DESIGNING AND CONTROLLING A SOURCE-FILTER MODEL FOR NATURALISTIC AND EXPRESSIVE SINGING VOICE SYNTHESIS

Olivier Bélanger, Caroline Traube, Jean Piché
LIAM (Laboratoire informatique, acoustique et musique)
Faculté de musique, Université de Montréal

ABSTRACT

In this paper, we describe a voice synthesis model developed for musical purposes. Based on a source-filter model, this synthesizer has been specifically designed to allow the synthesis of natural sounding singing voices by including pitch and amplitude variations and by careful tuning of consonant to vowel transitions. A particular attention is given to the reproduction of plosive consonants. The model covers all singing voice registers, from bass to soprano, and allows the control of several tone quality parameters such as vibrato depth and frequency, voice roughness and articulation speed. Its database is structured to synthesize whole consonant-vowel syllables. As a result, it is relatively easy to construct musically expressive phrases with just a few manipulations and control commands. The model uses Csound as the audio engine and can produce several voices at small cost in CPU.

1. INTRODUCTION

Because we know it so well and because it is characterized by many complex and subtle variations, speech has always been a challenging acoustic signal to synthesize. Pioneers is this field were Kelly, Lochbaum and Mathews, well-known for the Daisy Bell song digital synthesis, and Homer Dudley with his famous keyboard-controlled Voder. More recently, Perry Cook proposed the SPASM vocal tract editor, and Hui Ling Lu a singing synthesizer with vocal texture control [8]. In this project, because we wanted to develop a model that is natural-sounding but also precisely and easily controlled, we chose a classic source-filter model implementation. In this model, as one can expect, the source is a harmonic or noisy signal simulating the glottal excitation and the filter is made of a bank of bandpass filters in parallel simulating the vocal track (see Figure 1). The model itself is not new, but special care has been given to the tuning of the many parameters and modulation signals that make a difference between an artificial-sounding and a more natural-sounding singing voice synthesizer. The digital implementation of this synthesis model is realized in the Csound environment, while control parameters and notes sequences are generated in real time in Max/MSP.

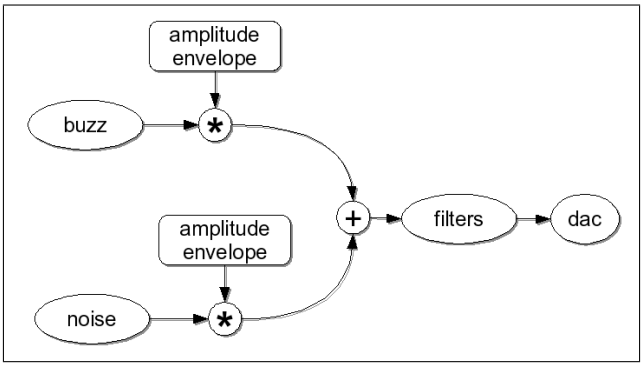

Figure 1. A classic source-filter model for voice synthesis.

2. MODELING THE GLOTTAL SOURCE

The excitation signal, which corresponds to the air flow from the singer's lungs passing through the vocal folds, is produced by two generators : an impulse train generator (*buzz*) and a noise generator. The impulse train generator is used to create a voiced (harmonic) glottal source, with an identifiable fundamental frequency, as needed in the production of vowels. The noise generator is used mainly to synthesize unvoiced sounds, principally during consonant articulation. Two independent amplitude envelopes are necessary to adequately mix these two signals in order to produce realistic vocal sounds. Naturalness of the synthesis is achieved with the addition of a finely tuned vibrato and a modulation producing voice roughness.

2.1. Harmonic source generation

In order to impart a natural character to the synthesized voice sounds, small random variations are added to the source, breaking the regularity of the impulse train. In fact, a voice synthetiser based on strictly periodic vibrations is usually perceived as a « singing-computer » since it lacks the micro-modulations and micro-variations that can be found in the glottal source produced by human vocal folds. In our model, random variations (jitters) are applied to fundamental frequency, amplitude and as well as brightness parameters of the impulse train. All these jitters have their own random range and speed (see Figure 2).

