Drawing melodies: Evaluation of chironomic singing synthesis

Christophe d'Alessandro,[a] Lionel Feugère, Sylvain Le Beux, Olivier Perrotin,
and Albert Rilliard

*Laboratoire de Mécanique et d'Informatique pour les Sciences de l'Ingénieur, LIMSI - CNRS,
Université Paris Sud, 91405 Orsay, France*

(Received 11 October 2013; revised 25 April 2014; accepted 29 April 2014)

Cantor Digitalis, a real-time formant synthesizer controlled by a graphic tablet and a stylus, is used for assessment of melodic precision and accuracy in singing synthesis. Melodic accuracy and precision are measured in three experiments for groups of 20 and 28 subjects. The task of the subjects is to sing musical intervals and short melodies, at various tempi, using chironomy (hand-controlled singing), mute chironomy (without audio feedback), and their own voices. The results show the high accuracy and precision obtained by all the subjects for chironomic control of singing synthesis. Some subjects performed significantly better in chironomic singing compared to natural singing, although other subjects showed comparable proficiency. For the chironomic condition, mean note accuracy is less than 12 cents and mean interval accuracy is less than 25 cents for all the subjects. Comparing chironomy and mute chironomy shows that the skills used for writing and drawing are used for chironomic singing, but that the audio feedback helps in interval accuracy. Analysis of blind chironomy (without visual reference) indicates that a visual feedback helps greatly in both note and interval accuracy and precision. This study demonstrates the capabilities of chironomy as a precise and accurate mean for controlling singing synthesis.

[http://dx.doi.org/10.1121/1.4875718]

PACS number(s): 43.75.Rs, 43.75.St, 43.75.Tv, 43.66.Hg [SAF] Pages: 3601–3612

I. INTRODUCTION

A. Aims of this research

Singing synthesis is an important domain in the field of musical acoustics. Chironomic singing instruments, i.e., real-time hand-controlled singing synthesizers, were recently developed and demonstrated by different research groups.[1–7]

It seemed appropriate to call this approach "chironomy," a Greek word employed since antiquity with the meaning "ruling (music or speech) with hand (motion)." Chironomic intonation stylization has recently been studied in the context of speech intonation.[8] Using an intonation imitation paradigm, the subjects' task was to copy the intonation of spoken sentences with the help of a stylus on a graphic tablet, using a system for real-time gesture-controlled intonation modification,[9] and with their own voices. The results obtained, in terms of distance measures and perceptual testing, were comparable for vocal imitation and chironomic imitation. Moreover, the best stylized contours using chironomy were not perceptually distinguished from natural contours. This demonstrated the effectiveness of chironomy for the control of synthetic speech intonation, and raised the question of its quality for a musical application.

Chironomic control in musical practice allows for a very subtle and expressive control (e.g., vibrato, glissando, and other continuous pitch inflections). Recent studies proved that singing proficiency appears to be relatively widespread.[10] Occasional singers appear to sing generally in tune, in spite of a lack of confidence among most of them. The main question addressed in this research is the assessment of chironomic singing proficiency.

Accuracy and precision are measures giving complementary views of melodic performance. They were used for assessment of the singing ability in a relatively large population.[11] Their meanings are the following. Accuracy is measuring a bias: The difference between the mean pitch realized for a series of trials, compared to the actual pitch target. Precision is measuring dispersion: The standard deviation within the series of trials. Accuracy and precision are measured in cents, perfect accuracy and precision corresponding to a bias of 0 cents and a dispersion of 0 cents. Intonation control can be considered satisfactory if it is both accurate (a mean close to the target) and precise (with a small standard deviation).

Chironomic control relies greatly on the motor ability acquired when learning hand writing. In a way, chironomic control takes advantage, for the task of intonation control, of the visuo-motor skill already acquired for another task. Chironomy can be considered as an audio-motor skill with a visual and a haptic feedback in addition to audio feedback, while singing only has haptic (proprioceptive) feedback (from the phonation system) in addition to audio feedback.

The aim of this research is to study pitch control performances of a group of subjects when using chironomic singing synthesis. This question is important, because it will assess the expected quality of chironomic control for singing synthesis. This article reports on formal evaluation of the subjects' ability in terms of intonation precision and accuracy. Comparison of chironomic singing synthesis and natural singing are also discussed. For this purpose, simple musical intervals and melodies are sung by a group of

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

subjects with the help of their voice or of a chironomic singing synthesizer and the results obtained are analyzed.

The main goals of this research are: (1) To measure the level of precision and accuracy reached by a group of subjects; (2) compare chironomic singing and natural singing abilities; (3) study the role of audio, visual, and proprioceptive modalities in chironomic singing.

After a short review of singing synthesis, the chironomic singing synthesizer used for the experiments is presented at the end of this section. Experiments are described in Sec. II. The results of precision and accuracy measurements are reported and analyzed in Sec. III. Section IV summarizes and discusses our findings in chironomic singing synthesis.

B. Singing synthesis and chironomic singing synthesis

Singing synthesis is as old as speech synthesis, but sound quality of the first trials was insufficient for professional musical use. Musical application of singing synthesis in contemporary music appeared in the 1980s, following the Chant program.[12] "Chant" was a rule-based formant synthesizer, as were the other systems at this time.[13–15]

Following the general evolution of speech synthesis, the next generation of singing synthesis systems were based on real voice samples that are concatenated and modified for producing the desired singing voice utterance. High quality voice can be produced in this way, with the expense of a large amount of post-production studio work. An example of successful application of such studio concatenation/modification singing synthesis is the movie Farinelli,[16] in which a virtual Castrato voice is synthesized by mixing a male and a female voice. But the main recent success in singing synthesis is the Vocaloid phenomenon.[17] This concatenative singing synthesis software, designed like a personal studio environment, reached an incredible popular success in the Japanese pop culture.

All the preceding approaches are studio-based, off-line synthesis systems. Another paradigm has been proposed, following the development of live-electronic music. Singing synthesis is considered an instrument, i.e., a real time gesture-controlled synthesis device.[18] In this case, the system encompasses two main components: A synthesis engine for parametric sound production, and a real-time human-machine interface for parameter control.[19] This "performative" singing synthesis paradigm can be traced back to the famous Kempelen's mechanical speech instrument.[20] An electrical speech synthesis machine, the Voder,[21] controlling speech using keyboards and pedals followed the same path. More recently speech synthesis and gestural control met in the glove-talk[22] system, a pioneering work using two data gloves and a foot pedal for controlling a formant synthesizer.

The Theremin is one of the earliest and most successful musical instruments based on free hand motion control. It allows for vocal-like intonation variations, like vibrato, glissando, portamento. In this way, it is somewhat similar to bowed string instruments, the Ondes Martenot, or the musical saw. However, mastering this instrument is very demanding, and it suffers many limitations (such as narrow frequency scale span, absence of tactile or visual cues for tones, limited amplitude control, poor sound). For similar reasons, gloves, pedals, wheels, and similar controllers proved not precise enough as singing synthesis interfaces.[23] Keyboards are precise, but discrete by nature, and then they do not allow continuous intonation control, i.e., expressive variation.

Following the success of the MAX and Pure Data programming environments,[24,25] singing synthesis control with the help of a graphic tablet has been introduced by Wanderley *et al.*[6] and Kessous *et al.*[3] This approach of performative singing synthesis proved successful for live music production, and has been demonstrated in several events by different groups.[1–7] The system used in the present research is the Cantor Digitalis.

C. The "Cantor Digitalis" chironomic singing synthesizer

The Cantor Digitalis is a singing synthesis system made of a digital formant synthesizer driven by one or several control interfaces. The synthesis engine is based on an improved version of the parallel formant synthesis design implementing the linear acoustic source-filter model of speech production.[26] The "filter" or "vocal tract" part of the system is computed using a parallel structure made of five digital second-order resonators. The parameters of these filters are their center frequencies, gains, and bandwidths. They are combined for controlling vowels, according to synthesis rules. The "source" or "glottal flow derivative" part of the system is computed using the Causal-Anticausal Linear Model (CALM).[2,27] The CALM parameters are combined for controlling four vocal dimensions: Voice tension, breathiness, roughness, and vocal effort. Compared to a classical parallel formant synthesizer, Cantor Digitalis is featuring several improvements, including presets for voice categories (baritone, tenor, alto, soprano), voice range profiles, source-filter interactions (formant and harmonic tuning), high $F0$ resolution, vocal tract size, vocalic space, and voice quality control.

