PHONODEON: CONTROLLING SYNTHETIC VOICES VIA MIDI-ACCORDEON

Anastasia Georgaki
Music Department,
University of Athens,
Athens, Greece
georgaki@music.uoa.gr

Ioannis Zannos
Audiovisual Arts
Department, Ionian
University,
Corfu, Greece
iani@otenet.gr

Nikolaos Valsmakis,
Technological Institute of
Crete, Department of Music
acoustics and Technology,
Rethymnon, Greece
vals@stef.teicrete.gr

ABSTRACT

This paper presents research on controlling synthetic voice via a MIDI accordion. As motivation for this research served the goal of reviving "lost instruments", that is of using existing traditional instruments "lost" to new music genres as interfaces to control new types of sound generation in novel and future genres of music. We worked with a MIDI-accordion, because it provides expressive control similar to the human breath through the mechanism of the bellows. The control interface of the two hands is also suited for characteristics of the singing voice: The right hand controls the pitch of the voice as traditionally used, while the left hand chooses combinations of timbres, phonemes, longer phonetic units, and different kinds of vocal expression. We use a simple model of phonemes, diphones and timbre presets. We applied this to a combination of sample based synthesis and formant based synthesis. Our intention is to extend the work to physical models. A further goal of this research is to bridge the gap between ongoing research in the domain of synthesis of the singing voice with that of experimentation in novel forms of artistic performance, by examining different ways of interpreting synthesized voices via the MIDI accordion.

INTRODUCTION

In previous papers [7,8], we have reviewed several fundamental domains of ongoing research in the synthesis of the singing voice (synthesis models, performance by rules, text-to- speech synthesis, controllers), and provided a taxonomy of existing approaches. We have also investigated technical problems regarding realistic simulation of the singing voice.

Though systems and software that synthesize singing voices exist, the work entailed in creating realistic imitations of the voice is complicated and onerous. Even after much fine-tuning, it is virtually impossible to synthesize vocal singing so as to be indistinguishable to listeners from real singing.

Research in this domain is primarily concerned with the refinement of techniques for the analysis - synthesis of the vocal signal, concentrating either on the development of new techniques addressing specific particularities of the vocal signal, or the refinement of established sound synthesis techniques. For example one of the techniques based on estimating the spectral envelope, while allowing the precise extraction of formants of high frequencies of female voices, poses many difficulties at the level of analysis [X. Rodet, 1992]. On the other task of automatic generation of control parameters is further complicated by the non-linearity of the vocal signal [P.Cook, 1996].

While research on the synthesis of the singing voice is well established and widely present in computer music and related fields, the aspect of controlling the synthetic voices through real instruments is yet understudied. Systems like Squeezevox[1] [2] and COWE present new ways on controlling synthetic voices from real instruments with special MIDI controllers[2]. These systems are yet under development and have not fully realized a "natural fit" to the multiple control parameters of the singing voice. Aim of the present research is to contribute in this line of research based on a different type of controller.

THE MIDI ACCORDION AS CONTROLLER

Tools can be viewed as extensions of human dexterities, developed with a specific intention. From this perspective, musical instruments can be viewed as extensions of the human voice and touch, conceived in order to combine the emotional and expressive characteristics of the singing voice with the dexterity of the fingers, hands or also other parts of the body. While the voice remains the most difficult instrument to master, making virtual voices "tangible" to the hand via instrument controllers will enable performers to experiment with manual agility on the numerous complex parameters of vocal timbre and expression.

---

[1] Squeezebox is an accordion device controller for controlling synthesized singing voices which allows sophisticated control of the vocal synthesis in real-time. With the right hand, pitch is controlled by the keyboard, vibrato with aftertouch, fine pitch and vibrato with a linear strip. Breathing is controlled by the bellows, and the left hand controls vowels and constants via buttons (presets), or continuous controllers such as a touch pad, plungers, or squeeze interface. (http://soundlab.cs.princeton.edu/research/controllers/squeezevox)

[2] How can a player perform and interact between different vocal techniques? The COWE model goes further the idea of accordion and treats the control of vocal models, including BelCanto singers, crying babies, Tibetan and Tuvan singers.