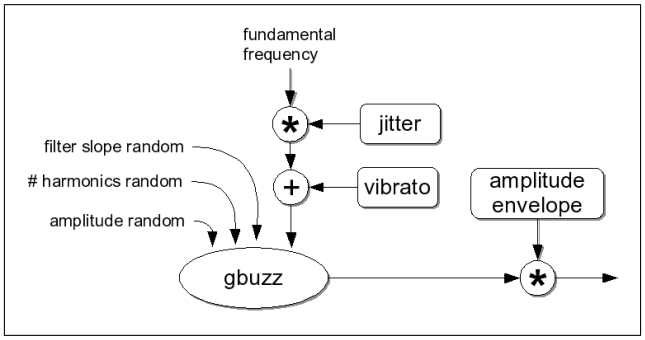

Figure 2. Harmonic source generation.

2.2. Vibrato

A carefully controlled vibrato applied to the harmonic excitation signal can also contribute to the perceived naturalness of the synthesized singing voice. Our experimentations show that a deviation of about 1% of the fundamental frequency with a triangular waveform gives the most natural results for all voice ranges. In our model, the vibrato's depth and frequency are modulated with distinct random variations. This produces a vibrato that is neither too regular nor too mechanical. Naturalness of the vibrato is also achieved by modulating the position of the formants (see section 3) and the amplitude of the overall signal, in compliance with a study on vibrato perception by Verfaille et al. [9]. Finally, we have also taken into account the fact that professional singers rarely produce vibrato from the start of a note. Therefore, as illustrated on Figure 3, our vibrato model includes an amplitude envelope allowing a gradual and smooth introduction of the vibrato after the start of the note.

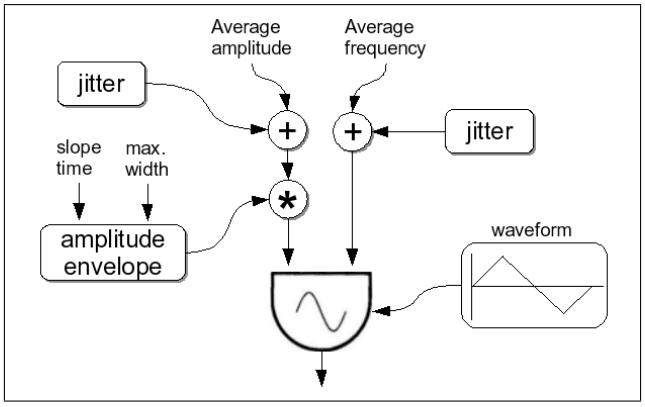

Figure 3. Structure of the vibrato module.

2.3. Noise source generation

The noise source is produced by a pink noise generator and is then modulated according to an amplitude envelope similar in its implementation to the amplitude modulation of the harmonic source. In this way, voiced and unvoiced sources can be mixed and cross-faded before being passed through a bank of bandpass filters simulating the vocal tract. In order to synthesize plosive consonants such as /d/ or /g/, the contours of the amplitude envelope modulating the noise source need to be finely tuned since a noise burst appears for only a brief time, often preceded and followed by an interruption of the source. The duration of this little gap of silence between the noise burst and the start of the voiced sound - lasting from 5 to 30 ms - is called the Voice Onset Time (VOT) and is crucial in the perception and categorization of the different plosive consonants [7]. For example, in the production of /k/, the VOT is much longer than in the production /b/.

2.4. Voice roughness

Another aspect contributing to the naturalness of voice synthesis is the low level noise, in the range 1 kHz to 4 kHz, always present in the production of vowels. This noise is caused by burst turbulence from the lungs when passing through the vocal folds. To simulate roughness, the voiced source excitation from the impulse train generator is multiplied by a filtered pink noise with a very low amplitude [2]. With our model, the user can vary the roughness amplitude in order to create different voice qualities, or to create a whispering effect, when the noise component completely masks the impulse train excitation.

