# GESTURAL CONTROL OF REAL TIME CONCATENATIVE SYNTHESIS

*Grégory Beller*

IRCAM, Paris, France
beller@ircam.fr

## ABSTRACT

This paper presented the researches and the developments realized for an artistic project called “Luna Park”. This work is widely connected, at various levels, in the paradigm of the concatenative synthesis, both to its shape and in the processes which it employs. Thanks to a real-time programming environment, synthesis engines and prosodic transformations are manipulated, controlled and activated by the gesture, via accelerometers realized for the experiment. This paper explains the sensors, the real time audio engines which include a new speech rate modifier, and the mapping that connects this two parts.

**Keywords:** Gesture, real time concatenative TTS, mapping, prosody, speech rate

## 1. INTRODUCTION

This work distinguishes itself from previous creation of G. aperghis, notably by use of gesture sensors and from the speech synthesis. Indeed, various sensors, (accelerometers, tactile ribbons and piezoelectric sensors) were developed and realized to allow the performers to control various audio engines, by the gesture. The mapping between the data stemming from these sensors and the various audio processings, realized within the real-time programming environment Max/MSP, is different from a sequence in the other one and can evolve in the time. That is why we present this article according to the following plan. In a first part, this article lists the various sensors realized for this creation and gives details of their developments, as well as the data which they produce. In a second part, the realized audio engines are described under the shape of real-time processes. Besides the concatenative synthesis engine, is presented an innovative engine of prosodic transformation allowing the real time modification of the speech rate. The third part proposes some examples of mapping between the sensors data and the audio engines parameters, notably used for the piece. Finally, the fourth part allows to conclude and to propose some perspectives.

## 2. GESTURE CAPTURE

### 2.1. Background

The gestural control of the speech synthesis constitutes nowadays a complete field of research. Controllers of various types were elaborated for various types of synthesizers of spoken voice, or sung voice. Among these mappings, we find the “Speech Conductor" [7], the “Glove-Talk" [8], the “SPASM" [6], the “OUISPER" project [10] and some others. We chose to use the movements of the hand for several reasons, besides crossing the scenographic reasons. First of all, the spontaneous speech can be naturally accompanied with a movement of hands. The idea to accompany the movements of hands by the speech, by the reversibility, seems thus natural. The percussive aspect of the movements fits the concatenative synthesis in which segments are activated in a discreet way in the time, so managing the segmental aspect of the speech. On the contrary, the continuous aspect of the movements of hands allows a control of the prosody, the suprasegmental aspect of the speech. If we consider the classic asymmetry right-left such as know it the conductors (for the right-handers, the left hand is rather connected with the expression, whereas the right hand is rather connected with the important temporal markers of the music), we can then create a gestural control of both hands of the synthesis, with for a right-hander, a right hand managing the segmental aspect and a left hand managing the suprasegmental aspect. It is one of the possible scenarios that we exploited for the creation of the piece (see section 4.).

### 2.2. Accelerometers gloves

The technology of gloves wireless accelerometers / gyroscopes used [9] allows to measure the accelerations of both hands according to 6 axes (3 in translation and 3 in rotation with gyroscopes). The raw data delivered by gloves are not necessarily easy to interpret. So a first stage of preprocessing allows to return more interpretable data.

#### *2.2.1. Preprocessing*

The data resulting from the wifi receiver are transmitted via UDP all 1 ms. To synchronize them to

the internal clock of Max/MSP, they are first median filtered (order 5) and sub-sampled by a factor 5. Thus we obtain a stable stream of data all 5 ms. Then various descriptors of the gesture arise from these preprocessed raw data.

#### *2.2.2. Variation of the Momentum*

The estimate of the immediate acceleration allows to know, at any time, the variation of Momentum relative to the gesture. This Momentum, according to the laws of the classical mechanics, is directly proportional in the speed. The raw data coming from the sensor are at first Òdenoised" thanks to the average on the last 4 samples. The root of the sum of the square of these six filtered values allows to obtain a proportional quantity in the variation of Momentum of the gesture.

