Comparing time domain and spectral domain voice source models for gesture controlled vocal instruments

Christophe d'Alessandro[1], Nicolas D'Alessandro[2], Sylvain Le Beux[1], Boris Doval[1]

[1]LIMSI-CNRS, BP 133 — F-91403 Orsay, France {cda;Sylvain.le.beux,doval}@limsi.fr

[2] TCTS - FPMs, Mons, Belgium, nicolas.dalessandro@fpms.ac.be

Abstract

Three real-time gesture controlled vocal instruments are presented. They are based on a time domain (LF) and a spectral domain (CALM) model of the glottal pulse signal. Gestural control is able to add expression to the synthetic voices, enabling simulation of various vocal behaviors. Expressive vocal instruments are demonstrated for musical and research purposes.

1. Introduction

Gesture controlled vocal instruments provide new tools for vocal studies. Taking advantage of the fast growing field of music technology, accurately controlled real time voice synthesis systems open new domains of investigation. On the one hand, analysis by synthesis has been recognized since several decades as a much powerful paradigm for speech and voice analysis. On the other hand, most of the research on voice source synthesis until now considered mainly fine grained voice source parameters rather than the domains of variation and co-variation of these parameters. The time span of voice source model is typically one pitch period, although the time span for voice quality perception encompasses several pitch periods or even a full speech sentence or musical phrase. When using real time vocal instruments, the questions of parameter variation and co-variation can no more be ignored or underestimated. Examples of such domains of variation are voice source mechanisms, the phonetogram, co-variation of voice open quotient and noise, vibrato and notes transitions and so one. Gesture controlled voice instruments are also useful for direct perceptive assessment of voice source model as the sounds produced are easily controlled by the "player" and evaluated by both the "player" and the "audience". Finally, real time voice instruments can serve as new musical instruments.

Following the "Speech Conductor" [1] project, this research aims at developing and testing gesture interfaces for driving ("conducting") voice and speech synthesis systems. Gestural control is able to add expression to the synthetic voices, enabling simulation of various vocal behaviours [8]. Then expressive vocal instruments can be designed, for musical and research purposes. In this communication, after a review of voice quality dimensions (section 2) two voice source models for real-time gesture–controlled voice synthesis are compared (section 3). Three vocal instruments are presented (section 4)and discussed (section 5).

2. Voice quality dimensions

No general agreement is currently available on the dimensions of phrase-level voice quality variations (i.e. variations of the vocal activity above the level of the pitch period). However dealing with these dimensions is of the utmost importance for voice analysis, synthesis and perception. A possible framework can be sketched using five main prosodic dimensions:

A. The voice register dimension: Voice registers depend on the underlying voice mechanisms. The different voice mechanisms are corresponding to different voice parameter settings. Changes between mechanisms are usually corresponding to voice "breaks", i.e. sudden voice parameter changes..

B. The noise dimension The noise dimension represents the relative amount of noise in the speech signal, an indication of breathiness or hoarseness. Noise is an important phrase level feature of the voice source.

C. The pressed/lax dimension The pressed/tense dimension corresponds mainly to changes in open quotient. It is a stylistic feature in speech and singing related to relaxed or strangled voices.

D. The effort dimension. The vocal effort dimension corresponds to the nuance *piano* or *forte* in singing. In speech it is used for signalling accentuation. Vocal effort seems relatively independent of vocal pressure.

E. The melodic dimension. The voice has very specific melodic patterns. These patterns are including in singing *vibrato*, *portamento*, note transitions and in speech intonation patterns.

3. Glottal pulse models

3.1 Augmented LF- model

The most widely used glottal flow model is the Liljencrants-Fant (LF) model [2]. This time-domain glottal flow model can be described by equivalent sets of 5 parameters: 1: A_v: peak amplitude of the glottal flow, or amplitude of voicing; 2: T_0: fundamental period (inverse of F_0); 3: O_q: open quotient, defined as the ratio between the glottal open time and the fundamental period. This quotient is also defining the glottal closure instant at time O_q*T_0. 4: A_m: asymmetry coefficient, defined as the ratio between the flow opening time and the open time. This quotient is also defining the instant T_m of maximum of the glottal flow, relative to T_0 and O_q ($T_m= A_m*O_q*T_0$). Another equivalent parameter is the speed quotient S_q, defined as the ratio between opening and closing times, $A_m = S_q / (1 +S_q)$; 5: Q_a: the return phase quotient defined as the ratio between the effective return phase duration (i.e. the duration between the glottal closure instant, and effective closure) and the closed phase duration. In case of abrupt closure $Q_a = 0$.

