Chironomic stylization of intonation[a]

Christophe d'Alessandro,[b] Albert Rilliard, and Sylvain Le Beux
LIMSI-CNRS (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur - Centre National de la Recherche Scientifique), BP 133, Orsay 91403, France

(Received 30 March 2010; revised 12 October 2010; accepted 17 November 2010)

Intonation stylization is studied using "chironomy," i.e., the analogy between hand gestures and prosodic movements. An intonation mimicking paradigm is used. The task of the ten subjects is to copy the intonation pattern of sentences with the help of a stylus on a graphic tablet, using a system for real-time manual intonation modification. Gestural imitation is compared to vocal imitation of the same sentences (seven for a male speaker, seven for a female speaker). Distance measures between gestural copies, vocal imitations, and original sentences are computed for performance assessment. Perceptual testing is also used for assessing the quality of gestural copies. The perceptual difference between natural and stylized contours is measured using a mean opinion score paradigm for 15 subjects. The results indicate that intonation contours can be stylized with accuracy by chironomic imitation. The results of vocal imitation and chironomic imitation are comparable, but subjects show better imitation results in vocal imitation. The best stylized contours using chironomy seems perceptually indistinguishable or almost indistinguishable from natural contours, particularly for female speech. This indicates that chironomic stylization is effective, and that hand movements can be analogous to intonation movements. *© 2011 Acoustical Society of America*. [DOI: 10.1121/1.3531802]

PACS number(s): 43.70.Fq, 43.71.An, 43.72.Ar, 43.66.Hg [MAH]   Pages: 1594–1604

I. INTRODUCTION

A new approach to intonation stylization is studied in the present research, using "chironomy" (coming from the Greek "cheir": hand and "nomos": rule), i.e., the analogy between hand gestures and prosodic movements. The aim is to explore our ability to control and copy speech prosody with the help of hand gestures.

Intonation stylization methods or models[1] often describe melodic patterns in terms of movements, described by "contours"[2,3] or trajectories specified by their "target points."[4,5] Addressing the question of prosodic representation in terms of controlled hand movements would certainly bring new experimental paradigms in prosodic studies. A hand-controlled prosodic modification tool makes an explicit link between the perception of prosody and its control. This may provide new research paradigms for addressing questions linked to the dynamic of pitch contours, e.g., expression of emotion, tonal alignement,[6,7] and intonation stylization. The first step of a research program addressing hand-gesture-controlled intonation concerns the effectiveness of hand-controlled intonation stylization. This paper aims at measuring precisely to what extent hand-controlled intonation contours are able to mimic natural intonation contours.

For implementing a hand-controlled prosody modification tool, one can take advantage of the resources developed in the context of new interfaces for musical expression. Real-time audio programming languages, control devices, and modification algorithms are available in the electronic music community, and can be used for voice processing.[8,9] With the help of this technology, intonation stylization by hand gesture can be effectively implemented and studied. The main question addressed herein is the ability and precision of handwriting movements in copying, or stylizing, intonation contours. An intonation mimicking paradigm seems appropriate for this goal. The task of the subjects is to copy the intonation pattern of a sentence with the help of a stylus on a graphic tablet, using a system for computerized chironomy (i.e., real-time manual intonation control), described in Sec. II. Gestural imitation is compared to vocal imitation of the same sentences. The experiments are described in Sec. III. Distance measures between copies and original sentences are used for performance assessment. A pair discrimination test is conducted in Sec. IV for assessing the quality of gestural copies. Section V summarizes the findings of this research, compares chironomic stylization to other types of stylization proposed in experimental intonation studies, discusses the perceptual and motor processes involved in this task, and proposes extensions and applications of this work.

II. CALLIPHONY: A SYSTEM FOR COMPUTERIZED CHIRONOMY STYLIZATION

A real-time system, called Calliphony, has been designed. This system aims at modifying the F0 of pre-recorded speech utterances with the help of gestural control.

Building on previous work on hand-gesture-controlled speech synthesis,[10] preliminary studies were conducted for selecting the most effective interface for controlling F0 with hand gestures. It appeared that the most accurate pitch control device was a stylus and a graphic tablet. Such a system is

[a] Portions of this work were presented in "Computerized chironomy: evaluation of hand-controlled Intonation reiteration" Proceedings Interspeech '07, August 27–31, 2007, Antwerp, Belgium.

[b] Author to whom correspondence should be addressed. Electronic mail: cda@limsi.fr

accurate enough for musical tasks, and can be effectively used in singing and musical synthesis.[8,11] Hand writing allows for the most accurate and intuitive drawing movement, and can effectively be used in musical F0 control. The F0 control device is based on a WACOM Intuos® (http://www.wacom.com/) graphic tablet. This device is made of a flat tablet and an independent stylus. The (*x*,*y*) position and the pressure of the stylus, when in contact with the tablet, are sent in real-time to the computer. In this version of Calliphony, we use only the *y* axis of the tablet, which is associated with the fundamental frequency on a semi-tone scale. The position of the stylus along the *y* axis drives a high quality real-time F0 modification system. This system is based on a real-time implementation of the pitch-synchronous overlap add [(PSOLA (Ref. 12)] pitch scaling algorithm. It is implemented in C within the Max/MSP programming environment.

Each sentence is processed according to the following steps. First, the sentence is pitch-marked. Second, a flat-pitched version of the sentence is computed off-line with the help of PSOLA. The pitch of the sentence is equalized to a constant value, taken as an average value for the speaker (e.g., 120 Hz for the male speaker). Note that only pitch is modified by this procedure, and not duration: The flattened sentence has the same timing as the original sentence. Finally, the sentence is replayed through the Calliphony system, with pitch values modified in real-time according to the Y position of the stylus on the graphic tablet. The latency of the system is about a pitch period, and goes unnoticed by the player, who is able to actually draw the intonation contour she/he wants to listen to.

As only the *y* axis of the graphic tablet is used for pitch control, similar intonation patterns can be realized using very different hand gestures. It was noted that subjects used a variety of motions: circular, sinusoidal, step-like, vertical gestures, and so on, to complete the task, according to their taste.

An illustration of the chironomic and vocal imitation process is displayed in Fig. 1 for the seven-syllable sentence “Nous voulons manger le soir” (/nuvulɔ̃mɑ̃ʒeləswaʁ/ “we want to eat in the evening”). The top panel shows the natural contour (middle contour), the best chironomic imitation (bottom continuous line which corresponds to the position of the stylus on the *y* axis over time), and the best vocal imitation contour (top line). In this example, a female subject performed imitation of male speech, resulting in an octave error for the vocal imitation, because of the difference in vocal register between the speaker and the impersonator. Vocal imitation and natural contour are time aligned (as described in Sec. III). The process of trial and error used for chironomic stylization is displayed in the middle panel. For seven or eight trials, the Y values of gestural copies over time are plotted against the natural contour. These trials are displayed in the bottom panel, but in the graphic tablet (*x*,*y*) coordinates, to show the actual stylus motions performed by the subject. The X dimension (arbitrary unit A.U. between −1 and 1) does not represent time: moving along X for a fixed Y has no effect. The active dimension, controlling pitch, is Y. The Y dimension represents arbitrary unit, between −1 and 1. $Y=0$ corresponds to no pitch modification (a flat contour at the average pitch of the sentence). More than 0 increases pitch, less than 0 decreases pitch.

