Hand gesture realisation of contrastive focus in real-time whisper-to-speech synthesis: Investigating the transfer from implicit to explicit control of intonation

Delphine Charuau, Nathalie Henrich Bernardoni, Silvain Gerber, Olivier Perrotin

Speech Communication 177 (2026) 103344

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Hand gesture realisation of contrastive focus in real-time whisper-to-speech synthesis: Investigating the transfer from implicit to explicit control of intonation

Delphine Charuau [a,b,*], Nathalie Henrich Bernardoni [b], Silvain Gerber [b], Olivier Perrotin [b]

[a] Sigmedia Lab, School of Engineering, Trinity College Dublin, Dublin, D02 PN40, Ireland
[b] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, F-38000, France

ARTICLE INFO

MSC:
0000
1111

Keywords:
Whisper-to-speech
Contrastive focus
Intonation
Manual control
Modality transfer

ABSTRACT

The ability of speakers to externalise the control of their intonation in the context of voice substitution communication is evaluated in terms of the realisation of a contrastive focus in French. A whisper-to-speech synthesiser is used with gestural interfaces for intonation control, enabling two types of gesture: an isometric finger pressure and an isotonic wrist movement. An original experimental paradigm is designed to elicit a contrastive focus on the /lu/ syllables of nine-syllable sentences by means of a read-question-answer scenario. For all 16 participants, focus was successfully achieved in speech and in both modality transfer situations by increasing the fundamental frequency and duration of the target syllable. Coordination of the articulation of the whispered syllables and the manual intonational control was acquired quickly and easily. Focus realisation by finger pressure or wrist movement showed very similar dynamics in intonation and duration. Overall, although wrist movement was preferred in terms of ease of control, both interfaces were judged to be equal in terms of learning, performance, emotional experience, and cognitive load.

1. Introduction

Speech production requires the precise coordination and synchronisation of respiratory, phonatory, and articulatory gestures, which are often accompanied by co-verbal and non-verbal facial, hand, and/or body movements (Wagner et al., 2014). One key aspect of speech is the control of intonation, which refers to the speaker's ability to modulate pitch over time for communicative purposes: to convey meaning, highlight contrasts, structure discourse (Mertens, 2008; Di Cristo, 2016), and express and emphasise attitudes (Ward, 2019) or emotions (Scherer, 2003). In most speaking situations, this modulation is produced unconsciously through the fine control of phonatory gestures, an ability acquired early during language development. However, intonational control can become effortful or degraded in certain contexts–such as second language learning or voice disorders–which means speakers must explicitly plan or relearn how to shape their pitch to fulfil communicative functions.

Following the recent development of Human Machine Interfaces, performative voice synthesis is an innovative research framework aiming to precisely control and study the transfer between implicit and explicit planning of speech gestures. It offers users real-time control over one or several parameters of the voice they generate. One configuration which has been studied in depth is the control of intonation by *chironomy* (from the Greek "ruled by hand (motion)"), where the synthetic voice intonation is "played" or controlled by the hands, like a (digital) musical instrument (d'Alessandro et al., 2005; d'Alessandro, 2022). Thus, by fetching intonation information from a non-vocal gesture (the hand), this approach relies on the strong assumption that the user is capable of transferring implicit voice intonation control when it has been naturally produced by the vibration of his/her vocal folds to an explicit control via hand gesture. This hypothesis has been validated in *imitation tasks*, with studies showing that a person can reproduce a given intonation contour even more precisely using hand gestures on a graphical tablet than with the human voice, whether through speech (d'Alessandro et al., 2011) or singing (d'Alessandro et al., 2014). The inverse transfer from explicit hand control to implicit natural control of intonation has also been observed and proved to be beneficial in learning typical intonation patterns in French (Xiao et al., 2022) and English (Xiao et al., 2023) as foreign languages. This high performance in chironomy for performative voice synthesis can be attributed to its intrinsic multimodal integration (vision, kinesthesis, and audition Perrotin and d'Alessandro, 2016), as well as to the existing

* Corresponding author at: Sigmedia Lab, School of Engineering, Trinity College Dublin, Dublin, D02 PN40, Ireland.
E-mail address: charuaud@tcd.ie (D. Charuau).

https://doi.org/10.1016/j.specom.2025.103344
Received 6 January 2025; Received in revised form 4 November 2025; Accepted 6 December 2025
Available online 15 December 2025

dexterity and precision of handwriting movements (in writing and drawing), which were repurposed for a new task.

Initially developed as tools for prosodic research and as a new family of digital musical instruments, performative voice synthesis could also foster key advances in vocal substitution. In the case of a degradation or absence of phonatory capacity in patients with laryngeal disorders, current medical solutions for replacing the defective or absent voice source involve injecting an artificial sound source into the vocal tract, either directly through the mouth or via the tissues of the neck, using an electrolarynx (Liu and Ng, 2007; Fuchs et al., 2016; Kaye et al., 2017; Ahmadi et al., 2018). This vibrator generates a substitute vocal source through which the user can articulate speech normally. Alternatively, a microphone can be used to pick up unvoiced speech (e.g., whispering (Toda et al., 2012; Perrotin and McLoughlin, 2020)), and the voicing can be reintroduced in real-time by voice synthesis played back on a loudspeaker. One of the main limitations of these systems is the lack of information available for intonation reconstruction, preventing the generation of the variations needed for speech structuring (Mertens, 2008; Di Cristo, 2016) or for expressing attitudes or emotions (Ward, 2019). To address this issue, we recently developed a performative synthesis solution based on real-time whisper-to-speech conversion. In addition to articulating the utterance with his/her own vocal tract, the user has the option of synchronously controlling their intonation via hand gesture (Perrotin and McLoughlin, 2020; Ardaillon et al., 2022).

In this paper, we aim to evaluate the ability of users to externalise the control of intonation for the realisation of prosodic functions in a voice substitution communicative context. This context relies on generating speech from whisper articulation and hand-controlled intonation using our performative whisper-to-speech system. In comparison to the application of performative synthesis to digital musical instruments or in foreign language acquisition, the voice substitution paradigm raises three research questions addressed in this study:

1. While the hypothesis of transfer between implicit and explicit control of intonation has been demonstrated in imitation tasks, these are rare in oral communication. In contrast, and to the best of our knowledge, the hypothesis of transfer of intonation control in more natural communicative conditions, where speakers must spontaneously generate prosodic patterns, has not yet been addressed. Therefore, *can a user produce a specific prosodic function with hand gestures following an elicitation from the communicative context only (i.e., without receiving any instruction or an intonation contour as example)?*
2. Most performative synthesisers developed for the previously-mentioned imitation tasks (Feugère et al., 2017; Locqueville et al., 2020) only require an intonation control, the phonetic content being predefined by the synthesiser. In voice substitution, the simultaneous control of articulation and intonation involve two prosodic dimensions: duration and pitch, respectively. Thus, *can a user coordinate an implicit duration control with an explicit intonation control to produce specific prosodic functions?*
3. These aforementioned performative synthesisers offer an intonation control with the position of an object (a stylus or finger) on a surface. This type of control heavily relies on the visual modality which, while beneficial for control accuracy (Perrotin and d'Alessandro, 2016), lacks ergonomics in oral communication situations. Different gestural interfaces which do not involve the visual modality will therefore be explored and compared. Specifically, *how do different types of manual and non-visual control compare in terms of accuracy and user experience?*

To address these questions, we centre our study on a well-documented prosodic function: contrastive focus. This function is particularly suitable for our investigation because it can be reliably elicited in communicative contexts (Dohen and Lœvenbruck, 2009). It involves variations in both duration and pitch (Dahan and Bernard, 1996), and has been shown to correlate with hand gestures (Leonard and Cummins, 2011). This echoes the broader literature showing that prosodic focus is often synchronised with co-speech gestures such as head nods, eyebrow raises, or hand gestures (Rochet-Capellan et al., 2008; Roustan and Dohen, 2010; Carignan et al., 2024). Beyond local effects, the realisation of contrastive focus can also influence prosodic organisation at the utterance level (Dohen and Loevenbruck, 2004). These multimodal correlations suggest that manual gestures can serve as a viable control modality for prosodic functions in voice substitution systems.

The rest of the article is structured as follows. In Section 2, we present the prosodic characteristics of contrastive focus, with particular attention paid to their manifestation in French, the language under investigation in this paper. We first describe its acoustic realisation, before examining the coordination between prosody and co-speech gestures as part of the multimodal nature of focus production. Then, we discuss existing gestural interfaces for intonation control, which leverage this multimodal coupling between manual gestures and prosodic realisation. Section 2 concludes with our three research hypotheses. Section 3 details our experimental protocol. The results are presented in Section 4 according to the three hypotheses. They are then discussed in Section 5, where we return to our three research questions listed above. Section 6 concludes the paper.

2. From natural to gestural control of intonation

2.1. Focus production

2.1.1. Acoustic realisation

The ability to highlight a specific element in an utterance, contrasting it with one or more alternatives, is a fundamental prosodic function known as contrastive focus. It is used to identify a specific item within a (theoretical) set of items, often to emphasise a specific choice or, in some cases, to correct a misunderstanding. For example, in (1), the response (B) places contrastive focus on the subject "John", correcting the assumption made in question (A).

(1) A: Did Paul eat Peter's cake?
B: **John** ate Peter's cake.

Across languages, the prosodic realisation of contrastive focus involves a set of acoustic adjustments that enhance the perceptual salience of the focused constituent and set it apart from the surrounding regions (Dahan and Bernard, 1996; Grice et al., 2017; Dohen and Loevenbruck, 2004). In French, focus marking is primarily conveyed through modulation of the fundamental frequency of oscillation (f_o), but also by changes in duration and intensity. Contrastive focus typically affects the entire focal constituent, which is marked by a characteristic pitch contour involving a sharp rise followed by a fall (Jun and Fougeron, 2000). This results in a prominent f_o peak that may extend across the full duration of the focused constituent. In French, these pitch movements are usually accompanied by increased intensity and by a significant lengthening of the focused word, particularly on its final syllable (Dahan and Bernard, 1996). These acoustic cues, f_o, intensity, and duration, do not operate independently; instead, they combine in a structured way at the syllable level to mark the entire focused constituent. Typically, f_o and intensity reach their peak early in the word, often on the initial or medial syllables, while increased duration occurs towards the final syllable (Astésano et al., 2004). This internal organisation distributes prosodic prominence across the constituent, reinforcing the salience of the focal marking as a whole.

Beyond the focal constituent itself, contrastive focus also extends to the adjacent regions: pre-focal and post-focal, which undergo prosodic adjustments. Understanding these peripheral patterns is essential, as they enhance the perceptual contrast around the focus and can be

further amplified in contexts in which primary cues such as f_o are unavailable (e.g., in whispered speech). In French, in the pre-focal region, the prosodic cues of focus are reduced. Specifically, the amplitude of f_o peaks is lowered, and overall variation in f_o is compressed (Dohen and Loevenbruck, 2004). Additionally, the duration of pre-focal regions is shortened (Astésano et al., 2004). These attenuations in the pre-focal region illustrate a prosodic balancing strategy, whereby prominence is downplayed on non-focused elements to enhance the contrast with the upcoming focused constituent. Similar tendencies have been reported in other languages; for instance, in German, Roessig (2023) describes comparable reductions as a "redistribution of prosodic resources". These cross-linguistic parallels suggest functional convergence, although the specific manifestations differ across languages. In the post-focal region, French usually exhibits a de-accentuation pattern, characterised by a flattening of tonal values and manifested as a low flat plateau, a gradual and delayed decline or a continuous decline in f_o extending to the end of the utterance (Delais-Roussarie et al., 2002; Jun and Fougeron, 2000). However, this reduction does not imply a complete prosodic neutralisation. Studies have shown that prosodic phrasing is preserved in post-focal regions, with boundaries between intonational phrases still clearly marked, particularly through temporal cues. For instance, both Delais-Roussarie et al. (2002) and Di Cristo and Jankowski (1999) demonstrate that final and initial lengthening are maintained in these regions, even in the absence of salient tonal movements. These results underscore the structural role of duration in prosodic phrasing, enabling speakers to preserve their organisation of the utterance while shifting the emphasis away from post-focus material.