Inspired by work on the "lost instrument" [3] our goal here is to analyse and experiment with diverse configurations of a MIDI controller capable of manipulating a large number of parameters, prerecorded voices and culturally specific vocal models. This will provide the composer with new means for investigation and experimentation (e.g. production of sound chimeras) and the performer with new expressive instruments. Many electronic musical instruments are based on existing designs of traditional musical instruments (e.g. MIDI keyboard or piano, MIDI guitar, Yamaha wind controller, percussion controller, Zeta violin, MIDI accordion etc., [see Pressing (1992).]). At early stages in the development of such instruments, their expressive capabilities were limited due to the absence of quasi-continuous real time control of parameters beyond the triggering of individual sound events. Due to this and other factors, performing musicians were often dissatisfied with the expressive capabilities of these instruments as compared to traditional acoustic instruments[3]. More recently, methods of control are becoming more refined. A new domain of research has emerged whose objective is to develop new mechanisms of control by mapping existing controllers onto synthesis algorithms.

In our case, the objective is a general performance model that is based on the control capabilities of the MIDI Accordion and that can be applied to several different types of synthesis techniques. As the principal author of the present paper is an accordion player we aim to make use of the bellows as a "breathing" mechanism similar to that of the human voice. Our first target was to control manually the time-variation transitions of the vowels and of diphones.

A MIDI Accordion has up to 4 categories of controls which are usually transmitted via MIDI to external devices. The type of MIDI messages sent and their resolution may vary depending on the type of MIDI-conversion device that is attached to the accordion. For the present experiments, we used an Accordion of the type "Victoria" with a MIDI unit of type Q-Select. This system is capable of sending the following messages:

a. The pressure of the bellows, which is controlled by pressing together or pulling apart the two parts of the accordion with the help of the upper left and right arms held between straps and the accordion parts, is sent as MIDI-control no. 11.
b. The control parameters of the piano-type choromatic keyboard played by the right hand are sent,in the form of MIDI note-on and note-off messages. In our case, the keyboard had 38 keys, covering 3 octaves and 1 major second.
c. The control parameters or the left hand playing on several parallel rows of buttons are sent as MIDI note-on and off messages, on a different MIDI channel than those of the right-hand keyboard. In our case, the accordion was equipped with 120 buttons arranged in 20 rows of 6 buttons.
d. Other controllers such a footswitch or expression pedal can combined with the accordion as additional controllers. At this stage we did not make use of additional controllers.

There are various implementations of the key switches and the bellows controls. Solid state magnetic or optical switches are used to ensure reliable and carefree operation. The bellows sensor is usually a sensor that measures the pressure inside the bellows relative to the outside environment or relative to an internal reference vacuum. Our MIDI interface did not transmit the velocity of either the keyboard or the buttons. The bellows pressure was transmitted reliably but without distinction of the direction of the movement of the bellows (pushed inward or pulled apart).

INITIAL EXPERIMENTS, APPROACH

At initial stages of this research we tested vocal sounds using a to create expressive vocal-like sounds by manipulating the breath controller and trying out several types of commercial voice-synthesis software, but it was very difficult to produce singing phrases with these. The best results were reached with *Vocaloid* and *Cantor*, which contains a 16 phoneme dictionary. With the latter it was possible to manipulate singing voice synthesis to a limited degree. Although phrases and phonemes could be produced by this approach the control of the parameters was limited, because the control interfaces of these packages are conceived for synthesis of predetermined entire phrase segments rather than for live music performance.

