

RAMCESS/HandSketch : A Multi-Representation Framework for Realtime and Expressive Singing Synthesis

Nicolas D'Alessandro, Thierry Dutoit

Laboratoire de Théorie des Circuits et Traitement du Signal, Faculté Polytechnique de Mons, Belgique

nicolas.dalessandro@fpms.ac.be, thierry.dutoit@fpms.ac.be

Abstract

In this paper we describe the different investigations that are part of the development of a new singing digital musical instrument, adapted to real-time performance. It concerns improvement of low-level synthesis modules, mapping strategies underlying the development of a coherent and expressive control space, and the building of a concrete bi-manual controller.

Index Terms: real-time expressive singing synthesis

1. Introduction

Expressivity is nowadays one of the most challenging topics studied by researchers in speech processing. Indeed, recent synthesizers provide acceptable speech in term of intelligibility and naturalness but the need to improve human/computer interactions has brought researchers to develop systems that present more "human", more expressive skills.

Speech synthesis research seems to converge towards applications where multiple databases are recorded, corresponding to a certain number of labelled expressions (e.g. happy, sad, angry, etc.). At synthesis time, the expression of the virtual speaker is set by choosing the units in the corresponding database, then used in well-known unit selection algorithms.

Recently remarkable achievements have also been reached in singing voice synthesis. We can highlight naturalness and flexibility of Bonada *et al.* [1] algorithms where singing frames are structured at a high performance level. These kinds of technologies seem mature enough to allow for the replacement of human vocals with synthetic, at least for backing vocals.

However existing singing systems suffer from two restrictions: they are aiming at mimicking singers rather than offering real creative voice timbre possibilities, and they are generally limited to note-based MIDI controllers.

In this context, we propose to investigate an original option. We postulate that, even if the use of databases is strategic in order to preserve naturalness, voice modelling has to reach a higher level. These improvements have to meet particular needs, such as more realistic glottal source/vocal tract estimation, manipulation of voice quality features at a perceptual and performance level, and strong real-time abilities. The other issue concerns mapping strategies that have to be implemented in order to optimize the performer/synthesizer relation.

In this paper we present the work of these last three years in the development of synthesis techniques, involving glottal source and vocal tract modelling in order to reach real-time and highly expressive singing. First we explain particular methods involved in the estimation of vocal tract parameters, in section 2. Then we describe our improvements in synthesis techniques, in section 3. We also present dimensional control of voice quality, in section 4. Finally results in the development of a bi-manual

singing voice controller is proposed, in section 5.

2. ZZT-Based Vocal Tract Estimation

As the source of our subtractive synthesis system aims at generating a full glottal signal (cf. section 3), we need to estimate vocal tract parameters without the source contribution. In our method we use Bozkurt *et al.* decomposition, based on the classification of zeros of the Z-Transform of signal frames [2]. If zeros outside and inside the unit circle are separated for a given GCI-synchronous speech frame, resynthesis of time-domain signal from separated zeros gives respectively anticausal and causal components of the frame.

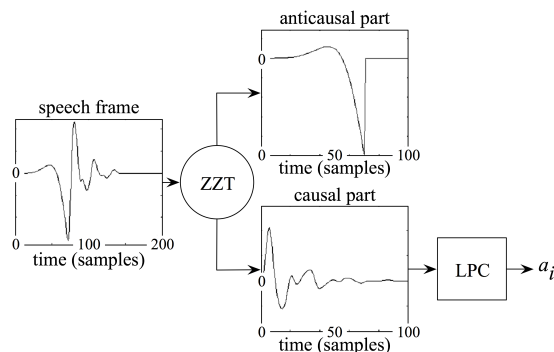


Figure 1: Different steps in the estimation of vocal tract LPC coefficients: ZZT-based causal/anticausal decomposition and LPC analysis of causal impulse response.

As we know that the glottal source mainly contributes to the generation of the anticausal component in speech [3], we can assume that the decomposed causal signal is close to the real vocal tract impulse response (only return phase of glottal signal has causal characteristics). Consequently LPC analysis of extracted causal contribution is computed and a set of a_i coefficients approximates the vocal tract for a given vowel. From this preset, manual refinement can also be achieved. This process is illustrated in Figure 1.

3. Real-Time Voice Synthesis Modules

In this section we describe how we modified usual synthesis modules involved in subtractive singing synthesis – glottal pulse generator and vocal tract filter – in order to improve real-time expressive behavior. It concerns first direct timing problems such as latency and reliability, but also more perceptual issues such as independence and coherence of control parameters. All this work of synthesis is part of the RAMCESS project [4].

3.1. The Glottal Source

Glottal pulse signal can be synthesized with many different models. In term of flexibility and quality, we can particularly highlight LF and CALM [3]. However none of them is really suitable for real-time processing. One the one hand, LF parameters are the solution of a system of implicit equations which is know to be unstable. On the other hand, CALM is linear filter processing but one of the filters has to be computed anticausally. This is possible in real-time but with a limited flexibility [4].

