Performative Voice Synthesis for Edutainment in Acoustic Phonetics and Singing: A Case Study Using the "Cantor Digitalis"

Lionel Feugère[1,2], Christophe d'Alessandro[1], and Boris Doval[3]

[1] LIMSI-CNRS, F-91403 Orsay Cedex, France
[2] UPMC Univ Paris 06, F-75005 Paris, France
[3] Equipe lutheries - acoustique - musique, UPMC Univ Paris 06, UMR 7190, Institut Jean Le Rond d'Alembert, F-75005 Paris, France

Abstract. A real-time and gesture controlled voice synthesis software is applied to edutainment in the field of voice pedagogy. The main goals are teaching how voice works and what makes the differences between voices in an interactive, real-time and audio-visual perspective. The project is based on "Cantor Digitalis", a singing vowel digital instrument, featuring an improved formant synthesizer controlled by a stylus and touch graphic tablet. Demonstrated in various pedagogical situations, this application allows for simple and interactive explanation of difficult and/or abstract voice related phenomena, such as source-filter theory, vocal formants, effect of the vocal tract size, voice categories, voice source parameters, intonation and articulation, etc. This is achieved by systematic and interactive listening and playing with the sound of a virtual voice, related to the hand motions and dynamics on the tablet.

Keywords: edutainment, voice synthesis, performative synthesis, graphic tablet.

1 Introduction

Like for any kind of knowledge to be taught, the learning process becomes easier and more entertaining when using various media as teaching materials. Besides, if the students can interact with the media in real-time, success is almost guaranteed. How voice works can be one of this knowledge to be taught.

Splitting a system into several subparts can help for understanding it. Concerning the voice, two observations can be noted. First, a part of the organs of the vocal apparatus is hidden from outside, and it may be dangerous to modify it further than what we can do naturally for understanding its behaviour. Second, despite it is of everyday usage, voice is a complex instrument which involves abstract concepts difficult to understand by the general public.

In this paper, an interactive and real-time application is presented, based on the Cantor Digitalis [1], a musical instrument for singing vowels synthesis implemented in Max/MSP [2].

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 169–178, 2013.

Owing to a signal-type approach and a physically meaningful mapping, it is easy to modify a large number of high level model parameters. The parallel formant [1] synthesis and the source filter model used in this synthesizer allow us to deconstruct the voice model to listen to abstract acoustic phenomena such as individual formants or vocal fold sounds.

A few voice models are available for teaching purposes, such as VocalTractLab [3] or Benoit Project [4]. Their synthesis method can not allow to listen to abstract phenomena, as they are often based on physical models of the vocal apparatus, then reflecting concrete physical phenomena. Also, the original feature of our application is above all the capability to play with the model in real-time and then to listen to the dynamic transformation of the vocal tract [2] and/or the glottal source [3]. The initial goal of the Cantor Digitalis is music, so we can easily use it in an entertaining way for pedagogical purposes by using its control interface while modifying the model parameters and listening to the effects.

After presenting the Cantor Digitalis instrument, we will deal with the decomposition of the source-filter model for a given voice and explain how we use it for pedagogical purposes. Then, still in a real-time interactive perspective, a voice is transformed from one to another through continuous and reactive transformation. We will finally conclude by the main contributions and the perspectives of our work.

2 Cantor Digitalis: Performative Singing Synthesis

Cantor Digitalis is a digital musical instrument allowing for control of pitch, vowel color, strength and quality of a synthetic voice model. Synthesis is controlled in real-time, like a musical instrument, hence the expression "performative synthesis". A general view of the instrument is given at the figure 1. It is based on an improved formant synthesis model, bi-manually controlled by the position and pressure of a stylus and a finger over a graphic tablet. It has been used in concerts within the Chorus Digitalis ensemble [1] [5].