For realistic voice synthesis, an aperiodic component must also be added to the periodic LF model (called then Augmented LF model, or A-LF). Two types of aperiodicities have to be considered: structural aperiodicities (jitter and shimmer) that are perturbations of the waveform periodicity and amplitude, and additive noise.

3.2 A-LF parameters and phonation dimensions

There is no one-to-one correspondence between voice qualities and glottal flow parameters. They can be sketched as follows: F_0 describes melody. A very low F_0 generally signals creaky voice and a high F_0 generally signals falsetto voice. O_q describes mainly the lax-tense dimension. O_q is close to 1 for a lax voice, and may be as low as 0.3 for very pressed or tense phonation. As A_v represents the maximum flow, it is an indication of flow voice, and it may help for analysis of the vocal effort dimension. Q_a correlates also with the effort dimension. When Q_a =0 the vocal cords close abruptly. Then the asymmetry A_m is generally high, and so is vocal effort. Conversely, large values of Q_a (0.05-0.2) give birth to a smooth glottal closure –the vocal effort is low. The asymmetry coefficient A_m has an effect on both the lax-tense dimension (asymmetry is close to 0.5 for a lax voice, and higher for a tense voice) and the vocal effort dimension (asymmetry generally increases when the vocal effort increases). Therefore some sort of mapping between raw voice source parameters and voice quality dimensions is needed.

3.3 Spectrum of the voice source

Modelling the voice source in the spectral domain is interesting and useful because the spectral description of sounds is closer to auditory perception. Time-domain and frequency domain descriptions of the glottal flow are equivalent only if both the amplitude and the phase spectrum are taken into account, as it is the case in this work.

The voice source in the spectral domain can be considered as a low-pass system. It means that the energy of the voice source is mainly concentrated in low frequencies and is rapidly decreasing when frequency increases. The spectral slope, or spectral tilt, in the radiated speech spectrum (which is strongly related to the source derivative) is at most -6 dB/octave for high frequencies. As this slope is of +6 dB/octave at frequency 0, the overall shape of the spectrum is a broad spectral peak. This peak has a maximum, mostly similar in shape to vocal tract resonance peaks (but different in nature). This peak shall be called here the “glottal formant'”. This formant is often noticeable in speech spectrograms, where it is referred at as the “voice bar”, or glottal formant below the first vocal tract formant.

Spectral properties of the source can then be studied in terms of properties of this glottal formant. These properties are: 1: the position of the glottal formant (or “frequency”); 2: the width of the glottal formant (or “bandwith”); 3: the high frequency slope of the glottal formant, or “spectral tilt”; 4:the height of the glottal formant, or “amplitude”.One can show that the frequency of the glottal formant is inversely proportional to the open quotient O_q [4]. It means that the glottal formant is low for a lax voice, with a high open quotient. Conversely, a tense voice has a high glottal formant, because open quotient is low.

The glottal formant amplitude is directly proportional to the amplitude of voicing. The width of the glottal formant is linked to the asymmetry of the glottal waveform. The relation is not simple, but one can assume that a symmetric waveform (a low S_q) results is a narrower and slightly lower glottal formant. Conversely, a higher asymmetry results in a broader and slightly higher glottal formant.

Around a typical value of the asymmetry coefficient (2/3) and for normal values of open quotient (between 0.5 and 1), the glottal formant is located slightly below or close to the first harmonic ($H_1 = f_0$). For O_q=0.4 and A_m=0.9, for instance, it can then reach the fourth harmonic

Up to now, we have assumed an abrupt closure of the vocal folds. A smooth closure of the vocal folds is obtained by a positive Q_a in time domain. In spectral domain, the effect of asmooth closure is to increase spectral tilt. The frequency position where this additional attenuation starts is inversely proportional to Q_a. For a low Q_a, attenuation affects only high frequencies, because the corresponding point in the spectrum is high. For a high Q_a, this attenuation changes frequencies starting at a lower point in the spectrum.