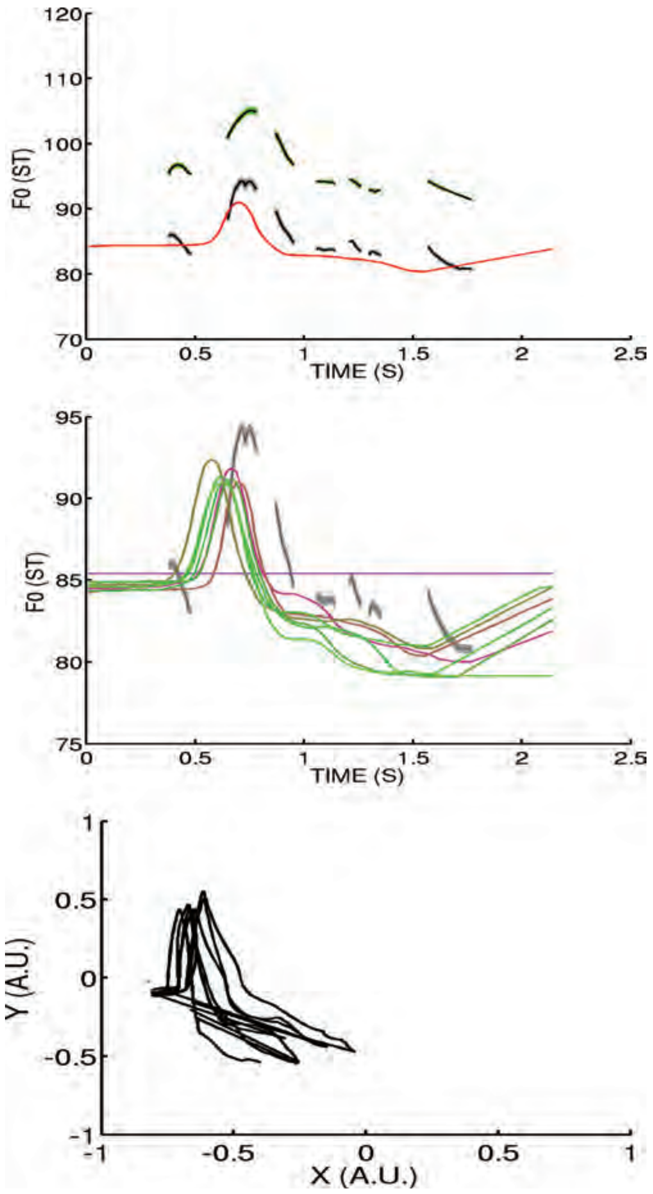

FIG. 1. (Color online) Example of chironomic and vocal imitations. Top panel: Natural contour (middle contour), best chironomic imitation (bottom continuous line, subject SG), the best vocal imitation contour (top line) in semitones and seconds. Middle panel: Examples of trials for chironomic imitation plotted over the natural intonation contour (thick discontinuous gray line). Bottom panel: Same trials in the graphic tablet XY coordinates (A.U.: arbitrary units).

III. AN EXPERIMENT IN CHIRONOMIC STYLIZATION

In this section, an experiment in gestural stylization of intonation is described. Gestural copies of intonation patterns are drawn in real time using computerized chironomy. The Calliphony system is used in an intonation stylization task, where subjects are asked to copy the intonation of target sentences starting from a flat F0 version of the same sentences. A preliminary experiment was presented in Ref. 13, but with a reduced set of subjects, and an evaluation of the gestural imitation of original speech only. This preliminary experiment, using trained subjects and a visual display of prosodic contours, showed the feasibility of prosody imitation by hand gestures. In the present article, a larger set of untrained subjects is used and no F0 display is provided to the subjects. In addition to gestural stylization, and for comparison, the subjects are also asked to produce vocal imitations of the target sentences.

A. Speech material

A dedicated corpus of seven sentences, ranging from two to eight syllables in length was designed. The corpus was built according to two criteria: most of the words have a consonant–vowel syllable structure and voiceless plosive consonants were avoided at the beginning of the sentences. These rules were applied in order to obtain easily comparable prosodic patterns amongst the sentences and to avoid large micro-prosodic effects due to plosive bursts. The seven sentences of the corpus are:

2 syllables: Salut
(hello)
/saly/
3 syllables: Répétons
(let's repeat)
/ʁepetɔ̃/
4 syllables: Marie chantait
(Mary was singing)
/maʁiʃɑ̃tɛ/
5 syllables: Marie s'ennuyait
(Mary was bored)
/maʁisɑ̃nɥijɛ/
6 syllables: Marie chantait souvent
(Mary was often singing)
/maʁiʃɑ̃tɛsuvɑ̃ /
7 syllables: Nous voulons manger le soir
(We want to eat in the evening)
/nuvulɔ̃mɑ̃ʒeləswaʁ/
8 syllables: Sophie mangeait des fruits confits
(Sophie was eating sugar fruits)
/sofimɑ̃ʒɛdefʁɥikɔ̃fi/

Two speakers (a female and a male, native speakers of French) recorded the corpus. They produced each sentence according to two different instructions: (1) emphasis on a specific word of the sentence (generally the verb) and (2) interrogative intonation. The purpose was to obtain varied intonation patterns. The sentences were recorded in a recording booth, and directly digitalized on a computer (44.1 kHz, 16 bits), using an AKG C414B microphone placed at 40 cm from the speaker's mouth. All sentences of the corpus were then analyzed in order to extract their fundamental frequency (in semitones), syllabic durations, and intensity.

B. Subjects and task

Ten subjects participated in the experiment. All subjects used the hand copying system for the first time. They can be considered as naive subjects with respect to hand-controlled intonation research. The subjects range from 24 to 37 yr old (mean = 27.3), none of them having known impairment in either auditory or motor functions. Seven subjects out of the ten have regular musical practice, or have been trained in music practice. Trained musicians have between 1 and 20 yr of training. The subjects (five males and five females) were recruited on a voluntary basis, all are members of the laboratory, and they were not paid for their participation in the experiment. Table I gives the subjects' age, gender, and musical experience.

TABLE I. Musical training for the ten subjects, expressed in years of either vocal or instrumental practice, together with their gender and age.

Subject	Gender	Age (yr)	Music (yr)
SA	M	37	3
SB	F	27	8
SC	M	24	4
SD	M	27	0
SE	M	26	0
SF	F	24	1
SG	F	26	20
SH	M	26	10
SI	F	28	10
SJ	F	28	4

The aim of the experiment is to investigate how close to the original F0 traces the hand-gestural copies can be. A dedicated computer interface has been developed under the Max/MSP platform. For each sentence in the corpus, the subject can listen to the original utterance by clicking on a button with the mouse pointer. The subject's task is to copy the prosody of the original sentence from a pitch-flattened version of the same sentence by drawing the prosodic contour using the stylus on the graphic tablet. A flat intonation contour is produced when the stylus is not in contact with the tablet, or when it stays in a constant Y position. Flattened sentences looped with a 500 ms silent interval. This is illustrated in Fig. 1, bottom panel, where the hand traces on the graphic tablet loop back to the beginning of the next trial.

Subjects are able to listen to the original sentence at any moment and to perform imitations until they are satisfied with their performances. Once satisfied, subjects stop the repetitions of the looping sentences. They can then listen to the modified sentences, in order to check whether it is a satisfactory copy of the original. They can then decide to have other trials or to continue the experiment. After performing a gestural copy of a given sentence, the subjects perform a vocal copy of the same sentence, with similar constraints: They can vocally imitate the original sentence as many times as they need, stop the recording, and listen to their performances, and eventually have other trials or continue the experiment. Unlimited number of copies can be recorded for each sentence and subjects can listen to their performance to judge whether it is a satisfactory copy of the original. After completing a given sentence, subjects go on to the next sentence. Figure 2 shows chironomic trials for four sentences. The number of repetitions varies quite a lot depending on the subject, and for a same subject on the sentence: it is ranging from two or three trials to more than 10 or 15 trials. For vocal copies also, the subjects can record a number of copies and listen to them until satisfaction.

