Computerized chironomy: evaluation of hand-controlled Intonation reiteration

Christophe d'Alessandro, Albert Rilliard, Sylvain Le Beux[1]

[1] LIMSI-CNRS, BP 133, F-91403, Orsay, France

{cda, rilliard, slebeux}@limsi.fr

Abstract

Chironomy means in this paper of intonation modeling in terms of hand movements. An experiment in hand-controlled intonation reiteration is described. A system for real-time intonation modification driven by a graphic tablet is presented. This system is used for reiterating a speech corpus (sentences of 1 to 9 syllables, natural and reiterant speech). The subjects also produced vocal imitation of the same corpus. Correlation and distances between natural and reiterated intonation contours are measured. These measures show that chironomic reiteration and vocal reiteration give comparable, and good, results. This paves the way to several applications in expressive intonation synthesis and to a new intonation modeling paradigm in terms of movements.

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

1. Introduction

Although various intonation models have been proposed for a variety of languages, the question of expressive intonation representation is still wide open. Phonological models of intonation are focusing on contrastive (often tonal) structures: they are not designed for description of expressive intonation variations. Phonetic description and stylization of intonation often describes melodic patterns in terms of "movements", "contours" or "target points". The approach defended in this paper is based on the hypothesis that intonation shares a lot of common features with other types of expressive human movements or gestures (like face and hand gestures). Then, addressing the question of intonation representation in terms of movements, like e.g. hand movements, could bring new insights in intonation research. The analogy between intonation and movements seems promising. A first application is direct hand-controlled expressive synthesis: this could be used for corpora enrichment in concatenative speech synthesis, or stimuli generation in expressive speech analysis. A second more fundamental application could be intonation modeling in terms of movement representation (movement speed, direction, height). A main advantage of such a modeling is that intonation and rhythm are dealt with in a unified framework.

Expressive intonation description in terms of hand gestures is known since antiquity under the term "chironomy" (cf. [9], part 1, p. 103). This term comes from the greek "chiro" (hand) and "gnomos" (rule). The term appears first in the fields of rhetoric for describing co-verbal hand movement that reinforce expression of the discourse [11]. Another meaning appears in medieval music, where chironomy is meant for the hand gestures of the conductor that indicates the tones to the choir in Gregorian chant ([9], part 2, p. 683).

Music and speech are forms of human communication by the mean of expressive sound control. Music, contrary to speech, developed the usage of external "instruments" for sound production and sound control. Instrumental music is produced by hand-, breath-, or feet-controlled "interfaces". As new interfaces for musical expression recently received a lot of attention, resources like real-time sound programming languages, control devices, modification algorithms are available in the electronic music community (cf. [4], [7]). Along this line a system for computerized chironomy, i.e. real-time melodic control by hand-driven movements is presented and evaluated in this paper. Among the devices available for controlling hand movement, hand-writing (graphic tablet) has been preferred. This is because hand writing allows for the most accurate and intuitive intonation control. The main questions addressed in the present experiment are:

1. How well can handwriting movements reproduce intonation movements?
2. How do handwriting and vocal intonation stylization compare?
3. In both cases, how close are natural intonation contours and stylized contours?

These questions are addressed using an intonation reiteration paradigm. The task of the subjects was to reproduce intonation patterns by vocal mimicking and hand-control movements. Both speech and reiterant (i.e. "mamama") speech sampled of various sizes were proposed. Distance measures between the original and reiterated speech are used as performance assessment.

The paper is organized as follows. The experimental apparatus, test paradigm and analysis procedures are described in Section 2. Results in terms of performance for intonation imitation are given in Section 3. Section 4 discusses the results obtained, and gives some conclusions.

2. Experiments

2.1. Prosodic control system

2.1.1. Gestural pitch shifter

A new system was developed in order to control the pitch of speech utterances by means of handwriting gestures. The system, to some extent similar to the one described in [1], is based on the Max/MSP programming environment, and use a real time version of the TDPSOLA algorithm. It deals with two inputs: (1) a recorded speech utterance with a flattened fundamental frequency (to e.g. 120Hz for a male speaker), and (2) the output of a gesture control device such as a graphic tablet. The value of one parameter of the graphic tablet (controlled by handwriting movements) is mapped to the pitch value of the spoken utterance, resulting in a direct control by the gesture of the output utterance pitch. Hence, this system allows one operator to precisely control the pitch of a previously recorded utterance, using only a pen on a graphic tablet.