In summary, the spectral envelope of glottal flow models can be considered as the gain of a low-pass filter. The spectral envelope of the derivative can then be considered as the gain of a band-pass filter. The source spectrum can be stylized by 3 linear segments with +6dB/octave, -6dB/octave and -12dB/octave (or sometimes -18dB/oct) slopes respectively. The two breakpoints in the spectrum correspond to the glottal spectral peak and the spectral tilt cut-off frequency

3.4 Causal/Anticausal Linear Model

For synthesis in the spectral domain, it is possible to design an all-pole filter which is comparable to e.g. the LF model. This filter is a 3rd order low-pass filter, with a pair of conjugate complex poles, and a simple real pole. The simple real pole is given directly by the spectral tilt parameter. It is mainly effective in the medium and high frequencies of the spectrum. The pair of complex-conjugate poles is used for modeling the glottal formant. If one wants to preserve the glottal pulse shape, and then the glottal flow phase spectrum, it is necessary to design an anticausal filter for this poles pair. If one wants to preserve the finite duration property of the glottal pulse, it is necessary to truncate the impulse response of the filter. The spectral model is then a Causal (spectral tilt) Anti-causal (glottal formant) Linear filter Model (CALM, see [3]). This model is computed by filtering a pulse train by a causal second order system, computed according to the frequency and bandwidth of the glottal formant, whose response is reversed in time to obtain an anti-causal response. Spectral tilt is introduced by filtering this anti-causal response by the spectral tilt component of the model. The waveform is then normalized in order to control the amplitude.

An aperiodic component is added to this model, including jitter, shimmer and additive filtered white noise. The additive noise is also modulated by the glottal waveform. Then the voice source signal is passed through a vocal tract formant filter to produce various vowels.

3.4 CALM parameters and phonation dimensions

This global spectral description of the source spectrum shows that the two main effects of the source are affecting the two sides of the frequency axis. The low-frequency effect of the source, related to the lax-tense dimension is often described in terms of the first harmonic amplitudes H_1 and H_2 or in terms of the low frequency spectral envelope. A pressed voice has a higher H_2 compared to H_1, and conversely a lax voice has a higher H_1 compared to H_2. The effort dimension is often described in terms of spectral tilt. A louder voice has a lower spectral tilt, and spectral tilt increases when loudness is lowering.

Then the vocal effort dimension is mainly mapped onto the spectral tilt and glottal formant bandwidth parameters (asymmetry), although the voice pressure dimension depends mostly on the glottal formant centre frequency, associated to open quotient.

Other parameters of interest are structural aperiodicities (jitter and shimmer) and additive noise.

4. Vocal instruments

Instrument 1 MIDI controlled A-LF

The real-time augmented LF model is implemented entirely in the Pure Data environment. The implementation is based on the normalized LF model worked out in [4].

A MIDI controller (MIDI master keyboard) is driving the A-LF model along three voice dimensions. The keys from (from left to right) define the vocal effort, and the velocity of the pressed key is linked to the glottal pressure.

In order to have dynamic mapping of these two dimensions we chose to have the possibility to change the parameters driving these dimensions. So that we could easily set the mid value and the span of asymmetry, open quotient and closing phase time, these parameters are each set by two knobs.

The Pitch Bend/Modulation wheel is respectively controlling Frequency and Volume in such a way that no sound is produced the wheel is released.

In addition to this, we used the pedal board to switch between the different presets of the vocal tract formants of different predefined vowels (a,e,i,o,u).

Finally, one expression pedal of this pedal board is used to add noise to the signal generated. This instrument could serve as a real time interface for the A-LF model, but it is not particularly easy to play.