#### *2.2.3. Hit energy estimation*

The hit energy estimation allows the immediate release from the observation of the variation of the Momentum of the gesture. Three values, delivered by the sensors of acceleration in translation, are stored in a circular buffer including all the time, 20 samples. Three standard deviation corresponding to these values are added, all the time, (norm I corresponding to the sum of the absolute values). This sum also allows to represent the variation of Momentum of the gesture. To detect variation of this value, corresponding to abrupt variations of the gesture, it is compared all the time with its median value (order 5). When the difference between these two values exceed certain arbitrary threshold, a discreet value appears to mean the presence of a fast change of the gesture. It allows, for example, to emit a regular click, when we beat a measure with the hand of bottom at the top, every time the hand changes direction. The hit energy estimation is a process allowing to generate discreet data from a gesture, by definition continuous. Indeed, of a continuous physical signal, it allows by thresholding, to define moments corresponding to the peaks of variation of the Momentum, which coincide, from a perceptive point of view for the user, in peaks of efforts (of acceleration). By this process, it becomes then possible to create precise air percussions either sounds activation at the moment when the hand of the user changes direction or accelerates surreptitiously.

#### *2.2.4. Absolute position of the hand*

The Earth´s gravitational field introduces an offset into the answer of the sensors which can be exploited to deduct the absolute position of the hands, as well as the presence of slow movements. This quasistatic measure brings a continuous controller to the performer. A playful example of the use of this type of data is the air rotary potentiometer in which the rotation of the hand can control the volume (or other) of a sound.

## 3. AUDIO ENGINES

### 3.1. Real time speech synthesis

A real time speech synthesizer has sense only if the text is generated in real time. The case appears (as shawn in section 4.) when the text is generated by statistical models (HMM, N-gram, K-NN) which transforms or generates one text on the fly. The real-time constraint does not allow us any more to use the phase “backward” of the Viterbi algorithm guaranteeing the optimality of the path on a whole sentence, because we do not know early, the end of the current sentence. The joint minimization of the distance in the target and the distance of concatenation can be made then only between every segment, in a local way and not in a global one. The advantage of this method lies in its ability to react, while its inconvenience is that it produces a sound result poorer than the classic batch TTS synthesis. In the piece, several paradigms are used to generate targets which guide the real time concatenative synthesis.

#### *3.1.1. Predefined sequences*

First of all, certain texts are fixed, a priori, and modeled under the shape of a sequence of words, syllables, phones or semi-phones. Segments constituting these sequences are then launched by the performers and can be chosen according to data stemming from sensors or from vocal analyses (see section 4.). It allows for expected syllables series, and to produce a clear semantic sense.

#### *3.1.2. Predefined sequences and N-gram*

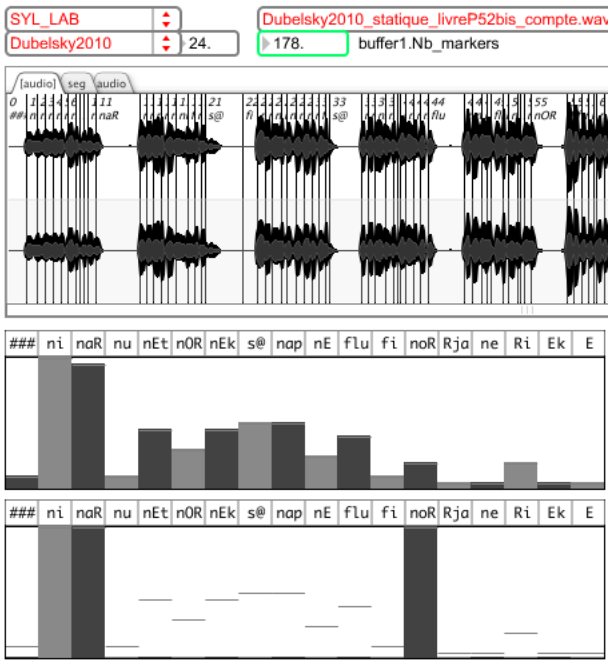
It happens that several segments possess the same symbol (several phones corresponding to the same phoneme, for example). We can then estimate the probability of transition from a symbol to the other one and make random series respecting more or less the initial sequence. Thanks to N-gram of order N variable (the more N is big, the more the sequence corresponds to the initial sequence and the more N is small, the more the sequence is random), we can control in real time the closeness of the generated material with regard to the predefined text. It notably allows the performers to control the semantic aspect of the output.