2.1.2. *Prosodic imitation interface*

In order to test whether the control of prosody by handwriting movements can realistically reproduce natural prosody, a specific computer interface has been developed (cf. figure 1) under the Max/MSP platform. It is intended to allow subjects of the experiment to imitate the prosody of natural speech either vocally or by handwriting movements. Each subject listens to a natural sentence by clicking on a button with the mouse pointer, and therefore has to imitate the prosody he has just heard by two means: vocally by recording its own voice, and by using the gestural controller of prosody.

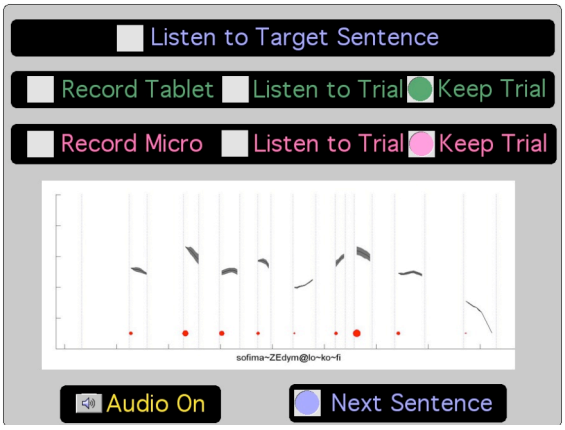

Figure 1: *interface of the experiment. Buttons allow to listen to the original sentence, record its own speech or the graphic tablet, listen to a performance and save it if it is satisfactory. The image represents the prosody of the current sentence.*

The interface displays some control buttons (cf. figure 1): (1) to record the voice or the graphic tablet imitation, (2) to replay the recorded imitations and (3) to save the satisfactory ones. It also displays a graphic representation of the prosodic parameters of the original sound, as it will be described latter.

As the aim of the experiment 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.

Finally, subjects go on to the next sound. As the test typically lasts several minutes per sentence, subjects are instructed to take rest from time to time.

2.2. Corpus

These experiments are 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 delexicalized version, replacing each syllable by the same /ma/ syllable, in order to obtain reiterant speech [8]. When constructing the corpus, words were chosen with respect to two criterions (use of CV syllable structure and no plosive consonant at the beginning of the words), in order to obtain easily comparable prosodic patterns amongst the sentences and to avoid important micro-prosodic effect due to plosive bursts.

Two speakers (a female and male, native speakers of French) recorded the corpus. They have to produce each sentence in a random order, and according to three different consigns: (1) using a declarative intonation, (2) performing an emphasis on a specific word of the sentences (generally the verb) and (3) using an interrogative intonation. The speakers were instructed to read the sentence and then to produce it using the current intonation style. Once the sentence is recorded in its lexicalized version, they have to reproduce it by using the same prosody, but in its reiterated version. Speakers were able to make as many trials as needed in order to obtain a satisfactory pair of sentences.

108 sentences were thus recorded and 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.

2.3. Subjects

Until now, 4 subjects have completed the experiment on a subset of 9 sentences ranging from 1 to 9 syllables, either lexicalized or reiterated, and with 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 imitate its own voice vocally and by handwriting movements.

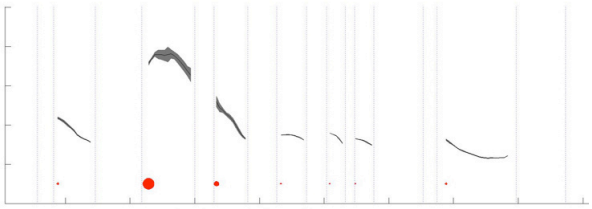

Figure 2: *prosodic parameters of a 7-syllable length sentence from our corpus.*

2.4. Prosodic contours measurements

All the sentences of the corpus were manually analyzed in order to extract their prosodic parameters: fundamental frequency (in semitones), syllabic duration, and intensity thanks to Matlab (the yin script [3]) and Praat [2] programs.

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.	[nu ʃɑ̃tɔ̃]
4	Marie chantait.	[maʁi ʃɑ̃tɛ]	Vous rigolez.	[vu ʁigole]
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.	[nu ʁəgaʁdɔ̃ ɛ̃ ʒoli tablo]

For all the sentences, graphics were displayed to depict the prosody of original sound in order to facilitate the task of subjects of the experiment (cf. figure 2). These graphics represents the smoothed F0 of the vocalic segments (manually aligned), with the line thickness representing the voicing strength. The voicing strength was calculated from the intensity (in dB) of the signal at the point of F0 analysis. The locations of the Perceptual Centers [10] are represented by red circles, the diameter of which is related to the mean intensity of the vocalic segment. Vertical dotted lines represent the phonemes' boundaries.