The realisation of contrastive focus relies on prosodic features, particularly the modulation of f_o. However, whispered speech presents a significant challenge for focus marking due to the absence of vocal fold vibration, which renders f_o unavailable as a cue. In such conditions, speakers rely on alternative compensatory acoustic strategies to preserve prosodic functions and their perceptual salience. The loss of prosodic information associated with intonation in whispered speech can be partially compensated for by secondary cues, such as variations in formant contours (Pérez Zarazaga and Malisz, 2023) or adjustments in spectral tilt (Heeren and Van Heuven, 2014). These acoustic features facilitate the identification of syntactic boundaries, although the ability to use them varies depending on speakers and contexts. Duration can also serve as a compensatory cue to offset the absence of intonation, especially to identify the modality of an utterance: the lengthening of final syllables has been observed as a marker in interrogative intonation in French whispered speech, compensating for the lack of pitch-based contrasts (Vercherand, 2011). Yet, the natural slowing of the speech rate in whispered speech, which is notably characterised by the elongation of vowel duration (Sharf, 1964; Schwartz, 1967; Houle and Levi, 2020), suggests that temporal adjustments need to be carefully considered when examining prosodic strategies in this speech mode, as this slowing could interfere with prosodic strategies or bias comparisons with normal speech. Overall, these compensatory strategies illustrate the adaptability of the prosodic system in the absence of pitch-based cues such as boundary signalling and utterance modality, by adjusting other acoustic parameters, especially duration and spectral features.

2.1.2. Co-speech gestures

As was previously mentioned, speech communication is an interaction between speech production gestures (respiratory, phonatory, and articulatory) with movements of the hands, facial features, head, torso, or body, also referred to as co-speech gestures (Wagner et al., 2014). Interestingly, successive works have demonstrated that most of these movements are involved in the production of focus, including eyebrow raising (Cave et al., 1996), head nodding (Carignan et al., 2024), and hyper-articulation of the jaw, tongue and lips (Dohen et al., 2004; Pagel et al., 2024). Last but not least, and most pertinent to our study, hand gestures accompany focus in multiple ways. Deictic or pointing gestures share the same function of drawing attention to a specific entity as a focus in speech. In fact, a precise temporal coupling between pointing gesture and articulatory movements have been observed in the production of stress and/or focus in terms of synchronisation (Rochet-Capellan et al., 2008) and duration (Krivokapić et al., 2017). Beat gestures are fast down-up movements of the hand which co-occur with specific prosodic events in speech. Leonard and Cummins (2011) show that beat gestures are expected to happen before or in synchrony with pitch accents, and demonstrated a strong synchronicity between the gesture apex and the peak of the pitch accent on the stressed syllable. Tapping gestures are by contrast non-communicative, but strong magnitude and temporal coupling have also been highlighted between articulatory and tapping gestures in the production of stressed syllables (Parrell et al., 2011), suggesting a coupling between the two motor domains. Roustan and Dohen (2010) compare pointing, beat, and tapping gestures (achieved by pressing a button) in the production of contrastive focus. While all gestures are well coordinated with the production of focus, the variation in temporal coupling is lowest with gestures that share the same communicative function as the focus, i.e., pointing, and highest with non-communicative gestures like tapping. Roustan and Dohen (2010) also show that the choice of gesture does not influence the acoustic and articulatory correlates of focus.

Overall, these studies provide evidence of a strong motor coupling between speech and co-speech gestures, which exhibit a high degree of synchronicity in the focus realisation, with a multiplicity of candidate gestures associated with the latter. In the development of performative synthesisers, our goal is to exploit these correlations between co-speech gestures and prosodic functions to induce a causal relationship. That is, prosodic functions are triggered by co-speech gestures through manual control of intonation.

2.2. Gestural interfaces for intonation control

Manual control in Human-Computer Interaction is extremely varied. An early categorisation of control interfaces depending on the type of input gestures stems from work on classic computer controllers (e.g., mouse, keyboard, joystick, touch screens), with a three-dimensional organisational structure (Card et al., 1991): number of degrees of freedom (among three rotations and three translations), the physical quantity involved (position, force, or their derivatives) and its resolution (continuous, discrete, binary). In the particular case of the gestural control of sound and/or building digital musical instruments, Vertegaal et al. (1996) theorise optimal links between the type of gestural control and the type of acoustic property to be controlled. They encourage using linear position for absolute modifications of an acoustic dimension, and linear force for relative modulation. In the latter case, they favour *isometric* force, i.e., with varying force and without displacement such as pressure, over *isotonic* force, i.e., involving a movement with constant force such as using an accelerometer. Wanderley et al. (2000) and Marshall and Wanderley (2005) confirm these recommendations in the control of singing synthesis. In both experiments, users expressed preferences for using the linear position of a stylus to control absolute pitch, and an isometric pressure button for vibrato control (relative pitch modulation) when presented with various control options.

While intonation control in singing can be considered as absolute (targeting notes on a scale), it makes more sense to consider intonation control in speaking as a relative control, as intonation is mostly described as falling/rising patterns rather than reaching specific f_o values in Hertz. Thus, based on the aforementioned studies, linear position control should be more suited to performative singing synthesis, while force control should be better adapted to performative speaking synthesis. Nevertheless, and interestingly, a series of works has tackled speaking tasks subsequently to singing tasks. These authors first adopted a linear position control for singing tasks (stylus position: (Kessous, 2004; d'Alessandro and Dutoit, 2009; d'Alessandro

et al., 2014; Perrotin, 2015; Feugère et al., 2017), hand position in space: (Pritchard and Fels, 2006; Locqueville et al., 2020)). They then continued it for speaking tasks (stylus position: d'Alessandro et al., 2011; Evrard et al., 2015, hand position in space: Fels and Hinton, 1998). In the few studies evaluating them, these controls have proved excellent at imitating intonation patterns (d'Alessandro et al., 2011, 2014). Since reproducing intonation contours requires a high degree of precision in terms of control of intonation, it is actually a task consistent with the chosen linear control paradigm. However, as mentioned in the introduction to the present study, imitation tasks are rare in oral communication, so we shall instead focus on interaction tasks, where a participant is immersed in a simulated interaction, producing spontaneous speech in response to a well-defined communicative context. This brings us to a second strand of work, which has seen the development of dedicated systems to performative speech synthesis, where force control has been preferred, i.e., adapted for a relative control of intonation. The near-century-old Voder used an isotonic pedal for intonation control (Dudley et al., 1939). More modern medical devices for speech rehabilitation use pressure buttons (e.g. the Trutone electrolarynx[1]) or accelerometers (Matsui et al., 2013). Nevertheless, to the best of our knowledge, none has been rigorously evaluated in the production of specific prosodic functions. Moreover, no prior research has discussed the relationship between the control gestures induced by these interfaces, and the co-speech gestures observed in the realisation of prosodic functions. It is all the more interesting that some of these co-speech gestures, such as pointing, beating, or tapping (Roustan and Dohen, 2010; Leonard and Cummins, 2011) are mostly described by their dynamics, and therefore adapted to force-based controls for their acquisition. Thus, we propose to leverage the hand movements of a variety of co-speech gestures which accompany the production of focus in natural speech, for the explicit control of f_o in performative whisper-to-speech conversion. Specially, we will compare and evaluate the realisation of contrastive focus with two force-based controls, one adapted to beat gesture-like movements (isotonic), and the other to button pressure (isometric).

In order to address our research questions (see Section 1) in light of these findings, and to assess how isotonic and isometric gestural control of f_o can convey contrastive focus using whisper-to-speech conversion, we test the following hypotheses:

- **H1** *Focus realisation:* participants can coordinate whisper and gesture-based interfaces to reproduce key prosodic features of contrastive focus in French–namely, syllable lengthening and a local rise in pitch;
- **H2** *Beyond the focused syllable:* this prosodic control extends beyond the focused syllable to the realisation of broader utterance-level patterns;
- **H3** *Participant feedback:* the participant experience depends on the user, with differences expected across the type of control (isotonic vs. isometric) in terms of perceived control, effort, cognitive load, or synthetic voice quality.

H1 and **H2** relate to the first two research questions, examining whether contrastive focus can be reproduced locally and at the utterance level, through duration and pitch variations using manual gesture-based control in our experimental paradigm. **H3** addresses the third research question by evaluating the user experience across different gesture-based controls.

The next section describes our performative system and the experimental protocol used to test these hypotheses.

[1] https://www.atosmedical.com/products/provox-trutone-emote-2

3. Material and methods

We designed a behavioural experiment combining speech production tasks, acoustic analyses, and subjective feedback questionnaires. To test the effectiveness of the external control of f_o with two types of gesture against the natural control of f_o, participants were asked to perform the same interaction speech task in three speaking modes: natural voice, whispered voice with isotonic manual f_o control (using wrist movement), and whispered voice with isometric manual f_o control (using finger pressure). The latter two modes are collectively referred to as chironomic control. In the chironomic control tasks, participants whispered into a microphone while simultaneously controlling f_o through hand movements, using our real-time whisper-to-speech conversion system. They heard a resynthesised speech signal synchronised with their production, built from the whisper articulation and the hand-controlled intonation. The interaction task was designed with contrasting scenarios that elicited prosodic focus in a controlled yet natural dialogue setting, enabling direct comparisons between prosody in natural voice and via chironomic control. The location of the prosodic focus within each utterance was systematically varied, creating a range of linguistic conditions to examine the effectiveness of external f_o control relative to natural voice. Finally, the subjective feedback questionnaires were designed to verify whether the interfaces were perceived as usable and effective, and whether the task was considered appropriate, to rule out any potential confounding effects which may have arisen due to discomfort or cognitive overload. The following subsections detail the participants, experimental conditions, materials, and methods used in this study.

3.1. Participants

We recorded 20 participants, but only 16 were included in the final analysis as four did not comply with the instruction to keep their arm resting on the armrest during the wrist movement phase, which affected the f_o control ranges. This aspect is further discussed in Section 3.4. All the participants in the study were native French speakers and did not report any speech, hearing, or arm or hand motor impairments. The median age of the participants was 25 years old (Q1: 24 y.o., Q3: 33 y.o.). All the participants received compensation in the form of a 15€ gift card valid at major retail stores. This experimental protocol was approved by the ethics committee of the Université Grenoble Alpes (CERGA, number: CERGA-AVIS-2023-21) and complies with the General Data Protection Regulation (GDPR). The experiment was conducted in an anechoic chamber at our laboratory.

3.2. Experimental protocol

3.2.1. Speech material

Scenario. We designed a scenario to induce contrastive focus without giving explicit instructions, following the work of Dohen and Lœvenbruck (2009). This involved presenting a speech task in the form of a simulated interaction between the participant and the experimenter. This interaction, summarised in Table 1a, consisted of three parts, with the text displayed sequentially on a screen in front of the participant. The participant started by reading the first utterance displayed on the screen. Then, a pre-recorded question asked by the experimenter was played back while also being displayed on the screen. This question repeated the participant's statement, changing one word to simulate a misunderstanding. Finally, the participant was instructed to repeat the initial utterance, again displayed on the screen. For the purposes of our analysis, we distinguish between the first and second realisation of the utterance, and call this the ***participant turn*** condition. The first realisation is labelled the baseline turn, while the second, produced in response to the experimenter's question, is called the corrective turn.