The synthesis engine is operated using a graphic tablet and a stylus (Wacom Intuos©). The data packets sent by the tablet are sampled at a rate of 200 Hz (a time resolution of 5 ms). Visual pitch references are printed on a transparent plastic sheet and attached onto the active surface of the tablet (see Fig. 1). Equally spaced vertical lines indicate the position of each semitone (ST) (with an A4 = 440 Hz, equally tempered scale). The note names are indicated, but there are no "keys" on the tablet: $F0$ variation is continuous. The target for a given $F0$ is a very thin line: Every bit of stylus deviation will make a deviation in pitch. The linear mapping of pitch is 1.4 cm per ST to provide a range of 20 STs, with a full octave centered on the tablet. The lowest printed note name is a C3 = 125 Hz [48 in the Musical Interface Digital Instrument (MIDI) notation] for male voices, and a C4 = 250 Hz (60 in MIDI notation) for female voices. The tablet offers 5080 lines per inch resolution, corresponding to 0.004 cents pitch resolution, to be compared to the difference limens for pitch perception (about 4 cents[28]). The tablet resolution is 2 orders of magnitude better than perceptually needed. For stylus pressure a range of 1024 levels is detected, which is much greater than

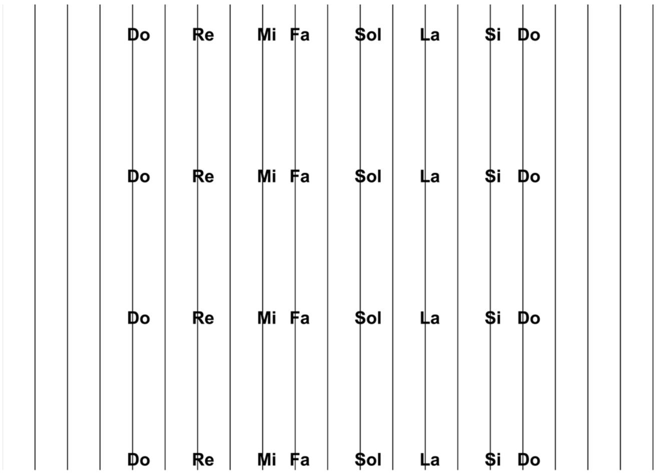

FIG. 1. Printed pattern attached onto the Wacom Intuos © graphic tablet for controlling the Cantor Digitalis.

the number of discernible intensity levels. For the stylus contact, a printed plastic sheet is used because it provides a more slippery surface than a sheet of paper or the raw surface of the tablet. Figure 1 gives a picture of the tablet and its printed pitch reference pattern.

Both the synthesis engine and the interface control of the Cantor Digitalis are programmed using the MAX real-time programming environment.[24] The Wacom tablet is operated using the *wacom.mxo* object.[29]

Chironomic control of intonation is achieved with gestures analogous to those of hand writing or hand drawing. Among all the available stylus parameters, two are retained for this experiment: The position of the stylus along the *X*-coordinates, and the pressure of the stylus on the tablet. Pressure on the stylus is associated to vocal effort control. Position of the stylus on the *X* axis is associated to pitch control. Other vocal dimensions (voice tension, breathiness, and roughness) are not used herein. As in real voice, the subject controls continuously the pitch variation: There are no "keys" with a given width and discrete pitch steps, like on, e.g., a keyboard. After some preliminary experiments, it seemed better to leave the *Y* axis free (i.e., motion of the stylus along the *Y* axis has no effect). This gives more freedom to the player's gestures, as it allows for playing with relaxed wrist motions, like waving motions for note transitions. Vocal effort is controlled by the stylus pressure, because it gives some expressivity to the voice, and this makes the player feel more comfortable with the synthetic voice. However, vocal effort variations are not further analyzed, as the experiments focus on intonation. A single /a/ vowel is used for all the experiments.

II. EXPERIMENTS IN CHIRONOMIC SINGING

A. Protocol and task

The aim of the experiments is to measure intonation accuracy and precision in chironomic singing synthesis. The protocol is identical for all the stimuli: A short synthetic singing example is presented to the subject together with the corresponding score and note names and she/he is asked to reproduce this example either by her/his voice or by drawing on the graphic tablet. In this task, for the sake of measuring pitch accuracy and precision, no vibrato is allowed. Tempo is imposed, using a visual and audio metronome.

A computer interface has been especially designed for this experiment. Before launching a stimulus, the user is asked to press the space bar of the keyboard, which lets her/him rest as much as she/he needs before pursuing the experiment. Then, the example to imitate is displayed and played, along with a metronome, waiting for the subject to perform. For the vocal tasks, she/he has to press the stylus on the tablet while singing and release it afterwards. For the chironomic tasks, the user is instructed not to release the stylus before the end of the pattern. The computer comes automatically to the next trial each time the stylus is released for both tasks.

The singing examples are /a/ vowels synthesized using a MIDI synthesizer (Instrument *choir Aahs 2* of the software MIDI SimpleSynth[30]), with a choral synthetic sound quality. Sound is played using a RME Fireface 400 soundboard (Audio AG, Haimhausen, Germany) and DTX900 Beyer dynamic headphones (Beyerdynamic, Heilbronn, Germany). Voice is digitally recorded using a DPA 4006-TL microphone (DPA Microphones, Alleroed, Denmark). The graphic tablet *X* position and pressure of the stylus are digitally recorded. A mute chironomic condition is also proposed. Subjects are asked to perform the sound imitation task with drawing only, without audio feedback. This results in three playing modalities: Chironomic singing with audio feedback (*Chironomy* modality), chironomic singing without audio feedback (*Mute Chironomy* modality), and vocal singing (*Voice* modality).

In each experimental session, a series of examples are imitated either by voice, chironomy, or mute chironomy. The examples are played in a randomized order among subjects. For each subject, the full experiment is split into three sessions, differing by the melodic material presented. All the recordings take place using high quality loudspeakers in an acoustically insulated and treated room designed for perceptual experiments. The subjects are allowed to take some rest whenever they want.

B. Musical patterns

1. Experiment 1: Intervals

Patterns in this experiment are ascending and descending diatonic intervals on a C major scale. The interval of major seventh in the diatonic scale is avoided, its intonation being more difficult. This musical material is displayed in Fig. 2(A), each bar corresponds to one example presented to the subject. The lowest note C of these intervals was the lowest C on the tablet.

The tempo is set to 120 beats per minute (b.p.m.). As the patterns are short and easy to memorize, the subjects need to listen to them only once. However, they still have the score with the name of the two notes displayed on the screen. They are instructed to perform three trials per pattern to reproduce the correct melody, and that only the best trial will be selected for each pattern. The 3 modalities (*Voice, Mute Chironomy*, and *Chironomy*) are recorded for the 12 patterns, resulting in 36 conditions recorded with 3 trials for each subject.

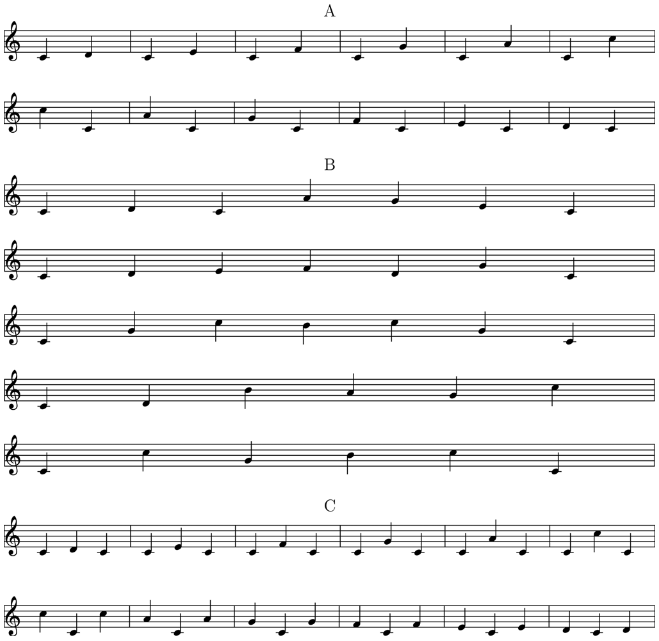

FIG. 2. Musical patterns used in the experiments. (A) Ascending and descending intervals (experiment 1). (B) Five melodies (experiment 2). (C) Double intervals (experiment 3). Different patterns are separated by bars.