#### *3.1.3. Predefined sets and HMM*

Like a text, some segment sets (syllables, phones, words) were also predefined. In the same way, segments belonging to these sets are activated by the

performers according to descriptors or in a random way. It notably allows to create textures (of phones or syllables) or still rhythms from a single syllable. An interface was create to allow to choose in real time, the probability of appearance of such symbols. It appears under the shape of a histogram of the available symbols. The modification of this histogram allows to modify the probability of transition from a symbol to the other one (HMM of order 1). Once the symbol is chosen, a corresponding segment can be activated according to various controllers / descriptors values. The figure 1 presents the interface that permits to generate in real time targets. At the top, a recording segmented in syllables thanks to the program IrcamAlign [11] which allows the speech segmentation in various units in batch mode), is the input. In the middle, the histogram presenting the relative rate of appearance of their symbols. Below, the same modified histogram allows to modify the probability of appearance of the activated syllables. For example, the generated output is composed of an equiprobable set of syllables “ni”, “naR” and “noR”.

**Figure 1:** Interface for the definition of the real time target by editing histogram.




### 3.2. Real time prosodic transformation

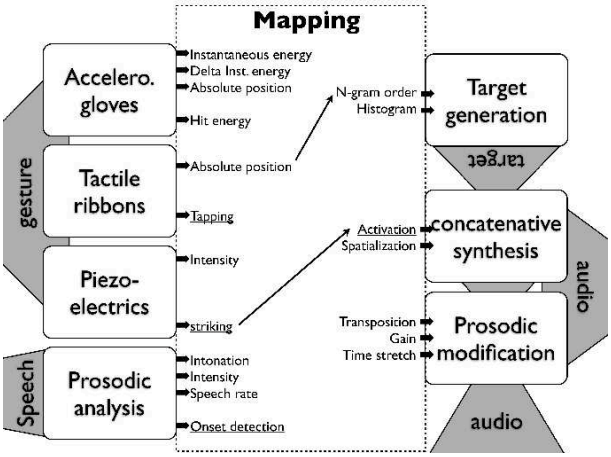
Real time prosodic transformation allows many applications [2]. It is possible thanks to speech analysis/synthesis paradigm allowing to estimate and to modify the prosodic dimensions of the speech. The prosodic transformation can apply to the voice of the performers or to the synthesis output in a equivalent way. The prosody or the way to speak can be described by a space in five dimensions: the intonation, the intensity, the speech rate, the vocal quality and the degree articulation [1]. If most of these dimensions are today measurable and modifiable in batch mode [3, 4], some of them are also measurable in real time. It is the case of the intonation (yin [5]) and of the intensity (loudness). We added a speech rate estimator (syllex). These three real-time speech descriptors allow to inform us about the way the performers utter. They can be used, in the same way as the sensors, to control the various audio engines. If the modification of the intonation by transposition and that of the intensity by variable gain are henceforth known and well enough mastered in real time, it does not also go away for the speech rate. Now we expose in this section, one new paradigm allowing the transformation of the speech rate in real time. All the prosodic transformations used are available in the SuperVP [12] audio engine, the IRCAM’s high quality phase vocoder in real time. In fact, the SuperVP library already implements a quality transposition, as well as some other modifications such as the spectral envelope transformation. One of the objects of this library, SuperVP.ring, allows to make these modifications on a circular buffer the size of which can be arbitrarily defined (3 seconds in our case). The advantage of the circular buffer is to keep the instantaneousness of the transformation, while enabling, at any time, to be able to move in short-term past (term equivalent to the size of the buffer). Thanks to it, we can locally stretch out certain portions of the signal as the vowels (using a real time voicing detection) and provoke at the listenerÕs the perception of a slowing down of the speech rate. If we cannot move to the future, the return in the immediate position can be made in a accelerated way, provoking, this time, the perception of an acceleration of the speech rate. As if the read head of the buffer behaved as an elastic which we stretch out and relax. Extremely, it is possible to freeze the read head of the buffer in a place that provokes a Òstop on sound" who can give interesting effects (extremely long vowels for example that makes speech sounds like sing).

