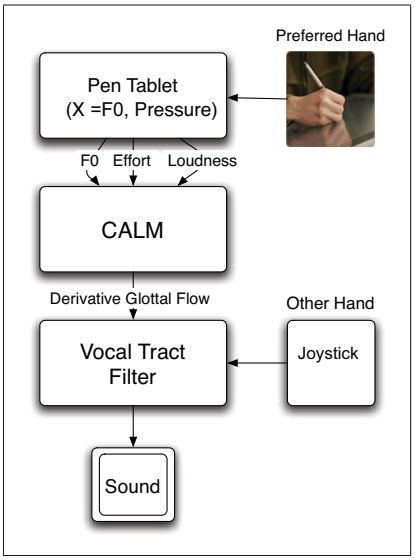

Figure 2: Controls of the synthesizer

and down-sampled so as to increase frequency resolution. Voice parameters pre-sets are defined for mimicking different voice registers. Pitch and loudness are linked using voice range profiles of real singers.

Specific voice types were derived from the formants values of speakers and singers, with some additional manual tuning to improved timbre homogeneity over the whole pitch range for a given register. Then most of the speakers and singers formants frequencies, bandwidths and amplitudes were further tuned so as to improve homogeneity and coherence between the three cardinal vowels of each individual voice and between the voices themselves.

A special attention is paid to the over-amplification of pitch harmonics by the resonant filters, in order to obtain homogeneous sound quality over the melodic range of each specific voice, as described below.

In singing synthesis, it is important to keep the voice intensity level rather constant when moving the pitch solely, and to avoid too strong resonance (that can lead to sound saturation or formant "whistling"). To avoid formant/harmonic whistling, the first six harmonics of the fundamental frequency are taken as references to modify formants amplitudes. Only formant amplitudes are modified (and not formant frequencies), so as not to alter vowels too strongly. Thus, when F_0 or one of its first six harmonics approaches the frequency of one of the first three formants, then the amplitude of this particular formant is decreased gradually, to reach a minimum when exactly matching with F_0 or one of its harmonic values. The two parameters - amplitude decrease and the frequency interval where the decrease occurs - depends on the formant index and continuously changes with F_0. These values were chosen empirically, so as to minimize the resulting audible amplitude variation due to the combination of harmonic/formant resonance and correction. The difference with and without the attenuation is illustrated on Figure 1.

2.2. Cantor digitalis: gesture controls

Cantor Digitalis, is driven by a modified Wacom graphic tablet. Wacom tablets of various size are used. They are augmented by a transparent keyboard, with lines and relief prints indicating the true pitch. The player is able to feel the true pitch on

Figure 3: Keyboard for controlling the virtual singer (top). The Chorus digitalis performing (bottom).

the continuous surface of the tablet. The pitch range is limited to about two octaves, as the aim is to implement specific voices (e.g. tenor or barytone). The X-axis represents pitch, including some aspects of the voice range profile. Each specific voice is associated to a specific voice range profile, i.e. an interdependency of loudness and pitch, as observed in singers. After some experiments, it appeared better to leave the Y-axis free (i.e. moving along the Y-axis has no effect). This gives more freedom to the player in his realization of pitch ornaments: it is possible to play with a relaxed wrist, making waving motion for instance for pitch vibrato (see Figure 3).

The stylus is played with the preferred hand, taking advantage for intonation control of the huge amount of training developed in the process of hand writing. Previous work showed that, almost without training, any subject was able to perform accurate pitch control with a stylus and a graphic tablet, at least for pitch intonation.

The stylus controls vocal effort and voice register. The two extremities of the Wacom stylus are equipped with different points (a fine point and a thick point) and springs. The player feels the spring reaction when pressing on the point, and this effect is associated with vocal effort. Vocal effort is a mixture of sound intensity and voice source spectral tilt: increasing the effort increases the sound intensity and decreases the source spectral tilt (it enhances spectral richness).