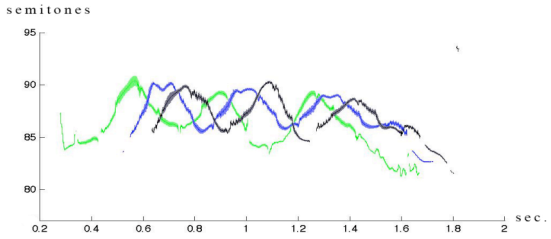

Figure 3: *raw F0 value (in tones) for an original sentence (gray) and the two vocal imitations of one subject. Stimuli are not time-aligned.*

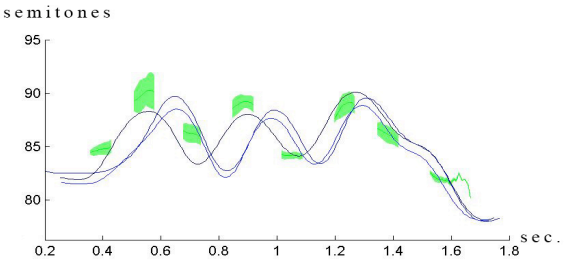

Figure 4: *stylized F0 of an original sentence (the same as in fig. 3 – gray curve, smoothed values for the vocalic segment expressed in tones), and the value of the pitch parameter controlled by the graphic tablet for all the imitations performed by one subject. Stimuli are time-aligned.*

2.5. Prosodic distances and correlation

In order to evaluate the performance of the imitation (either vocal or gestural), two physical measures of the distance between the fundamental frequency extracted from the imitation and the one from the original sentence were set of, on the basis of the physical dissimilarity measures introduces by Hermes [6]: the correlation between the two F0 curves, and the root-mean-square difference between theses two curves. As already noted by Hermes, the correlation is a measure of the similarity of the two sets of F0 parameters, whereas the RMS difference is a dissimilarity measure, but both give an idea of the similarity of the compared F0 curves. However, correlation test the similitude between the shapes of the two curves, without taking into account their mean distances: e.g. one can reproduce an F0 curve an octave lower than the original, if the shape is the same, the correlation will be very high. On the contrary, the RMS distance will give an idea of the area between the two curves, and is sensitive to differences between the two mean F0 levels.

Using a similar procedure as the one described in [6], the two prosodic distances were applied with a weighting factor in order to give more importance to the phonemes with a higher sound level. The weighting factor used is the intensity, as a measure of the local strength of voicing.

These two dissimilarity measures were automatically calculated for all the gestural imitations recorded by the four subjects for each of the 54 sentences. Then only the closest gestural imitation (according to first the weighted correlation and then the weighted RMS difference) was kept for the result analysis.

This part of the work can be completely automated, as there is no change in the duration of the output of the gestural controller of speech (only F0 is controlled). This is not the case for the oral imitations, which have to be manually labeled in order to calculate such distances. 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.

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. 3 & 4).

3. Results

3.1. Prosodic distances and correlation

The physical distances between stimuli produced by handwriting movements are summarized in table 2. In analyzing the results of the experiment, the relative influence of each controlled parameter will be detailed.

Table 2: *mean distances for each subject and for all 54 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

3.1.1. Effect of subjects

There is no important difference between the results obtained by all subjects: all correlations are comparable and around .9, showing that subjects are able to perceive and reproduce the shape of the intonation curve by means of handwriting movement. The only noticeable difference is the RMS distance obtained by subject AR (4.7) compared to the score of other subject (around 3.1). This difference indicates an F0 curve closer to the original one for the three other subjects than for AR. This can be explained by the fact that AR is the only subject without a musical formation, and therefore he is not trained to reproduce a given melody as the other are. However, as the correlations are quite the same, it does not imply difficulty to reproduce the pitch variation, but only the pitch height.

3.1.2. Effect of sentence length

The sentence length has a more noticeable effect on the distances. As shown in the figure 5, the dissimilarity measures increase as the sentences length grows: correlation continuously decrease when sentence length increase, and except for some accident for the 3 and 7-syllable length sentences, RMS difference grows according to sentence length. The two accidents could be explained by high RMS distances obtained by two subjects for this stimulus, and by the fact that this measure is particularly sensitive to small differences between curves. The effect of sentence length could be an artifact, 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.