## 4. EXAMPLES OF MAPPING

In this part, we give some examples of mapping between the control data and the parameters of the audio engines. The figure 2 lists the various available controllers (to the left), as well as the various parameters of the audio engines (to the right).

The mapping consists in connecting the controllers with the parameters (by some linear or non-linear scales) and it can vary in the time, as it is the case in the piece. Two types of connections are possible: the discreet connections and the continuous connections. Indeed, the discreet controllers (underlined on the figure 2), giving only a value from time to time, as the hit energy estimator, correspond to the control of type percussive and are going to serve for controlling the discreet parameters of the audio engines, as the activation of a segment for the concatenative synthesis (highest arrow). On the contrary, a continuous connection connects a continuous controller, as the linear absolute position on a tactile ribbon in a continuous parameter of audio engines such as the transposition, for instance (lowest arrow).

**Figure 2:** The mapping is at the interface between the data stemming from controllers (to the left) and the parameters of the audio engines (to the right).




## 5. FUTURE PERSPECTIVES

Of this period of research can be deduced several perspectives concerning the gesture capture, the real time speech synthesis, as well as their connections. Concerning the gesture capture, the used sensors possess the advantage to be rather light to be integrated into gloves, as well as good sensitivity in the movement (capture of the variation of the Momentum is accurate). On the other hand, they present a rather low energy autonomy, and do not offer the measure of the absolute static position. It would be beneficial to add to accelerometers, another technology allowing to access the absolute position. As regard the audio engine, it would be interesting to bend over the other speech synthesizers based on parametric (articulatory), semi-parametric (HMM) or hybrid models (concatenative/HMM). Indeed, the concatenative synthesis in batch mode has for advantage its degree of realism, which becomes difficult to maintain in real time. Finally the mapping between the gesture and the speech synthesis is a rich subject of research. As a research track, we can imagine more complex mappings where become interleaved the temporal and the semantic aspects of both the hand gesture and the vocal gesture.

## 6. ACKNOWLEDGMENTS

Author would like to thank IRCAM's research and production teams for their helps.

## 7. REFERENCES

[1] G. Beller. *Analyse et Modèle génératif de l'expressivité : Application à la parole et à l'interprétation musicale*. PhD thesis, Université Paris XI, IRCAM, June 2009.

[2] G. Beller. Transformation of expressivity in speech. *Linguistic Insights*, 97:259–284, 2009.

[3] G. Beller, N. Obin, and X. Rodet. Articulation degree as a prosodic dimension of expressive speech. In *Speech Prosody 2008*, pages 681–684, Campinas, 2008.

[4] G. Beller and X. Rodet. Content-based transformation of the expressivity in speech. In *Proceedings of the 16th ICPhS*, pages 2157–2160, Saarbruecken, August 2007.

[5] A. D. Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *JASA*, 111:1917–1930, 2002.

[6] P. Cook. Spasm: a real-time vocal tract physical model editor/controller and singer: The companion software synthesis system. *Computer Music Journal*, 17(1):30–34, 1992.

[7] C. d'Alessandro, N. D'Alessandro, S. L. Beux, J. Simko, F. Cetin, and H. Pirker. The speech conductor: gestural control of speech synthesis. In *eNTERFACE*, 2005.

[8] S. Fels and G. Hinton. Glove-talk 2: A neural network interface which maps gestures to parallel formant speech synthesizer controls. *IEEE Transactions on Neural Networks*, 9(1):205–212, 2004.

[9] E. Fléty and C. Maestracci. Latency improvement in sensor wireless transmission using ieee 802.15.4. In *NIME*, 2011.

[10] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone. Towards a segmental vocoder driven by ultrasound and optical images of the tongue and lips. In *Interspeech*, pages 2028–2031, 2008.

[11] P. Lanchantin, A. C. Morris, X. Rodet, and C. Veaux. Automatic phoneme segmentation with relaxed textual constraints. In *LREC2008*, Marrakech, Morocco, 2008.

[12] A. Roebel, F. Villavicencio, and X. Rodet. On cepstral and all-pole based spectral envelope modeling with unknown model order. In *PRL*, 2006.