Table 1
Example of scenario (top) and the full corpus of utterances (bottom) used in our experiment.

	Scenario	Participant turn
– Participant :	Le loup doux a suivi le beau loup.	baseline
– Experimenter:	Le loup doux a suivi le beau chien?	
– Participant:	Le loup doux a suivi le beau loup.	corrective
Word-by-word translation: The wolf gentle followed the beautiful [*wolf./dog?/wolf.*]		

(a) Scenario of the experiment. Both the underlining which shows the syllable targeted by the experimenter's question, and the colours that indicate syllable status (baseline non-target (light red); baseline target (dark red); corrective non-target (light green); corrective target (dark green)) are only displayed in this Table for explanatory purposes–they were not visible to the participants.

#	Syllable location		Utterance			Contrast
	target	non-target	*Subject (S)*	*Verb (V)*	*Object (O)*	*Word changed in the question*
U1	S1	O2	*Lou* du Mans	a suivi	le *loup* doux.	Jean
U2	S2	O3	Le *loup* doux	a suivi	le beau *loup*.	chat
U3	S3	O1	Le beau *loup*	a suivi	*Lou* du Mans.	chien
U4	O1	S3	Le beau *loup*	a suivi	*Lou* du Mans.	Jean
U5	O2	S1	*Lou* du Mans	a suivi	le *loup* doux.	chat
U6	O3	S2	Le *loup* doux	a suivi	le beau *loup*.	chien

(b) Testing corpus utterances. /lu/syllables are in italic (note that *"Lou"* and *"Loup"* are both pronounced /lu/), and the underlined syllable is targeted by the experimenter's question.

Corpus. The testing corpus utterances used in the scenario are summarised in Table 1b (as opposed to the training corpus, presented in Section 3.2.2). It is composed of three sentences of nine syllables following a Subject-Verb-Object (SVO) structure, with three syllables per constituent. The verbal constituent is identical for all sentences and the subject and object constituents are each made of three fully voiced consonant-vowel monosyllabic words (to avoid voicing decision errors in whisper-to-speech conversion). Each of these two constituents contains the syllable /lu/, which may be spelled either as "Loup" or "Lou", and each sentence appears twice in the corpus where either the first or the second /lu/is the target syllable, i.e., that which was changed by the experimenter. We call the other /lu/the non-target syllable. In total, the corpus is composed of six utterances, each with the target syllable occurring in a different position within the utterance. We call this the ***syllable location*** condition, with levels: S1, S2, S3, O1, O2, O3.

Interaction task. During each interaction scenario, the syllable /lu/was pronounced four times by the participant. We define each production with the ***syllable status*** condition following two dimensions: the /lu/syllable position in the dialogue (baseline *vs.* corrective), and whether it is expected to be focused or not (target *vs.* non-target). This results in four syllable status levels highlighted in Table 1a: baseline non-target (light red); baseline target (dark red); corrective non-target (light green); and corrective target (dark green). Since no instruction other than reading each utterance was given, we could hypothesise that the participant would naturally produce a contrastive focus in the corrective target condition only. In the following section, we call the performing of three repetitions of the six scenarios corresponding to each utterance the "interaction task". The presentation of the 18 stimuli was in random order, i.e., the repetitions were not consecutive.

3.2.2. Experimental tasks and training

The experiment was divided in three phases, summarised in Fig. 1. Each corresponded to a level of the speech ***production mode*** condition (natural voice, finger pressure, and wrist movement) and started with one or several familiarisation procedures to the protocol and/or control. During the first phase, the participant used his/her natural voice and began with a training interaction task, also including three repetitions of six scenarios, but with different utterances to those shown in Table 1b and reported in Appendix A.1. Then, the participant performed the interaction task as described in Section 3.2.1. During the second and third phases, the participant used the whisper-to-speech conversion system with a wrist movement then finger pressure control, or inversely. The order of these two phases was randomly assigned for each participant, while ensuring that half were performed with wrist movement first, and vice versa. For each of these two phases, the experimenter presented the system and conducted a brief system check with the participant to ensure that the interface worked properly. Then, the familiarisation phase with the interface began, with a reading task of the *MonPage 2* text (Pommée, 2021, p. 114,reported in Appendix A.2) with free control of intonation. This initial task allowed the participant to freely explore the interface and the synthesised voice output without constraints, facilitating an initial first intuitive grasp of the system before more controlled training. The second training step was a training imitation task to practice the intonation control, in which the participant was asked to imitate three sentences, each produced with three different intonation patterns (declarative; interrogative; with focus, see Appendix A.3). These nine utterances were pre-recorded by d'Alessandro et al. (2011) and include a male and a female speaker. Stimuli with the same gender as the participant were presented in the experiment, to better match his/her f_o range. The nine utterances were presented in random order, with three consecutive repetitions each (27 stimuli in total). We included this task to ensure that participants had a concrete guided opportunity to experience how gesture-based input maps onto a variety of f_o modulations before engaging in more open interaction. As last training step, the participant was given a training interaction task (the same protocol and training corpus as phase 1). Finally, both second and third phases ended with an interaction task.

3.2.3. Feedback questionnaires

After the second and third phases, the participant filled out a questionnaire presented as an online form to provide feedback on the use of each gestural control he/she had just performed with. The questions are inspired by the Intrinsic Motivation Inventory (IMI), a multidimensional self-assessment of participants' subjective experience related to a laboratory experiment (Ryan, 1982; Ryan et al., 1983).[2] We selected 23 questions to collect participant feedback regarding his/her experience with the interface, the experimental task, and the voice synthesis. Each question used a five-level absolute categorical rating scale, and all are reported in Appendix B.

[2] https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

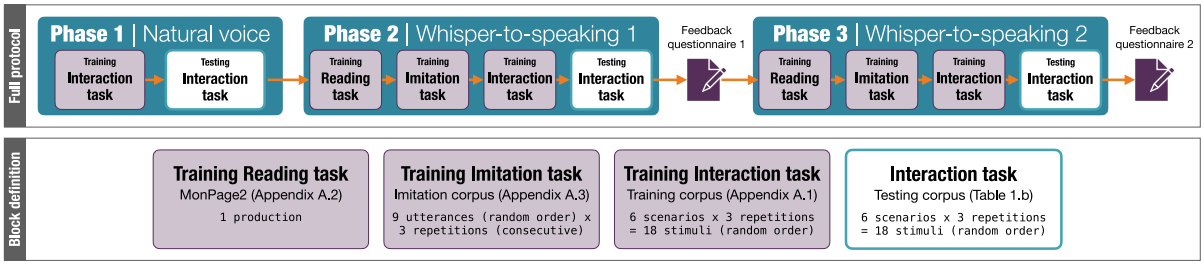


Fig. 1. Top: block diagram of the experimental protocol. Bottom: description of each experimental block. The training tasks are shown in purple, while the interaction tasks (as the focus of the present study) are shown in white.

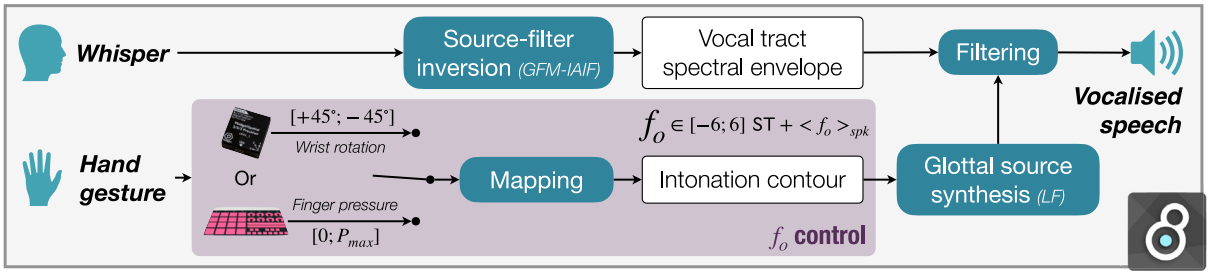


Fig. 2. Diagram of the voice substitution system with manual intonation control. The gesture control range (±45° rotation for wrist movement; 0 to maximum finger pressure P_{max} for finger pressure) is linearly mapped to an octave (±6 ST) around the speaker's average fundamental frequency, noted as < f_o >$_{spk}$. The resulting f_o is used for the resynthesis of the speech signal.

3.2.4. Experimental conditions

Each participant was seated in front of a table on which a screen displayed instructions and scenarios/utterances for various tasks. A Beyerdynamic DT979 headset with microphone was used to record the participant's voice and to provide continuous real-time auditory feedback from their natural voice in Phase 1, and from their synthesised voice in Phases 2 and 3. The same headset was also used to play back the pre-recorded experimenter's audio stimuli during the interaction tasks. In all phases, the participant had to press a key on a keyboard to pass to the next stimuli, and they could rest as long as they wanted before doing so. The full experiment lasted 1h15 on average.

3.3. Whisper-to-speech voice substitution system

The system used in Phases 2 and 3 (presented in Fig. 2) comprised a whisper-to-speech conversion module and gestural interfaces for intonation control. The conversion module generates speech from both the input whisper articulation and the intonation controlled by hand, and the synthetic speech is played back to the participants in synch with their whisper productions to create a closed speech production loop. The system is detailed below.

Whisper-to-speech conversion. The system is an extension of the method proposed by Perrotin and McLoughlin (2020), which consists of: (1) the source-filter decomposition of the whisper input by the Global Flow Model-based Iterative Adaptive Inverse Filtering (GFM-IAIF) method (Perrotin and McLoughlin, 2019), to isolate the vocal tract spectral envelope from the coloured noise corresponding to the whisper sound source; (2) the generation of a glottal source signal via the Liljencrants-Fant (LF) model (Fant et al., 1994), with a fundamental frequency f_o controlled by hand gestures through the interface; (3) the filtering of the glottal source signal by the vocal tract spectral envelope. These three steps are implemented in real-time on the Max/MSP sound processing platform.[3]

Gestural interfaces. The f_o used in synthesis is linearly controlled on a semitone scale (ST) over an octave (±6 ST) around the speaker's average f_o, noted < f_o >$_{spk}$. The latter was measured during the training interaction task of Phase 1, with natural voice (see Fig. 1). We proposed two types of gesture for intonation control: an isometric finger pressure and an isotonic wrist movement. Intonation control by thumb pressure is proposed in the commercialised and widely-used Trutone electrolarynx solution. In our experiment, the finger pressure gesture was performed on a Morph tablet from the Sensel brand,[4] which measures the pressure of the index finger of the user's preferred hand, from zero pressure to a maximum pressure P_{max}, respectively associated with −6 ST and 6 ST around < f_o >$_{spk}$. The Sensel Morph was placed flat on the table. The wrist rotation gesture was inspired by the beat gestures that can co-occur with focus (Leonard and Cummins, 2011). The wrist movement was measured with a 1044_1B accelerometer from the Phidget brand,[5] held in the primary hand of the participant, whose forearm lay horizontally on the chair's armrest. The accelerometer measured the top/down wrist movement in rotation degrees, 0° being the horizontal position of the wrist corresponding to < f_o >$_{spk}$; 45° and −45° being mapped to −6 ST and 6 ST around < f_o >$_{spk}$, respectively. Thus, moving the wrist down corresponds to an increase of f_o. For the wrist movement phase, the participant must keep his/her forearm on the chair armrest to ensure that the wrist movement remained in the correct [45°;−45°] interval, and we excluded from our analyses those participants who failed to follow this instruction.