2. Experiment 2: Melodies

Experiment 2 is similar to Experiment 1, but for longer musical material. The proposed patterns are 5 melodies composed of 6 or 7 notes, displayed in Fig. 2(B). These melodies are especially composed with all the intervals of Experiment 1. As melodies are harder to memorize than simple intervals, the subjects can listen to each voice example as often as they wish. As singing an unknown melody can be challenging for occasional singers, they record as many trials as they wish, with a minimum of three. They are informed that the best trial will be selected for each pattern. The tempo is set to 120 b.p.m. The 3 modalities are recorded for each pattern, resulting in 15 conditions recorded with a minimum of 3 trials for each subject.

3. Experiment 3: Tempo

The aim of this experiment is to study the possible influence of tempo in chironomic singing. The 12 patterns, displayed in Fig. 2(C), are made of double intervals, ascending/descending, or descending/ascending, beginning and ending with the same note. To focus on the tempo, only the *Chironomy* modality is used. As the patterns are short and easy to memorize, the subjects need to listen to them only once. They are instructed to perform only three trials per pattern to reproduce the correct melody, and that only the best trial will be selected for each pattern. The metronomic tempi are 120, 179, and 240 b.p.m., resulting in 36 conditions recorded with 3 trials for each subject.

C. Subjects

A group of 20 subjects took part in experiments 1 and 2 (average age 31 yrs; 6 females, 14 males). In this group 14 subjects received formal musical training, and/or have a regular musical practice. The mean musical practice experience was 18 yrs. None of them reported any known auditory impairment, but 12 subjects self-evaluated their singing ability as poor in terms of accuracy. Sixteen subjects were right-handed and four were left-handed.

A group of 28 subjects took part in experiment 3 (average age 29 yrs; 11 females, 17 males). In this group 18 subjects received formal musical training, and/or have a regular musical practice. The mean musical practice experience was 16 yrs. None of them reported any known auditory impairment, but 15 subjects self-evaluated their singing ability as poor in terms of accuracy. Twenty-three subjects were right-handed and five were left-handed.

All the subjects were members of the laboratory and participated in the experiment on a voluntary basis, without being paid.

Controlling a synthetic voice with a stylus was a new experience for all the subjects, except three. Therefore before recording data, a training session was offered in order to get familiar with the device and protocol. Subjects performed a similar task as in the recorded experiments. To avoid learning effects, different patterns that were not used in the main experiments were presented.

D. Accuracy and precision analyses

Pitch identification for each note follows a procedure similar to that of Pfordresher *et al.*:[11] Compute raw pitch, identify steady-state phases of each note, compute pitch for each note as the average value of this steady-state phase.

The fundamental frequency ($F0$) for each recorded trial is directly available for the tablet recordings. For voice recordings, $F0$ is obtained using the STRAIGHT[31] pitch detection algorithm. $F0$ values are converted from Hertz (Hz) to STs relative to a 440 Hz reference according to the MIDI convention $\mathrm{ST} = 12\ \log_2(\mathrm{Hz}/440) + 69$. Examples of $F0$ traces are plotted in Fig. 3, with the corresponding target $F0$ of the reference to be imitated.

After $F0$ detection, the steady-state phase of each note is identified. The steady-state phase durations are less than

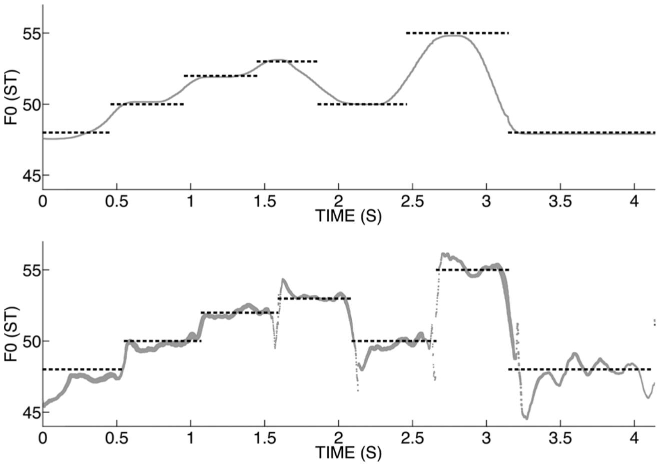

FIG. 3. Examples of raw $F0$ curves (plain) and pitch references (dashed) for the second B melody. Top panel: Chironomy condition; bottom panel: Voice condition.

500 ms (interval between metronome beats). For such short duration tones, the pitch for each interval corresponds to the time average of its $F0$ contour.[32,33]

A semi-automatic stylization procedure is used for speeding up the process of steady state phase identification. It can be sketched as follows. The time axis is divided into short (10 ms) intervals. Two consecutive intervals whose average pitch difference is under a given threshold are merged into a larger interval. Typical thresholds are 50 cents for the voice recordings and 10 cents for tablets recordings. This process is repeated until all intervals are separated from their neighbors with at least one threshold. Then the $F0$ time-average on each interval is computed and it is the pitch assigned to the identified note. All the note identification and pitch values are visually checked for possible artifacts.

Figure 4 shows an example of the pitch extracted from a vocal imitation of the third melody (smooth curve). The steps represent the stylized curve. The × symbols indicate the pitches associated to the segments detected by the algorithm.

Once the extraction is done, each trial of each condition of each subject is associated to a list of detected notes to be compared to the list of targeted notes for subsequent analyses. All the trials incorrectly performed by the subjects (e.g., with an incorrect number of notes, or incomplete) are discarded.

Note and interval accuracy and precision are computed according to Eqs. (1)–(3) of Pfordresher *et al.*[11] Accuracy and precision are expressed in frequency units (cents). Good accuracy (respectively, precision) means an accuracy (respectively, precision) measure close to 0 cents. Accuracy can be positive or negative, while precision is always positive. Note accuracy and interval accuracy are computed for each trial. Only one trial, the best trial, is kept for each experimental condition. The trial for which the sum of note accuracy and interval accuracy is minimum is considered as the best trial. It is selected and used for all the subsequent analyses.

III. RESULTS

A. Data grouping and analysis

Accuracy and precision being statistical measures, they depend on the specific set of notes chosen. In the study of Pfordresher *et al.*,[11] different sets are used for measuring accuracy and precision: Accuracy is computed on all the notes played by each subject, whereas precision is computed taking all the notes with similar pitches among the data of each subject. On the contrary Ternström and Sundberg[34] computed both accuracy and precision for all the notes with similar pitches among the data of each subject. In the present study, three groups of data sets are considered: The group "Subject" (this set contains all notes of each subject), the group "Pattern" (this set contains all notes of each musical pattern), and the group "Interval" size (this set contains all notes preceded by a given interval). The factors and data sets for each group are summarized in Table I.

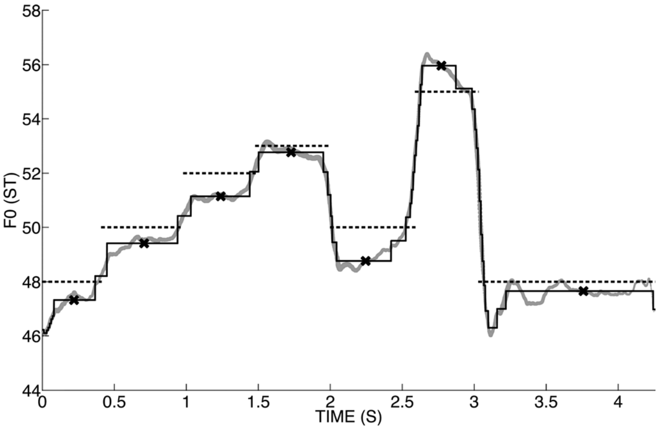

FIG. 4. Pitch estimation for a trial of the second B melody. Stylized segments (thin line) are superimposed on the detected $F0$ (thick lines) and target $F0$ (dashed lines). The pitch for each note is marked by an ×.

The Subject group contains two factors: The subject factor (20 levels) and the "modality" factor (3 levels), resulting in 60 sets of measures in this group. The Pattern group contains two factors: The pattern factor (17 levels) and the modality factor (3 levels), resulting in 51 sets of measures in this group. The Interval group contains 2 factors: The "interval size" factor (15 levels) and the modality factor (3 levels). There are 15 different intervals and 3 modalities resulting in 45 sets of measures for interval accuracy and precision. For note accuracy and precision, the unison interval (first note of a pattern) is also taken into account to give 16 intervals and 48 sets of measures in this group.