2.1 Formant Synthesis Using the RT-CALM Source Model

The production model of Cantor Digitalis uses the source-filter theory: the glottal source flow is modeled using the RT-CALM [6] and the vocal tract resonances by parallel bandpass filters.

RT-CALM is a real-time version of the CALM model [7]. The spectral properties of the glottal flow model (GFM) are described by the shape of the spectrum derivative: a resonance in low frequencies, and a spectral slope for mid and high frequencies. A white noise modulated by the shape of the GFM represents voice

[1] A formant is the result of one or several vocal tract resonances that contribute to the perception of a vowel.

[2] The shape of the vocal tract changes with the location of the articulators (lips, tong, jaws, ...)

[3] The glottal source is the sound signal created just above the vocal folds.

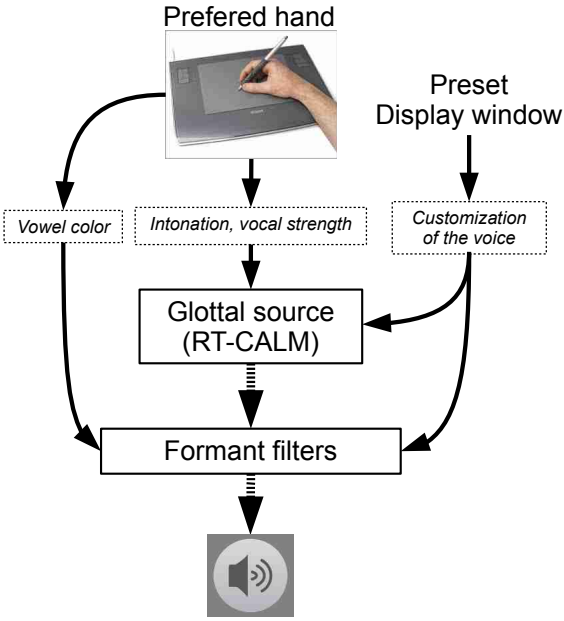

Fig. 1. Functional diagram of Cantor Digitalis

aspiration. Voice strength is realized by an increase in signal intensity, a decrease in the spectral slope of the GFM derivative, and a position shift of the spectrum low frequency maximum toward higher frequencies.

Five bandpass filters model the vocal tract resonances and a bandstop filter the anti-resonance of the piriform sinus [8]. The bandpass filters represent the formants of the vowels. Thus, a database of formants defines the vowels by a set of amplitude / bandpass / frequency values. A scale factor is applied to the formant central frequencies to model the vocal tract size. Then, the different singer types (bass, tenor, alto, soprano) are only characterized by two parameter shifts in our model: this global scale factor and the pitch range.

2.2 Source-Filter Interactions and Automatic Source Perturbation

For singing, a number of rules has been added to the source-filter model, concerning source-filter interactions and automatic source perturbation.

Source-filter interactions are modeled by specific dependencies between:

– Fundamental frequency F_0 and formant central frequencies F_i (i indicates the rank of the formant from the lowest to the highest central frequency).

The literature on acoustics shows that singers modify the frequencies of their formants to adapt them with F_0, in particular F_1 and F_2 are increased

with F_0 in the upper part of their range so that they remain greater than F_0 [9]. Indeed, formants are the key of the voice sound level.

– Voice effort and the first formant central frequency F_1

  It has been demonstrated [10] that F_1 increases by 3.5 Hz/dB between soft and loud voice, or approximately 50 Hz over 15 dB range. For $F_1 = 600$ Hz, a scaling factor proportional to vocal effort can be applied to F_1 , in order to get a 10% increase of F_1 from soft vocal effort to maximum vocal effort (voice effort parameter approximately evolves linearly with sound level).

