Calliphony: A system for real-time gestural modification of intonation and rhythm

Sylvain Le Beux[1], *Christophe d'Alessandro*[1], *Albert Rilliard*[1], *Boris Doval*[2]

[1]LIMSI-CNRS, Université Paris-Sud XI, Orsay, France
[2]LAM-IJLRA, UPMC Université Paris 06, France
{slebeux,cda,rilliard}@limsi.fr , boris.doval@upmc.fr

Abstract

This paper presents new achievements done with our Calliphony software [1, 2] which allows for real-time modification of the intonation and rhythm of speech, driven by manual control of both fundamental frequency and time scaling either independantly or conjointly. We will present here the main features of the software and discuss possible use for prosody research.
Index Terms: intonation modification, rhythm modification, speech synthesis, prosody analysis, gestural control.

1. Introduction

Whereas there have been plenty of digital instruments built for musical purpose [3], until recently there was only few speech synthesizer or speech modification algorithms monitored thanks to gestural interfaces [4, 5, 6]. We argue that this new paradigm, borrowed from computer music community, can lead us towards new methodologies in analysis of intonation and rhythm. The general process of Calliphony can be sketched as follows: it consists of a piece of software, running on a computer where a graphic tablet is connected, in order to monitor the modification of pitch and duration of speech. Prosody modifications (considered here as intonation and rhythm) are controlled interactively by manual gestures that give the player the real-time feedback of the speech being processed. It must be noted that the system we present here is restricted to the use of a graphic tablet, which offers both high tracking definition and temporal sampling resolution. However, any other gestural tracking device can be used within this scope, either being manufactured or handmade with various sensors, provided it has at least two degrees of freedom and sufficient resolution and sampling rate.

Hence, Calliphony system offers the possibility to deal with production, perception and communicative function [7] within the same framework. Perception and communicative function, because inherently the modifications achieved on the speech are treated by the player in a similar way he would do it during a standard conversation. In particular, no assessment is made *a priori* on the output prosody. Only the player decides what would be the correct intonation and rhythm to give to speech. Even if further evaluation is to be made for duration modification, it has been shown that it is possible to accurately reproduce a given intonation, through gestural control of pitch in real-time [1].

The originality of our system, contrary to several comparable systems [8, 9, 10], lies in the possibility to achieve *variable* pitch and duration modifications in real-time while preserving formants. In other terms, the pitch and duration modifications applied to the original speech signal can be sampled as low as every synthesis period, as will be explained further. Moreover, none of the systems cited before allows to control these modifications in terms of gestural control. The name *Calliphony* was thus coined in analogy with calligraphic drawings, as our system, driven by a graphic tablet, is monitored by manual gestures comparable to the action of scripting or drawing. We will also discuss further the opportunity to describe and analyse intonation and rhythm in terms of movement or dynamic evolution.

To put it simple, while there are great debates about the best way(s) to model intonation and rhythm, Calliphony software gives the opportunity to tackle these questions in a new manner. Its interactive analysis-by-synthesis procedure let us modify speech in any way we want, and then only evaluate the modifications either objectively, through classical speech signal processing of phonetics parameters, or subjectively, via perceptive listening tests.

This paper is organized as follow: in the next section 2, we will present possiblities offered by Calliphony for intonation and rhythm modification, based on a modified TD-PSOLA algorithm. After giving some details about sound and gestures recordings, we will present some applications in section 3 and finally have a discussion about our system in section 4.

2. Calliphony in details

The modification of both pitch and duration rely on the TD-PSOLA algorithm [11] modified to work in real-time. Nevertheless, there are some adaptations that have been made in order to conform both with speech signal characteristics and real-time constraints per-se.

2.1. Offline pre-processing

As noted by E. Moulines & Laroche [11], there are two main tunings of the algorithm that need to be taken care of. The first one deals with the need for windowing the short-term signals at the highest energetic points in the speech signal. The other one has to do with the treatment of unvoiced parts of speech. Thus, an offline pre-processing has been set up in order to take into account these two problems and obtain the best performances possible of the real-time PSOLA algortithm.
Concerning windowing, as it is well known from speech signal analysis, the glottal closure instant (GCI) represents the instant when the more energy is present within a single period of speech. The most efficient way to calculate GCIs was to use the Matlab script from DYPSA algorithm [12].
By the way, [11] recommends using a constant rate for unvoiced parts of speech, whereas DYPSA algorithm does not include a voicing detector. Thus, in conjunction with the detection of GCIs, we then used another Matlab script for discriminating voiced from unvoiced parts of speech. "Swipe" algorithm from