3. MODELING THE VOCAL TRACT

Filtering induced by the vocal tract is simulated with a bank of bandpass filters in parallel, producing perceptually salient amplitude peaks in the magnitude spectrum. Five formants are defined for each vowel of the international phonetic alphabet. The first two are used for the identification of the vowel, the third and fourth formants contribute, with the first two, to the perception and categorization of the consonants. The fifth formant plays a role for tone quality adjustment. Another set of formants can be used to simulate the nasal cavity, as nasal vowels and consonants are produced by the interaction of the oral and nasal cavities, which implies the weakening or strengthening of certain frequency regions in the magnitude spectrum [4]. Each formant is specified with three parameters, stored in tables : central frequency, amplitude and bandwidth of the bandpass filter. These parameters were obtained by spectrum analysis of voice samples and are automatically called depending on the register of the chosen voice and the produced vowel. To make sure that the vowel production is not perceived as too stable or mechanical, the formants central frequencies are randomly varied in a 2% range of their nominal values in order to produce different tone qualities each time a given vowel is sung.

We also had to address a specific problem which arises with voices in the high registers. One peculiarity of the soprano voice is that the central frequency of the first formant often falls below the fundamental frequency. If the formants are not adjusted, we obtain a synthesized voice with a very thin and poor quality. Research by Joliveau et al. [6] shows that the soprano singers modify the shape of

their vocal tract to tune the first formant on the fundamental frequency when the later is higher. This tuning gives a rounder quality to the voice. This was confirmed by our model. It should be noted that vowel perception is altered by formant tuning, but in any case, at a fundamental frequency between 700Hz and 1kHz, the spectral envelope is so sparsely sampled that vowel recognition is always problematic.

4. CONTROLLING FORMANT TRAJECTORIES FOR THE SYNTHESIS OF CONSONANTS

One of the original features of this project is the implementation of formant trajectories for the synthesis of consonants, which is often lacking in source-filter based vocal synthesizers. A stop or plosive is a consonant sound produced by stopping the airflow in the vocal tract. When a plosive consonant (such as /b/ or /d/) is produced, the vocal tract undergoes a series of transformations ending on the shape that produces the targeted vowel. In our model, these transformations are modeled by sweeping the central frequency of each formant in a manner that reproduces the articulation of the consonant. Eq.1 presents the formula we used to generate the trajectory from a starting frequency f_1 to a frequency f_2, where α is the curvature index and N the total number of samples in the transition.

$$F(n) = f_1 + (f_2 - f_1)\frac{1 - e^{-\alpha n/N-1}}{1 - e^{-\alpha}} \qquad (1)$$

Tuning the amplitude and the bandwidth of the filters is also important to reproduce the phenomenon correctly. The transitions during a stop consonant are very fast, about 50 to 100 ms, but the auditory system is extremely sensitive to this kind of changes as it is the principal cue in the perception of stop consonants [1], [5]. There are two other important cues to extend the range of perceived consonants : the presence of noise and the voice onset time (VOT). A noise component is present in all unvoiced or semi-voiced sounds, such as sibilants or unvoiced stop consonants. In the case of unvoiced stop consonants, one needs to precisely adjust the voice onset time. In our implementation, the parameters determining the contours of formant trajectories for a given plosive consonant are stored in a table and are automatically called when a syllable needs to be synthesized. In order to produce the syllable /da/ for example, the formant trajectory starts on locus values for a /d/, then gradually drifts to the target vowel /a/, and finally holds these values for the whole duration of the vowel. This is illustrated on Figure 4 showing the first three formants' central frequencies as a function of time for the synthesis of the syllables /da/ (in blue) and /ba/ (in red).