3.4. Data processing

Although data from all experimental blocks were recorded, only the data generated by the interaction tasks are analysed below. We relied on the participants' voice recordings in all conditions and on the f_o contour measured by the interfaces in the chironomic control condition.

[3] http://cycling74.com

[4] https://morph.sensel.com

[5] https://www.phidgets.com/docs/Accelerometer_Guide

3.4.1. Data extraction

The speech-to-text alignment of the audio files was performed using the Astali application (Loria, 2016). The segmentation into syllables and their annotation (target/non-target) was manually done using Praat. For each syllable, we reported its relative duration (D_r) in relation to the utterance's duration, excluding pauses. With each utterance consisting of nine syllables, the average relative duration of a syllable was 11%. The articulation rate was calculated as the number of syllables divided by the duration of the utterance without pauses. The fundamental frequency (f_o) of natural voice was measured automatically using Praat's `To pitch` function, while the f_o controlled by gesture was provided directly by the system. These values are expressed in semitones (ST). To remove the variability in speaker and production mode, we subtracted from each f_o contour the median f_o calculated for all productions of the corresponding speaker and production mode. We refer to the resulting centred fundamental frequency as f_{oc}, also expressed in ST. In addition to full contours, we also report the peak of f_{oc} for the syllables of interest (target and non-target). Last, for all focused syllables (corrective target), we measured the peak position relative to the boundaries of the syllable, with 0 % being the beginning of the syllable and 100 % the end of the syllable. To account for possible early or late alignment, we also include values below 0 % (when the peak occurs in the preceding syllable) and above 100 % (when it occurs in the following syllable). When reporting the results in the next Sections, SD refers to Standard Deviation.

3.4.2. Statistical analyses

Regression models on extracted data and feedback questionnaires. The impact of syllable status (baseline non-target, baseline target, corrective non-target, corrective target; see Table 1), syllable location (S1, S2, S3, O1, O2, O3; see Table 1b) and production mode (natural voice, finger pressure, wrist movement; see Fig. 1) was investigated on relative duration D_r, f_{oc} peak and its position. As relative duration is bounded in the interval [0;1], a beta regression with random effect was applied using the `glmmTMB` function from the R `glmmTMB` library. For f_{oc} peak, a mixed-effects model was employed using the `lme` function from the R `nlme` library. In both cases, the participant and repetition number were included as random effects in the model. Multiple comparisons were conducted using the `glht` function from the R `multcomp` library, from which the *p*-values given below were derived, allowing us to test the significance of the results pairwise. When the results are presented as significant, this means that all tested pairs were significant. Otherwise, we detailed the significant and non-significant pairs. The overall significance level was set to $p < 0.05$. In the particular case of the analysis of the absolute duration of utterances (Section 4.1.1), a mixed-effects model was employed with participant turn (baseline, corrective), production mode and utterance (U1, U2, U3, U4, U5, U6) as fixed factors, and participant and repetition number as random factors.

The impact of the chironomic control on responses to the feedback questionnaire was also studied. The questionnaire consisted of 23 Likert-scale items with ordinal data ranging from 1 to 5. For each question, we analysed whether the distribution of the responses differed depending on the interface. An ordinal regression with random effect was consequently applied using the `clmm` function of the R `ordinal` package. We included both the production mode and order (whether the interface had been performed first or second by the participant) as fixed factors, and participants as random effects in the model. A pairwise comparison was conducted with estimated marginal means using the `emmeans` function from the R `emmeans` library. The overall significance level was again set to $p < 0.05$.

For each statistical model described above, the effects of individual factors and their interactions were tested by removing them one at a time from the full statistical model, and assessing if the removal of each factor had a significant impact on the model. We started with the random factors, then proceeded to interactions between factors. Only if the latter were non-significant were the factors involved in those interactions removed. For this, we used both likelihood ratio tests with the R `anova` function and AIC criterion with the `dredge` function from the R `MuMIn` library.

Generalised additive mixed models (GAMMs) on intonation contours. GAMMs were used for a pairwise comparison of the intonation contours (f_{oc}) between two levels of a given fixed factor, with participants as a random factor. This was done on time-aligned intonation contours, i.e., interpolated trajectories so that the syllable centres would be equidistant, as shown in Fig. 4. GAMMs fit time-dependent data, and thus not only provide a global significance of one factor on the overall data trajectory, but also indicate when the difference between the trajectories is significant. The procedure detailed by Sóskuthy (2021) was followed, in fitting one model with the effect of the factor in consideration and another without. The factor has a significant impact on f_{oc} if the difference between the two models is significant. When the factor had a significant effect, we reported the time location of the significant differences between intonation contours (see, for example, Fig. 4). The R `mgcv` and `itsadug` packages were used for this test. The overall significance level was also set to $p < 0.05$ after Bonferroni correction. We tested two factors: the effect of utterance, by comparing one pair of utterances at a time among baseline turns for each production mode; and the effect of the participant turn (baseline vs. corrective) for each utterance and production mode.

4. Results

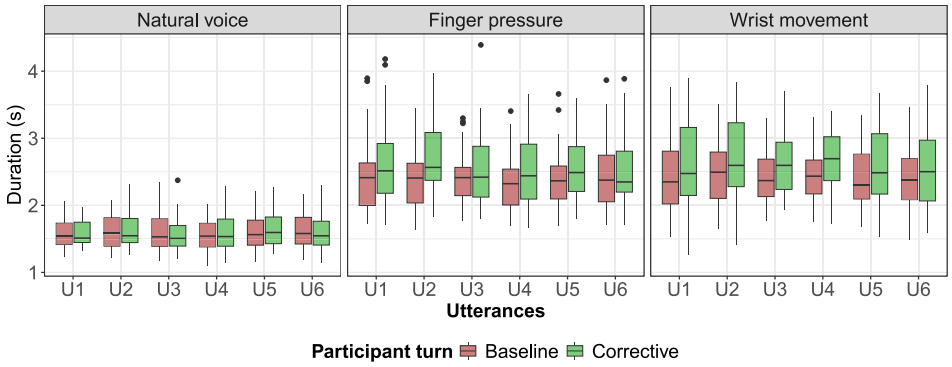
In this section we present an analysis addressing our three hypotheses set out in Section 2.2. First, Section 4.1 examines the overall impact of the external control context on utterance-level prosody. Before engaging in a more specific analysis in the following sections, we investigate whether the use of chironomic control and whisper-to-speech conversion affects global utterance properties, such as absolute duration and f_o variation, compared to natural voice. Section 4.2 investigates the use of gesture-based interfaces to reproduce the key prosodic features of contrastive focus in French (H1). This includes a quantitative analysis of the relative duration D_r and the f_{oc} peak on /lu/syllables. Section 4.3 extends the scope to utterance-level patterns (H2), analysing the impact of the contrastive focus context on the dynamics of the full intonation contour. Finally, to evaluate the participant experience with each interface (H3), we conclude with an analysis of the feedback questionnaires in Section 4.4.

4.1. Global effect of the production mode

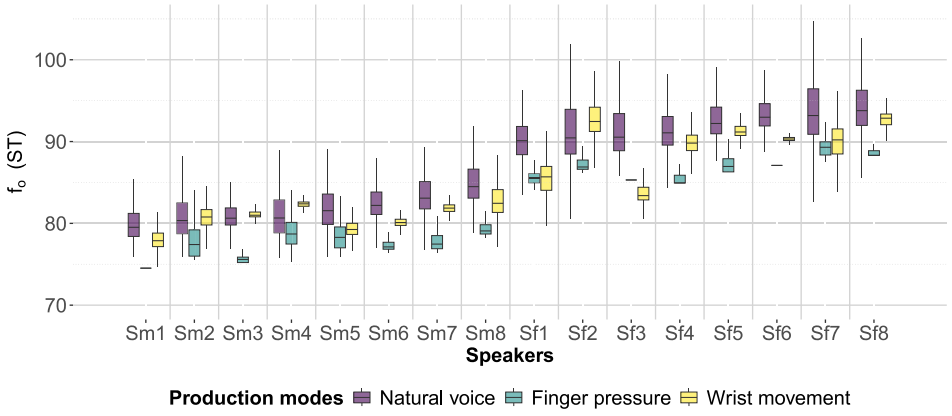
4.1.1. Effect of duration

Fig. 3(a) displays the distribution of the utterance duration in both the baseline and corrective participant turns for the three production modes. One striking observation is the significant lengthening ($p < 0.05$) of utterance duration in both chironomic control conditions compared to natural voice, for each utterance and both baseline and corrective participant turns. No significant difference was observed for any utterance between the two chironomic controls ($p > 0.05$). Among the baseline participant turns, the average utterance is 1.60 s (SD = 0.24 s) with natural voice, 2.40 s (SD = 0.45 s) with finger pressure, and 2.45 s (SD = 0.45 s) with wrist movement. Among corrective participant turns, the average utterance duration is 1.61 s (SD = 0.25 s) with natural voice, 2.57 s (SD = 0.52 s) with finger pressure, and 2.64 s (SD = 0.54 s) with wrist movement. This increase in utterance duration is consistent with the decreased speaking rate observed in whisper speech (Sharf, 1964), which is used in both chironomic control conditions. We measured an average of 5.4 syllables/s (SD = 0.8 syllables/s) with natural voice in comparison to an average of 3.8 syllables/s (SD = 0.7 syllables/s) with finger pressure and 3.7 syllables/s (SD = 0.8 syllables/s) with wrist movement.

These results also indicate a significant increase in utterance duration in the corrective participant turns compared to the baseline participant turns for both chironomic controls and all utterances ($p < 0.05$), but no significant difference was observed for any utterance in the natural voice condition ($p > 0.05$). This may reveal different



(a) Duration distribution (in seconds) in baseline and corrective participant turns per utterance and production mode.



(b) Fundamental frequency distribution (ST) per speaker and production mode

Fig. 3. Global distributions of duration per utterance and production mode (top) and f_o per participant and production mode (bottom).

strategies in the focus realisation in the corrective condition between *voice* and *chironomy*, which will be investigated further in Section 4.2. The absence of significant duration differences between the baseline and corrective participant turns in the natural voice modality may reflect a compensation mechanism between pre-focal shortening and post-focal lengthening, resulting in a globally stable utterance duration despite the presence of focus, a characteristic effect of French prosody.

4.1.2. Effect of f_o

Fig. 3(b) ranks the distributions of produced f_o (i.e., before centralisation) per speaker and production mode, in ascending order. We observe two continua of distributions spaced about an octave (12 ST) apart with natural voice, corresponding to male (resp. female) distributions on the left (right) of the Figure, and identified by *m* (resp. *f*) in the speakers' indices. Distributions of both chironomic controls follow natural voice distributions for each speaker (as a consequence to the centring of each chironomic control f_o range around the speaker's average $< f_o >_{spk}$), but these are not perfectly aligned. This reveals that the full octave range offered by each chironomic control was not systematically exploited by the participants. In particular, finger pressure distributions are skewed towards the minimum f_o allowed by the interface, which corresponds to an absence of pressure: i.e., the resting position of the interface. In contrast, the resting position of the wrist movement (horizontal hand) is mapped to $< f_o >_{spk}$, explaining the better alignment of the wrist movement median f_o with the natural voice median f_o. For three participants (Sm1, Sf3, Sf6), the finger pressure distributions appear as flat boxplots, displaying only the median. These speakers applied no pressure on the interface except when producing focus, resulting in a quasi-constant f_o in non-focus positions. This dominant value implies that limited variation during focus production remains invisible in the boxplot. In contrast, participants who maintained continuous pressure on the interface generated more varied f_o values, leading to more balanced and spread distributions.