The subjects start with a short familiarization phase performed on six declarative sentences, during which they were given

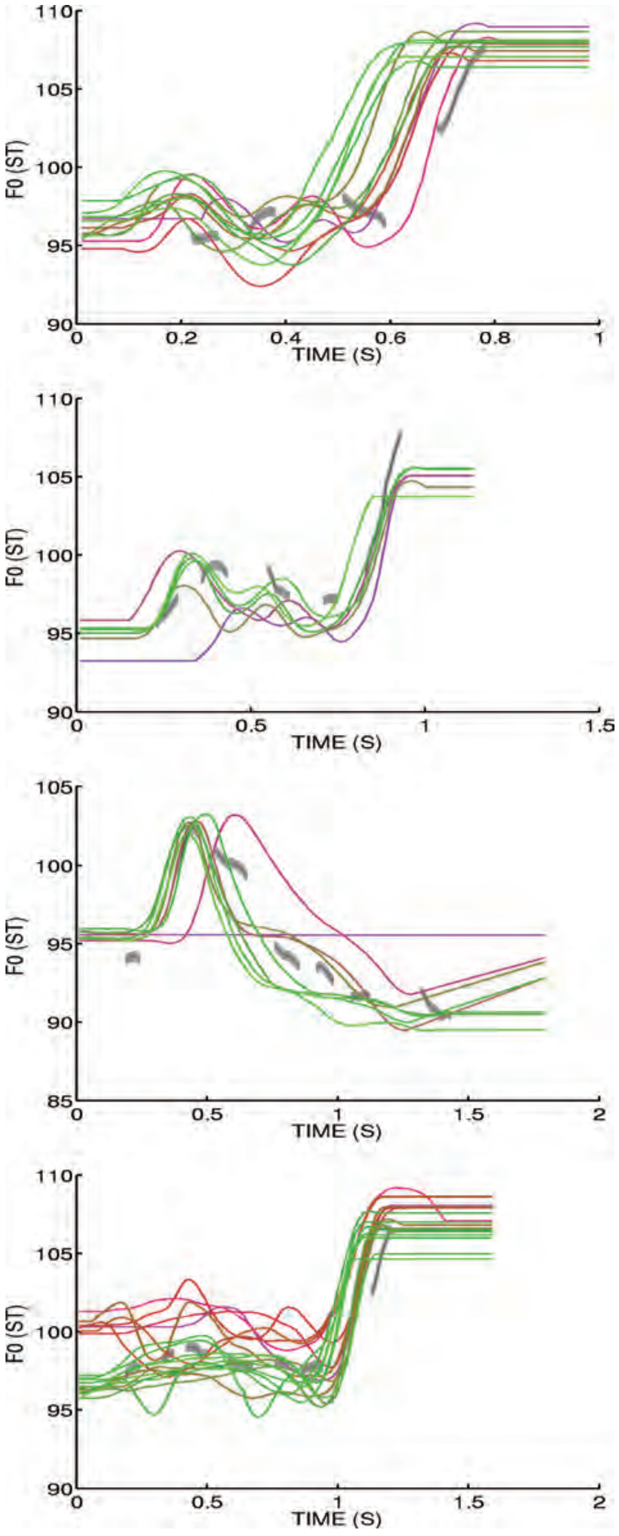

FIG. 2. (Color online) Examples of trials for chironomic imitation (seconds, semitones). The natural contour (thick line) and chironomic contours (thin lines), for four-syllable (top, subject SH), five-syllable (second from top, subject SH), six-syllable (second from bottom, subject SG), and seven-syllable (bottom, subject SH) sentences.

a graphic representation of the pitch contour. After the familiarization phase, no visual information is provided to subjects.

As the test can typically lasts from one to several minutes per sentence, subjects are instructed to take rest from time to time, about every 20 min.

It must be noted that this prosodic imitation task was easily learned by the subjects, and that performing the task does not seem particularly difficult nor tiring. Vocal imitation might even be more arduous than gestural imitation for some subjects. In the situation of a female speaker copying male intonation, the vocal imitation is often one octave apart from the original, as can be seen in Fig. 1.

C. Comparing original sentences and gestural copies

The subjects' performances in copying intonation, either using their own voice or using the gestural modification system, must be assessed by objective, signal-based measures. Two F0 similarity measures introduced in Ref. 14 seemed appropriate in this case. These measures are based on F0 contours, using a weighting factor in order to give more importance to phonemes with a higher sound level.

A measure of similarity between two F0 contours is given by the energy-weighted correlation. Let $f_1(i)$ and $f_2(i)$ represent fundamental frequencies for the utterances to be compared, m_1 and m_2 the average fundamental frequencies for these utterances. $f_1(i)$ and m_1 refer to the original sentence, while $f_2(i)$ and m_2 refer to the gesturally or vocally imitated sentence. These frequencies, expressed in semitones (calculated with 1 Hz as the reference value), measured using STRAIGHT,[15] are sampled every 10 ms and weighted using the signal power $w(i)$, from the original sentence, averaged on the corresponding 10 ms (expressed in dB). The weighted correlation between two F0 contours is defined by

$$r_{f_1 f_2} = \frac{\sum_i w(i)(f_1(i) - m_1)(f_2(i) - m_2)}{\sqrt{\sum_i w(i)(f_1(i) - m_1)^2 \sum_i w(i)(f_2(i) - m_2)^2}}.$$

This equation represents the correlation between the F0 contours, weighted by the time-varying power of the signal. Note that the means m_1 and m_2 are subtracted to F0 contours. Such a correlation normalizes the two curves with respect to their mean, and therefore compensates for a difference of register. Such a behavior is interesting in our case, as differences of register are systematic, e.g., the octave difference when a woman vocally imitates male intonation contours. For further statistical analysis, as the distribution of correlations does not follow a Gaussian distribution, the Fisher's Z transformation is used for obtaining a Gaussian distribution of correlations.[14]

The root-mean-square (RMS) difference between two contours is a measure of their dissimilarity, and behave like a distance measured between the two curves. The RMS difference is given by

$$R = \sqrt{\frac{\sum_i w(i)((f_1(i) - m_1) - (f_2(i) - m_2))^2}{\sum_i w(i)}}.$$

The RMS difference is always positive, with a value of 0 for identical contours. This distance represents the average difference in semitones between two F0 contours. For further statistical analyses, as the distribution of RMS differences obtained does not follow a Gaussian distribution, log(R) is used instead.

To quote Hermes, the correlation is a measure of the similarity of the two sets of F0 parameters, whereas the RMS difference is a dissimilarity measure. Correlation tests the similitude between the shapes of the two curves. On the contrary, the RMS distance will give an idea of the area between the two curves.

These two measures were automatically calculated for all gestural copies. Comparison of natural sentences and gestural copies is fully automated, because the pairs of signals are perfectly aligned: only F0 differs. Only the closest gestural copy was stored for further analysis. The best imitation is chosen first according to the weighted correlation. For stimuli with similar correlations, the weighted RMS difference is also taken into account.