Following this, we decided to develop our own software for integrated real-time control and synthesis of the singing voice, based on existing modular platforms. We started on *Max/MSP* and examined the external objects *MBROLA* (unit concatenation based on pre-sampled phonemes and diphones), *fof* (formant synthesis) and *csound* (in particular the *fof* object within the *Max*/*MSP* port of *csound*). While these objects allow real-time control of sound at the sub-phoneme level, this control is limited to a small number of parameters. Moreover, trying to combine several types of synthesis at once or to extend the models of syntheses posed unsurmountable problems due to the increase in complexity of the resulting systems. This called for an environment that includes a fully fledged programming language. So the next choice as a platform was *SuperCollider*. The *Formant* unit generator of *SuperCollider* can be layered to produce 5 or more formants, by adding several formant units driven by the same fundamental frequency. The results are convincing, even though the basic wave-shape of

[3] On the one hand this dissatisfaction can be attributed to limited resolution, accuracy and responsiveness of the gestural interface, on the other hand the sound synthesis system that is driven by the gestures, usually via MIDI, is not adequate to satisfy the auditive needs of the performer.

the formant itself cannot be changed. Furthermore, given the capabilities of *SuperCollider* for combining any number of unit generators with sample-accurate synchronization and control, we could enrich the initial synthesis model by adding sample-based and spectral modules. So the decision was to develop a generic control model that would enable us to drive our current developing synthesis methods, plus to add new synthesis methods as these become available to us. The resulting model has a three-layered architecture: (1) Performance and Control, (2) Mappping / Drivers and (3) Synthesis Engines:

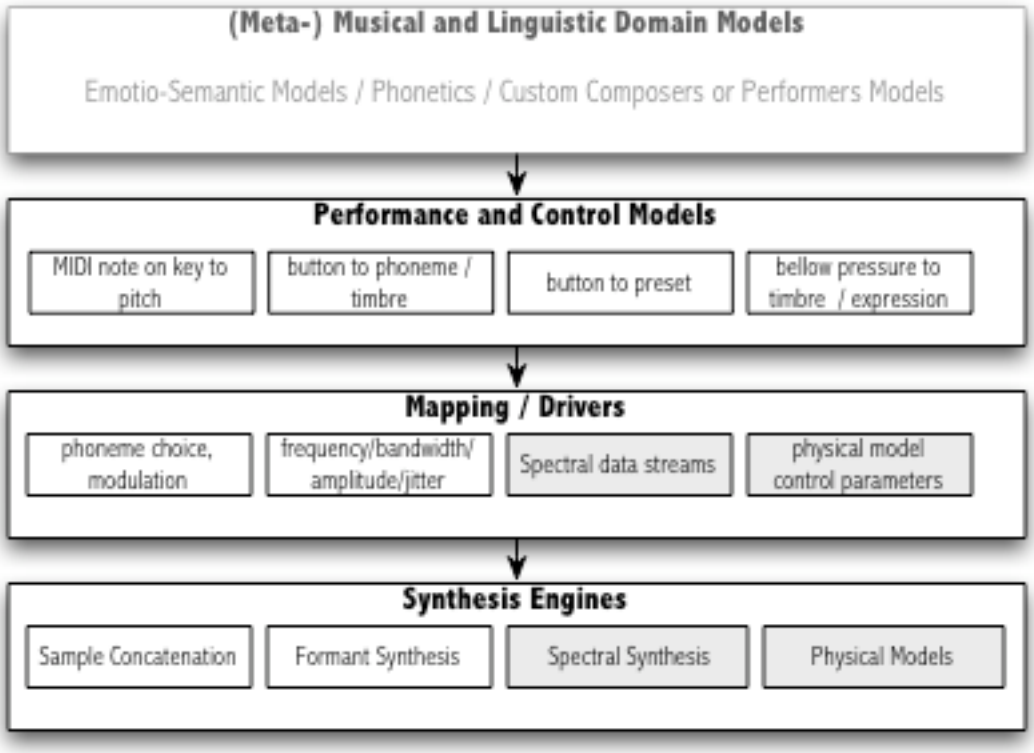

Figure 1: Control-To-Synthesis Architecture of the Phonodeon

The top-level layer, Performance and Control Models, is formed according to choices made during the project based on the characteristics of the specific instrument interface at our disposal. We were aware that a layer above that could be developed to generalize and guide the design based on meta-musical and linguistic domain models, but this was felt as going beyond the scope of the project at its current stage.