Overall, the results highlight differences of global use between natural voice and chironomic control which are intrinsically linked to the different speaking modes (modal vs. whisper), and intonation control (implicit vs. explicit; differences between mapping). These differences also exhibit substantial inter-speaker variability. We will explore whether this affects the focus realisation in the next section.

4.2. Focus realisation

Fig. 4 displays smooth representations of f_{oc} contours per utterance (rows), production mode (columns) and participant turn (red: baseline; green: corrective). Each contour is obtained by, first, normalising the time of each f_{oc} production so that syllable centres are equidistant; and second, by extracting the median (plain line) and the 2nd to 3rd quartiles (the shaded area) of all productions across the participants and repetitions.

Validity of the protocol. We first examine the difference between the intonation contours of similar utterances among the baseline participant turns (U1 vs. U5, U2 vs. U6, U3 vs. U4; see Table 1b). If the GAMM statistical test on the utterance factor showed significant differences between each pair of utterances for all production modes, then the difference between intonation contours is never higher than 1.4 ST, which is below the threshold of perceptual salience in dynamic pitch variations in speech, measured at 2 ST for subjects able to discriminate pitch differences ('t Hart, 1981). This indicates that participants were extremely consistent in reproducing similar intonation contours across

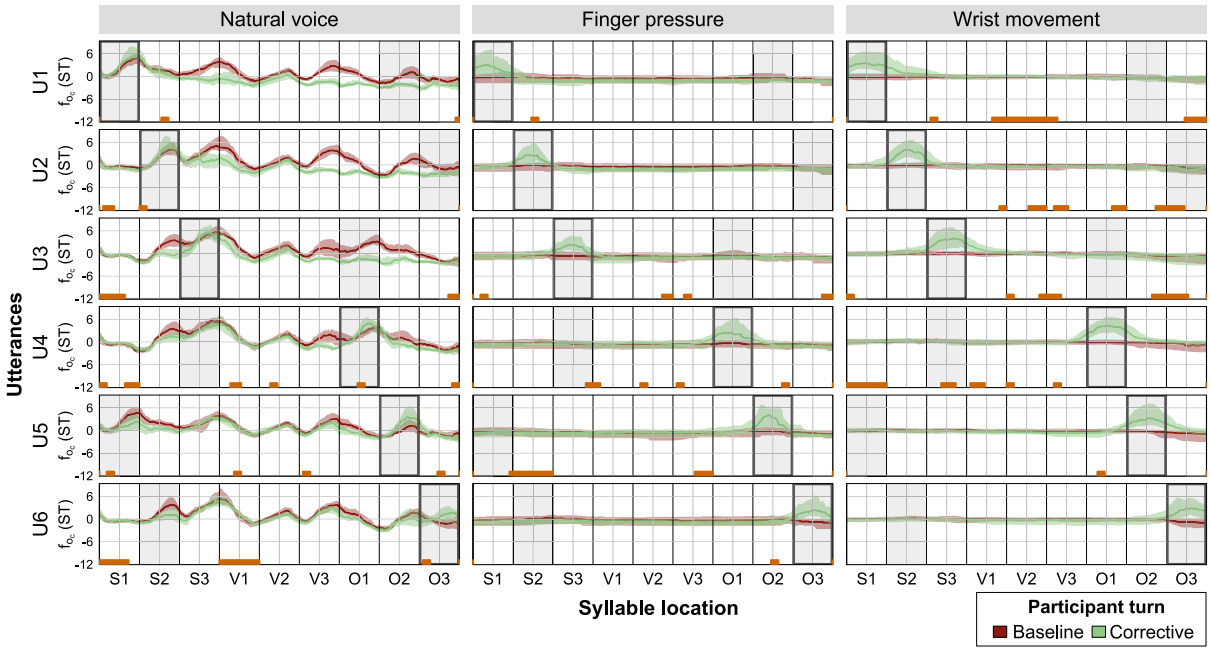


Fig. 4. The median (plain line) and 2nd to 3rd quartiles (shaded area) of syllable-aligned f_{0c} contours per utterance (rows), production mode (columns) and participant turn (red: baseline; green: corrective) across participants and repetitions. The solid grey rectangles with thick and thin edges on each utterance indicate the target and non-target /lu/syllables, respectively. The thick orange lines on the x-axes indicate portions of f_{0c} contours where baseline and corrective realisation of the same utterance are NOT significantly different according to the testing of the participant turn factor with GAMMs (similar utterances: U1 *vs.* U5; U2 *vs.* U6, and U3 *vs.* U4).

scenarios when presented with the same text content, without any anticipation of the syllable targeted in the following question. When comparing baseline and corrective realisations of the same utterance, i.e., within a scenario, the GAMM statistical test on the participant turn factor showed globally significant differences for all the utterances and production modes. Fig. 4 displays the local portions of NON-significance between baseline and corrective intonation contours with orange thick lines on the x-axes. The latter are significantly different for all the targeted syllables highlighted with thick grey rectangles, with the exception of small portions (< 30%) in syllable O1 in utterance 4 and in syllable O3 in utterance 6 with natural voice. Therefore, a contrast was effectively performed by the participants on the target syllable between the baseline and corrective participant turns, as is further detailed in the next sections. The absence of anticipation of focus production on the baseline condition between scenarios on the one hand, and the significant differences observed between the baseline target and corrective target syllable within scenarios on the other hand, demonstrate that we correctly introduced a relation of dependence between the experimenter's question and the realisation of contrastive focus in the corrective condition only.

In the remainder of this section, we perform a quantitative analysis of the extent to which the participants were able to reproduce the expected prosodic markers of contrastive focus in French while using the manual interfaces, namely a lengthening of the focused syllable and a local rise in f_0 contour. Based on previous descriptions of contrastive focus in French, we consider three key prosodic cues as dependent variables: (1) the duration of the syllable, (2) the height of the f_{0c} peak, and (3) its alignment with the syllable boundaries. While duration and f_0 rise are well-established markers of focus, the alignment of the f_0 peak can indicate whether the prosodic gesture is properly synchronised with the segmental structure, as observed in natural speech. Fig. 5 displays the distribution of relative duration D_r (top) and f_{0c} peak (middle) of all /lu/syllables, depending on their syllable location in the utterance, their syllable status and production mode. It should be made clear that for the same syllable location, the target and non-target syllables belong to distinct utterances and scenarios.

4.2.1. Focus realisation with articulation: effect on duration

The main difference between the duration of the baseline target (dark red) and corrective target (dark green) /lu/in the top of Fig. 5 is an elongation of the corrective target syllable in all conditions, and with a larger lengthening with the chironomic controls than with natural voice. The global evolution of the average relative duration of /lu/syllables from baseline target to corrective target across all syllable locations is: from 12.3 % (SD = 1.9 %) to 16.1 % (SD = 2.9 %) with natural voice, from 12.1 % (SD = 3.1 %) to 18.1 % (SD = 4.5 %) with finger pressure, and 12.3 % (SD = 2.9 %) to 18.3 % (SD = 4.4 %) with wrist movement. The differences between the baseline target and the corrective target /lu/are significant for all the production modes and syllable locations ($p < 0.05$). These results show that focus is well achieved at the articulatory level, as natural articulation was involved in all production modes.

4.2.2. Focus realisation with manual control: effect on f_0

In line with the effect of duration, Fig. 5 (middle) displays a systematic increase of f_{0c} peak on the /lu/syllables bearing focus, i.e., between the baseline target (dark red) and corrective target (dark green) syllable status, in all the production modes. With natural voice, the participants display a global rise of f_{0c} peak of 1.64 ST from the baseline target to the corrective target /lu/across all syllable locations, and the baseline target to corrective target differences are significant ($p < 0.05$) only when the focus is on the second syllable of the object constituent (O2). With chironomic control, a larger global rise of f_{0c} peak is observed between the baseline target and the corrective target /lu/across all syllable locations: 3.05 ST with finger pressure and 4.59 ST with wrist movement. The differences between the baseline target and the corrective target /lu/are significant for all syllable locations ($p < 0.05$). In light of these results, the local rise in f_{0c} peak to mark a focus is well achieved with chironomic control.

4.2.3. Coordination between articulatory and manual gesture

Now that we have demonstrated the focus realisation in terms of duration with natural articulation and f_0 with either natural or

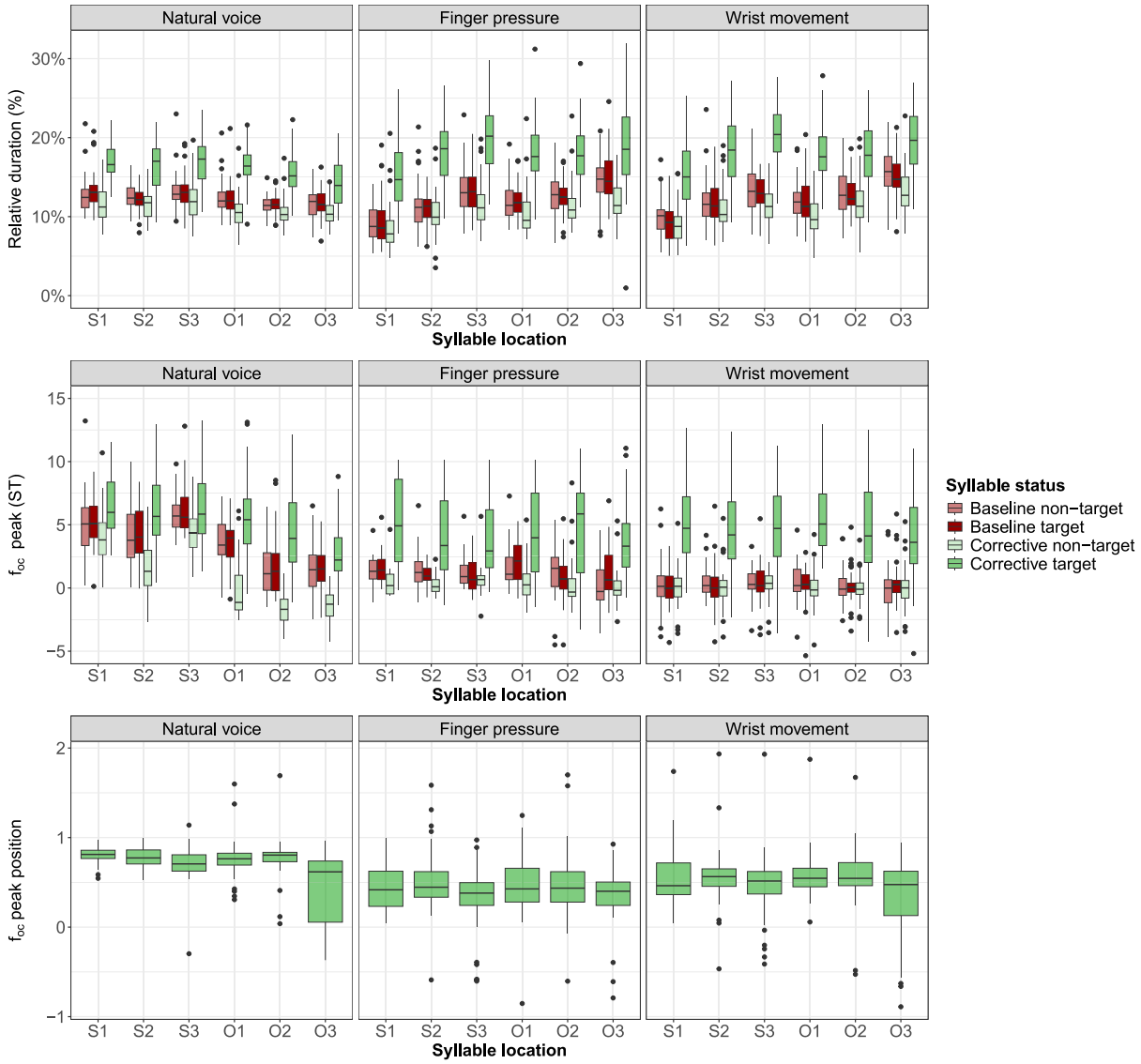


Fig. 5. Relative duration D_r (top) and f_{oc} peak (middle) of the syllables /lu/, and f_{oc} peak relative position to the syllable (bottom), according to their syllable location in the utterance, syllable status, and production mode.

hand control intonation, the remaining question is whether their coordination has been affected by the introduction of manual control. The bottom of Fig. 5 displays the f_{oc} peak position relative to the boundaries of the /lu/syllables bearing focus (corrective target syllable status), according to their syllable location in the utterance, and production mode. With natural voice, the f_{oc} peak tends to be located towards the end of the syllable, at an average of 70.3 % across all syllable locations, with a small dispersion except for the last syllable of the object constituent (O3). Also, the peak appears slightly earlier when the focus falls on the last syllable of the constituent (S3 and O3). Chironomic control production modes are characterised by an earlier but stable f_{oc} peak position on the focused syllables, with an average of 42.7 % with finger pressure and an average of 51.9 % with wrist movement, across all syllable locations. The dispersion of peak position is also greater than with natural voice. Interestingly, we also observe with chironomic control an anticipation of peak position on the last syllables of the constituents (S3 and O3). In this case, the average peak positions are as follows: 31.5 % and 33.5 % with finger pressure, and 47 % and 35.4 % with wrist movement at S3 and O3, respectively.