4.1. Place of articulation and acoustic loci

According to a study by Delattre et al. [3], it is possible to minimize the database by setting the starting point of the second formant transition at the same value for all

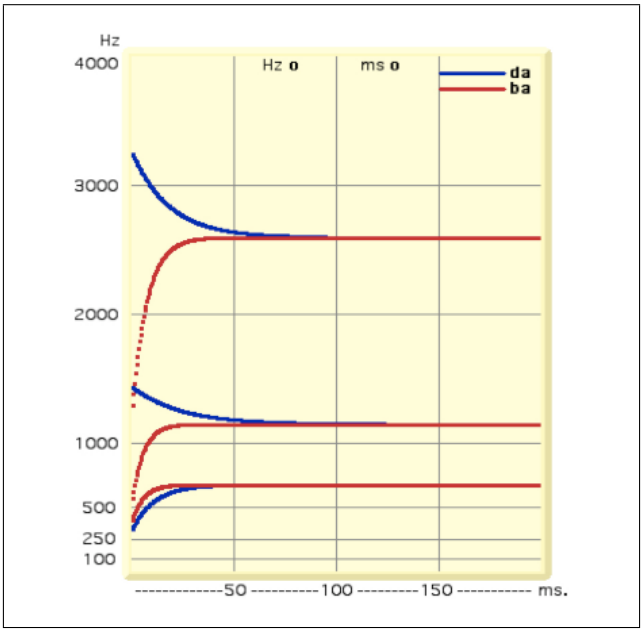

Figure 4. Formant trajectories for syllables /da/ and /ba/.

consonants which have the same place of articulation, independently of the subsequent vowel. For example, the labial consonants /b/, /p/ and /m/ are all articulated at the lips and can be synthesized by starting the transition at the same *acoustic locus* (starting point of the transition). In the same way, alveolar consonants /d/, /t/ and /n/ are all articulated with the tongue against or close to the superior alveolar ridge (behind the superior teeth) and /g/, /k/ and /gn/ are articulated at the velum. Amplitude trajectories have to be finely tuned to respect the locus theory, as shown on Figure 5.

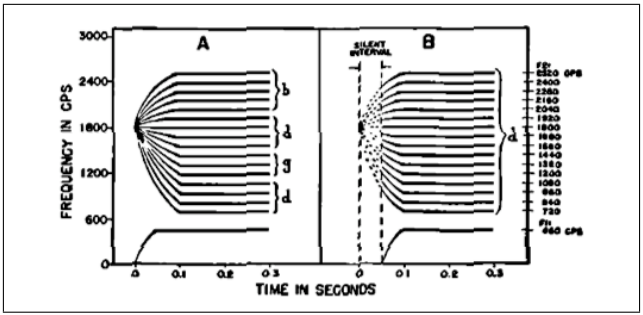

Figure 5. Second-formant trajectories according to place of articulation, with (B) and without (A) silence in transition [3].

5. CONTROL PARAMETERS

An efficient voice synthesis model has to be flexible to allow the production of many kinds of articulated sounds, ideally with just a few control parameters. In this model, most of the parameters that need to be specified to synthesize a syllable are loaded in tables and called in groups to set the consonant, the subsequent vowel and the articula-

tory behavior. Only five control parameters are left to the user : the duration of the note, the fundamental frequency, the consonant, the vowel and the register. There are about 12 other parameters that are optional and that can be used to tuned the behavior of the model.

Articulation is an important aspect of voice production since there is often a continuity between two articulated vocal sounds. It is essential to reproduce this phenomenon in order to synthesize realistic vocal streams. In this model, when calling a new note in Csound, the user can specify that the event is tied to the previous one, by giving a negative duration time. The program then bypasses all initialization commands and starts from the point where the previous event had ended (the set of steady state formants values corresponding to a vowel for example).