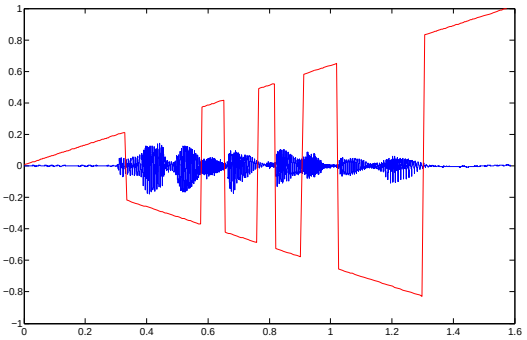

Figure 1: *Analysis of glottal closure instants and voicing (red) from the natural sentence (blue) "Marie chantait souvent" pronounced by a male speaker. On the figure, GCIs samples values are normalized by the latest value (here* $\approx 1.6 \times 44100$*).*

A. Camacho [13], was thus used for this purpose. Swipe algorithm was tuned to analyse the speech utterances within the range from 75 to 600 Hz, corresponding to a maximum fundamental period of 13.5 ms. Therefore, if DYPSA algorithm was giving intervals greater than this value, new points were introduced regularly to enhanced PSOLA performances. Of course, voicing detection is prone to errors, but the algorithm was tuned so that it gave more false detection of unvoiced parts than detection of an unvoiced part as voiced. It is indeed less critical keeping a voiced part at a constant rate, than stretching excessively an unvoiced part of speech.
As a result, in order to be able to modify both pitch and duration of speech sentences in real-time, we had to create a separate text file, together with the audio file so as to be able to know at which exact sample positions were lying the successive GCIs. In order to differentiate between voiced GCIs and unvoiced time instants, a positive value for voiced parts and negative one for unvoiced parts was given. The figure 1 depicts the result of the analysis of GCIs and voicing for a given sentence.
After this pre-processing procedure, we then have two files at our disposal (i) the original audio file from a recorded speaker, and (ii) the analysis of glottal closure instants, specified by sample numbers at the sampling rate of the input file, whether positive or negative depending on its belonging to a voiced part or not.

Once these two files are loaded into Calliphony software, the original speech file can be looped and modified, taking into account the pitch and time scaling factors, according to the position of the stylus on the graphical tablet. To conform with our usual representation of intonation and time, we mapped the absolute pitch in semitones on the vertical axis of the tablet and the time stretching factor on the horizontal axis, resembling that of a musical score.

2.2. Real-time constraints

Given that our implementation of TD-PSOLA algorithm needed to be performed in real-time, there were two main difficulties related to this framework, namely during the overlap-add step and during the initialization of the audio process.
According to TD-PSOLA scheme, the windowing of short-term analysis signals is done over two periods. In other terms, windowing goes from one GCI to the GCI two fundamental periods further (the GCI of interest being in the middle). Moreover, within the resulting synthesis speech signal, two consecutive frames can belong to two portions of the original speech signal quite far apart, depending on the values of pitch and duration modifications. Thus, these two short-term synthesis signals can possess relatively different widths.
The problem then lies in the fact that when the second short-term signal needs to be copied, its N first samples can belong to time instants located before the synthesis time instant from the preceding frame. This is typically the case when the next short-term signal is longer than the preceding one, as illustrated on figure 2.

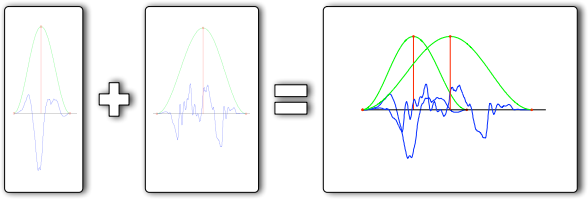

Figure 2: *Overlap-add procedure where the succesive short-term signals have different widths. Windows are indicated in green and the resulting short-term analysis signal in blue, whereas the GCIs are indicated by a red stem.*

In order to cope with this problem, we decided to process the signal thanks to a circular buffer, where the synthesis time instants are located later in time, so as to be able to copy any short-term signal whatever pitch and duration modifications are applied, either maximum or minimum. By taking windows running over two fundamental periods and with pitch modifications allowed from half to double current pitch value, and for duration modification from half to double speed, the constant latency of the system is of 27 ms. This value correspond to two times the largest period, which in our case is defined by the GCI analysis done down to 75 Hz. Hence, when the audio process is turned on, there is a silent initialization of 27 ms. As a result, the gestural commands applied on the speech signal are delayed by the amount of time given by this latency. Hopefully, this delay is not critical for the scope of our application. The overall procedure is illustrated on figure 3.