Automatic source perturbation are divided into deterministic and non deterministic perturbations. We implemented the deterministic heart pulse perturbation on amplitude and frequency of the glottal source vibration, as demonstrated by Orlikoff [11]. They showed that vocal sound pressure and fundamental frequency across a heart cycle has a deterministic perturbation component and that its deviation depends on vocal effort: 14% (soft), 8% (moderate), 3% (loud) for the amplitude variation; 1.4% (soft, moderate), 0.8% (loud) for frequency variation. Along a heart cycle, the perturbation looks like very coarsely a damped sinusoid around the average value and then was modeled as such. The sound result is a small perturbation of pitch and sound level giving more naturalness to the synthetic voice.

Among the non deterministic perturbation are jitter and shimmer which are modeled by a white noise perturbation over the amplitude and frequency of the GFM fundamental pattern.

2.3 Chorus Digitalis: Choral Performative Synthesis

Cantor Digitalis is controlled by one or two graphic tablets. A pen is used to control the pitch very accurately along the X-dimension with the preferred hand, taking advantage of our writing skills. The stylus pressure over the tablet is mapped to voice effort. The tablet is visually augmented with a continuous pitch keyboard to help playing accurately.

The vowel color has been mapped in several ways. By using a second tablet or a part of the main tablet, we can control the vowels in a 2D space with the second hand (figure 2), where four vowel formants are interpolated to get a continuous vocalic space. An alternative to bi-manuality is to control the vowel color along the Y-axis of the single tablet with the preferred hand.

The Chorus Digitalis musical ensemble has performed several times, using 1 to 6 Cantor Digitalis, each controlled by a musician and with a customized voice. The figure 3 represents the Chorus Digitalis publicly performing in May 2012.

3 Edutainment: The Source-Filter Model of Vocal Production

After having designed an application for gestural control of voice synthesis for some years, we realized that in order to teach how the voice works, a good way was to *deconstruct* what we have done.

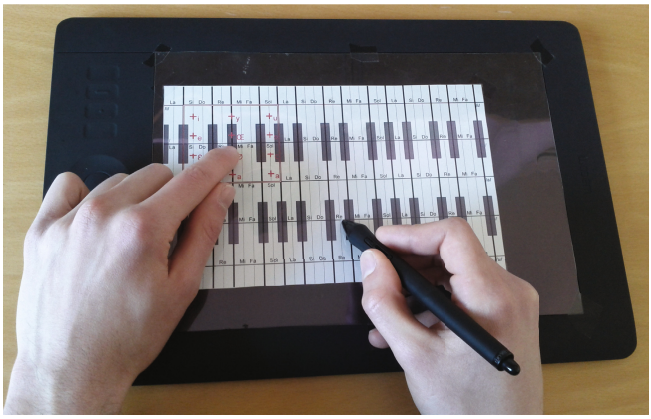

Fig. 2. The bi-manual controller

Fig. 3. The Chorus Digitalis in concert

We implemented a software interface to easily listen to certain parts of the voice model. The figure 4 is a screen capture of the software interface. It is separated into a clickable area and a visualization area. The clickable area allows one to build or deconstruct a voice using the source-filter approach. The visualization area displays a real-time spectrum of the output, i.e. what is heard. Below are detailed possible uses of the software for teaching purposes.

3.1 Playing with the Glottal Source

Pitched voice source is an acoustic effect of the vibration of the vocal folds. However, this intrinsic sound is never heard separately, as the voice sound is the convolution of the source and the vocal tract. Besides, it is impossible to remove