The distributions obtained for each of the data sets defined above are plotted in Fig. 5. Each box-plot represents the distribution of note and interval accuracy (respectively, precision) for one modality, computed on one group: Subject, Pattern or Interval. The bold line represents the median of values, whereas the box includes 50% of the values. Statistical significance of the differences in accuracy (respectively, precision) between the three modalities for each group are studied by pairs using a Wilcoxon rank-sum test,[35] with the help of the "wilcoxst" procedure of the R environment.[36]

B. Main effects

The main result is the very good precision and accuracy obtained with the *Chironomy* and the *Mute Chironomy* modalities, and the higher bias and dispersion obtained for the *Voice* modality. Precision values of both note and interval are low for the tablet modalities, and are always significantly higher for the *Voice* modality ($W \leq 36$, $p < 0.05$, whatever the grouping set). It shows the wider dispersions in the results of the *Voice* modality compared to the two tablet modalities. As the subjects are for the most part untrained singers, the *voice* modality appears more difficult than the others.

The picture for accuracy measures is more complex. The note accuracy is low, close to zero, for both tablet

TABLE I. Summary of experiments, groups of data sets, factors, and number of data sets.

Experiment	Group	Factors	# Sets
1 and 2	Subject	Subject and modality	20 × 3
1 and 2	Pattern	Pattern and modality	17 × 3
1 and 2	Interval	Preceding interval and modality	15 × 3
3	Tempo	Subject, pattern, and tempo	28 × 12 × 3

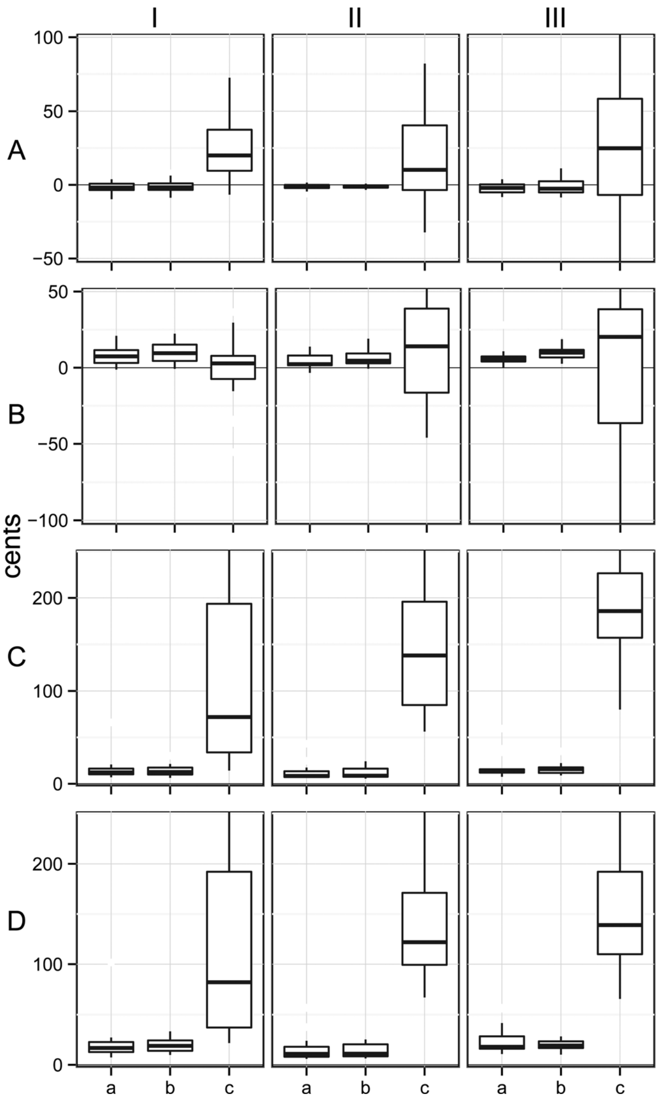

FIG. 5. The four rows depict the distributions of (A) note accuracy, (B) interval accuracy, (C) note precision, and (D) interval precision, as measured in the modalities (a) Chironomy, (b) Mute Chironomy, and (c) voice, based on the following grouping sets: (I) Subject, (II) Pattern, and (III) Interval.

modalities and is always lower than the one observed for the *Voice modality*, but it is only significantly lower when considering the Subject grouping set ($W=49$, $p<0.05$, and $W=50$, $p<0.05$ between the Voice modality and the *Chironomy*, and the *Mute Chironomy* modalities), and the interval grouping set for the difference between the *Voice* and the *Chironomy* modalities ($W=75$, $p<0.05$).

For interval accuracy measures, the two tablet modalities still score below 25 cents, but they never significantly outperform the *Voice modality*, whatever the grouping set. On the contrary, *Voice modality* is even significantly more accurate at the interval level than the *Mute Chironomy* modality in the case of the subject grouping set ($W=286$, $p<0.05$): *Voice* has a median of 3 cents, while *Mute Chironomy* has a median of 10 cents. Considering the threshold of pitch perception, this difference is negligible.

Whatever the measure and the grouping set, both chironomic modalities do not show any significant difference in their performances. The subjects perform as well with or without audio feedback. Because of the generally high visuo-motor ability, the results of the *Mute Chironomy* modality are excellent. The medians of accuracies are between −3 and 10 cents whereas the medians of precisions are between 8 and 19 cents. Considering a threshold of pitch perception of about 4 cents, better results could hardly be expected. When audio is present, one can suspect a different behavior for doing the task, listening also to pitch accuracy or interval ratio, but this cannot be measured here, due to the level of performance of the *Mute Chironomy* modality.

C. Effect of singing and musical training

A striking effect observed in the preceding result is the average poor performance of the *Voice* modality, compared to the tablet modalities. As the subjects' profiles are varied in terms of musical training and experience, it is interesting to examine individual performances. Figure 6 displays absolute value of accuracy and precision as a function of subjects and modalities, in experiments 1 and 2. The subjects are ranked according to interval precision (see top panel) for *Voice*. Interval precision has been chosen because it appeared that this measure was the most representative of musical quality. Precision is preferred to accuracy because it is an indication of reliability. Precise subjects are almost always accurate. Interval and note precisions are almost always highly correlated.

Figure 6 shows that for the best subjects, very good accuracy and precision are obtained for all modalities, voice included. The *Voice* modality obtains the best results for some subjects. But all the subjects, regardless of their natural singing proficiency, show a high proficiency in chironomic singing.

The same data for accuracy and precision are plotted in Fig. 7. Each + sign represents a subject, for note accuracy versus note precision (top), and interval accuracy versus interval precision (bottom). The closer the measures from the origin (0,0), the better the singing ability. Although the best subjects with *Voice* have similar precisions than subjects with *Chironomy* or *Mute Chironomy*, half of the subjects with *Voice* are less precise than the worst subjects with *Chironomy* or *Mute Chironomy*. Moreover, the dispersion of precision values for *Voice* is higher than for the *Chironomy* or *Mute Chironomy*.