The other hand controls a joystick, using the following mapping:

1. vowel pre-sets are available on the hat switch, and controlled by the thumb.
2. sound volume is controlled by the throttle
3. the voice type (barytone, tenor, alto, soprano) is set by extra buttons on the base
4. the lax-tense dimension, aspiration noise, jitter and shimmer are controlled by the stick motion.

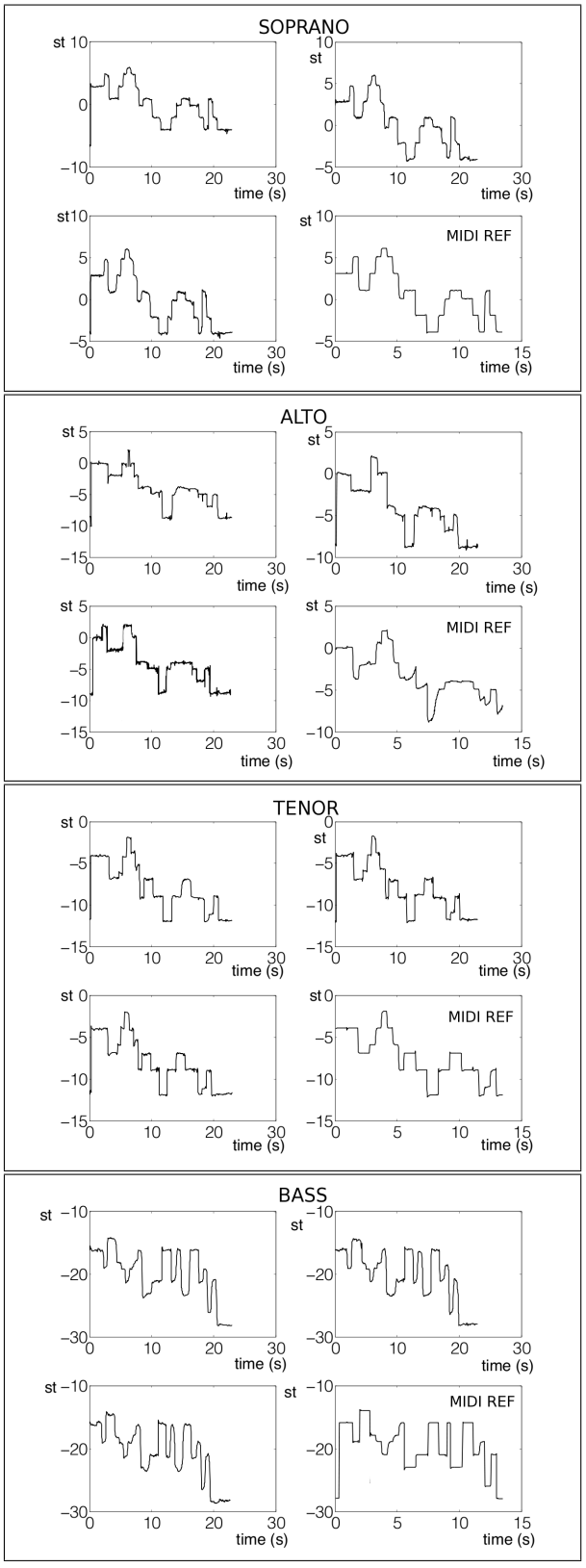

Figure 4: Comparison of 3 gestural and automatic MIDI performances (Pitch in semitones Re 440 Hz, time in seconds)

2.3. Chorus Digitalis

Orchestra of laptop have recently developed in various places [8]. Each player control a personal computer. The Chorus Digitalis project differs from this approach, as a single computer is used for multi-user music performances. This is possible thanks to a software environment, The Méta-Mallette, developed by Puce Muse [9]. This environment targets the general public by proposing various computer music instruments (e.g. transformation, manipulation of samples, synthesis) on a same platform. Several instruments on a same computer are played by several users with the help of several USB interfaces. Each instrument/player has a dedicated number of Audio & video I/Os that can be linked and shared together. Cantor Digitalis has been integrated in this environment, giving birth to a choir, Chorus Digitalis. Each of the chorus members controls one voice synthesizer, with a particular singer characterization.

