Calliphony: A real-time intonation controller for expressive speech synthesis

Sylvain Le Beux, Albert Rilliard, Christophe d'Alessandro,

LIMSI-CNRS, BP 133, F-91403, Orsay, France

{slebeux, rilliard, cda}@limsi.fr

Abstract

Intonation synthesis using a hand-controlled interface is a new approach for effective synthesis of expressive prosody. A system for prosodic real time modification is described. The user is controlling prosody in real time by drawing contours on a graphic tablet while listening to the modified speech. This system, a pen controlled speech instrument, can be applied to text to speech synthesis along two lines. A first application is synthetic speech post-processing. The synthetic speech produced by a TTS system can be very effectively tuned by hands for expressive synthesis. A second application is data-base enrichment. Several prosodic styles can be applied to the sentences in the data-base without the need of recording new sentences. These two applications are sketched in the paper.

Index Terms: prosodic modeling, prosodic perception, gestures, prosodic synthesis

1. Introduction

As speech synthesizers attain acceptable intelligibility and naturalness, the problem of controlling prosodic nuances emerges. Expression is made of subtle variations (particularly prosodic variations) according to the context and to the situation. In daily life, vocal expressions of strong emotions like anger, fear or despair are rather the exception than the rule. Then a synthesis system should be able to deal with subtle continuous expressive variations rather than clear cut emotions.

Expressive speech synthesis may be viewed from two sides: on the one hand is the question of expression specification (what is the suited expression in a particular situation?) and on the other hand is the question of expression realization (how is the specified expression actually implemented). The first problem (situation analysis and expression specification) is one of the most difficult problems for research in computational linguistics, because it involves deep understanding of the text and its context. In this paper, only the second problem is addressed. The goal is to modify speech synthesis in real time according to the gestures of a performer playing the role of a "speech conductor" [1]. The Speech Conductor adds expressivity to the speech flow using Text-to-Speech (TTS) synthesis, prosodic modification algorithms and gesture interpretation algorithms.

This work is based on the hypothesis that human expressivity can be described in terms of movements or gestures, performed through different media, e.g. prosodic, body or facial movements. This question is closely related to musical synthesis, a field where computer based interfaces are still subject of much interest and development [2]. It is not the case for speech synthesis, where only a few interfaces are available for controlling in real time expressivity of spoken utterances. Existing gesture controlled interfaces for speech production either are dealing with singing synthesis (cf. [3], [4]) or with full speech synthesis [5], but with a sound quality level insufficient for dealing with expressivity.

In this paper a new system for real-time control of intonation is presented, together with application to text-to-speech synthesis. This system maps hand gestures to the prosodic parameters, and thus allows the user to control prosody in a cross-modal way. As a by-product, the cross-modal approach of prosody generation represents a new way to generate and describe prosody and may therefore shed a new light on the fields of prosody systems and prosody description.

The paper is organized as follows. The real-time intonation controller is described in Section 2. The performances of the controller for real-time intonation modification are evaluated in section 3. Applications to expressive text-to-Speech synthesis are sketched in section 4. Section 5 discusses the results obtained so far, proposed future work and gives some conclusions.

2. Real-time intonation controller

2.1. Principle

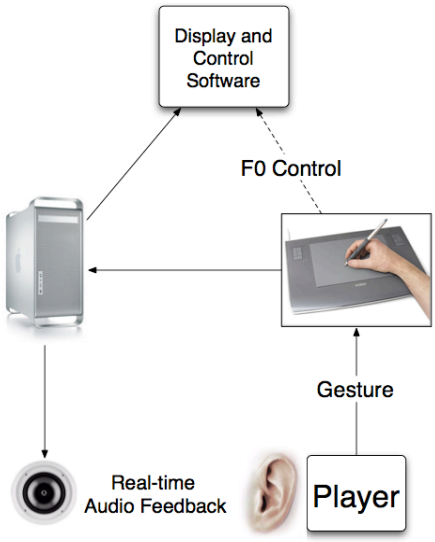

Figure 1: *Generic diagram of the system*

The real-time intonation controller operates in principle like a musical instrument. The loop between the player and the instrument is depicted in Figure 1. The player's hand movements are captured using an interface, and these movements are mapped on the input controls of the synthesizer. The sound is modified accordingly, played, and this audio feedback is perceived by the player who modifies his gestures as a function of the perceived and intended sounds.