D. Comparing original sentences and vocal copies

The measures described above deal with segments of the same length, a condition not met for vocal imitations. Vocal imitation of intonation has been extensively studied using the so-called reiterant speech paradigm. In this situation also, differences in timing of natural speech and vocal imitation are observed.[16]

Comparing natural sentences and vocal copies requires one more step, because the subjects were not able to copy the exact timing of the original sentences. Then, natural and copied sentences are differing both in timing and intonation. Only differences in intonation are investigated in the present study and then timing differences have to be neutralized. The mean difference in syllabic duration between original sentences and vocal imitations is 32 ms (median 24 ms) for all subjects (mean and median of the absolute value of the syllabic length differences for all imitations by all subjects). The mean difference in sentence duration between original sentences and vocal imitations is 98 ms (median 85 ms).

Natural sentences and vocal copies have exactly the same phonetic content. They can be easily aligned using a timing compensation procedure like dynamic time warping. When the two signals are aligned, the above mentioned distance measures can be applied in the same manner as for the natural sentences and gestural copies. Figure 3 displays four examples of natural contour, aligned vocal imitation and chironomic imitation, for five-, six-, seven-, and eight-syllable length sentences.

E. Results

Median values for correlation and RMS difference for the ten subjects are reported in Fig. 4, using box plots [showing the median, first, and third quartile (25%–75%, i.e., 50% of the data) and first and ninth decile (10%–90%, i.e., 80% of the data)].

The correlations are high, generally over 0.9 for vocal copies, and over 0.8 for gestural copies. The average RMS distances are between 1 and 2.5 semitones, a rather small difference for speech F0. Again, vocal copies are closer to the original intonation contours. Some subjects (SE, SH, SG, SI) performed with a median correlation above 0.9 for gestural imitation. In this experiment, many subjects performed poorer than in Ref. 13 where trained subjects performed a

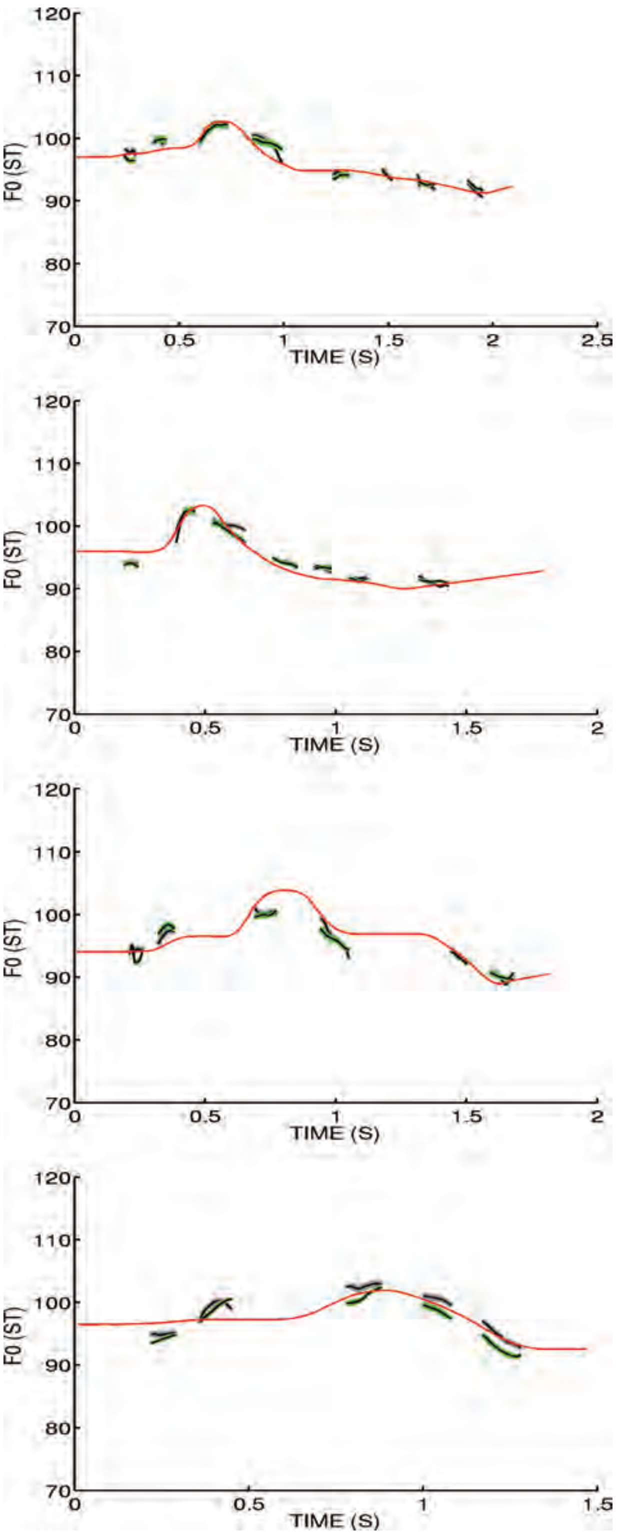

FIG. 3. (Color online) Examples of natural contour (light thick line), best chironomic imitation (continuous line), and best vocal imitation (dark thick line) (seconds, semitones, all for subject S), for eight-syllable (top), seven-syllable (second from top), six-syllable (third from top), and five-syllable (bottom) sentences.

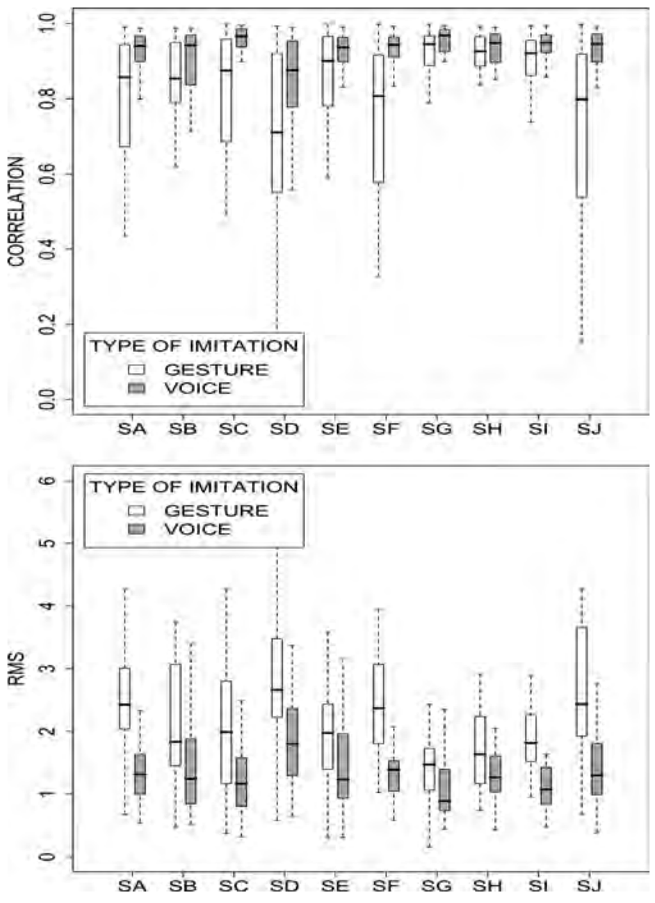

FIG. 4. Imitation experiment: Correlation (top panel) and RMS distance (bottom panel) for vocal and gestural imitations as a function of the SUBJECT factor. Median, first, and third quartile, first and ninth decile.