Overall, f_{oc} peaks are mostly realised within focused syllables, regardless of the production mode and syllable location, thereby demonstrating a successful coordination between articulation and intonation control in both implicit and explicit control of intonation.

Conclusion. In view of the results presented in this section, we can conclude that the successful transfer of focus production through the variation of duration and f_o on the target syllable was achieved, hence validating H1. In chironomic control conditions, all participants clearly realised focus with a raise of f_o peak at the expected location. This demonstrates that the participants not only perceived the importance of f_o in achieving focus, but also managed to use the interfaces to emphasise the target syllable. This behaviour was observed in all our participants. In parallel, a significant lengthening of the corrective target syllables was observed, similar to focus production with natural voice.

4.3. Beyond the focused syllable: realisation of the utterance intonation pattern

In this section, we aim at qualifying the participants' performance in the production of prosodic variations across full utterances (H2).

We first analyse the influence of the utterances' prosodic variations on focus realisation, i.e., the effect of syllable location on both duration and f_o variations on the target syllable from baseline to corrective participant turn. Second, we examine the effect of the focus realisation on the full intonation contours, i.e., the impact of the target syllable on pre- and post-focal regions in the corrective condition.

4.3.1. Impact of utterance-level prosody on the production of focus

The effect of the utterance-level prosodic variations on focus realisation is reflected in the variations of D_r and f_{oc} on the /lu/syllable with baseline target syllable status, i.e. without bearing focus, according to the syllable location. It should be noted that we could equivalently study baseline non-target syllables, as we showed in Section 4.2 that the difference between the baseline target and the baseline non-target is below the threshold of perceptual salience. The question is whether the prosodic values of one non-focused syllable (baseline target) has an effect on its focus (corrective target). Understanding this interaction can provide insights into how local prosodic contexts modulate focus marking within the overall prosodic contour of an utterance.

Duration. The duration of baseline target syllables with natural voice (top of Fig. 5), remains relatively stable across all constituents (i.e., there are no significant differences between syllable locations), which translates into a similar stability of the duration of corrective target syllables across the utterance, but with a slight shortening of syllables at the end of the object constituent (the only significant differences are between O1 and O2, and O1 and O3). In general, this confirms an isochronous production of /lu/syllables in all syllable statuses and syllable locations with natural voice. Inversely, we observe an increase in duration of the baseline target syllables within both subject and object constituents for both chironomic control conditions. This evolution is significant between all syllable locations with wrist movement, and between all syllable locations of the subject constituent, and only between O1 and O3 within the object constituent with finger pressure. This behaviour is mirrored in the corrective target syllables with both *chironomic controls*, the increase of duration being statistically significant within the subject constituent but not within the object constituent.

Intonation. Pairwise comparison of the six f_{oc} peak distributions depending on their syllable location in the baseline target condition (middle of Fig. 5) reveals 11/15 significantly different pairs, including all pairs in the object constituent and the S2-S3 pair. In the absence of focus (baseline target), this shows a strong and consistent dependence of f_{oc} variation with the syllable location, with a marked drop of f_{oc} at the end of the object constituent, matching a decreased intonation at the end of the utterance. By contrast, the syllable location has little effect on f_{oc} variation on corrective target syllables: only 6/15 pairs are significantly different, none of which except O2-O3 are within the same constituent. This reduced variation in the corrective target syllable f_{oc} suggests both an influence of the global decrease of intonation on the focus realisation at the end of the utterance, and a ceiling of f_{oc} peak in the production of focus at the beginning of the utterance. With chironomic controls, there is no statistically significant effect of the syllable location on f_{oc} peak, either in baseline target or in corrective target syllable status. Therefore, apart for the focus realisation, there is little intonation variation within utterances, meaning that the participants prioritised the production of focus to the detriment of other prosodic functions.

Conclusion. Overall, these observations show, first, that in the baseline target condition, i.e., without focus, /lu/syllables tend to be isochrone with varied intonation according to the syllable location with natural voice, and inversely isotone with varied duration according to the syllable location with chironomic control. In all cases, and for both duration and intonation, we highlight a strong correlation between their variations in the baseline target condition with those of the corrective target condition. In particular, the duration of the corrective target syllables is isochrone with natural voice and increasing within constituents with chironomic control. The f_{oc} peaks of the corrective target syllables are isotone with chironomic control and slightly decreasing at the end of the utterance with natural voice. Moreover, we notice a saturation in the values of f_{oc} peak on corrective target syllables for all production modes. These observations reinforce the previous findings by showing that although variations of duration and intonation differ between natural voice and chironomic control in the absence of focus, the influence of syllable location on focus realisation is effectively transferred from implicit to explicit control of prosodic variations.

4.3.2. Impact of focus realisation on the utterance-level prosody

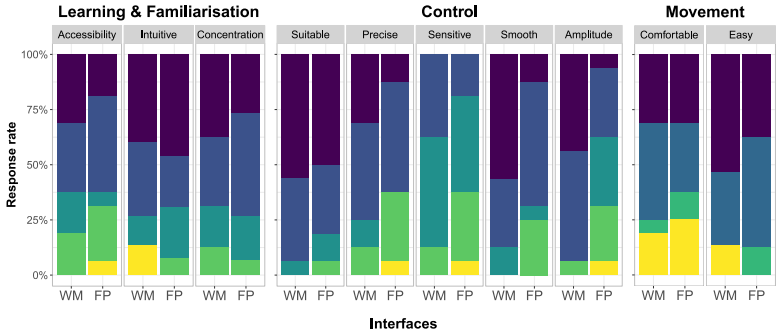
The effect of the focus realisation on the remainder of the utterance can be observed by comparing the baseline and corrective intonation contours displayed in Fig. 4, as well as on the corresponding GAMM fits assessing the effect of the participant turn factor, as provided in Fig. C.7. As has been mentioned, the thick orange lines on the x-axes in the figure display portions where baseline and corrective intonation contours are not significantly different.

The first striking result is that with natural voice in the corrective participant turn, there is a decrease in the height and variability of f_{oc} in the post-focal region when the focus is within the subject constituent (utterances U1 to U3). The GAMM test shows that the baseline and corrective intonation contours are significantly different during the full post-focal region. We thus find the typical de-accented post-focal intonation pattern, characterised by compressed values until the end of the utterance (Jun and Fougeron, 2000; Delais-Roussarie et al., 2002). In the finger pressure condition, the GAMM test shows that the baseline and corrective intonation contours are significantly different during the full post-focal region in utterances U1 and U2, and in the majority of the post-focal region in utterance U3. The wrist movement condition shows less significant differences between the baseline and corrective contours in the post-focal region for utterances U1 to U3. In both chironomic control conditions, the difference between the baseline and corrective contours is below 1 ST, i.e., under the threshold of perceptual salience.

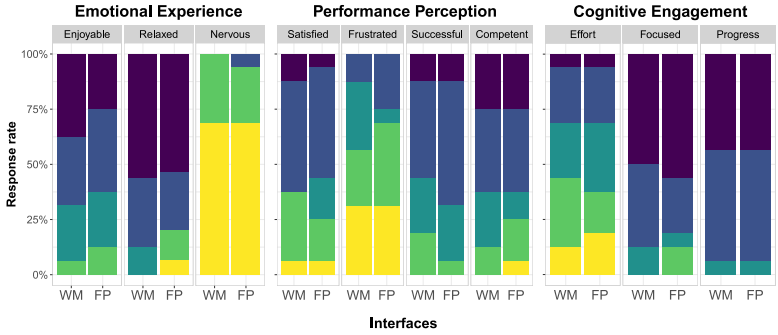
Turning to the pre-focal region, we observe a reduced f_{oc} amplitude on the syllable that precedes the focused one in the natural voice condition on utterances U3 to U5, as described by Dohen and Loevenbruck (2004) and Roessig (2023). The GAMM test shows that baseline and corrective intonation contours are significantly different on the pre-focal syllable in those utterances, with a difference higher than 2 ST, the threshold of the perceptual salience of pitch. The same difference is only observed with finger pressure for utterance U6. For all the other utterances with finger pressure and all the productions with wrist movement, if the baseline and corrective intonation contours are sometimes significantly different on the pre-focal syllable, the difference is always below 1 ST, i.e., still under the threshold of perceptual salience.

Overall, this scenario allowed us to reproduce the pre- and post-focal phenomena with natural voice. When comparing the two chironomic controls, we observed that the participants produced statistically significant post-focal f_o variations with finger pressure. While these variations bear a resemblance to the post-focal compression observed in natural voice, they may result from the mechanical aspects of the control gesture, such as the relaxation of finger pressure on the interface, rather than from an intentional prosodic modulation. Overall, no perceptible pre-focal or post-focal f_o variations appeared in either of the chironomic control conditions.

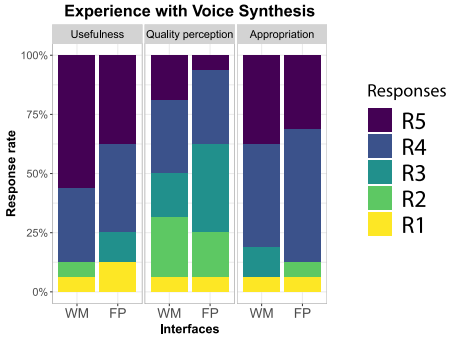
Conclusion. We observed a similar effect of the syllable location on productions without (baseline) and with focus (corrective) for all production modes, suggesting an impact of utterance-level prosody on the production of focus. Nevertheless, beyond the marking of contrastive focus, when using chironomic control the participants did not reproduce other intonative patterns typically observed in natural speech on



(a) Experience with the interface



(b) Experience with the task



(c) Experience with voice synthesis

Fig. 6. Participants' responses to feedback questionnaires (see Appendix B). Interfaces: WM = wrist movement, FP = finger pressure. Answers are R1 - Strongly disagree, R2 - Disagree, R3 - Neutral, R4 - Somewhat agree, R5 - Strongly agree, except for the *Sensitive* question (see Appendix B).

pre- and post-focal regions. This suggests that intonation encoding with chironomic control remains selective in our task, primarily emphasising contrastive focus. Thus, H2 does not hold regarding the realisation of utterance-level intonation variations with chironomic control.