2.2. Gesture interface: writing movements

Many devices, among which MIDI keyboard, Joystick and data glove, have been tested for capturing gestures with intonation control in mind.

Keyboards are not well fitted because it allows only discrete scales, although in speech a continuous control is mandatory. An additional pitch-bend wheel proved not very convenient from an ergonomic point of view.

As for the joystick and data glove, the precision in terms of position seemed insufficient: it proved too difficult to reach accurately a given target pitch. Such devices seem better suited for giving directions (as in a flight simulator) than precise values.

The graphic tablet has been chosen because it presents a number of advantages: its sampling frequency is high (200 Hz) and its resolution in terms of spatial position is sufficient for fine-grained parameter control (5080 dots per inches). Moreover, all the users are trained in writing since childhood, and are 'naturally" very much skilled in pen position control. Scripture, like speech, is made of a linguistic content and a paralinguistic, expressive content (in this case called "calligraphy"). There is a remarkable analogy between pitch contour and scripture. This analogy between drawing and intonation is very effective and intuitive from a performance point of view. Untrained subjects proved to be surprisingly skilled for playing with intonation using the pen on the graphic tablet, even at the first trial. For intonation control, only one axis of the tablet is necessary. The vertical dimension (Y-axis) is mapped on the F0 scale, expressed in semi-tones. The x-scale is not used: it means that very different gestures can be used for realizing a same intonation pattern: some players were drawing circle like movements, when others preferred vertical lines or drawing similar to pitch contours. The second spatial dimension of the tablet will be used later for duration control in a second stage. Other degrees of freedom are still left in the tablet (pressure, switch) and will be use for controlling additional parameters, e.g. parameters related to voice quality.

Taking these observations into account, we decided to opt for a Wacom graphic Tablet, A4 size and we based our platform on a Power PPC Apple G5 Mac, 2.3 GHz bi-processor.

2.3. Real-time software

Real-time processing of information is a key point of the Calliphony system: as the user adapts his hand movement to perceived pitch at the output of the system, the delay has to remain inaudible. Calliphony is elaborated under the Max/MSP[1] software ([6], [7]), which is a graphical development environment intended to processes sound in real-time and which has already proven several years of reliable experience in real-time sound processing. Concerning the modification of speech pitch, we used a TD-PSOLA [7] Pitch-Shifter external provided by Tristan Jehan for Max/MSP environment [9].

As described on Figure 2, Calliphony takes as inputs the Y-axis position of the pen on the graphic tablet, and a recorded sound to be modified. It then maps the pitch value of the sound output to a value corresponding to the Y-axis value. This mapping is done on a logarithmic scale, such as the metric distance of each octave is the same. This corresponds analogously to the perception of the pitch by the human ear.

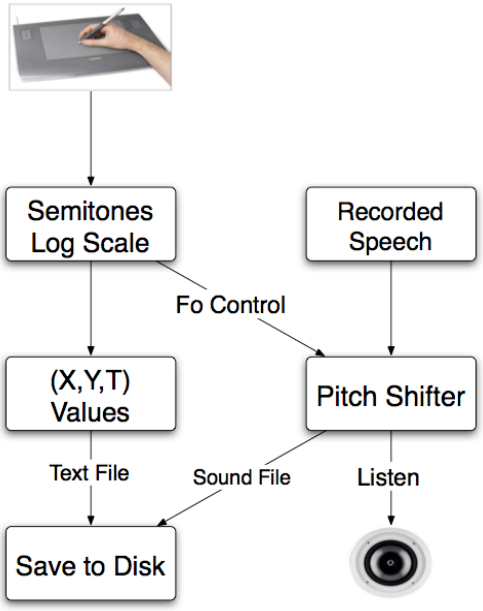

Figure 2: *"Calliphony" system description*

3. Evaluation of the controller

The use of handwriting movement to control pitch is not a priori straightforward. An evaluation procedure has therefore been developed, in order to assess the ability of a human to perform real-time control of speech prosody. The principle of this evaluation procedure is to measure the ability of the Calliphony player to imitate as closely as possible the prosody of an original sentence. The handwriting imitation performances are compared to the oral ability of the same user to imitate the same sentences. This work is described in more detail in a companion paper (cf. [10])

3.1. Prosodic imitation interface

A specific interface (cf. fig. 3) was developed to allow the subjects of the experiment to easily perform their imitation task. This interface encapsulate the Calliphony system, so that the user can listen to an original sentence, and then imitate the prosody both on a F0 flattened version of the sentence and vocally by recording his own voice.