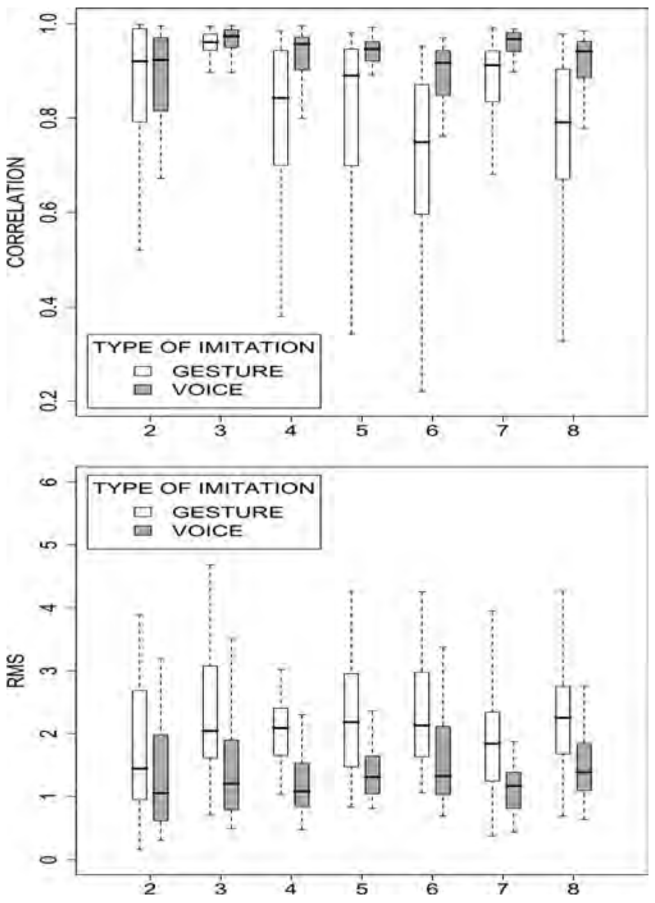

FIG. 5. Imitation experiment: Correlation (top panel) and RMS distance (bottom panel) for vocal and gestural imitations as a function of the SENTENCE factor (number of syllables). Median, first, and third quartile, first and ninth decile.

similar experiment. The difference between the two experiments could be explained by two main factors. On the one hand, for most subjects, this experiment was their first use of a graphic tablet; therefore, they had no implicit training. On the other hand, no visual description of the intonation was provided. It seems that a visual presentation of the pitch contour can help in drawing an equivalent contour on the tablet. However, the best subjects in the present experiment performed as well or even better than the subjects in Ref. 13.

Two analyses of variance (ANOVAs) were carried out on these ratings, with a significance level set at 0.01. Three fixed factors were used: the type of imitation (TI, two levels: gesture and voice), the SPEAKER (two levels: female and male), and the STIMULI (seven levels); and the random factor was the SUBJECT (ten levels) who participated in the experiment.

The first ANOVA is based on the Fisher-transformed correlation. The three factors have significant effect on the results (TI: $F_{1,526} = 74.88$, $p < 0.001$; SPEAKER: $F_{1,526} = 91.92$, $p < 0.001$; STIMULI: $F_{6,526} = 25.96$, $p < 0.001$). All interactions except the interaction between TI and SPEAKER are significant. However, the factors that have the main effect size, according to the partial η^2, in decreasing order are the STIMULI, the interaction between STIMULI and SPEAKER, the SPEAKER and the TI.

The coherence of the ten subjects during their performance is evaluated using Cronbach's α, showing a high coherence ($\alpha = 0.814$).

The second ANOVA is based on the log-transformed RMS distance. Here also, the three factors have significant effect on the results (TI: $F_{1,526} = 125.63$, $p < 0.001$; SPEAKER: $F_{1,526} = 125.63$, $p < 0.001$; STIMULI: $F_{6,526} = 25.96$, $p < 0.001$). But no interaction has any significant effect.

The coherence of the ten subjects for the RMS measurement is also evaluated using Cronbach's α, showing a high coherence ($\alpha = 0.931$).

The results for the different stimuli are reported in Fig. 5. There are significant differences between stimuli. However, these differences seem linked to the peculiarities of the sentences, because no simple systematic explanation (e.g., sentence length) has been found.

The female voice seemed easier to copy compared to the male voice (the results are reported in Fig. 6). The interaction between STIMULI and SPEAKER (significant for correlations, see Fig. 7) is mainly due to the two-syllable long stimuli: The shortest stimulus of the male speaker receives clearly lower scores than the corresponding female one. Again, it is difficult to explain why. Visual inspection of pitch contours showed that they are very comparable, except for the difference in register. A more systematic study of differences between high-pitched and low-pitched

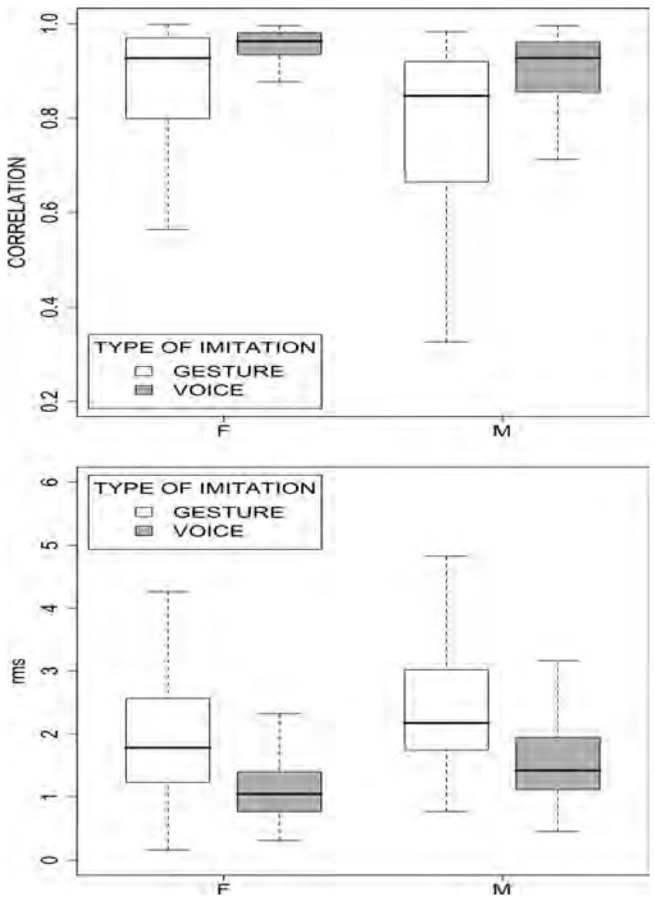

FIG. 6. Imitation experiment: Correlation (top panel) and RMS distance (bottom panel) for vocal and gestural imitations as a function of the SPEAKER factor. Median, first, and third quartile, first and ninth decile.

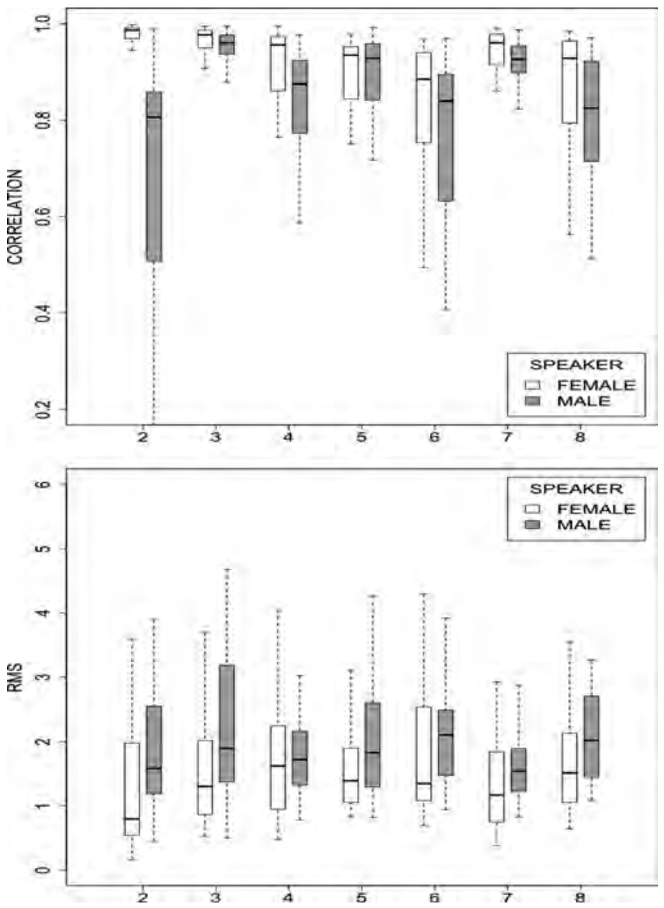

FIG. 7. Imitation experiment: Correlation (top panel) and RMS distance (bottom panel) for gestural imitations as a function of the STIMULI and SPEAKER factors. Median, first, and third quartile, first and ninth decile.

sentences would be needed to better understand this effect, but it is out of the scope of the present study.