2.3. Rhythm modification

Duration modifications of the speech can offer opportunities to use our system in order to achieve rhythm modification, provided that some additional features are given to the input procedure.
It has already be shown that Calliphony is capable of reproducing accurately a target intonation pattern [1]. In the current section, we will present other new possibilities offered by Calliphony for dealing also with rhythm modification.
It is indeed rather versatile to be able to adapt to several situations whether one wants to study more specifically the effect of segmental durations modification or needs to build rules between intonation modifications and rhythmical anchor points.

2.3.1. Global vs. local stretching

Together with GCIs and voicing information given in the pre-processing text file, it is possible to add other features relative to temporal characteristics. By default, duration modifications

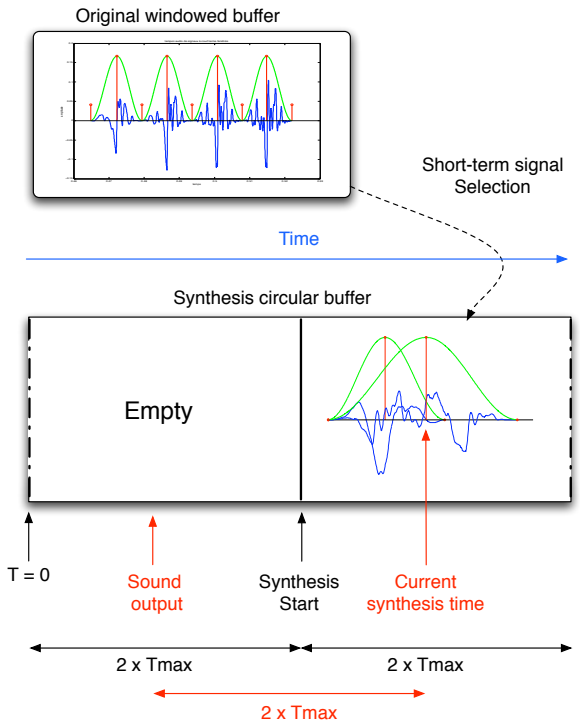

Figure 3: *Processing scheme of Calliphony core treatment, from the original windowed speech to the sound card output. On top figure is represented the concatenation of short-term analysis signals, windowed from the original input speech. After selection, these short-term signals are overlapped and added into the circular buffer with a constant delay of* $2 \times Tmax$.

are accomplished on a continuous scale, or more precisely sampled at the beginning of every new synthesis period. This means that duration modification (as well as pitch) is a piecewise linear function of time instants. Thus, it is totally feasible to specify time instants correponding to any syllable, stress, p-centers in the pre-processing text file so as to keep the duration factor constant over longer chunks than just fundamental periods.
Several modification schemes are then possible, like for example : (i) continuous, sampled every new period, (ii) phoneme-based, sampled at the beginning of every new phoneme or (iii) syllable-based, sampled at every new syllable. Of course, these examples are somewhat arbitrary, but show the possibility to test different strategies, by choosing different temporal anchor points. These various segmentations can easily be done using Praat software [14] by manual annotation, provided that the corpus is not too big.
For the sake of simplicity, these sample values were added at the beginning of the text file with the very first value stating the number of new anchor points, followed by the samples values of the anchor points at $44.1\ kHz$ (i.e. the audio file sampling rate in our case). It is thus possible to replace those values by any others while respecting this file format.
Calliphony indeed gives the ability to modify the original speech signal along various ways, either based on phonetic constraints or not.