Figure 3: interface used for the handwriting imitation of prosody. Buttons allow to listen to the original sentence, record its own speech or the graphic tablet, listen to a recorded performance and save it. The current sentence's F0 is displayed.

[1]It is noticeable however that Max/MSP software is not multithreaded and consequently did not allows to take full advantage of the multi-processors architectures.

Table 1: *The 18 sentences of the corpus, from 1 to 9-syllable length.*

Nb syllable	Sentence	Phonetic	Sentence	Phonetic
1	Non.	[nɔ̃]	L'eau	[lo]
2	Salut	[saly]	J'y vais.	[ʒi vɛ]
3	Répétons.	[ʁepetɔ̃]	Nous chantons.	[nuʃɑ̃tɔ̃]
4	Marie chantait.	[maʁi ʃɑ̃tɛ]	Vous rigolez.	[vuʁigole]
5	Marie s'ennuyait.	[maʁi sɑ̃nɥijɛ]	Nous voulons manger.	[nu vulɔ̃ mɑ̃ʒe]
6	Marie chantait souvent.	[maʁi ʃɑ̃tɛ suvɑ̃]	Nicolas revenait.	[nikola ʁəvənɛ]
7	Nous voulons manger le soir.	[nu vulɔ̃ mɑ̃ʒe lə swaʁ]	Nicolas revenait souvent.	[nikola ʁəvənɛ suvɑ̃]
8	Sophie mangeait des fruits confits.	[sofi mɑ̃ʒɛ de fʁɥi kɔ̃fi]	Nicolas lisait le journal.	[nikola lizɛ lə ʒuʁnal]
9	Sophie mangeait du melon confit.	[sofi mɑ̃ʒɛ dy məlɔ̃ kɔ̃fi]	Nous regardons un joli tableau.	[nuʁəgaʁdɔ̃ ɛ̃ ʒoli tablo]

As the aim of the evaluation is to investigate how close to the original the imitations can be, subject are able to listen the original sound when they need to, and to perform imitation until they are satisfied. Several performances can be recorded for each original sound.

3.2. Evaluation paradigm

3.2.1. Corpus

The evaluation procedure is based on a dedicated corpus constructed on 18 sentences, ranging from 1 to 9 syllables length (cf. table 1). Each sentence was recorded in its lexicalized version, and also in a reiterant delexicalized version, replacing each syllable by the same /ma/ syllable. Constraints on the corpus construction were: the use of CV syllable structure and no use of plosive consonant at the beginning of the words. Such constraints aimed at obtaining easily comparable prosodic patterns amongst the sentences and at avoiding important micro-prosodic effect due to plosive bursts.

Two native speakers of French recorded the corpus (a female and a male), according to three consigns: (1) to perform a declarative prosody, (2) to make an emphasis on one specific word of each sentence (generally on the verb) and (3) to perform an interrogative prosody. This results in 108 sentences, directly digitalized on a computer (41kHz, 16bits) for each speaker, using an USBPre sound device connected to an omnidirectional AKG C414B microphone placed 40 cm to the speaker mouth, and performing a high-pass filtering of frequency under 40Hz plus a noise reduction of 6dB.

3.2.2. Calliphony players

4 users have completed the experiment on a subset of 9 sentences ranging from 1 to 9 syllables, either lexicalized or reiterated, and using the three prosodic conditions (declarative, emphasized, interrogative), for the male speaker. All subjects are involved in this work and completely aware of its aims and are therefore familiar with prosody. Three out of the four subjects are trained musicians. One of the four subjects is the male speaker of the original corpus, who has therefore imitated its own voice vocally and by handwriting movements.