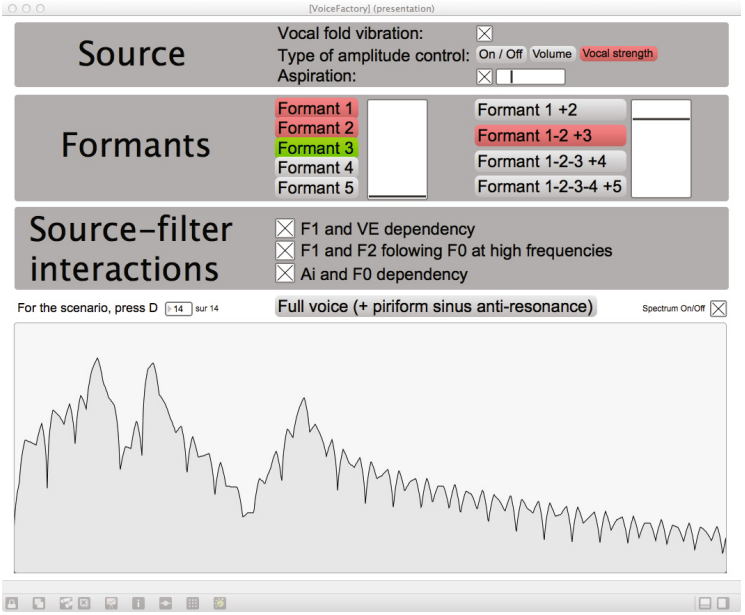

Fig. 4. Screen capture of the "voice factory" interface

the vocal tract, whereas by analogy it would be possible to do it with a saxophone for example, where the mouthpiece can be removed from the instrument body.

With a voice model using the source-filter theory, it is possible to listen to the source, as if the vocal tract had been removed. Then, by using the same interface as with the Cantor Digitalis, the player can listen to how the glottal source sounds. The glottal control is limited to pitch and vocal effort. Then changing vocal tract configuration (vowels) does not affect the sound.

The software interface allows for choosing the mapping between the stylus pressure and the voice intensity, and then to test it while listening to the different mapping: an on/off mapping depending on stylus contact with the tablet; a volume mapping, by linearly controlling the signal amplitude with the stylus pressure; a more natural control, vocal effort mapping, which acts on amplitude and spectral slope in high frequencies.

Finally, the source noise, due to turbulent flow around the vocal fold, can be added to the harmonic source sound or can be listened independently, an effect hardly achieved with a real voice.

3.2 Listening to Individual Formant Motions along Articulatory Trajectories

Using the glottal source sound, the interface allows for continuously moving each formant to listen to its filtering effect. Formant is a common term in language science, but it can be difficult to understand, above all for students who are not

familiar with signal processing. Moreover, an isolated formant never occurs in speech, as the global formants pattern results of the global shape of the vocal tract. Then it is difficult to identify each formant in the sound of a real voice.

While listening to one of the 5 formants, the vocal tract can be modified by exploring the vowel space and then the contribution of this formant to the chosen vowel can be identified. The real-time control of the center frequency of the formant allows to roughly identify articulatory movements: jaw aperture to formant 1, tongue position to formant 2, lip aperture to formant 3.

3.3 Combining Formants to Make Vowel Identification Emerging

After having listened to each formant independently from each other, the formants can be added one by one from the one with the lowest to the highest central frequency. Each additional formant is continuously controlled owing to a slider to highlight the sound effect. The addition of the 2^{nd} formant to the 1_{st} one allows for identification of almost all vowels. The addition of the 3_{rd} formant mainly resolves ambiguities between a few high vowels. The 4^{th} and 5^{th} formant improve voice quality, but without changing the vowel quality.

It is very interesting to listen to the emergences of human voice and vowels, while looking at the spectrum evolution. It really enables to understand the link between sound and spectrum, through concept of resonances/formants.

The harmonic contribution of the glottal source sound can be removed, allowing to listen to the effects of the formants on the turbulent noise of the source, while moving into the vowel space.

3.4 Synchronizing the Voice Source and Vocal Tract Motions

This software has been used several times with general public (science festival, science & music day, ...) and in classes with scientific students.