IMPLEMENTATION, EXPERIMENTS, RESULTS

The current system contains the following components:

a) MIDI configuration and input. This component connects to the existing MIDI interface for input, and determines the action to be taken when a certain type of input is received. This is done by calling a custom function for each type of midi input, which sends the appropriate messages to the mapping/drivers modules.
b) Phoneme choice module. This module choses a phoneme (or by extension a more complex phonetic unit) from a dictionary of phonemes upon the receipt of a message. The phoneme module then calls a sample-phoneme or formant-phoneme object depending on the specific case.
c) Phoneme sample and concatenation module. This plays back prerecorded phonemes at different speeds, blending them in with the formant module. It is used particularly for the production of consonants.
d) Expression and timbre presets. These are presets of parameters including jitter and/or vibrato of fundamental frequency as well as formant frequency, transition time intervals between individual phonemes or states (for example short times for abrupt transition or longer times for gradual transitions).
e) A state-management/decision module that keeps track of the input from the button combinations chosen by the player in order to decide which phonemes to play with which presets based on the received sequences of button combinations.

We started by using the formant data sets already given by available sources such as documentation on csound's *fof* generator. In further develolping our model, we introduced mechanisms for modifying these basic sets in order to take into account some well known characteristics of the voice such as:

- the unique identity of formants for every individual
- the constant evolution of the articulation during singing speech
- the dependence of the formant frequency to the fundamental frequency and the intensity
- the dependence of the transfer function to the dimension of the vocal tract

The next step in the development of our synthesis model concerned the control of the evolution of the fundamental and formant frequencies. This involved two aspects: First, introducing quasi-regular vibrato as well as more chaotic *jitter*[4] variations in the frequency to increase the expressivity and naturalness of the vowels by imitating the natural instability or controlled variation of the vocal chords tension. Second, introducing a dynamically controllable timing factor for the transition between parameter values when changing from one phoneme to the next one. This makes it possible to control the speed of transition and thus to introduce variation in the manner of articulation. Both of these factors have been integrated in the synthesis model and are controlled by additional parameters that are included in the preset data sets.

At a higher level, it is necessary to consider also the shapes of the individual transition trajectories and the differences in the manner of transition when this happens between phonemes sung in different registers. While some tools have been developed for doing this with the use of interpolated break-point envelopes, this will require further refinement and expansion of our data models before it can be integrated in the current implementation.

The results gained from our experiments so far can be summarized as follows:

a) the synthesis of vowels is satisfactory (especially in intermediate pitches) even if we perceive sometimes a synthetic attribute due to the trajectories of formants. The only problem that has to be solved, concerning the vowels, is the dynamic simulation of the signal and, especially, the control of the transition from one vowel to another.
b) Integrating consonants in singing phrases is problematic with the current approach because of the heterogeneity of the spectra of the sources coming from different synthesis models. While a :
c) Regarding musical applications, it is necessary to spend more time cooperating with composers and performers in order to the construct data sets for quasi-vocal timbers, and more extensive as well as intensely colored timbres with formants for voiced sounds, or of the noisy timbres for consonants.

[4]The *jitter* is a non-periodic modulation of the phonation frequency which is referred to the aleatoric variations of the waveform. The vocal jitter is minimal in a professional singer's voice but it affects the quality of the perceived voice.

CONCLUSION, FURTHER WORK

We succeeded in creating an initial prototype for synthetic singing voices that is controlled in real time from a MIDI accordion. More than the "realistic" quality of the results, it was important for us to develop an interface that would feel flexible and manageable as well as expressive to the performer. The architecture described above is both general enough to allow extensions and simple enough for realizing fairly "direct" control of the sound. In the effort to further develop our work, there are a number of tasks to take on and possibilities to consider::

a) Ameliorate the coherence between manipulation of parameters in the frequency domain (frequencies, bandwidths, amplitudes of formants) and the direct perception that is the recognizability and identification with real voice phonemes of the results.
b) Use rules of evolution of the formants and of interactions between the parameters in order to improve the current simple mechanism of presets and make it possible to group these into perceptually coherent categories.
c) Introduce models of the oscillation of the glottis, of respiration muscles and other acoustic mechanisms such as vocal cords' tension during singing, and of the interaction between vocal effort (muscles, vocal cords etc..) and vocal result (articulators).