Pfordresher *et al.*[11] propose several thresholds above which a person can be considered an inaccurate (respectively, imprecise) singer. In occidental music, the ST being the smallest musical interval in a scale, a threshold of 50 cents is chosen for accuracy and precision. According to this threshold, as for the *Voice* modality, it appears that 85% of the subjects are accurate and that 40% are precise, whereas Pfordresher *et al.* reported 69% accurate and 27% precise subjects.[11] Moreover, 40% of our subjects are both accurate and precise, to be compared to 25% in the study of Pfordresher *et al.*[11] The present results are clearly better than previously reported results. This may be explained by the fact that more than half of our subjects received musical training. On the contrary, none of the subjects were musicians in the study by Pfordresher *et al.*, which targeted

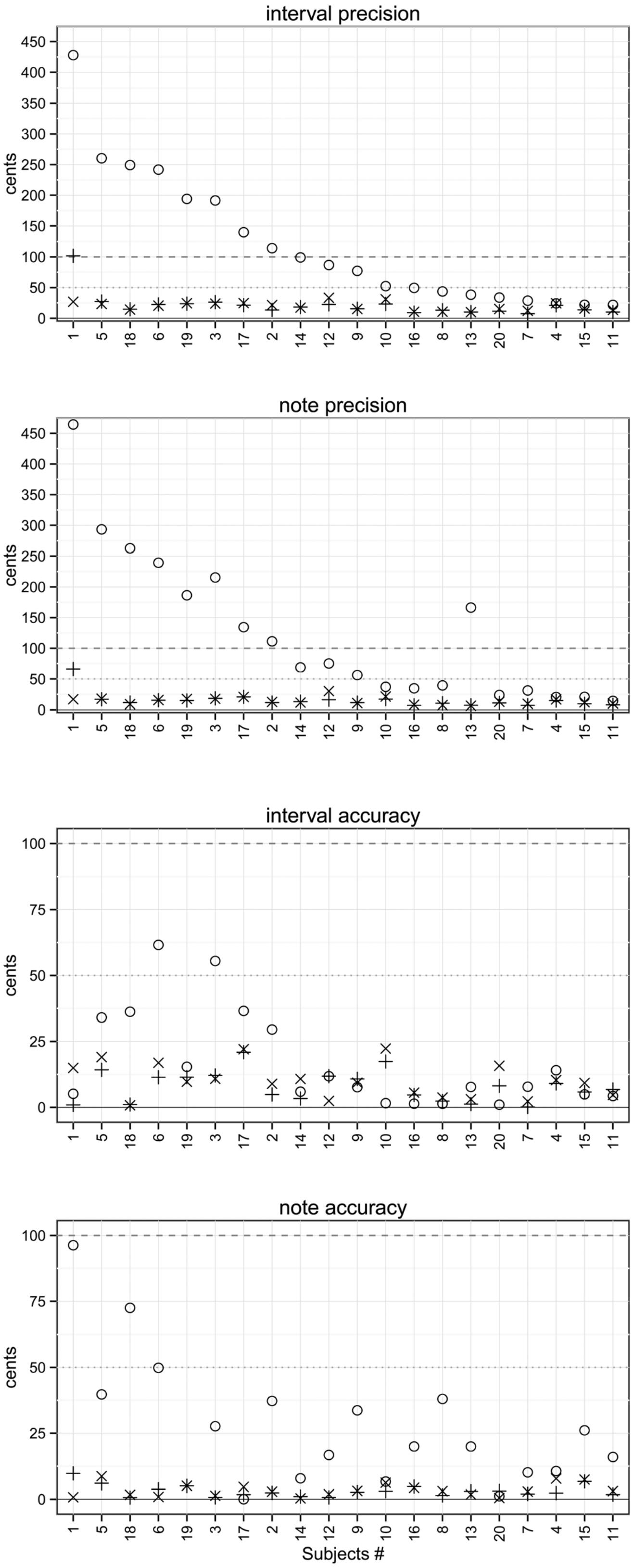

FIG. 6. Mean values, for each subject, of (from top to bottom) interval precision, note precision, interval accuracy, and note accuracy. Subjects numbers are ordered according to their mean level of interval precision in the voice modality (see text). The three modalities are represented by (+) for chironomy, (×) for mute chironomy, and (○) for voice. The y axis scale is restricted to [0,450] cents for the two precision measures, and [0,100] cents for the two accuracy measures, the absolute values of which are plotted. This restriction leaves subject #19 out of the note accuracy graph in the voice modality.

singing ability in the general population, to the exclusion of specialized musicians.

All but one subject (respectively, all) are accurate and precise with *Chironomy* (respectively, *Mute Chironomy*). This is all the more impressive because almost all the subjects were not used to this type of task at all. These results show that, unlike for the *Voice* modality where performances of subjects are scattered and may depend on the singing experience of the subjects, chironomy allows almost all of them to sing accurately and precisely, whatever their musical background.

Figures 6 and 7 show that although all the subjects but three have better results with *Chironomy* or *Mute Chironomy* than *Voice* regarding note accuracy, nine subjects have better results for *Voice* regarding interval accuracy. This can be explained by the presence of visual references on the tablet, favoring note targeting rather than interval accuracy. On the contrary accurate intervals are easier to sing than accurate pitches for singers, at least for singers without perfect pitch.

D. Effect of patterns

Both intervals and short melodies were proposed to the subjects. Imitating melodies was possibly more demanding, because longer patterns impose a more cognitive load to the subjects. Note accuracy, note precision, interval accuracy, and interval precision are plotted separately for intervals and melodies in Fig. 8, for the three modalities.

There is no noticeable difference between intervals and melodies for note accuracy. A small tendency appears for interval accuracy and precision, with slightly better results for simple intervals. Overall, the effect is very small, and one feels entitled to conclude that melodies are sung with similar accuracy to simple intervals.

As for interval precision, a special attention must be paid to the panel (A)-IV of Fig. 8. The *voice* modality shows better scores for intervals accuracy in melodies. It is also the condition where the worst results are observed for the two chironomic modalities. This can be explained because singing melodies is the closest situation to a musical performance, but the most complex gestural task in our experiments.

The difference between interval accuracies in intervals and melody patterns show a significant increase in both modalities (Wilcoxon rank sum test: $W = 2$, $p < 0.01$ for the chironomy modality; $W = 2$, $p < 0.01$, for the mute chironomy modality). Conversely, the interval accuracies observed for voice are lower for melodies than for 2-notes singing (mean improvement of 20 cents, not significant: $W = 38$, $p = 0.43$). The mean interval accuracy for voice is even closer to zero than for both chironomic modalities (not significant differences).

E. Effect of interval size

A peculiarity of chironomic singing is that larger intervals correspond to larger hand displacements, that are possibly more demanding in terms of motor control. Therefore, another parameter which is worth studying is the effect of the melodic motion, which corresponds to hand motion in the chironomic experiments. To study a possible effect of the direction of the movement preceding a note on its

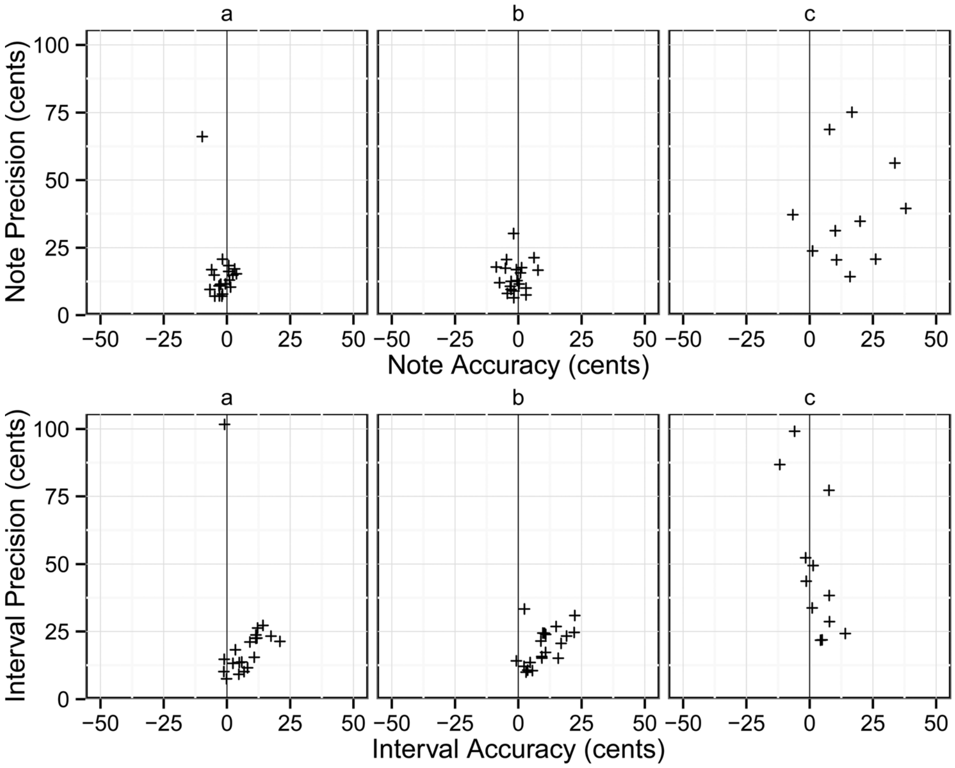

FIG. 7. Mean note accuracy vs note precision (top) and interval accuracy vs interval precision (bottom), for each subject (each subject corresponds to a +) in experiment 1 and 2 (mixed), and for the three playing modalities: (a) Chironomy, (b) mute chironomy, and (c) voice. The scale is restricted to [−50, 50] cents for the x axis and [0,100] cents for the y axis, leaving the nine worst singers out of the graph for the voice modality.

precision and accuracy, note accuracy, and precision are computed at the ends of intervals for downward melodic movements, unisons, and upward melodic movements. The results are presented in Fig. 9.