4.4. Participants' feedback

This section summarises the participants' subjective feedback on their experiences regarding the two chironomic control interfaces across the three dimensions: experience with the interface, experience with the task, and experience with voice synthesis (see Fig. 6). The figure presents the distribution of responses for each question and interface. No major differences were observed between wrist movement and finger pressure overall. Additional statistical analyses sometimes revealed significant effects related to the order of interface use. When such effects were found, they are explicitly mentioned in the text.

Experience with the interface. Both interfaces were generally found to be accessible and intuitive, with over 60 % of participants giving high agreement ratings (R4+R5; see Fig. 6(a)). The statistical analysis revealed that a clear learning effect appeared in that the second interface used was rated more accessible and intuitive, reflecting increased familiarity. Conversely, the first interface was perceived as requiring more concentration, indicating the initial cognitive load. Regarding control criteria, wrist movement consistently outperformed finger pressure: it was judged more suitable (93.7 % *vs.* 81.2 %), more precise (75 % *vs.* 62.5 %), smoother in movement (87.5 % *vs.* 68.7 %), and as possessing more appropriate movement amplitude (83.7 % *vs.* 37.5 %). The respective sensitivity ratings were similar, but with a tendency to perceive wrist movement as more sensitive. Comfort and ease of movement were comparable for both interfaces, with no effect of order.

Experience with the task. The participants enjoyed the task and felt relaxed using both interfaces, with wrist movement slightly favoured for relaxation and the absence of nervousness (see Fig. 6(b)). While the figure illustrates overall trends, our statistical analyses showed a notable order effect, as the task performed second was rated more enjoyable, suggesting that user experience improved with growing familiarity. Satisfaction with performance, perceived competence, and feelings of success were positive for both interfaces and improved in the second task. Cognitive engagement was stable, with similar effort levels reported; however, effort tended to increase in the second task, possibly

due to sustained focus demands. The participants remained focused and reported a strong sense of progress regardless of the interface used.

Experiment with voice synthesis. The results (see Fig. 6(c)) showed that the majority found the interfaces useful for intonation control in voice synthesis (87.5 % for wrist movement; 75 % for finger pressure). However, the statistical results revealed an interaction between interface order and production mode: wrist movement was rated higher when used first, while finger pressure was preferred when experienced second, suggesting that the order influenced perceived usefulness. Satisfaction with the synthetic voice was moderate, with 50 % positive for wrist movement and 37.5 % for finger pressure. Most participants felt capable of appropriating the synthesised voice as their own (81.2 % for wrist movement; 87.5 % for finger pressure). Preference was consistently given to the second interface used for satisfaction and appropriation, indicating habituation effects. These results support the potential of chironomic control for voice synthesis but also point to a need for quality improvements.

To conclude, while the results show a preference for wrist movement on the control-related criteria, both of the chironomic controls were perceived as suitable for voice substitution on all other criteria including learning and familiarisation, movement, emotional experience, performance perception, cognitive engagement, and experience with voice synthesis, thus suggesting that H3 is not strongly supported by the findings. Last but not least, an effect of order was often observed, indicating a preference for the second interface used, regardless of the interface, demonstrating a high degree of adaptability on the part of the participants, while also revealing the potential of both interfaces for voice substitution applications.

5. Discussion

5.1. From implicit to explicit control of focus

A successful paradigm for focus realisation. The first remarkable result of this study is the success of the experimental paradigm to elicit a contrastive focus in a modality transfer situation. Inspired by prior studies investigating possible correlations between verbal and gestural modalities (Dohen and Lœvenbruck, 2009; Roustan and Dohen, 2010), a new paradigm has been elaborated here to explore a possible transfer between these two modalities in terms of controlling intonation in the case of a whisper-to-speech conversion device. Our results show that with the two proposed chironomic controls there is no anticipation of a focus in the baseline target conditions, and that the focus is placed on the target syllable, as expected in the corrective target condition. Similarly to what is found in speech (Dohen et al., 2004; Grice et al., 2017), the production of a contrastive focus results in: (i) an increase in f_o, and (ii) an increase in the duration of the target syllable. A second notable result is the speed and ease with which the participants took control of the two interfaces tested here. This ease was reported almost unanimously in the participant's questionnaires responses, for both types of control (wrist movement and finger pressure). The relevance of using both of the interfaces to control a whisper-to-speech conversion system has therefore been confirmed.

Coordination between implicit duration and explicit intonation control. The resulting coordination between the articulation of whispered syllables and the manual intonational control evidenced on f_{oc} peak position within the target syllables show that the participants' manual gestures synchronised naturally with their articulatory gestures, independently of the interface and type of control. In a study asking participants to produce voluntary pointing, beat, or pressing gestures with a natural vocal focus realisation, Roustan and Dohen (2010) have previously demonstrated a synchronisation between articulation and the three types of hand gestures. We can therefore assume that this ability has been usefully transferred to our causal gesture-to-focus relation paradigm. Moreover, their experiment revealed lower synchronisation dispersion with communicative beat gestures than with non-communicative pressure gestures, which we also found to be the case between our wrist movement (similar to beat gestures) and finger pressure conditions. Regarding our natural voice condition, the participants spontaneously placed the focus close to the end of the syllable, while they favoured a more central position when using either chironomic controls, with even a slight advance for finger pressure control. This observation echoes the anteriority of pointing and beat gestures over articulatory gestures when producing focus (Rochet-Capellan et al., 2008; Leonard and Cummins, 2011). Whether this temporality was transferred to our causal control paradigm deserves to be explored in greater depth in a future study. Another example of anticipation is the earlier achievement of focus at the end of the sentence, i.e., on the O3 target syllable. This result, which suggests an anticipation linked to the proximity of the end of the sentence, was observed for both natural voice and wrist movement control. Finally, like (Roustan and Dohen, 2010), we did not observe a significant effect of the type of chironomic control on articulation dynamics, based on duration measures.

Comparison between interfaces. Focus realisation with finger pressure and via wrist movement displayed very similar durations and f_{oc} variations, thereby illustrating that both interfaces are equally suitable candidates for this task, despite some fundamental differences in the control they offer. In addition to finger pressure and wrist movement being *isometric* and *isotonic* controls, respectively, the resting position of finger pressure (no pressure) was mapped to the minimum of f_o range, while the wrist movement resting position (horizontal hand) was mapped to the mid-f_o range. Because speech intonation is a modulation of f_o around its mid-range, the wrist movement then allowed a more consistent control for all participants, with less dispersion in terms of f_o contours. The feedback questionnaires confirmed this observation, with the participants preferring wrist movement for control-related criteria such as smoothness and precision. Inversely, the asymmetric control of finger pressure requiring an effort for raising f_o and a movement release for lowering f_o might have led to the realisation of post-focal f_o lowering with this interface only.

The literature on Human-Computer Interaction supports a stronger preference for finger pressure. As detailed in Section 2.2, Vertegaal et al. (1996) and Wanderley et al. (2000) have already demonstrated that an *isometric* force is more suitable than an *isotonic* force and is preferred by participants to control a relative variation of an acoustic parameter. In computer gaming, Pirker et al. (2017) compare hand gestures in space (using a Leap Motion interface) and pressing buttons (using a keyboard) for the control of actions on a screen, and report that although hand gestures in space are more engaging than using a keyboard at first, they are more tiring and less precise in the long term. Overall, they reveal a larger preference for keyboard control among all users, which is even more pronounced among expert users. Although our experiment revealed no overall preference for finger pressure, the participants felt that finger pressure would be more useful than wrist movement in a voice substitution application when they could compare it to wrist movement. This is perfectly in line with the implementation of finger pressure control in existing on-the-market medical aids such as the Trutone electrolarynx.

Overall, if participants generally found the wrist movement easier to control, they used it for a short period of time (about 25 min). It is possible that a longer period of use, leading to progressively greater expertise, might lead some the participants to switch their preference to a finger pressure control to gain in ergonomics. This is what we observed for the single question of *Usefulness* in a voice substitution application, where finger pressure was preferred when tested after wrist movement. As a result, we cannot yet make a definitive choice between the two interfaces. Even if wrist movement shows a slightly higher synchronisation dispersion with articulation, both interfaces successfully and equally enabled the achievement of focus.

5.2. Beyond focus realisation

Local and global variations of duration. The speech production with chironomic control reflects data from the literature on French, which shows that the last syllable of a rhythmic group is lengthened (Astésano, 2001). An increase in syllable duration was found in synthetic speech as the /lu/syllable progresses within the constituent, whose boundaries may coincide with those of a rhythmic group, whether subject or object. However, this was not observed with natural voice, where /lu/syllables tend to be isochrone across syllable locations, even to the point of revealing an inverse trend, with a slight shortening within the object constituent. A complementary analysis, considering the relative duration of all syllables (not only the target syllable /lu/) according to their position within the constituents, revealed a lengthening of the final syllable within each constituent. This discrepancy between the behaviour of /lu/in particular and syllables more generally could be attributed to the fact that the analysed words were monosyllabic, so syllable lengthening corresponded to the lengthening of the entire word. This potentially competed with the lengthening of the nucleus of the constituent, represented by the monosyllabic word *Lou* or *Loup* (wolf)–i.e., the syllable /lu/in all cases–which retains its role as the prosodic nucleus regardless of its position in the sentence. In addition to local syllable lengthening, a global slowdown in speech rate was observed in synthesised speech with chironomic control. There are two possible explanations for this increasing duration of whispered speech with an interface. On the one hand, there is a natural slowing down of the speed of articulation while whispering when compared to speech (Sharf, 1964; Houle and Levi, 2020). On the other hand, the speech rate can be slowed down by an increase in cognitive load. Whispering accompanied by chironomic control requires awareness, explicitness, and an externalisation of manual control in order to produce the desired intonation. Our data do not allow us to draw further conclusions on the importance of these two possible impacts.

Chironomic control of intonation. Concerning the f_o contour, our natural voice condition displayed similar patterns to those previously described by Dohen and Loevenbruck (2004): the three constituents have default tonal patterns of Accentual Phrases (Jun and Fougeron, 2000), i.e., [LHiLH*] for subject and verb, and [LHiL%] for object, which includes a final boundary tone. By contrast, chironomic controls display highly monotonous f_o contours, with only a small decreasing trend at the end of the utterance.

This clear difference between natural voice and chironomic control conditions raises the question of the naturalness of the generated speech, with respect to both local focus and global sentence realisation. From our objective measures alone (i.e., without performing perceptive tests), the large disparity observed between the two conditions suggests that the chironomic control does not yet allow the generation of natural speech output. Regarding focus realisation, the f_{oc} excursion is two (resp. three) times higher than natural voice with finger pressure (resp. wrist movement), and this contrast is all the more salient as the rest of the sentence is monotonous with chironomic control. This lack of f_o variation in the rest of the sentence, which normally evolves on the temporal scale of the phonological structure of the utterance, and is also known as micro-prosody, brings us back to the problem of current substitution solutions that offer little to no intonation control, and whose resultant voices often sound unnatural or robotic.

Finally, the benefit of this paradigm lies more in the successful production of the *prosodic function* of focus, i.e., producing a local contrast in duration and intonation that is perceptible, rather than the *naturalness* of focus, i.e., producing variations of f_o that are in the same order of magnitude as a natural voice. This is already extremely encouraging as (i) increasing comprehensibility by providing a control on prosodic functions is more beneficial in terms of improving communication with voice substitution than in improving naturalness (Zieliński and Rączaszek-Leonardi, 2022); and (ii) very little training had to be undergone by the participants to reach such a goal. Furthermore, they reported a feeling of constant progress during the task suggesting that there remains a definite margin for progress with more practice time.