3.2.3. Prosodic parameters and distances measures

In order to evaluate the objective distance between the original and the imitated sentences, their pitch values have to be carefully extracted and computed. All the sentences of the corpus were manually analyzed. Their prosodic parameters were automatically extracted: fundamental frequency for vocalic segments (in semitones) and the corresponding voicing strength (calculated from intensity), syllabic duration and intensity thanks to Matlab (the yin script [11]) and Praat [12] programs.

The objectives distances between the prosody of the original sentence and the imitated prosody were calculated on the basis of the physical dissimilarity measures introduces by Hermes [13]: the correlation between the two F0 curves, and the root-mean-square (RMS) difference between theses two curves. The voicing strength was used (as suggested by [13]) as a weighting factor in the calculation of these two distances measures.

Objectives distances between the original sentence and each repetition at the output of the Calliphony system were automatically calculated by using 10 ms spaced vector of F0 values for each vocalic segment. Then only the closest imitation, according to the weighted correlation measure and then the weighted RMS distance, was kept for the result analysis. This part of the work can be completely automated, as there is no duration change between the output of Calliphony and the original sentence. This is not the case for the oral imitations, which have to be labeled prior to extract F0 values for vocalic segments.

Moreover the distance computation supposes segments of the same length, a condition not met for vocal imitations. Therefore, only the distances between the original sentences and the gestural imitations have been calculated so far.

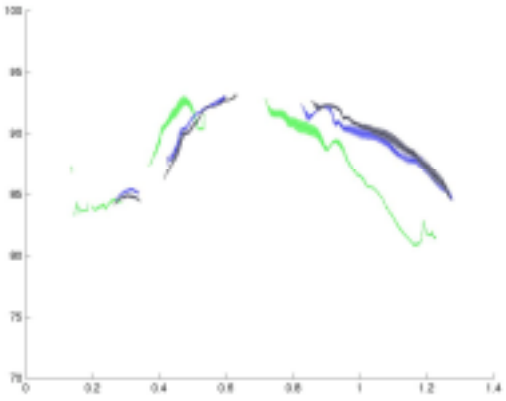

Figure 4: raw F0 value (in tones) for an original sentence (gray) and the two vocal imitations of one subject. Stimuli are not time-aligned.

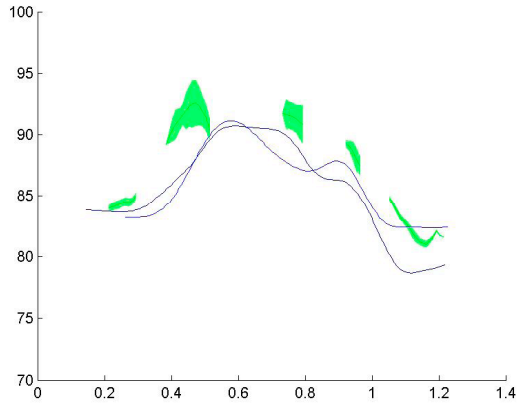

Figure 5: stylized F0 of an original sentence (the same as in fig. 4 – gray curve, smoothed values for the vocalic segment expressed in tones), and the value of the pitch parameter controlled by the graphic tablet for all the imitations performed by one subject. Stimuli are time-aligned.

Graphics with the raw F0 value of both the original and the vocal imitations have been produced in order to visually compare the performances of gesture vs. vocal imitations. Graphic with the stylized F0 of the original sentences (smoothed F0 for the vocalic segments) superimposed with the course of the pen on the graphic tablet were also produced in order to compare the two imitations modalities (fig. 4 & 5).

3.3. Results

The mean objective distances are summarized in Table 2. There is no major difference between the four users, except for a higher RMS distance for AR, the only non-musician amongst the users (for a discussion about this issue cf. [10]).

Table 2: *mean distances for each subject and for all sentences imitated by handwriting movements.*

Subject	***R***	***RMS***
CDA	0.866	3.108
BD	0.900	3.079
SLE	0.901	3.091
AR	0.898	4.728
Total	0.891	3.502

The prosodic condition (declarative, emphasized, interrogative prosody) did not have a significant impact on the users' performances. The reiterant speech condition neither.