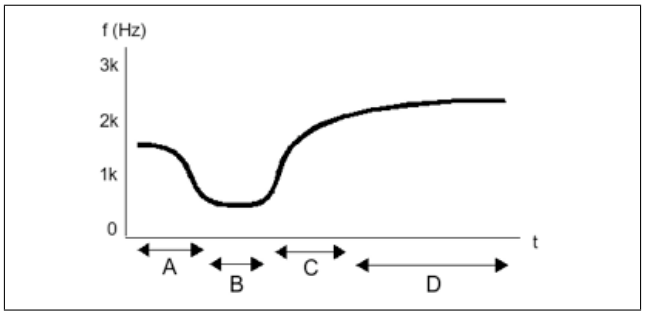

Figure 6. Typical trajectory for one formant central frequency over one event.

Figure 6 shows a typical trajectory for one formant central frequency as a function of time, starting with a sung vowel followed by a new note called with a voiced stop consonant. On this figure, section A corresponds to the falling part of the preceding vowel, with a duration of about 30 ms. Section B represents a gap of silence or a low level resonance during the consonantal portion (important for the perception of voiced stop consonants), just before the attack of the new speech sound. This duration is variable, from 20 to 100 ms, depending on the desired consonant. Section C corresponds to the trajectory of the desired consonant and section D to the steady state fixed by the parameters of the new vowel. All formants parameters (central frequency, amplitude and bandwidth) follow similar trajectories.

6. CONCLUSION AND FUTURE WORK

The source-filter based singing voice synthesis model we have presented in this paper has been specifically designed to improve the synthesis of natural sounding singing voices by including pitch and amplitude variations and by the careful tuning of consonant to vowel transitions. A great advantage of this model is that all the information needed for the production of vocal sounds is included in a single relatively small text file. It is not necessary to download any voice library to make the synthesis works, as is the case with synthesis by concatenation of diphones. We still need to complete the parameter database for the production of some consonants that are not yet implemented, in order to allow the synthesis of whole sentences. Also, we would like to provide the possibility to choose among various voice types and timbres, all defined by the tuning of the tone quality parameters, a feature that will be essential when the synthesizer is used for musical purposes.

7. REFERENCES

[1] Blumstein, S., Stevens, K., "Acoustic invariance in speech production : evidence from measurements of the spectral characteristics of stop consonants", *Journal of the Acoustical Society of America*, Vol. 66, Issue 4, pp. 1001-1017 (1979).

[2] D'Alessandro, N., d'Alessandro, C., Le Beux, S., Doval, B., "Real-time CALM Synthesizer, New Approaches in Hands-Controlled Voice Synthesis", *Proc. NIME'06 - Int. Conf. on New Interfaces for Musical Expression*, Paris, France (2006).

[3] Delattre, P., Liberman, A., Cooper, F., "Acoustic Loci and Transitional Cues for Consonants", *Journal of the Acoustical Society of America*, Vol. 27, Issue 4, pp. 769-773 (1955).

[4] Delvaux, V., Demolin, D., Soquet, A., Kingston, J., "La perception des voyelles nasales du français", *Proc. JEP'04 - Journées d'Etude sur la Parole* - , Fès, Maroc (2004).

[5] Jackson, P., "Acoustic cues of voiced and voiceless plosives for determining place of articulation", *Proc. CRAC'01 - Workshop on Consistent & Reliable Acoustic Cues for sound analysis* (2001).

[6] Joliveau, E., Smith J., Wolfe J., "Vocal tract resonances in singing : The soprano voice", *Journal of the Acoustical Society of America*, Vol. 116, Issue 4, pp. 2434-2439 (2004).

[7] Niyorgi , P., Ramesh, P., "Incorporating voice onset time to improve letter recognition accuracies", *Proc. ICASSP'98 - IEEE International Conference on Acoustics, Speech, and Signal Processing* - (1998).

[8] Smith, J., "Virtual Acoustic Musical Instruments : Review of Models and Selected Research", *WASPAA'05 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2005).

[9] Verfaille, V., Guastavino, C., Depalle, P. "Perceptual Evaluation of Vibrato Models", *Proc. CIM'05 - Conference on Interdisciplinary Musicology*, Montreal, Quebec, Canada (2005).