It seems that the direction of the movement (up or down) has some influence on the sign of note accuracy, showing an overshoot effect. In both chironomic modalities, downward movements tend to produce notes below the target, while upward movements tend to produce notes above the targets. For the *Chironomy* modality, the observed difference in accuracy between downward and upward movement is about 5 cents, and is significant (Wilcoxon rank sum test: $W = 9$, $p < 0.05$); for the *Mute Chironomy* modality, the mean difference is of more than 9 cents, and is significant

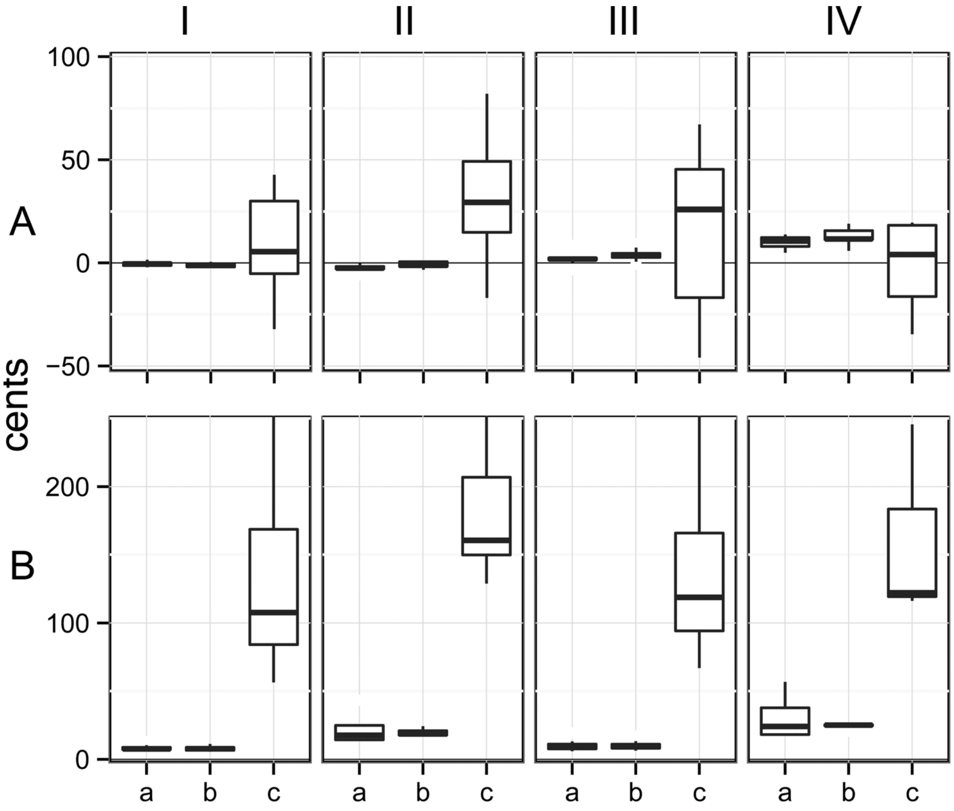

FIG. 8. Distribution of (A) accuracy and (B) precision measures based on the sets of patterns, for notes (I and II) and intervals (III and IV), in the three modalities: (a) Chironomy, (b) mute chironomy, and (c) voice. The results for simple intervals (experiment 1) are in I and III, the results for melodies (experiment 2) are in II and IV.

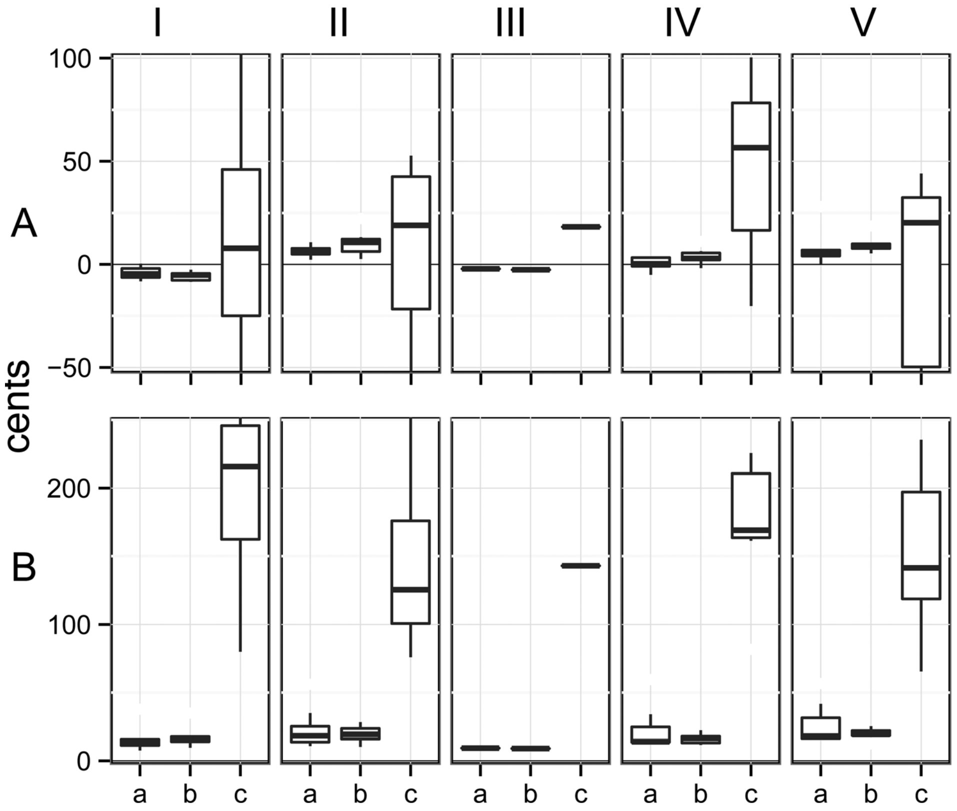

FIG. 9. Distribution of (A) accuracy and (B) precision measures based on the sets of intervals, for notes (I, III, and IV) and intervals (II and V), in the three modalities: (a) Chironomy, (b) mute chironomy, and (c) voice. Observations are separated according to the direction of the melodic movement leading to the considered note: (I and II) Downward movement, (III) no movement, and (IV and V) upward movement.

($W = 0$, $p < 0.05$). Listening to sound tends to help subjects to reduce this effect of movement direction on accuracy: The *Mute Chironomy* modality shows stronger deviations linked to the preceding movement direction, which are reflected in the significantly higher interval accuracies observed in the *Mute Chironomy* modality compared to the *Chironomy* modality, for both upward and downward movements ($W = 56$, $p < 0.05$).

An unexplained fact is that the results obtained for the *Voice* modality accuracy are always positive (higher than the reference), from 3 to 25 cents. On the contrary, the chironomic modalities are equally biased toward both directions. The pitch estimation algorithm has been thoroughly tested using synthetic data, and showed a good accuracy (0 ± 5 cents) and no bias toward positive values. Therefore, the small bias observed in the data cannot be explained by a pitch estimation bias.

F. Effect of tempo

Experiment 3 focused on the effect of tempo. In all the experiments, the subjects were asked to synchronize their notes with metronome beats. The difficulty of the task is supposed to increase for faster tempi. Accuracy and precision computed for each tempo, using all the notes played by each subject for each pattern in Experiment 3, are displayed in Fig. 10.

Although some variation can be observed in Fig. 10, the differences are no more than a few cents, close to or under the pitch difference limens. At a tempo of 240 b.p.m., the subjects can play 3-note patterns using the tablet with high accuracy and precision. This corresponds to semi quavers played at a tempo of 60 quarter notes per minute, which can be considered already as a fast tempo for vocal music. In this experiment, it seems that a critical tempo after which the performances fall down has not been reached. The question of critical tempo would certainly deserve a specific study, which was out of the scope of this initial work.