The most influential factor in the experiment is the length of the sentence, as correlations continuously decrease while the number of syllable increase (cf. figure 6). This result can be explained either by an increasing difficulty of the user's task, or by an artifact due to the sentence length, because computation of correlation does not take into account any weighting for length compensation. More analyses would be needed before concluding on a sentence length effect.

Finally, the most important result of this evaluation procedure is the high overall correlation and low RMS distance obtained by all users. This result generally validates the ability of human user to imitate very closely an original prosody by using handwriting movements. Moreover, the observation of the imitated F0 curves shows a complete smoothing of any micro-prosodic variations: this indicates that users only reproduce prosodic movement at the level of the syllable or above, and that the task adequately matches prosody imitation and generation purposes.

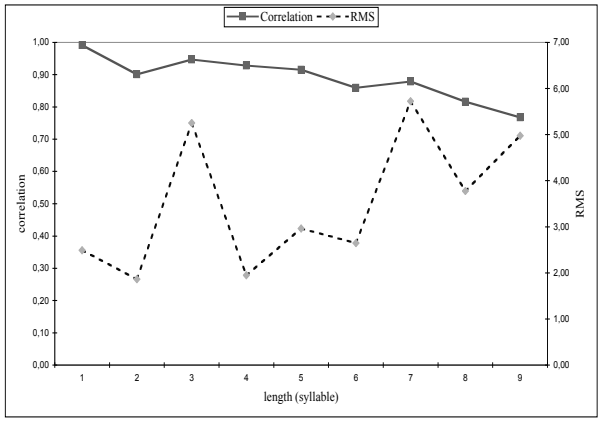

Figure 6: evolution of the two distances measures with the sentence's length. X-axis : length of stimuli, left Y-axis: correlations (plain line), right Y-axis: RMS difference (dotted line).

4. Application to expressive speech synthesis

Since the adequacy of a hand-driven interface to control speech prosody is validated, this section will explore some possible applications of this interface.

4.1. Intonation post-processing

A first application of the Calliphony system is directly derived from the scheme developed for the evaluation of the system: to allow a user to directly change the pitch of a spoken utterance. Such application can be useful in the field of speech synthesizers: as such devices have already reach a high degree of naturalness, they are know seeking for expressivity. The major problem is then to record and adequately model the huge corpora needed to be able to face any kind of expressivity for any sentences.

Our proposal is to give the end user the possibility to directly add the expressivity he needs on the output of his speech synthesizer thanks to the Calliphony system. This system is easy to use and only need a few practice. Someone could then easily add e.g. a focalization on a desired word.

4.1.1. Assessment procedure

In order to assess the ability of our system to add such kind of expressivity to synthetic speech, a validation procedure has been set up, and is reported hereafter. It is based on exactly the same principle as the validation of the Calliphony system for prosody imitation reported above, with the only difference being that flattened speech (the input of the Calliphony system) has been replaced here with a synthetic sentence, produced with the Selimsi TTS system [14]. The player of Calliphony hears an original sentence from our corpus, carrying either a focalization on one word or an interrogative prosody. He has then to reproduce the pitch contour of the original sentence on the synthetic sentence, on a similar task that the one described above.

The major difference between the two experiments concerns the segmental duration of the modified stimuli: for the preceding evaluation, the segmental durations are

exactly the same as the original, as it is only a flattened version of the natural stimulus, whereas the synthetic sentences has his proper durations. It induces two major differences between the two protocols. The first one concerns the modification procedure: it is harder to perform an imitation when important lengthening are present in the original sentence. The second one concerns the distance measure between the original and the reproduced pitch contours. As the pitch values are compared for vowel only, and as synthetic and natural vowel doesn't necessarily have the same duration, instead of extracting one value of pitch for each 10 ms, 10 values for each vowel were calculated, regularly spaced along the vowel. These 10 values per vowel are then used to calculate correlation and RMS distance using the same formulae as those presented above.

Table 3: mean distance scores obtained for focalized sentence, interrogative sentences and for all sentences.

	Correl	RMS
Focalization	0,92	3,18
Interrog.	0,86	4,14
Mean	0,89	3,66

4.1.2. Results and analyses

The results obtained for this assessment are quite similar to those already exposed.