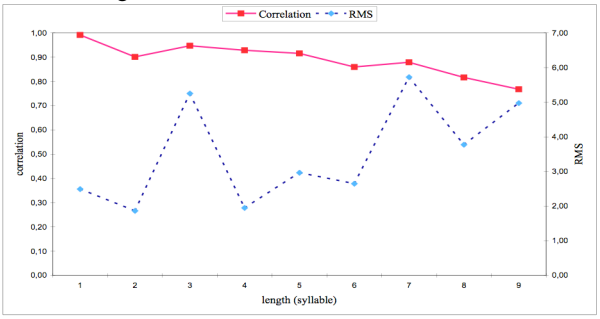

Figure 5: *evolution of the two distances measures with the sentence's length. X-axis : length of stimuli, left Y-axis: correlations (plain line), right Y-axis: RMS difference (dotted line).*

3.1.3. Effect of the prosodic style and type of stimuli

Declarative, emphasized or interrogative sentence modalities give similar results according to the correlation measure, but RMS distance is smaller for declarative curves (2.0) than for emphasized or interrogative ones (respectively 4.2 and 4.4). It can be linked to the preceding result: subjects are able to reproduce the shape of all intonation contours, but the precise perception the pitch level is harder when the curve present a glissando (e.g. during emphasis or interrogation) than a more flat curve, like for declarative intonation.

Finally, to imitate a reiterant sentence is nor easier nor harder than to imitate a lexicalized one: distances are the same for both kinds of stimuli.

4. Discussion and conclusions

4.1. Performance level and feasibility

Good performance levels are achieved in terms of correlation and distances between original and reiterated intonation contours. Of course, it must be pointed out that the best reiterated utterance has been selected for each sentence. However, the amount of training of each subject was not very heavy. The task seemed not particularly difficult, at least compared to other intonation recognition tasks, like e.g. musical dictation.

4.2. Gestures and vocal modalities

A remarkable and somewhat striking result is that the performance levels reached by hand written and vocal reiterated intonation are very comparable. This could suggest that intonation, both on the perceptual and motor production aspects, is processed at a relatively abstract cognitive level, as it seems somehow independent of the modality actually used. This fact was already inferred by orators in the ancient world, because description of the expressive effect of co-verbal gestures (i.e. multimodal expressive speech) has been remarked even in early roman rhetoric treatises [11]. Then one can hypothesize that intonation control can be achieved by other gestures than pitch gestures with comparable accuracy.

4.3. Intonation and gestures

It seems that micro-prosodic variations have been neglected for almost all sentences. Writing is generally slower than speaking, and then hand gestures are not able to follows fine grained intonation details like micro-prosody [5]. Moreover, results for delexicalized speech and normal speech are comparable, although micro-prosody is almost neutralized in delexicalized speech. Then the hand gestures correspond rather to prosodic intonation movements. The specific gestures used by different subjects for achieving the task at hand have not been analyzed in great detail for the moment. Some subject used rather circular movements, other rather linear movements. This point will be addressed in future work.

4.4. Conclusion and future work

This paper presents a first evaluation of computerized chironomy, i.e. hand-driven intonation control. The results show that vocal intonation reiteration and chironomic intonation reiteration give comparable intonation contours in terms of correlation and RMS distance. Applications and implications of these finding are manifold. Chironomic control can be applied to expressive speech synthesis. It can also be used for expressive speech analysis, as expressive contours can be produced and represented by the hand-made tracings. Future work will address the question of gesture control of rhythm and voice quality parameters. An auditory evaluation of the reiterated intonation contours is also planned. Finally, this work can also serve as a basis for intonation modeling in terms of movements. This could form a unified framework for expressive gesture representation, using common features like velocity, target position, rhythmic patterns etc.

5. References

[1] D'Alessandro, N., d'Alessandro, C., Le Beux, S. & Doval, B. (2006). "Real-time CALM synthesizer: new approaches in hands-controlled voice synthesis". Proc. of NIME2006, 266-271, Paris, France, June 4-8.

[2] Boersma, P. & Weenink, D. (2006): Praat: doing phonetics by computer (Version 4.5.05) [Computer program]. Retrieved 12/2006 from http://www.praat.org/

[3] de Cheveigné, A. & Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music". JASA, 111, 1917-1930.

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

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

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

[8] Larkey, L.S. (1983). "Reiterant speech: an acoustic and perceptual validation". JASA, 73(4), 1337-1345.

[9] Mocquereau, A. (1927). "Le nombre musical grégorien". Desclée, Paris.

[10] Scott, S.K. (1993). P-Centers in speech – an acoustic analysis. PhD thesis, University College London.

[11] Tarling, J. (2004). "The Weapons of Rethoric", Corda Music, Pub. London.