Vocal imitations receive statistically significant higher scores than gestural ones. Some subjects performed very well in gestural copies, but other performed poorer. This indicates that gestural imitation may require some training for some subjects: remind that the experiment was conducted with a limited training phase and that some subjects had never used this type of interface prior to the test.

Comparing Fig. 4 and Table I, it appears that three (SG, SH, SI) out of the four (SE, SG, SH, SI) best subjects are the most trained musicians, while the fourth has no musical training at all. However, this subject reported significant experience in playing with computer interfaces. This could indicate that instrumental and musical training (musical instruments or interfaces) can improve ability in chironomic prosodic control.

IV. PERCEPTUAL ASSESSMENT OF GESTURAL COPIES

A. Subjects and tasks

In addition to signal-based measures, a perceptual experiment was designed for assessing the perceptual proximity between original sentences and gestural copies. The question addressed is the perceptual similarity of gestural copies and natural sentences. As the results showed a high correlation between gestural copies and natural sentences, a mean opinion score (MOS) 5-point scale paradigm seemed appropriate. Each stimulus consisted of a pair of two sentences, made of two versions of the same utterance, with the same content and timing, separated by a 100 ms silent interval. The task of the subjects was to discriminate within a sentence pair on the basis of prosodic (intonation) similarity. A pair is composed of (A) the original stimulus and (B) either the same natural stimulus or a chironomic imitation of this sentence. Similarity ratings are made on a 1–5 MOS scale: 5: identical stimuli; 4: almost identical stimuli; 3: different but similar stimuli; 2: different stimuli; 1: very different stimuli.

The sentences were chosen among the available gestural copies. Copies with various distances from the natural sentences were chosen, in order to introduce some clearly different stimuli among the set. This helps prevent a bias in testing: If all stimuli were very similar, subjects could be inclined to answer randomly, just to avoid having to answer too often "identical." Stimuli ranging from two to eight syllables in length were selected among the copies obtained in the experiment with the following criteria for each sentence. Although the A stimulus is the natural sentence (labeled Nat), the B stimulus is selected among the following four categories: (1) the natural sentence (identical pairs); (2) the two stimuli selected as the two highest ranked stimuli (according to their correlation with the original stimulus; r generally > 0.9) (stimuli labeled SynA and SynB); (3) stimulus (labeled SynC) selected among the medium range of correlation (0.5–0.7), although differences are noticeable, the copy is still close to the original; (4) a stimulus (labeled SynD) selected as a rather poor copy of the original (either a flat sentence or a clearly failed copy—correlations ranging from about 0 to -0.8). This method results in five different possibilities. Similar sets of stimuli were selected from both speakers (male and female). Then 140 pairs of stimuli (seven sentences, AB and BA presentation order, five pairs per sentence, two speakers) were presented to the subjects using the usual randomization procedure. Stimuli were presented monaurally over the headphones in a quiet room, and each test session lasted about 20 min.

A total number of 15 subjects (three females, 12 males, mean age = 31), members of the laboratory, native speakers of French, and without known hearing loss, participated in this experiment on a voluntary basis. A first set of ten training trials was used to familiarize the subject to the task.

B. Results

An ANOVA was carried out on these ratings, with the significance level set at 0.01. The three fixed factors were the length of the stimuli (LENGTH, seven levels: from two to eight syllables), the five stimuli (STIM, five levels: natural, synthetic A to D), and the speaker (SPK, two levels: female and male). The ANOVA used the similarity level as the dependent variable. All factors have a significant effect on the results: The LENGTH has a significant effect ($F_{6,1890} = 6.897$, $p < 0.001$), which is mainly linked to the five-syllable long sentences, and that may be interpreted more as an effect of stimuli than an effect of length. Scores decrease significantly with the STIM from natural to the worst imitation ($F_{4,1890} = 415.055$, $p < 0.001$). All interactions (LENGTH * STIM; LENGTH * SPK; STIM * SPK; LENGTH * STIM * SPK) are also highly significant ($F_{24,1890} = 8.26$, $p < 0.001$; $F_{6,1890} = 16.463$, $p < 0.001$; $F_{2,1890} = 12.323$, $p < 0.001$; $F_{24,1890} = 7.426$, $p < 0.001$).

Though the effect size of the three factors are clearly different (for LENGTH: $\eta^2 = 0.009$; for STIM: $\eta^2 = 0.381$; for SPK: $\eta^2 = 0.055$), indicating a greater influence of the quality of imitation on the results (either an identical stimuli, a good imitation or a bad one) than of the speaker producing the original stimuli itself greater than the stimuli's length.

Subsequent *post hoc* comparison (Tukey's HSD test), with an α level of 0.01, for both STIM and LENGTH factors and for the interaction showed the following results (displayed in Fig. 8). For STIM, four groups emerge: (1) the natural stimuli, which receive a mean similarity score of 4.93; (2) the two best imitations, which receive similar mean scores between 3.54 and 3.42; (3) the average imitation, which is ranked around 3.06; and (4) the worst imitation, which receives a mean score of 2.39. Remember that stimuli were chosen according to their distances from the original: This distance seems well reflected in the MOS reported by the subjects in the perceptual experiment. For the LENGTH factor, the five-syllable length stimuli receive higher scores similar to six and eight-syllable long sentences, but significantly higher than others. All other lengths received comparable scores. This effect is difficult to interpret, and seems only linked to peculiarities of the F0 contours studied rather than to a systematic effect. The stimuli from both speakers received different mean ratings, stimuli produced by the female speaker being significantly better ranked than the stimuli from the male speaker.