Besides the above presented applications, it was sometimes used to illustrate the coordination task of the glottal source and the vocal tract to produce small words with a given expression. A person was asked to control the glottal source (pitch and vocal effort) while an other had to control the vocal tract (vowel color). They were asked to reproduce short expressive words like *Oh yeah* or *Oui-oui* (*Yes-yes* in French): first by analysing the evolutions of the pitch and the vowel color; second by trying to reproduce the appropriate gesture; last by synchronizing their gestures to produce the targeted expressive word. This allows people to understand how their own voice is produced by controlling a synthesized voice. The enthusiasm was clear and the users understand fast how to improve their first trial and to generalize the method to produce other short words and different expressions.

The source-filter interaction presented in the section 2.2 can be switched on or off, in order to appreciate its effect.

4 Edutainment: Voice Settings and Individualization

In the preceding section, we presented the way of building / deconstructing a given voice to teach how the vocal apparatus works. Now, we are talking about how to use voice individualization for teaching purposes.

For this sake, we use the voice individualization window, a part of the Cantor Digitalis application. The window interface, as shown in figure 5 is divided into presets and manual settings. The upper part concerns voice types and the bottom part concerns formant values of the vowels. All these settings apply immediately to the voice sound.

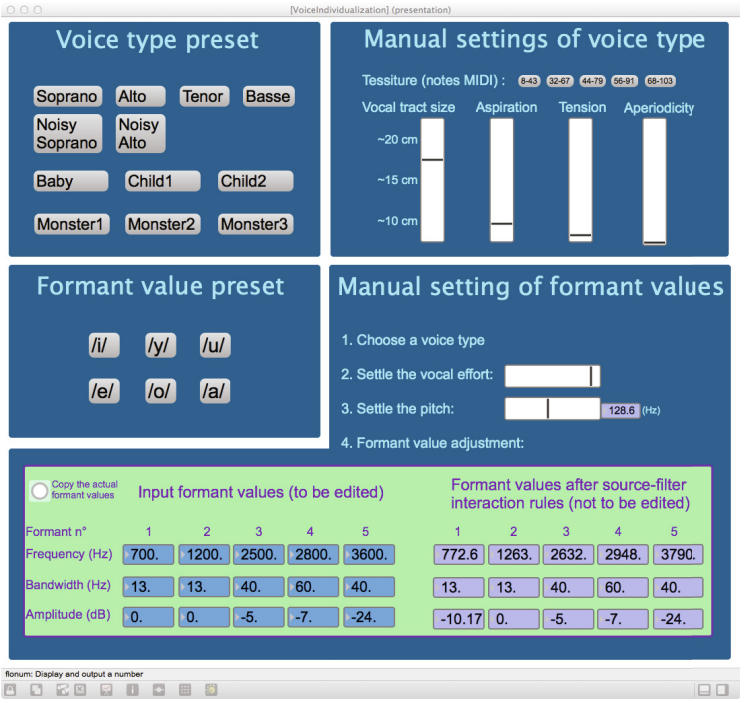

Fig. 5. Screen capture of the individualization voice interface

4.1 Adjustment of Formants

A basic set containing 6 vowels is available. By interpolations of these six vowels, continuous intermediate vowels are obtained, using gestural control over the 2-dimensional plan of the graphic tablet.

Each bandpass filter, representing a formant, can be manually adjusted through its central frequency, bandwidth, and amplitude. Then 3 values for the 5 formant filters can be set independently at the bottom left of the interface window, and the effective values (after the source filter interaction) are given at bottom right. The modification is realized in real-time but it is not possible

for the user to play this new vowel with the tablet (pitch and vocal effort are controlled via a slider). This can be used to demonstrate that different vowel colors (i.e. different formant configurations) can be perceived as a same vowel, an effect demonstrating the categorical perception of vowels.

4.2 Vocal Tract Size

The smaller the vocal tract, the greater the frequency resonances, and then the greater the central frequency of the vocal tract formants. Then the central frequency of its formants will be larger. Thus, we can change the apparent vocal tract size of our synthesized voice by a common multiplier applied to the central frequencies of the formant filters. In real-time, we can pass from a baby voice (small vocal tract) to an adult voice (large vocal tract). This is even more effective if the pitch range consistently decreases with an increase of the vocal tract size.