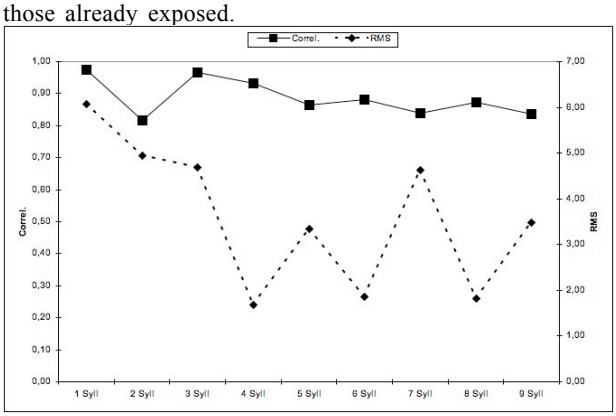

Figure 7: mean distances (correlation and RMS distance) obtained for sentences of all length, from 1 to 9 syllables.

Mean correlation and RMS distance are good (cf. tab 3), and indicate a close stylization of the pitch curve on the synthetic stimuli, even if there are duration differences. Mean score obtained for focalization vs. interrogation sentences are quite similar, with slightly better score for focalization. About the effect of the sentence's length (cf. fig. 7), the effect is a bit more complicated than the one observed with natural speech: if correlation decreases gradually with the sentence length, as it has already been observed, the RMS distance did not have any particular tendency, except for the 1-, 2- and 3-syllables length sentences, that receive high RMS distances scores, unlike for natural speech.

The objective distance between modified prosodic parameter at the output of Calliphony and the original natural prosody are close together, giving a very good idea of the system's performances at producing expressive speech.

However, it must be noted that, has the duration parameter is not dealt with in this first version of Calliphony. This is not satisfactory for high quality expressive synthesis, were durations modification is mandatory. In addition, the sound quality is better for natural speech modification compared to synthetic speech modification. In our current implementation Calliphony results in two successive modification of the signal (concatenation and PSOLA modification), a situation that is not optimal indeed. More work is still needed before to obtain a better sounding system, but we think that the ability of players to add expressivity to synthesizers has been convincingly demonstrated.

Considering the data-bases that are not previously tagged, the system can still be used online in a slightly different manner. When the purpose is only to produce some expressive sentences (for various perceptive experiments for example) then one is able to modify online the synthesized sentences and to record them directly after modification.

This gives the opportunity for someone not necessarily familiar with speech synthesis and processing, to produce expressive sentences in a convenient manner, without having to buy an expensive system or to acquire deep knowledge in speech processing. Moreover, one can use synthesized sentences from any TTS engine publicly available or can directly record sentences on its owns with a simple microphone and recording software, before achieving its modification thanks to our system.

As a extent, thanks to the good quality of expressive modifications on synthetic speech, it could find applications in various research and development situations, going from advertising to industrial mass media, or even animated cartoon characters voice synthesis.

4.2. Data-base enrichment

Another application of the Calliphonic system to TTS concerns specifically data-driven speech synthesis. Synthesis systems based on selection/concatenation of non-uniform units need large corpora of recorded speech. Our system can be used to for enrichment of the speech data-base, prior to synthesis. In this case, natural speech is modified, and a same sentence can be given several prosodic variations, as depicted in Fig.8

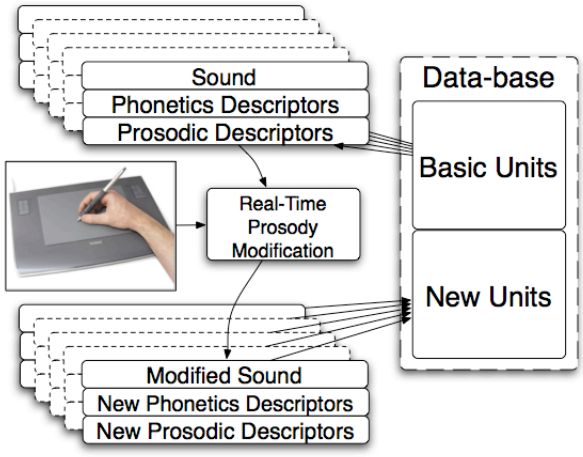

Figure 8: Enrichment of Data-Base with Non-Uniform Units

There are several steps to achieve this enrichment and it can be applied to various types of data-bases. There are no constraints on the content of the data-base. The system can then be used to add new expressions that were not recorded, or to have more utterances of a less represented expression.