Instrument 2: CALM, keyboard and glove

The second class of instruments explores hand movements in space and hand closure/opening gestures. This application is written in the MAX environment [9]. Both types of gestures seem well suited for accurate control of voice quality dimensions like melody, effort, voice pressure, hoarseness and breathiness. For this synthesizer, a P5 data glove is used. This input device allows driving 8 continuous variable parameters at once: 3 spatial positions x, y, z associated with the movement of the glove relative to a fixed device on the table and 5 parameters associated with bending of the five fingers. Several keys on the computer keyboard are controlling vowel presets. The glove is driving the CALM. Only the two horizontal spatial dimensions (x,z) are used as follows: the x variable is linked to intensity E and the z variable is linked to fundamental frequency. All the fingers but the little finger are used to control respectively (beginning from the thumb) noise ratio, Open Quotient, Spectral Tilt and Asymmetry. This mapping is most reliable and effective (compared to the keyboard used in the first experiment). Only a short training phase seems sufficient to obtain very natural voice source variations. The computer keyboard is used for changing values of the formant filters for synthesizing different vowels, and then basic vocal tract articulations.

However this instrument seems not easily playable for “classical style” singing. The melodic transitions allowed by the keyboard do not have a typically singing quality. Hand closure/opening gestures are somewhat comparable to opening closure gesture in the vocal tract and adduction/abduction of the vocal folds. This analogy is potentially useful for synthesis and will be pursued in our future work.

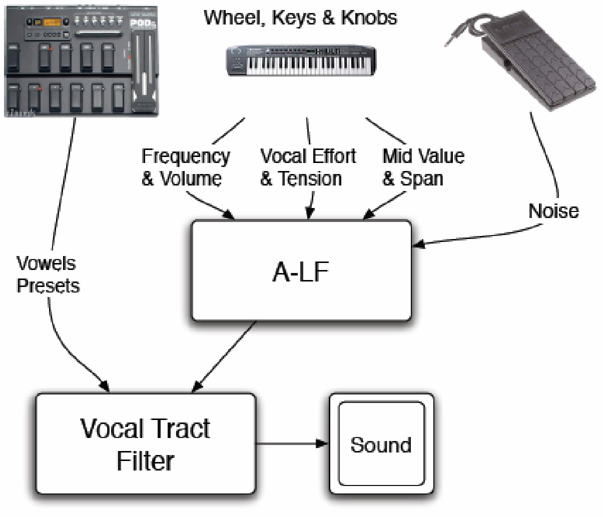

Figure 1: Instrument 1. Midi master keyboard and augmented LF model

Instrument 3: CALM calligraphic singing

The third type of instruments makes use of writing-like gestures with the help of a graphic tablet. It is also written in the MAX environment [9] The graphic tablet is organised along two spatial dimensions and a pressure dimension. The X-axis corresponds to the melodic axis. It is organized in the left to right direction from bass to treble in the same way as a musical keyboard or a guitar. The Y-axis corresponds to vocal effort. The dimensions of vocal effort and voice pressure are mapped from bottom (piano) to top (forte) along this axis. Pressure of the pen is driving the general volume of the voice. This instrument is surprisingly easy to play and expressive. Many vocal effects are possible, including *vibrato*, *portamento*, *messa di voce*, *staccato* and *legato*.

5. Discussion

Key points of this research are the number and nature of voice quality dimensions. Five main dimensions have been identified and controlled by specific gestures: vocal effort (related to spectral richness and amplitude), vocal tension (related to the glottal formant, voice open quotient and asymmetry), fundamental frequency, breathiness and hoarseness. Several mappings between control devices and model parameters have been proposed for implementing these voice quality dimensions, depending on the underlying voice source model.

The results obtained by the two voices source models seem very close in terms of sound quality. However the CALM model appears less demanding in terms of computational load and is more intuitively controlled. This is because spectral parameters are perceptually close to voice quality dimensions. The spectral model also gives a simple framework for vocal tract modelling (using formant synthesis) and source-filter interactions.

Their sound quality and their playability render these instruments usable for musical purposes [5] [6] [7]. We plan to use these types of instruments for expressive speech synthesis and for expressive prosody research. Our future work will include implementation of a low-dimension physical model of the vocal folds (e.g. a 2-mass model) in the same real-time synthesis environment.