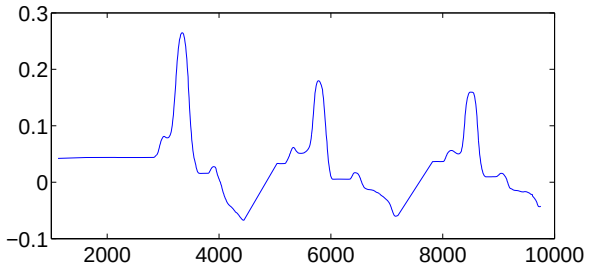

Figure 4: *Raw continuous data extracted from the tablet.* Y *axis is mapped to pitch (arbitrary unit). Time units in* ms.

2.3.2. Inter-dependencies

Obviously, there is not any restriction for applying some dependency rules between pitch and duration modifications. For example, it is possible to modify the range of duration modification factor depending on the velocity of pitch modification, or to state that duration should increase at the end of sentence (or breath group).
This kind of interrelated modifications paves the way for achieving different speaking styles. If one wants to gives a given speech sentence a journalistic style, then a certain kind of rules can be applied to Calliphony, so that it conforms to the prototypical pronounciation of many journalists.

2.4. Sound and data recordings

Calliphony software allows for recording both audio and gestural productions.
Concerning audio it is simply done by creating a new wave file once the real-time modification process has been launched. For gestural data, a text file is created in the simple form of raw data, organized in columns for each parameters. In the case of the graphical tablet, this text file contains both (X, Y) positions values, together with pressure exerted on the stylus. Time stamps are also recorded, given within ms precision. If necessary, the (X, Y) tilt values can be recorded for further gestural analysis. A last important point for post-processing is the following fact: once the real-time modification and recording have been launched, the audio speech file is played continuously *ad-libitum* until the user stop the recording. As a matter of fact, the recorded audio file contains consecutive repetitions of the modified audio file. In order to be able to match gestural data together with the corresponding audio repetitions, a simple $(0/1)$ value is added to the output text file, alternating for each consecutive repetition. Therefore, audio repetitions can be automatically segmented by combining time stamps and repetition boolean value.
Figures 4 and 5 described this two steps procedure, depicting respectively raw gestural data extracted from the tablet and the alignment of gestural traces with original speech sentence.

3. Applications

On first thoughts, Calliphony can be used to either correct the output of any speech, either natural or from a concatenative speech synthesizer, or even to change its attitude to an other one. When the selection procedure of the concatenative

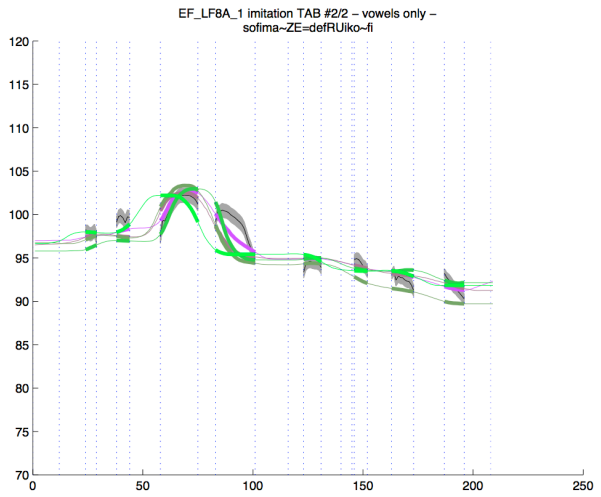

Figure 5: *Visualization of gestural recordings (thin lines) together with the extracted pitch from original sentence "Sophie mangeait des fruits confits" (thick gray line). The best gestural reiteration in terms of correlation is indicated by a purple line. Pitch in semitones, time in seconds*

synthesizer fails to choose two consecutive units that join smoothly, Calliphony can correct these errors conveniently. There are other situations where the simple modification of pitch and duration of speech of a given sentence can raise new attitudes that are not present in the database of units.

As should be clear from now on, the purpose of this present article is to give the opportunity to researchers dealing with prosody synthesis and analysis to be able to address their problems in a new manner. Whether one wants to hear the effects of pitch and duration modifications on the speech signal instantly for giving new insights on these phenomenons, or if one just wants to build series of stimuli recordings, Calliphony can be considered as a suitable solution.

4. Discussion

Calliphony system was first developped as a real-time pitch shifter, and then evolved to a combined pitch shifting and time stretching system controlled by hands movements. It has been proved that it was possible to control these dimensions accurately with gestures.