The improvement that we propose can be seen as a compromise between both LF and CALM models. In order to avoid the resolution of implicit equations, only the left part of the LF model is used. It is computed using the left part (cf. equation 1) of the normalized GFM model described in [5].

$$n_g(t) = \frac{1 + e^{at} (a \frac{\alpha_m}{\pi} \sin(\pi t / \alpha_m) - \cos(\pi t / \alpha_m))}{1 + e^{a\alpha_m}} \quad (1)$$

where t evolves between 0 and 1, and is sampled in order to generate the $O_q \times \frac{F_s}{F_0}$ samples of the opened phase (O_q : open quotient, F_0 : fundamental frequency, F_s : sampling rate); α_m is the asymetry coefficient and $a = f(\alpha_m)$ is the pre-processed buffer of solutions of the equation 2.

$$1 + e^a (a \frac{\alpha_m}{\pi} \sin(\frac{\pi}{\alpha_m} - \cos(\frac{\pi}{\alpha_m})) \quad (2)$$

Then the right part (the return phase) is generated in spectral domain, which means that the left LF pulse is filtered by the spectral tilt low-pass first order filter presented in [3]. This option is also preferred because a long filter-generated return phase smoothly overlaps with following pulse, thus avoiding discontinuities. The complete process, also integrating the derivation of the pulse and the normalization, is illustrated in Figure 2.

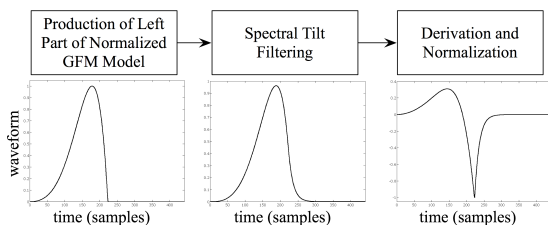


Figure 2: Synthesis of the glottal pulse by combination of LF left part time-domain generation and spectral tilt filtering.

3.2. The Vocal Tract

The vocal tract is computed with a simple tube model. LPC coefficients a_i are converted into reflection coefficients k_i , and then into area coefficients A_i , defining geometrical properties of vocal tract. A complete coefficient conversion framework have been developed in order to jointly manipulate multiple representations (spectral and physical) of the vocal tract. This approach is powerful in order to create typical voice quality effects: vowel interpolation, obstructions, singer formant, etc [4].

4. Dimensional Study of Voice Quality

On the top of the synthesis parameters that we have highlighted (F_0 , O_q , α_m , T_l , spectral and geometrical features of vocal tract), it is interesting to build a layer which is able to provide more perception-based control to the performer. This dimensional study of voice quality have been achieved, resulting in a set of dimensions and their corresponding mapping equations

with synthesis parameters [4]: *melody*, *vocal effort*, *tenseness*, *hoarseness*, *breathiness* and *registers*.

5. Bi-Manual Voice Controller

Realization of an inspiring control surface is a challenge in itself and we have recently started to work in this field, by building a first bi-manual controller, focused on musical and voice quality control issues. This project, called HandSketch, structures the mapping of continuous and precise dimensions (pitch, tenseness, vocal effort, etc.) around pen-based control, and the achievement of articulations (vowel transitions, pitch jumps, etc.) with a force sensing resistor network. Details can be found in [6] and a performing example is illustrated in Figure 3.

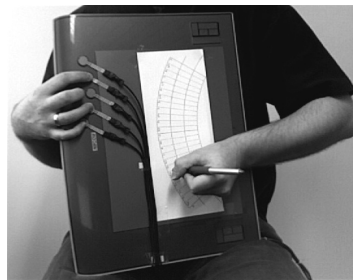


Figure 3: HandSketch playing position, connected to RAMCESS synthesizer for real-time singing voice performance.

6. Conclusions

In this paper, we presented the different steps that are structuring the development of a full singing instrument, from synthesis layer improvements to perception-based control space, and building of a concrete bi-manual controller. We can produce highly expressive voice quality modifications. Further work concerns the extension of our "vowel synthesis" to spoken contents.

7. Acknowledgements

Authors would like to thank LIMSI-CNRS and past eINTERFACE workshops teams and organizers for their great involving in the RAMCESS synthesizer development.

8. References

- [1] J. Bonada and X. Serra, "Synthesis of the Singing Voice by Performance Sampling and Spectral Models", IEEE Signal Proc., 24(2):67–79, March 2007.
- [2] B. Bozkurt, L. Couvreur and T. Dutoit, "Chirp Group Delay Analysis of Speech Signals", Speech Comm., 49(3):159–176, 2007.
- [3] B. Doval and C. d'Alessandro, "The Voice Source as a Causal/Anticausal Linear Filter", Proc. of ISCA VOQUAL'03, 2003.
- [4] N. D'Alessandro, B. Doval, S. Le Beux, P. Woodruff, Y. Fabre, C. d'Alessandro and Thierry Dutoit, "Realtime and Accurate Musical Control of Expression in Singing Synthesis", JMUI, 1(1):31–39, March 2007.
- [5] B. Doval, C. d'Alessandro and N. Henrich, "The Spectrum of Glottal Flow Models", Acta Acustica, 92:1026–1046, 2006.
- [6] N. D'Alessandro and T. Dutoit, "HandSketch Bi-Manual Controller", Proc. of NIME'07, 2007.