4.3 Voice Quality Parameters

Several voice qualities parameters are available for controlling the glottal source model, such as aspiration noise rate, voice tension (inversely proportional to the frequency of the glottal flow spectrum maximum) or aperiodicity (jitter/shimmer of the vocal fold vibration).

4.4 Beyond Human Voice

By modifying all these parameters in real-time, it is easy to individualize the synthetic voice by trial and error. It is possible to demonstrate the vocal apparatus similarities between humans and some animals, by changing the vocal tract size, aspiration noise and vocal fold tension. An impressive example is obtained by the transformation of human voice into a beast roar by increasing a lot the vocal tract size, adding aspiration and vocal tension.

5 Conclusion

We presented an application derived from the Cantor Digitalis instrument and intended for acoustic phonetics teaching. It allows to use gestural control to interact with a real-time building and individualization of a voice model.

Two main paths can be explored: 1. How does the voice basically work using source-filter model approach? 2. What are the differences and similarities between human voices (or even some animal voices)?

Abstract phenomena like formants, vowel identification, voice decomposition into source and filter can be understood with the help of audio and gestural real-time interaction.

This application has already been demonstrated in several edutainment contexts like general public science festivals or during university classes. But no proper objective or subjective evaluation has been done yet.

6 Perspectives

Being first of all intended for voice teachers and in order to help as many teachers/students in voice education as possible, the next step would be to make the application available as free or open-source software. Now, the software is not yet available.

New features could be added to the present prototype, according to the needs expressed by users like voice or music teachers.

Finally, this type of performative synthesis is ideally suited to the creation and design of new voices ("human" or "monster-like") in a fast way, because of the real-time response of the system to any parameter modification. A specific and powerful feature of performative synthesis is its ability to play with vocal dynamics and vocal motion, that are often the key for natural sounding voice synthesis.

References

1. Feugère, L., Le Beux, S., d'Alessandro, C.: Chorus digitalis: polyphonic gestural singing. In: 1st International Workshop on Performative Speech and Singing Synthesis, Vancouver (2011)
2. Max/MSP, `http://cycling74.com/products/max/` (visited on February 28, 2013)
3. Vocal Tract Lab, `http://www.vocaltractlab.de/` (visited on February 28, 2013)
4. Benoit Project, `http://benoit.susannefuchs.org/tutorial3.html` (visited on February 28, 2013)
5. Le Beux, S., Feugère, L., d'Alessandro, C.: Chorus digitalis: experiment in chironomic choir singing. In: Proceedings of the Conference on 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Firenze, Italy, August 27-31, pp. 2005–2008 (2011) ISSN: 1990-9772
6. d'Alessandro, N., d'Alessandro, C., Le Beux, S., Doval, B.: Real-time calm synthesizer: new approaches in hands-controlled voice synthesis. In: Proc. of New Interfaces for Musical Expression, Paris, pp. 266–271 (2006)
7. Doval, B., d'Alessandro, C., Henrich, N.: The voice source as a causal/anticausal linear filter. In: Proceedings of Voqual 2003: Voice Quality: Functions, Analysis and Synthesis, ISCA (2003)
8. Dang, J., Honda, K.: Characteristics of the piriform fossa in models and humans. J. Acoust. Soc. Am. 101(1), 456–465 (1997)
9. Henrich, N., Smith, J., Wolfe, J.: Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones. J. Acoust. Soc. Am. 129(2) (2011)
10. Liénard, J.-S., Di Benedetto, M.-G.: Effect of vocal effort on spectral properties of vowels. J. Acoust. Soc. Am. 106(1), 411–422 (1999)
11. Orlikoff, F.R.: Heartbeat-related Fundamental Frequency And Amplitude Variation In Healthy Young And Elderly Male. Journal of Voice 4(4), 322–328 (1990)