3. Experiments in polyphonic Singing

Experiments were conducted with the Chorus Digitalis quartet, composed of 4 musicians, 4 graphic tablets and 4 joysticks (See Figure 3). The members of the quartet are the 3 authors of this paper and an additional member of the laboratory. With a limited amount of training (a few hours per weeks during about 3 months) the choir was able to play polyphonic choral music and improvisations.

In this section, performances of the choir playing an anonymous XVth century four parts polyphony, “Alta trinita beata” are analyzed. Each part of the polyphony is played with a different voice. Only intonation is analyzed herein.

Three chironomic versions and a MIDI version of the tune are analyzed. They are displayed in Figure 4. For each part, F_0 is analyzed and compared to the theoretical F_0 values, as written in the score, using a MIDI representation of the score. Figure 5 shows the score, the chironomic performance and the automatic MIDI performance of a section of the tune (from bottom to top: bass, tenor, alto, soprano). The tempo of the chironomic and MIDI performances are different. Gestural performances include melodic ornament like vibrato.

On average, the difference between the score and played pitches is about 5-10 % of semi-tone, including vibrato. This difference is a bit higher for the bass singer on average, but this player played with more legato than others, which can explain that the notes were less steady. One can conclude that pitch accuracy is acceptable.

Concerning vibrato, the bass player had on all trials a vibrato in the same order of magnitude as real singers, around half a semi-tone. All other players have a vibrato of approximately a quarter of semi-tone. This can also explain that the bass player was less accurate than the other ones, as the greater extent of the waving movement can lead to a less accurate pitch.

Moreover, for all players, the average pitch was less accurate for the last trial. At the same time, this trial was achieved with an overall increased tempo thus highlighting the trade-off between speed and accuracy.

Finally, we observed the tuning of the players with one another for both octaves and fifths (except for the alto player for which too many notes were missing). It revealed that on average the accuracy was the same than for the each of them solely, around 5-20 % meaning that although not at the perfect pitch, all players tend to be tuned with one another.

Synchronization of the singers is also acceptable. As a matter of fact, the Chorus Digitalis is facing the same types of difficulties than a real choir: synchronization, pitch accuracy, common nuances, etc.

The synchrony of the virtual singers in the choir is evaluated by note tracking and temporal comparison of notes onsets. A note onset is defined according to pitch changes. Thus, two consecutive notes with the same pitch and without pause between each other present only one onset, the one of the first note.

For each singer, the delays between the reference MIDI score onsets and the singer onsets all along the musical phrase are measured. The three chironomic versions as well as each singer in a version are studied. The first and last onsets of each part are aligned using translation and time stretching.

From the choir recording, the three versions for each singer, and also the four singers together for each version are analyzed. First, while the onset delay evolution is not constant with time, one can observe a correlation between the onset delays of the three versions for a same singer. Second, tempo variations of the four singer in a same version are well correlated. This indicates that the players listen to each other to get a common tempo.

Note that onset detection based on pitch only may not be enough. Amplitude changes are also important, and are not necessarily synchronized with pitch changes: one can reinforce the onset by an amplitude change only after having actually reached the target pitch.

3.1. Conclusions and Future work

The Chorus Digitalis made its first public performance in March 2011 at UBC, Vancouver (P3S workshop concert, with Boris Doval playing the tenor part). With a relatively modest amount of training, reasonable musical results were obtained, compared to the training needed to play other musical instruments. A short video of this performance is joined to this paper.

The instrument seems playable and viable. The results obtained showed that intonation, ornamentation and synchronization between players achieved good levels of accuracy. The difficulties encountered in virtual choral singing are essentially the same as those encountered in real choral singing.