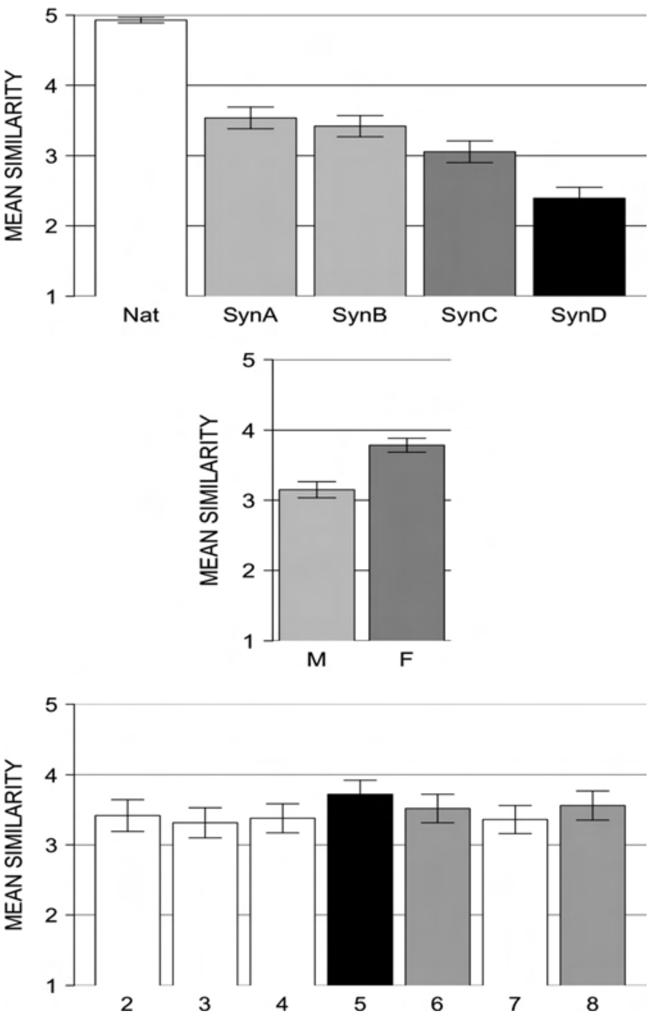

FIG. 8. Perceptual experiment. Mean similarity ratings for the five types of STIM (top panel), male vs female SPK (middle panel), and each levels of the LENGTH factor (bottom panel).

Analysis of the interactions was done separately for the stimuli of the female and the male speakers. *Post hoc* analysis of the interaction between STIM and LENGTH, for female stimuli, shows that a homogeneous subset of stimuli regroups all original stimuli for all length, together with 12 synthetic imitations (including a supposedly bad imitation) that did not differ significantly from the natural sentence they try to imitate. The stimuli forming this subset are indicated by an asterisk symbol in Fig. 9. The results show that for the female speaker, for six sentences over seven, the best gestural copies are not significantly different from the natural stimuli.

Note that the similarity scores of SynC for two-syllable stimuli and SynD for three-syllable stimuli are higher than their SynA and SynB counterparts, which is counterintuitive. These differences are not statistically significant. They can

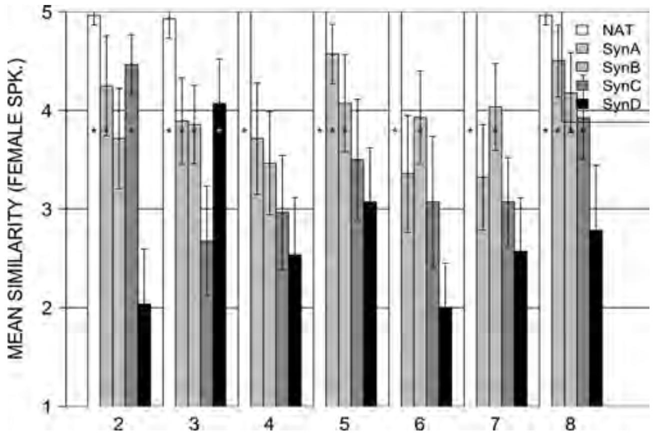

FIG. 9. Perceptual experiment, female speaker. Mean similarity scores for the interaction STIM * LENGTH. Homogeneous subset comprising all natural stimuli and comparable synthetic imitation are marked with an asterisk.

result from audible artifacts in the analysis-synthesis process, which are more often encountered for female voices. Nevertheless, these two outliers do not influence much the overall trend observed in Fig. 8 (top panel) where a significant decrease is observed between conditions NAT, SynA and SynB, SynC and SynD stimuli. According to this perception test, for most of the best gestural copies are rated between "identical stimuli" and "almost identical stimuli."

Post hoc analysis of the interaction between STIM and LENGTH, for male stimuli, shows that a homogeneous subset of stimuli regroups all original stimuli for all length, together with only one synthetic imitation that did not differ significantly from the natural sentence they try to imitate. The stimuli forming this subset are indicated by an asterisk symbol in Fig. 10. The results show that for the male speaker, for only one sentence over seven, the best gestural copies are not significantly different from the natural stimuli. According to this perception test, most of the best gestural copies are rated between "almost identical stimuli" and "different but similar stimuli."

This difference between female and male results seems to reflect the differences in correlation and RMS distance observed in the stylization experiment analysis. It is difficult to interpret on the basis of intonation contours, which seems rather similar by visual inspection. Copying the male speech intonation contours seemed more difficult for the subjects, and these contours seemed also perceptually farther from the original than for the female contours.

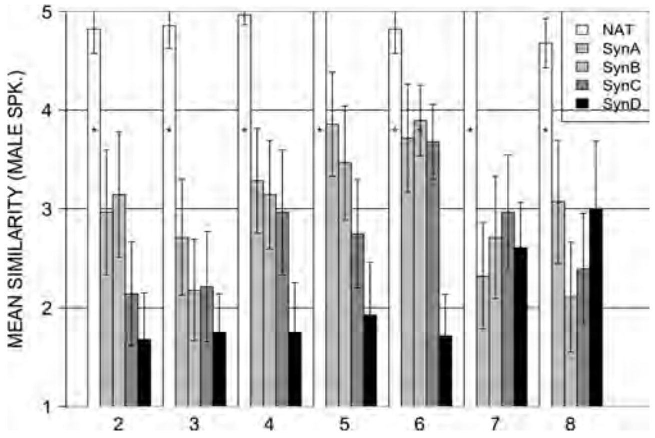

FIG. 10. Perceptual experiment, male speaker. Mean similarity scores for the interaction STIM * LENGTH. Homogeneous subset comprising all natural stimuli and comparable synthetic imitation are marked with an asterisk.

In a preliminary perceptual experiment (not reported in details in this article for saving space), the stylized data obtained in Ref. 13 have been used. These data are obtained for the male speaker only, by trained subjects (three were the authors of this research), and a visual display of intonation during stylization. They are generally closer to the original than the data reported in Sec. III. A total number of 15 subjects (4 females, 11 males, mean age = 28.5) participated in the preliminary perceptual experiment. For the male speaker, a better perceptual evaluation is obtained: for six sentences over seven, the best gestural copies are not significantly different from the natural stimuli. Then it seems that training can improve chironomic stylization.

V. DISCUSSION

A. Summary

The results of the chironomic stylization experiment indicate that intonation contours can be stylized with accuracy by hand-drawing gestures. The subjects were offered no special training. The task does not seem particularly difficult, at least compared to other intonation recognition tasks, e.g., musical dictation. The results of vocal imitation and chironomic imitation are very comparable. However, the subjects show better results in vocal imitation than in gestural imitation, at least when mean pitch differences (e.g., octave differences) are compensated for.

Perceptual evaluation shows that stylized contours are, for the best of them, indistinguishable from natural contours. For most of them they are almost indistinguishable from or similar to natural contours. This indicates that chironomic movements can somehow be analogous to intonation movements.

B. Intonation stylization and chironomic stylization

There is a long tradition of intonation stylization in prosodic studies. Straight-line stylization gives the so-called close-copy stylization,[3] that are perceptually similar to natural contours (but often distinguishable). Stylization based on target points and more elaborated interpolation procedure,[4] can under certain conditions, produce almost indistinguishable stylized contours. Automatic stylization based on syllabic decomposition and a model of tonal perception (short-term integration of F0 variations) can also produce perceptually indistinguishable intonation contours.[2]

Chironomy brings interesting information to the question of intonation stylization. Like for other types of stylization, it seems that micro-prosodic variations are integrated in the process of stylization. In the case of gestural imitation, writing uses generally slower gestures than speaking, and then hand gestures are not able to follow fine grained F0 details like micro-prosody. Note that micro-prosody is not under conscious control for speakers. Therefore, it seems that hand gestures correspond to prosodic intonation movements rather than to details of fundamental frequency contours.