IV. DISCUSSION AND CONCLUSION

A. Effect of visual references

The role played by visual references on the tablet is an important question to be addressed in order to understand

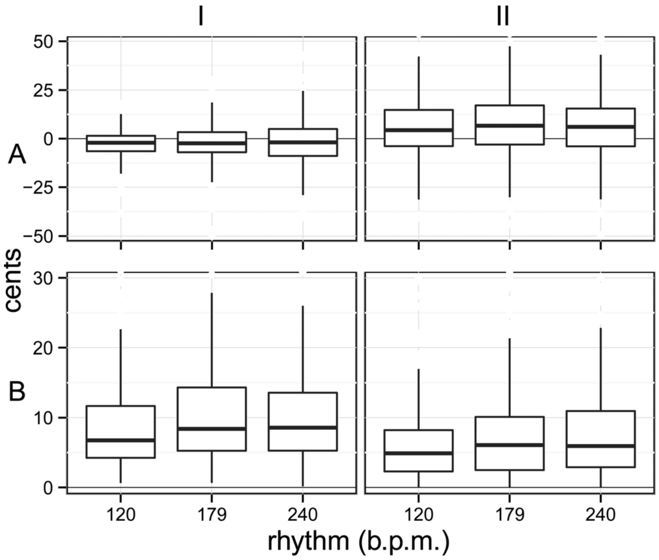

FIG. 10. Distribution of the measures of (A) accuracy and (B) precision, for (I) notes and (II) intervals, for each of the three tempi of experiment 3: 120, 179, and 240 b.p.m.

the differences in performance between the voice and chironomic conditions. To better assess the role played by the visual modality, it would be interesting to test also a chironomic imitation experiment without visual reference on the tablet, or "*blind chironomy*." For untrained subjects the *Voice* and *Chironomy* modalities would possibly be more comparable, as subjects would make use only of the audio and kinesthetic modalities. The visual reference on the graphic tablet is clearly a technical advantage for the chironomic condition compared to singing. Most subjects in this study were not perceiving absolute pitch, and had no fixed reference for imitation of the sung examples, except their short term memory.

For addressing this question, it is interesting to reinterpret blind chironomic experiments reported in previous work[8] on chironomic speech intonation stylization. Although the conditions were differing to some respect, the experiments are somewhat comparable and it can provide us with an indication of the results for blind chironomy.

In d'Alessandro *et al.*[8] subjects were asked to reproduce the intonation of given sentences under two conditions: With their own voice, and by drawing the melody on a tablet, without any printed visual reference. The protocol was similar to the protocol in the present study: A sentence was presented to the subjects, and they were asked to imitate its melody (intonation) either using their voice or blind chironomy. For the chironomic condition, a real-time pitch modification algorithm was used for playing the modified sentences according to the subjects' gestures recorded on the tablet.

The pitch of each vowel in each syllable was measured, converted to ST and compared with the pitch of the reference sentence. Examples of pitch contours obtained in this experiment are plotted in Fig. 11. For each vowel in a syllable, the root-mean-square (rms) distance and correlation between the voice (respectively, chironomic) imitation and the natural voice were computed.

For the sake of comparison with the present study, the following analogies can be used: Syllables are the basic melodic units, analogous to notes; rms distance between the trial and the target is the distance, analogous to the average pitch distance to a pitch target; the mean of rms distances (respectively, standard deviation) is a measure of pitch accuracy (respectively, precision).

Extrapolating the results in d'Alessandro *et al.*,[8] the mean and standard deviation of the rms distances obtained for all subjects and all sentences are computed. This gives for the voice condition (respectively, blind chironomy condition) an accuracy of 1.42 ST (respectively, 2.17 ST) and a precision of 0.78 ST (respectively, 0.95 ST).

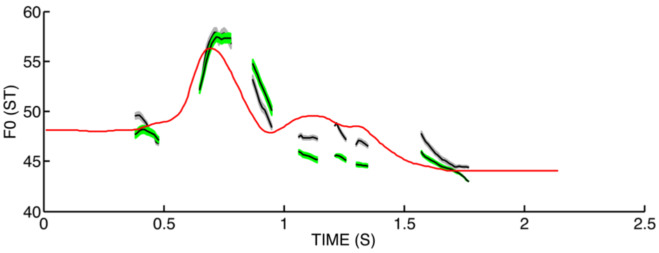

FIG. 11. (Color online) Examples of trials for blind chironomic imitation (seconds, STs) of intonation. Natural contour (thick dark line), blind chironomic imitation (thin lines), and voice imitation (thick light line). For five-syllable (top), and seven-syllable (bottom) sentences.

Both accuracy (respectively, precision) is better, by a factor of 1.5 (respectively, 1.2), for the voice condition than for the blind chironomy condition. However, a perception experiment in Ref. 8, using resynthesis of stylized contours, ultimately showed that voice or blind chironomic stylization was perceptually equivalent.

This indicates that, without visual reference, the subjects obtained similar accuracy and precision, within the limits of pitch perception, for the vocal imitation condition and the blind chironomic imitation condition.

These data do not correspond to a musical task. This can explain why the results for accuracy, and to a lesser extent precision, are clearly worse compared to the musical data.

As the conditions and subjects were different among the musical and speech experiment, no definitive conclusion can be reached. However there is a clear indication that although chironomy is still effective without visual reference, it is much improved when the subjects are provided with it. This is further supported by recent data. Vocal imitation of song and speech has been studied in a recent work.[37] According to the authors, "Results in general support the view that vocal imitation is integrative rather than modular, and that imitation abilities in one domain (e.g., song) predict imitation in another domain (e.g., speech)."

One can assume that the visual feedback of chironomy would be of a lesser importance with practice. As a matter of fact, actual musical training with chironomic singing synthesis (several public concerts were actually performed) indicated that more experienced performers rely much less on visual clues, as is the case for other musical instruments. Learning chironomic singing would certainly follow the same path as learning fretless musical instruments, like the violin or, in the electronic domain, the Ondes Martenot or the Theremin.

For real world use of chironomic singing there is no reason to deny visual cues: The singing instrument is always equipped with a kind of visual fretting or keyboard-like reference. The *blind chironomy* modality has not been tested in the present work because it would have been too artificial to totally blind our subjects.

B. Writing, drawing, singing

All the subjects performed surprisingly well with chironomic singing. For natural singing, the results are more scattered. Both singing ability and experience differ among subjects. Some subjects are trained musicians or amateur singers (performing regularly in public), although others have no practice nor interest in singing. The top singers in our experiments have comparable performance levels for all the modalities. On the contrary the poor singers are still able to perform well using chironomy. The chironomic singing proficiency for all the subjects in these experiments appears much better than the singing proficiency reported for the general population.[10]

This result can be explained by the learned ability of the subjects regarding stylus manipulation for writing and drawing. All the subjects spontaneously used their predominant hand, the hand they use for writing. All subjects were trained hand writers, and therefore used their skills in the chironomic task. This visuo-motor skill was enhanced by visual references on the tablet, which enabled accurate and precise imitations by most of the subjects, whatever their musical or singing training. These clues transposed the musical task into a drawing task, which set the subjects on equal footing.

Differences in the *Mute Chironomy* and *Chironomy* modalities are very small, on the order of magnitude of pitch difference limens. This situation is somewhat comparable to other musical situations, where performers train effectively on mute instruments (like mute keyboards or stringed instruments). The visual and kinesthetic controls can be used alone for learning musical control gestures. However, the audio feedback seems useful for compensating the overshoot effect due to larger hand displacements for larger musical intervals. The *Chironomy* modality improved for interval accuracy, compared to the *Mute Chironomy* modality. In this case, musical ratios are appreciated and corrected through the audio modality. It may indicate that the tablet helps to focus on absolute positions of notes, whereas subjects rather focus on the intervals while hearing with their own voices, either in real or chironomic singing. Although one can play simple melodies without audio feedback, sound is evidently needed for expressive musical performance and fine adjustments in real musical situations. In singing, kinesthetic control also plays an important role, but masking of the audio feedback leads to a significant degradation in intonation accuracy.[38]

Chironomic control seems well suited for pointing to targets. In the tests, this corresponded to note accuracy and precision. It is noticeable that the worst measure for *Mute Chironomy* is interval accuracy. On the contrary, *Voice* obtained comparatively good results for interval accuracy. It may indicate that subjects rather focus on the intervals while singing with their own voices. For intervals, the task consists not only of pointing at targets, but appreciating musical ratios, which is clearly a task relying on the audio modality. Chironomy does not give special visual or kinesthetic clues for intervals accuracy, although it is prominent in singing proficiency.