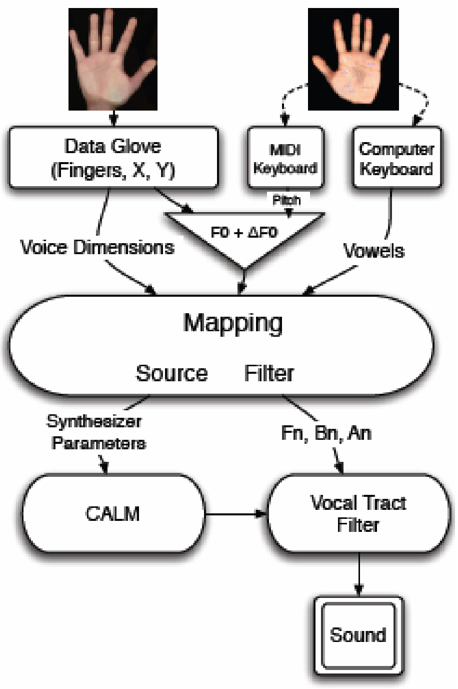

Figure 2: Instrument 2. Data glove controlled CALM synthesizer.

References

[1] C. d'Alessandro, N. D'Alessandro, S. Le Beux, J. Simko, F. Çetin, H. Pirker (2005) "The Speech Conductor: Gestural Control of Speech Synthesis", Proc. of the SIMILAR-Enterface'05 workshop, Presses univ. de Louvain, ISBN : 2-87463-003-9, p. 52-61.

[2] G. Fant, (1995) "The LF-Model Revisited. Transformations and Frequency Domain Analysis," STL–QPSR, 2–3, 119–56, 1995.

[3] B. Doval, C. d'Alessandro and N. Henrich, (2003) "The Voice Source as an Causal-Anticausal Linear Filter," Proc. ISCA ITRW VOQUAL'03, Geneva, Switzerland, pp. 15–19.

[4] B. Doval, C. d'Alessandro and N. Henrich, (2006) "The spectrum of glottal flow models" Acustica united with Acta Acustica in press.

[5] N. D'Alessandro, C. d'Alessandro, S. Le Beux, B. Doval, "Real-time CALM Synthesizer New Approaches in Hands Controlled Voice Synthesis", Proc. Int Conf. on New Interfaces for Musical Expression, NIME 2006, Paris, June 2006, p 266-271.

[6] P. Cook, (2005) "Real-Time Performance Controllers for Synthesized Singing," Proc. NIME 2005, Vancouver, Canada, May 2005, pp. 236–237

[7] L. Kessous, (2004) "Gestural Control of Singing Voice, a Musical Instrument," Proc. of Sound and Music computing 2004, Paris, October 20-22, 2004.

[8] M. Wanderley and P. Depalle, "Gestural Control of Sound Synthesis", Proc. of the IEEE, 92, 2004, p. 632-644.

[9] D. Zicarelli, G. Taylor, J. K. Clayton, jhno, and R. Dudas, Max4.3 Reference Manual, MSP4.3 Reference Manual. cycling'74/Ircam, 1994-2004.

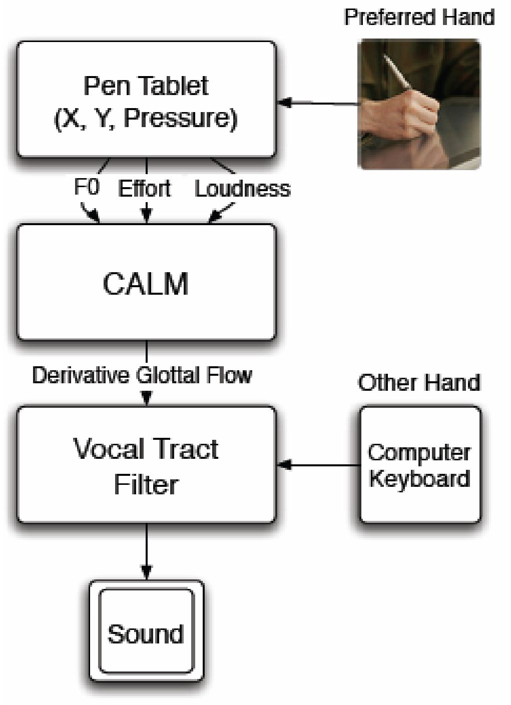

Figure 3: Instrument 3. graphic tablet controlled CALM synthesizer.