Calliphony framework, as explained in this paper, allows for studying both intonation and rhythm either independantly or conjointly, thanks to real-time gestural control of prosody modifications. Furthermore, both modified speech recordings or gestural data can then be analyzed both with classical speech analysis methods, or in a new manner by studying the cinematics and dynamics of manual gestures, as far as these movements represent good candidates for a fine prosodic stylization.

This view also conforms with recent studies [15] considering prosody and rhythm as an oscillation and an entrainment. This is mainly due to the fact that movements of the hands and limbs can be considered physiologically as a second-order system, thus revealing the same properties as a comparable physical analog system.

5. Perspectives

As today Calliphony seems pretty stable for both pitch and duration modifications, it would then be interesting to test the possibilities to add voice quality and vowels reduction modifications. By the way, there are no conceptual barriers for using other speech signal modification algorithms, provided that they can be processed in real-time. For now on, we only tested Calliphony upon french language. Hence, it would be interesting to experience its limitations and potentials on any other languages than french. Finally, Calliphony would also greatly benefit from a more refined coding of pre-processing data, as the number of information should increase to include other phonetical or acoustical characteristics.

6. References

[1] C. d'Alessandro, A. Rilliard, and S. Le Beux, "Computerized chironomy: evaluation of hand-controlled intonation reiteration," in *Proceedings of Interspeech 2007*, (Antwerpen, Belgium), pp. 1270–1273, ISCA, 2007.

[2] S. Le Beux, A. Rilliard, and C. d'Alessandro, "Calliphony: A real-time intonation controller for expressive speech synthesis," in *6th ISCA Workshop on Speech Synthesis*, (Bonn, Germany), ISCA, 2007.

[3] "New interfaces for musical expression," Retrieved October 31 2009, from http://www.nime.org/.

[4] S. Fels and G. Hinton, "Glove-talk ii: A neural network interface which maps gestures to parallel formant speech synthesizer controls," *IEEE Transactions on Neural Networks*, vol. 9, no. 1, pp. 205–212, 1998.

[5] N. D'Alessandro, P. Woodruff, Y. Fabre, T. Dutoit, S. Le Beux, B. Doval, and C. d'Alessandro, "Realtime and accurate musical control of expression in singing synthesis," *Journal on Multimodal User Interfaces, Springer Berlin/Heidelberg*, vol. 1, pp. 31–39, mars 2007.

[6] D'Alessandro, C. N., d'Alessandro, S. Le Beux, and B. Doval, "Real-time calm synthesizer: new approaches in hands-controlled voice synthesis," in *Proc. of New Interfaces for Musical Expression 2006*, (Paris, France), pp. 266–271, June 2006.

[7] K. Kohler, "Rhythm in speech and language," *Phonetica*, vol. 66, pp. 29–45, April 2009.

[8] F. Pesce, "Realtime-stretching of speech signals," in *Proceedings of Digital Audio Effects Conference (DAFx'00)*, (Verona, Italy), pp. 189–194, 2000.

[9] K. Sreenivasa Rao and B. Yegnanarayana, "Duration modification using glottal closure instants and vowel onset points," *Speech Communication*, vol. 51, pp. 1263–1269, 2009.

[10] D. Barry, D. Dorran, and E. Coyle, "Time and pitch scale modification: A real-time framework and tutorial," in *Proc. of Digital Audio Effects Conference, DAFX 08*, vol. 978-951-22-9517-3, (HUT, Espoo, Finland), pp. 103 – 110, 2008.

[11] E. Moulines and J. Laroche, "Non parametric techniques for time-scale and pitch scale modification of speech," *Speech Communication*, vol. 16, pp. 175–205, 1995.

[12] A. Kounoudes, P. A. Naylor, and M. Brookes, "The dypsa algorithm for estimation of glottal closure instants in voiced speech," in *Proceedings of IEEE International Conference on Acoustics Speech Signal Processing*, vol. I, pp. 349–352, May 2002.

[13] A. Camacho, "Detection of pitched/unpitched sound using pitch strength clustering," in *Proceedings of the Ninth International Conference on Music Information Retrieval*, (Philadelphia), pp. 533–537, 2008.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.20) [computer program]," Retrieved October 31 2009, from http://www.praat.org/.

[15] F. Cummins, "Rhythm as an affordance for the entrainment of movement," *Phonetica*, vol. 66, pp. 15–28, April 2009.