Then the prosodic content of the database can be extended and/or improved without the need of new recordings. This is independent of the TTS system itself, because it is only a matter of data-base pre processing.

This application is in a preliminary stage: no formal evaluation of the synthetic speech obtained is available for the moment.

5. Discussion and conclusion

Speech instruments have been an important part of the history of speech synthesis, but have played only a marginal role in speech synthesis application or research. We think that high quality real-time speech modification algorithms and new high precision interfaces have the potential for dramatically changing the current situation.

In this paper, we explore the ability of hand-writing movement for expressive speech synthesis. The system has been called "calliphony", i.e. expressive speech beyond phonemes by analogy with "calligraphy", i.e. expressive writing beyond graphemes. The results indicate even untrained players are almost as skilled for vocal imitation as for written imitation of expressive prosody.

Then, the system can be applied to TTS post processing and database pre-processing. TTS post-processing can be a useful extension of a TTS system for tuning synthetic speech utterance output without the need of deep engineering or expensive recordings. The quality obtained is basically the quality of the TTS system itself.

We are currently exploring the quality reached by data-base enrichment, a pre-processing for augmenting the prosodic content of a selection/concatenation TTS system, without recording new sentences.

Future work will be devoted to duration and tempo modifications. Our experiments show (or confirm) that changing intonation without changing duration or tempo is not enough in many situations. Changing voice quality is also required for more realistic prosodic modifications. Additional control parameters will then be needed.

Another path of research for future work is the interface itself. We are currently pursuing the study of the range of possibility offered by an ad-hoc controller called the Meta-Instrument. This controller offers up to 54 continuous controllers simultaneously, supervised by the fingers and the arms (see [15]).

6. References

[1] D'Alessandro, C., et al. (2005) "*The speech conductor : gestural control of speech synthesis*." in eNTERFACE 2005. The SIMILAR NoE Summer Workshop on Multimodal Interfaces, Mons, Belgium.

[2] http://www.nime.org/

[3] Cook, P., (2005) *"Real-Time Performance Controllers for Synthesized Singing,*. Proc. NIME Conference, 236_237, Vancouver, Canada.

[4] Kessous, L. , (2004) *"Gestural Control of Singing Voice, a Musical Instrument"*. Proc. of Sound and Music Computing, Paris.

[5] Fels, S. & Hinton, G. , (1998) *"Glove-Talk II: A Neural Network Interface which Maps Gestures to Parallel Formant Speech Synthesizer Controls."* IEEE Transactions on Neural Networks, 9 (1), 205_212.

[6] Puckette, M. (1991). *"Combining Event and Signal Processing in the MAX Graphical Programming Environment"*. Computer Music Journal 15(3): 68-77.

[7] http://www.cycling74.com/

[8] E. Moulines and F. Charpentier, (1990) *"Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,"* Speech Communication, vol. 9, pp. 453–467.

[9] http://web.media.mit.edu/tristan/

[10] d'Alessandro, C., Rilliard, A. & Le Beux, S. (Submitted). *"Computerized chironomy: evaluation of hand-controlled intonation reiteration."* Proc. of InterSpeech 2007.

[11] de Cheveigné, A., Kawahara, H., (2002) *"YIN, a fundamental frequency estimator for speech and music"*, J. Acoust. Soc. Am. 111, 1917-1930.

[12] Paul Boersma & David Weenink, (2001)*"PRAAT, a system for doing phonetics by computer."* Glot International 5(9/10): 341-345.

[13] Hermes, D.J. (1998). *"Measuring the Perceptual Similarity of Pitch Contours"*. Journal of Speech, Language, and Hearing Research, 41, 73-82.

[14] Prudon, R. and C. d'Alessandro. (2001) "*A selection/concatenation text-to-speech synthesis system: databases development, system design, comparative evaluation*." in 4th ISCA/IEEE International Workshop on Speech Synthesis.

[15] Serge de Laubier, Vincent Goudard, (2006) *"Méta-Instrument 3 : a look over 17 years of practice",* in Proc. of the NIME Conference, IRCAM, Paris, France