Future work will be devoted to perceptual and performance analyses using mimicking experiments. The strategies and abilities in learning to play the instrument for subjects with different musical backgrounds will be studied. Extension of the choir to more voices is also planed.

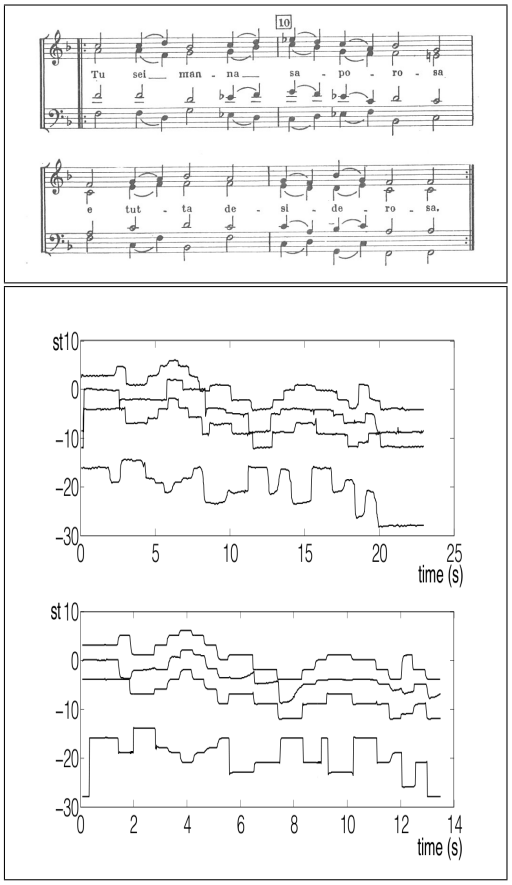

Figure 5: Score of (part of) Alta Trinita Beata, top, example of gestural performance (middle), automatic MIDI performance (bottom). (Pitch in semitones Re 440 Hz, time in seconds)

4. Acknowledgements

This work has been partially funded by the FEDER-Région Ile-de-France Cap Digital project ORJO.

5. References

[1] Dudley, H. (1939). "Remaking speech". The Journal of the Acoustical Society of America, 11(2) :169-177.

[2] Fels, S. and Hinton, G. (1993) "GloveTalk: A neural network interface between a DataGlove and a speech synthesizer", IEEE Trans. Neural Networks 4, 2-8.

[3] Zbyszynski, M., Wright, M. , Momeni, A. and Cullen, D. "Ten years of tablet musical interfaces at CNMAT", in Proceedings of International Conference on New Interfaces for Musical Expression, New York (2007), pp. 100-105.

[4] D'Alessandro, N., Doval,B. , d'Alessandro,C., Le Beux, S., Woodruff, P., Fabre,Y. , Dutoit, T. (2007) "RAMCESS: Realtime and Accurate Musical Control of Expression in Singing Synthesis ", Journal on Multimodal User Interfaces, 1(1), p 31-39.

[5] Cook, P. (2005). "Real-time performance controllers for synthesized singing". In Proc. NIME Conference, pages 236-237, Vancouver, Canada.

[6] d'Alessandro, C., Rilliard, A. and Le Beux, S. (2011) "Chironomic stylization of intonation" J. Acoust. Soc. Am., 129 (3), 1594-1604.

[7] DAlessandro, N. , dAlessandro, C. , Le Beux,S. , Doval, B. (2006) "Real-time CALM Synthesizer New Approaches in Hands-Controlled Voice Synthesis", Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06), Paris, France, pp. 266-271.

[8] Trueman, Cook, P., Smallwood, S. and G. Wang. (2006) "PLOrk: Princeton Laptop Orchestra, Year 1." In Proceedings of the International Computer Music Conference (ICMC'06), New Orleans, U.S.A.

[9] De Laubier, S. , Goudard, V. (2008) 'Puce Muse - La Méta-Mallette," *Journée d'Informatique Musicale (JIM'08)*.