In a recent study,[39] it appeared that the accuracy of vocal pitch matching improves for pitch imitation when natural voice examples are given instead of synthetic voice examples. In the present study, synthetic voice of relatively poor quality has been used. One can speculate that using real voice examples might have improved the results obtained for the *voice* modality, and may have reduced the difference with the chironomic modalities.

C. Conclusion

The aim of this research is to study the melodic precision and accuracy in chironomic singing synthesis. Cantor Digitalis, a chironomic singing system, has been designed. The system uses a stylus to control pitch on a graphic tablet equipped with printed patterns. Accuracy and precision, i.e., bias and variance, in performance of various melodic patterns have been measured. The recorded material included natural voice sound and the stylus traces on the tablet with and without audio feedback. The main result of our study is the high accuracy and precision obtained by all the subjects for chironomic control of singing synthesis. The mean chironomic note accuracy obtained is less than 12 cents and mean chironomic interval accuracy less than 25 cents, for all the subjects. For some subjects, natural singing and chironomic singing are equally accurate and precise. For other subjects, with less experience and interest in singing, chironomic singing is much easier than natural singing. Chironomic singing relies much on the writing and drawing ability acquired since childhood. Visual, audio, and kinesthetic modalities are used, taking advantage of the high skills developed in target pointing with the help of a stylus. The audio modality in this case helps for interval accuracy and precision. Comparing the results with previous experiments demonstrates the important role played by the visual pitch reference patterns printed on the tablet. This visuo-motor advantage explains the high level of performance reached in chironomic singing, with only minimal training. Of course the mute chironomy condition is only a laboratory condition. The part played by audio is evident in real world musical performance, as it is almost impossible to achieve expressive control or to synchronize with other players without audio feedback. The chironomic approach appears as a choice candidate for designing performative singing synthesis instruments.

ACKNOWLEDGMENTS

This work was supported by the Agence Nationale de la Recherche ANR, through the ChaNTeR project (ANR-13-CORD-0011, 2014-2017), and the Région Iles-de-France through the ORJO project (2009-2013).

[1] M. Astrinaki, N. d'Alessandro, and T. Dutoit, "MAGE—A platform for tangible speech synthesis," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, Ann Arbor, Michigan (2012), pp. 353–356.

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

[3] L. Kessous, "Gestural control of singing voice, a musical instrument," in *Proceedings of Sound and Music Computing Conference*, http://smcnetwork.org/files/proceedings/2004/P39.pdf (Last viewed October 9, 2013).

[4] S. Le Beux, L. Feugère, and C. d'Alessandro, "Chorus digitalis: Experiments in chironomic choir singing," in *Proceedings of the International Conference on Speech Communication*, Firenze, Italy (2011), pp. 2005–2008.

[5] D. Trueman, P. Cook, S. Smallwood, and G. Wang, "PLOrk: Princeton Laptop Orchestra," Year 1, in *Proceedings of the International Computer Music Conference*, New Orleans, LA (2006), pp. 164–167.

[6] M. M. Wanderley, J. Viollet, F. Isart, and X. Rodet, "On the choice of transducer technologies for specific musical functions," in *Proceedings of the International Computer Music Conference*, Berlin, Germany (2000), pp. 244–247.

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

[8] C. d'Alessandro, A. Rilliard, and S. Le Beux, "Chironomic stylization of intonation," J. Acoust. Soc. Am. **129**(2), 1594–1604 (2011).

[9]S. Le Beux, A. Rilliard, and C. d'Alessandro, "Calliphony: A real-time intonation controller for expressive speech synthesis," in *Proceedings of the International Speech Communication Association Speech Synthesis Res. Workshop*, Bonn, Germany (2007), pp. 345–350.

[10]S. Dalla Bella, J.-F. Gigure, and I. Peretz, "Singing proficiency in the general population," J. Acoust. Soc. Am. **121**, 1182–1189 (2007).

[11]P. Q. Pfordresher, S. Brown, K. Meier, M. Belyk, and M. Liotti, "Imprecise singing is widespread," J. Acoust. Soc. Am. **128**, 2182–2190 (2010).

[12]X. Rodet, Y. Potard, and J. B. Barriere, "The CHANT project: From synthesis of the singing voice to synthesis in general," Computer Music J. **8**(3), 15–31 (1984).

[13]G. Bennett and X. Rodet, "Synthesis of the singing voice," in *Current Directions in Computer Music Research*, edited by M. V. Mathews and J. R. Pierce (MIT Press, Cambridge, MA, 1989), pp. 19–44.

[14]P. Cook, "Singing voice synthesis: History, current work, and future directions," Computer Music J. **20**(3), 38–46 (1996).

[15]J. Sundberg, "Synthesis of singing by rule," in *Current Directions in Computer Music Research*, edited by M. V. Mathews and J. R. Pierce (MIT Press, Cambridge, MA, 1989), pp. 45–56.

[16]P. Depalle, G. Garcia, and X. Rodet, "A Virtual Castrato (!?)," in *Proceedings of the International Computer Music Conference*, San Francisco, CA (1994), pp. 357–360.

[17]H. Kenmochi and H. Ohshita, "VOCALOID-commercial singing synthesizer based on sample concatenation," in *Proceedings of the International Conference on Speech Communication* (2007), pp. 4009–4010.

[18]E. Miranda and M. M. Wanderley, *New Digital Musical Instruments: Control and Interaction Beyond the Keyboard* (A-R Editions, Middleton, WI, 2006), pp. 1–25.

[19]P. Cook, "Real-time performance controllers for synthesized singing," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, Vancouver, Canada (2005) pp. 236–237.

[20]H. Dudley and T. H. Tarnoczy, "The speaking machine of Wolfgang Von Kempelen," J. Acoust. Soc. Am. **22**(2), 151–166 (1950).

[21]H. Dudley, "Remaking speech," J. Acoust. Soc. Am. **11**(2), 169–177 (1939).

[22]S. Fels and G. Hinton, "Glove-Talk II—a neural-network interface which maps gestures to parallel formant speech synthesizer controls," IEEE Trans. Neural Networks **9**, 205–212 (1998).

[23]C. d'Alessandro, N. D'Alessandro, S. Le Beux, J. Simko, F. Cetin, and H. Pirker, "The speech conductor: Gestural control of speech synthesis," in *Proceedings of eNTERFACE Summer Workshop on Multimodal Interfaces*, Mons, Belgium (2005), pp. 52–61.

[24]MAX programing environment, http://cycling74.com/ (Last viewed March 9, 2013).

[25]M. S. Puckette, "Pure data," in *Proceedings of the International Computer Music Conference*, International Computer Music Association, San Francisco, CA (1996), pp. 269–272.

[26]J. N. Holmes, "Formant synthesizers: Cascade or parallel?," Speech Commun. **2**(4), 251–273 (1983).

[27]B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," in *Proceedings of the International Speech Communication Association Voqual'03: Voice Quality: Functions, Analysis and Synthesis*, Geneva, Switzerland (2003), pp. 15–20.

[28]B. C. J. Moore, "Frequency difference limens for short-duration tones," J. Acoust. Soc. Am. **54**, 610–619 (1973).

[29]http://www.jmcueyeti.fr/download.html (Last viewed October 8, 2013).

[30]http://simplesynth.sourceforge.net/ (Last viewed October 9, 2013).

[31]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).

[32]C. d'Alessandro and M. Castellengo, "The pitch of short-duration vibrato tones," J. Acoust. Soc. Am. **95**(3), 1617–1630 (1994).

[33]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).

[34]S. Ternströ and J. Sundberg, "Intonation precision of choir singers," J. Acoust. Soc. Am. **84**, 59–69 (1988).

[35]D. F. Bauer, "Constructing confidence sets using rank statistics," J. Am. Stat. Assoc. **67**(339), 687–690 (1972).

[36]R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org (Last viewed October 9, 2013).

[37]J. T. Mantell and P. Q. Pfordresher, "Vocal imitation of song and speech," Cognition **127**, 177–202 (2013).

[38]D. Mürbe, F. Pabst, G. Hofmann, and J. Sundberg, "Significance of auditory and kinesthetic feedback to singers pitch control," J. Voice **16**(1), 44–51 (2002).

[39]R. Y. Granot, R. Israel-Kolatt, A. Gilboa, and T. Kolatt, "Accuracy of pitch matching significantly improved by live voice model," J. Voice **27**(3), 390.e13–390.e20 (2013).