Stylization procedures are generally based on detection of target points and some sort of interpolation between these points. They are basically local procedures taking into account a narrow time span. On the contrary, chironomic stylization is based on the memory trace of intonation movement, and its reproduction by gestural planning. It is rather a planned motor action based on the kinematics of intonation contours. This paves the way for kinematic and dynamic studies of prosodic movements.

Note that in the present experiment, the subjects received no special training prior to performing stylization, nor were they familiar with intonation research. Chironomic stylization can certainly benefit of more training, as reported in a preliminary study.[13]

C. Process of chironomic stylization

Copying intonation with the help of hand gestures is a complex task, involving several modalities and a mixture of motor and perceptual functions. It is an asynchronous task, with the following sequential steps:

(1) Listening to the original acoustic stimuli, focusing on the intonation contour.
(2) Memorization of the intonation contour. The acoustico–phonetic trace of the intonation contour or its motor equivalent is stored.
(3) Planning and realization of the equivalent hand gesture, or vocal gesture, using the short-term memory trace of the intonation contour.
(4) Comparison of the utterance produced with the original contour, or its memory trace.
(5) The process is repeated until satisfaction of the subject.

The first step is limited by thresholds for pitch and differential pitch perception,[3] and depends on some F0 time integration.[17–19] These thresholds result in smoothed pitch contours. It is probably at this stage that micro-prosody is lost.

One can hypothesize that the resulting memory trace of these smooth pitch contours can be represented by perceptual target pitches anchored at specific time points or by perceptual representation of these trajectories.

Planning and realization of the equivalent hand gesture involves motor action, similar to writing. The kinematics of chironomic contours is constrained by the law of gestures for hand writing 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 subjects used rather circular movements, other rather linear movements. The specific shape of the hand drawing seems not very important, as long as the pitch target points are reached with correct timing. This indicates the importance of the kinematics of the intonation contour.

Finally, the decision process, for accepting a contour or not, is also based on the memory of the original contour or on a mental representation of this contour. This second call to memory is likely to introduce again some sort of loss in stylization accuracy.

A remarkable and somewhat striking result is that the performance levels reached by hand written and vocal intonation imitation are comparable (although vocal imitation appeared better). This could suggest that intonation, both on its perceptual and motor production aspects, is represented and embodied at a relatively deep cognitive level, as it seems somehow independent of the modality actually used to reproduce it. The present work addressed gestural intonation stylization at a phonetic, language independent level. The linguistic relevance of gesturally stylized intonation contours is certainly worth studying.

D. Conclusion

Chironomy provides a promising analogy between intonation contours and manual movements. The results obtained indicate that stylized contours can in some situations be perceptually equivalent to natural contours. Applications and implications of these findings are manifold. Chironomic control can be applied to expressive speech synthesis, for instance for corpora enrichment in concatenative speech synthesis.[20] Chironomic control can also be effective in the context of real-time singing synthesis. Analyses of the traces produced during intonation stylization can be used for expressive speech analysis.

Chironomic stylization brings a new experimental paradigm for the question of intonation modeling in terms of movements. Not only the shape, direction, and size of intonation movements, but also their kinematics, i.e., their links with rhythm and their development in time, can be studied within this paradigm. Then intonation and rhythm can be dealt within a unified framework for expressive gesture representation, using common features like velocity, target position, and rhythmic patterns. In addition to kinematics, the question of prosodic dynamics (recently addressed, e.g., in Refs. 21 and 22) and intonation planning could also benefit from chironomic experiments.

[1] D. J. Hermes, "Stylization of pitch contours," in *Methods in Empirical Prosody Research*, edited by S. Sudhoff, D. Lenertov, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schlieer (Walter de Gruyter, Berlin, New York, 2006), pp. 29–61.

[2] C. d'Alessandro and P. Mertens, "Automatic pitch contour stylization using a model of tonal perception," Comput. Speech Lang. **9**, 257–288 (1995).

[3] J. T. Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation* (Cambridge University Press, UK, 1990), pp. 10–66.

[4] D. J. Hirst, P. Nicolas, and R. Espesser, "Coding the F0 of a continuous textin French: An experimental approach," in *Proceedings of the International Congress of Phonetic Sciences*, Aix en Provence, France (1991), pp. 234–237.

[5] D. J. Hirst, "Form and function in the representation of speech prosody," Speech Commun. **46**, 334–347 (2005).

[6] K. Kohler, "Timing and communicative functions of pitch contours," Phonetica **62**, 88–105 (2005).

[7] Y. Xu, "Timing and coordination in tone and intonation—An articulatory-functional perspective," Lingua **119**, 906–927 (2009).

[8] N. d'Alessandro, C. d'Alessandro, S. Le Beux, and B. Doval, "Real-time CALM synthesizer: New approaches in hands-controlled voice synthesis," in *Proceedings of International Conference on New Interfaces for Musical Expression*, Paris, France (2006), pp. 266–271.

[9] N. d'Alessandro, B. Doval, T. Dutoit, C. d'Alessandro, Y. Favre, and S. Le Beux, "Realtime and accurate musical control of expression in singing synthesis," J. Multimodal User Interfaces **1**, 31–39 (2007).

[10] S. S. Fels and G. Hinton, "GloveTalk: A neural network interface between a DataGlove and a speech synthesizer," IEEE Trans. Neural Networks **4**, 2–8 (1993).

[11] M. Zbyszynski, M. Wright, A. Momeni, and D. Cullen, “Ten years of tablet musical interfaces at CNMAT,” in *Proceedings of International Conference on New Interfaces for Musical Expression*, New York (2007), pp. 100–105.

[12] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” Speech Commun. **9**, 453–467 (1990).

[13] C. d’Alessandro, A. Rilliard, and S. Le Beux, “Computerized chironomy: Evaluation of hand-controlled Intonation reiteration,” in *Proceedings of Interspeech 2007*, Antwerpen, Belgium (2007), pp. 1270–1273.

[14] D. J. Hermes, “Measuring the perceptual similarity of pitch contours,” J. Speech Lang. Hear. Res. **41**, 73–82 (1998).

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Commun. **27**, 187–207 (1999).

[16] L. S. Larkey, “Reiterant speech: An acoustic and perceptual validation,” J. Acoust. Soc. Am. **73**, 1337–1345 (1983).

[17] C. d’Alessandro, S. Rosset, and J. P. Rossi, “The pitch of short-duration fundamental frequency glissandos,” J. Acoust. Soc. Am. **104**, 2339–2348 (1998).

[18] M. Rossi, “Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole or (the glissando threshold, or perceptual threshold for tonal variations for speech sounds),” Phonetica **23**, 1–33 (1971).

[19] M. Rossi, “La perception des glissando descendants dans les contours prosodiques or (perception of falling pitch glissando in prosodic contours),” Phonetica **35**, 11–40 (1978).

[20] S. Le Beux, A. Rilliard, and C. d’Alessandro, “Calliphony: A real-time intonation controller for expressive speech synthesis,” in *Proceedings of the 6th ISCA Speech Synthesis Workshop*, Bonn, Germany (2007), pp. 345–350.

[21] D. Byrd and E. Saltzman, “The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening,” J. Phonetics 31, **2**, 149–180 (2003).

[22] D. Byrd, J. Krivokapic, and S. Lee, “How far, how long: On the temporal scope of prosodic boundary effects,” J. Acoust. Soc. Am. **120**, 1589–1599 (2006).