In terms of further synthesis techniques to explore, three techniques may be named here: First, use of different types of filters as well as of direct manipulation of the spectrum ("Phase Vocoding"). Second, more extened use of the the unit selection and concatenation method[5] including real-time control of the specific characteristics of the performance (pitch, duration, timbre, etc.) (see Rodet, 2002). Finally, it seems to us that the most promising type of synthesis technique for future work may be based on physical modeling of the voice. Unfortunately, we could not yet connect any of the existing physical model packages such as SPASM/Singer by Perry Cook as well as its STK variants for to our own software for real-time control[6]. This will require time- consuming porting and development work. While this work is under way, our plan is to continue the experiments with the current synthesis modules in order to attempt a more seamless combination of the various synthesis techniques.

[5] This method was firstly been applied on speech synthesis

[6] Some parst may be realizable with already existing Ugens ins SuperCollider. Also, experiments similar to Phonodeon have been made by Perry Cook and Colby Leider. See http://www.cs.princeton.edu/~prc/CHI01Web/chi01.htm and http://www.cs.princeton.edu/~prc/NewWork.html#STK

REFERENCES

[1] Cook, P. R., and C. Leider (2000) "Making the Computer Sing," *Proceedings of the XIII Colloquim on Musical Informatics*, L'Aquila, Italy, September,

[2] Cook, P. R., and C. Leider (2000) "Squeeze Vox: A New Controller for Vocal Synthesis Models," *International Computer Music Conference* , Berlin, August,

[3] Dufourt Hugues(1996) L'instrument philosophe. Entretien avec p. Szendy, *Cahiers de l'Ircam, no7,*Paris.

[4] Gael Richard (1990) *Rules for fundamental frequency transition in singing synthesis*. Dept of Speech Communication and acoustics, Royal Institute of Technology, Stockholm.

[5] Georgaki Anastasia (1998a) *Problémes techniques et enjeux esthétiques de la voix de synthèse dans la recherche et création musicales.* Thèse de doctorat, EHESS/IRCAM, Paris.

[6] Georgaki A (1999). "Towards the conception of a vocal synthesizer", *Proceedings of the VIth symposium of Brazilian Computer Music,* Rio de Janeiro, Brazil

[7] Georgaki A.(2004a) "Virtual voices on hands". Prominent applications on the synthesis of the singing voice, Sound and music computing Proceedings,SMC05, IRCAM, Paris 2005.

[8] Georgaki A.(2004b) "New trends on the synthesis of the singing voice ", ICMC'04 Proceedings, Miami , Florida, 2004.

[9] Kim Y. E. "Structured Encoding of the Singing Voice using Prior Knowledge of the Musical Score", *Proc. 1999 IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 17-20, 1999.

[10] Meron Y. (1999) *High Quality Singing Synthesis using the Selection-based Synthesis Scheme*. PhD thesis, University of Tokyo

[11] Sundberg J. (2001) "Sounds of singing. A matter of mode, style, accuracy, and beauty", invited paper, *4th Pan European Voice Conference,*Stockholm,

[12] Pressing, J. (1992). *Synthesizer performance and real-time techniques*. Madison, Wisconsin, USA: A-R editions.

[13] Virsyn's "Cantor"(2004) Vocal synthesis software: http://www.kvr-vst.com/get/984.html

[14] Yamaha Corporation Advanced System Development Center. *New Yamaha VOCALOID Singing Synthesis Software Generates Superb Vocals on a PC*, 2003 (.http://www.vocaloid.com)