6. Conclusion

This study investigated the possibility of transferring the control of vocal intonation to a manual gesture, within the controlled framework of focus production. An experimental paradigm was used to elicit focus without explicitly instructing the participants to do so, and to allow comparative situations without elicited focus. A single target syllable /lu/, placed at different positions in a carrier sentence, was used. Two types of interfaces and controls were assessed: finger pressure with a Sensel Morph pressure-sensitive tactile tablet, and wrist movement with a hand-held accelerometer. Our results clearly demonstrate a successful transfer of modality from natural voice to chironomic control to produce a contrastive focus through an increased intonation and duration of the target syllable. Not only did the participants perceived the importance of intonational modulation in producing a focus, but they were also able to explicitly plan the corresponding intonation contour, and then use both interfaces to emphasise the target syllable accordingly. This behaviour was observed across all our participants.

This work opens up a wide range of prospects for voice substitution using human-machine interfaces. First, the variations observed between the participants' performance and feedback highlight the importance of offering a follow-up to voice rehabilitation adapted to the patient's preferences and tuned to his/her expressive possibilities, both in terms of the type of gesture and interface used, and the parameters of the voice synthesis such as the f_o range. Second, the remarkable speed of task learning highlighted in this study should be set against the difficulties encountered in the prosodic use of current electrolarynxes. One could envision the use of our proposed experimental paradigm as a learning method for manipulating existing voice substitution devices to achieve the precise control of individual prosodic functions. Naturally, this calls for an extension of our study to other prosodic functions in speech, such as tones or lexical stress found in languages other than French, the demarcation of syntactic units, sentence modalities, and/or the expression of attitudes and emotions, and their interactions. While each has different temporal and dynamic constraints, all are essential for spoken communication.

CRediT authorship contribution statement

Delphine Charuau: Writing – review & editing, Writing – original draft, Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Visualization, Validation . **Nathalie Henrich Bernardoni:** Writing – review & editing, Writing – original draft, Conceptualization, Methodology, Supervision. **Silvain Gerber:** Formal analysis. **Olivier Perrotin:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Conceptualization, Methodology, Vizualisation, Software, Formal analysis, Resources, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Olivier Perrotin reports financial support was provided by French National Research Agency. No additional relationships or activities to declare. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by ANR GEPETO - GEsture for the PEdagogie of InTOnation (ANR-19-CE28-0018). We thank the reviewers for their thorough and thoughtful reviews of our manuscript. We also thank the participants of this study, whose involvement was invaluable to the success of this work.

Table A.2
Full corpus of utterances used for the training interaction task with *natural voice*.

Syllable target	Utterance			Contrast
	Subject (S)	*Verb (V)*	*Object (O)*	*Word changed in the question*
S1	*Jean*	a mangé	le riz de mamie.	Paul
S2	Les *daims*	ont suivi	Jean dans les bois.	chats
S3	Les chats *doux*	ont volé	le beau jeu.	gris
O1	Les daims	ont suivi	*Jean* dans les bois.	Tom
O2	Jean	a mangé	le *riz* de mamie.	choux
O3	Les chats doux	ont volé	le beau *jeu*.	bol

Appendix A. Training material for the experiment

A.1. Training interaction task corpus

See Table A.2.

A.2. Training reading task text

Source: *MonPage 2*, Pommée (2021, p. 114).
The full text was displayed and read at once by the participants.

Lundi, le chat, le loup et Papa vont à Bali. Les copains sont tout contents.
Mardi, Papy y va aussi. Il dit : ≪ Je n'ai pas un sou ! Qui va prendre soin de moi ? ≫ ≪ Moi ! ≫ dit le chat, ≪ moi ! ≫ dit le loup. ≪ Vous ? ≫, Papy réfléchit.
Mercredi, Papy dit : ≪ Toi, le chat, tu es doux, tu es chou, tu n'as pas de poux ! Mais pas ce loup : il a une cape rouge et je n'aime pas ce gars-là ! ≫
Jeudi, le chat et Papy se baladent à Bali. Papa glisse ! Aïe ! Ouille ! Son cou craque, son coude claque, c'est la débâcle !
Vendredi, Papa a mal. Il pleure, il crie ! ≪ Toi, Papy, aide-moi, trouve le nain ! ≫ ≪ Un nain ? On n'en a jamais vu par ici ?! ≫
Samedi matin, le chat va voir son ami le loup et lui dit : ≪ Aide-moi à soigner Papa ! ≫
Samedi soir, le loup lui donne sa recette magique : ≪ Coupe un oignon, cache-le sous la souche, et lorsque le lilas fleurira, Papa sera guéri ! ≫ Abracadabra, ça y est, on a réussi !
Dimanche, le chat tout doux, le loup magicien, Papa et Papy quittent Bali. Les copains sont tout contents.

A.3. Training imitation task corpus

Source: d'Alessandro et al. (2011).
We used the three sentences below from d'Alessandro et al. (2011), each pronounced with three intonation patterns (declarative; interrogative; declarative with focus), making a total of nine training utterances. These utterances were recorded by a male and a female speaker, and for each of the participants we used the recordings corresponding to the same self-reported gender, to best match his/her intonation range.

Declarative (falling intonation)	**Interrogative** (rising intonation)	**Focus** (on the underlined word)
Marie chantait souvent.	Marie chantait souvent ?	Marie chantait souvent.
Nous voulons manger le soir.	Nous voulons manger le soir ?	Nous voulons manger le soir.
Sophie mangeait des fruits confits.	Sophie mangeait des fruits confits ?	Sophie mangeait des fruits confits.

Appendix B. Feedback questionnaire

The following 23 questions were submitted in French to the participants after the second and third phases of the experiment. The titles of each question in bold refer to Fig. 6 and were not displayed to participant. All questions except for question *Sensitive* (marked with an *) used the following absolute categorical rating scale (R1 to R5):

Pas du tout d'accord	Pas d'accord	Neutre	Plutôt d'accord	Tout à fait d'accord
Strongly disagree	*Disagree*	*Neutral*	*Somewhat agree*	*Strongly agree*

Part I: Experience with the interface

Accessibility: Je trouve cette interface facile à prendre en main
I find this interface easy to use

Intuitive: Je trouve que l'utilisation de cette interface est intuitive
I find this interface intuitive to use

Concentration: Je trouve que l'utilisation de cette interface demande de la concentration
I find that using this interface requires concentration

Suitable: Je trouve cette interface adaptée à cette activité
I find this interface suitable for this activity

Precise: Je trouve cette interface précise
I find this interface precise

Sensitive: *Je trouve cette interface : Trop sensible Très sensible Sensible Peu sensible Pas du tout sensible
**I find this interface: Too sensitive Very sensitive Sensitive Not very sensitive Not at all sensitive*

Smooth: Avec cette interface, le mouvement est fluide
With this interface, movement is smooth

Amplitude: Avec cette interface, l'amplitude du mouvement est adaptée à la tâche
With this interface, the amplitude of movement is adapted to the task in hand

Comfortable: Je trouve que l'utilisation de cette interface est confortable au niveau du mouvement sur toute la durée du test
I found the interface comfortable to use, in terms of movement, throughout the test

Easy: Avec cette interface, le mouvement est facile à produire
With this interface, movement is easy to produce

Part II: Experience with the task

Enjoyable: J'ai trouvé la tâche agréable
I found the task enjoyable

Relaxed: J'étais détendu· e en faisant cette activité
I was relaxed while doing this activity

Nervous: Je me suis senti· e de plus en plus nerveux· se au fur et à mesure de l'avancée dans la tâche
I felt more and more nervous as the task progressed

Satisfied: Je suis satisfait.e de ma performance dans cette activité
I am satisfied with my performance in this activity

Frustrated: Je suis frustré.e par ma performance dans cette activité
I am frustrated by my performance in this activity

Successful: Je pense avoir réussi cette activité
I think I have succeeded in this activity

Competent: Après avoir travaillé avec cette interface pendant quelques minutes, je me suis senti· e assez compétent.e pour cette activité
After working with this interface for a few minutes, I felt quite competent for this activity

Effort: J'ai fait beaucoup d'effort pendant cette activité pour réussir
I put a lot of effort into this activity to succeed

Focused: J'ai pu rester concentré· e durant cette activité
I was able to stay focused during this activity

Progress: J'ai eu l'impression de progresser au fur et à mesure de la tâche
I had the impression that I was making progress as I went along

Part III: Experience with voice synthesis

Usefulness: Si je perdais ma voix, je trouverais cette interface de contrôle de l'intonation utile
If I were to lose my voice, I'd find this intonation control interface useful

Quality perception: Si je perdais ma voix, je pense que je serais satisfait· e de la voix de synthèse obtenue à l'aide de cette interface
If I were to lose my voice, I think I'd be satisfied with the synthesised voice obtained using this interface

Appropriation: Si je perdais ma voix, je pense réussir à m'approprier la voix de synthèse à l'aide de cette interface
If I were to lose my voice, I think I'd be able to appropriate the synthesised voice using this interface

Appendix C. GAMM modelling of f_{oc} contours

See Fig. C.7.

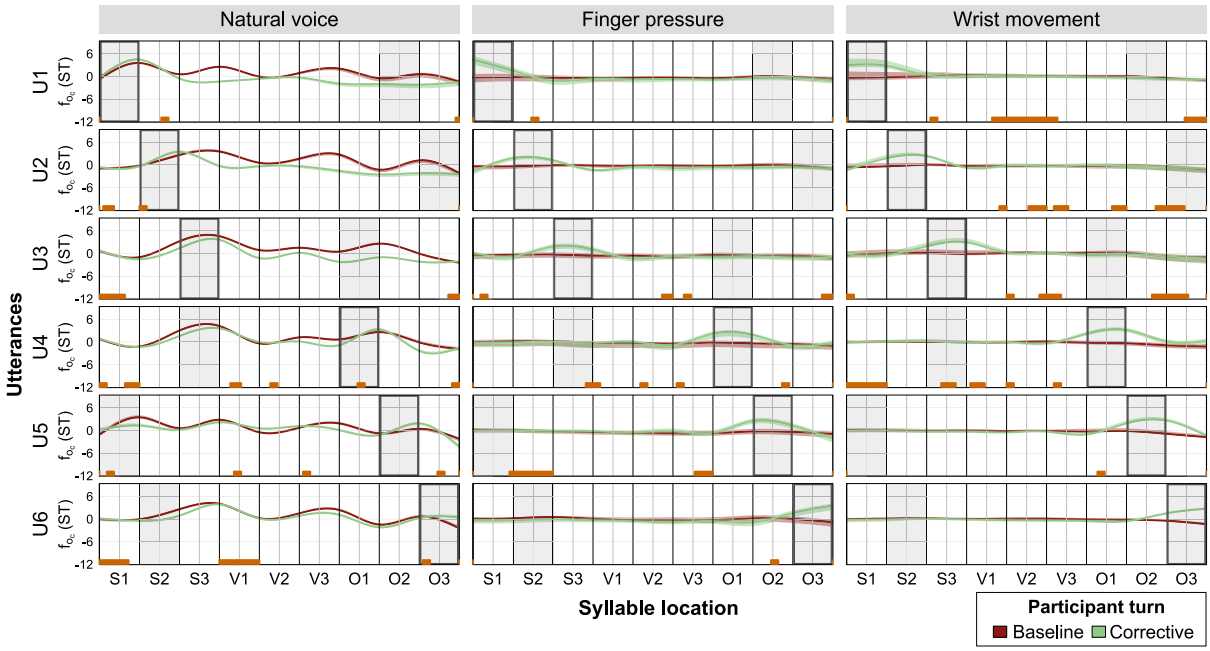


Fig. C.7. The median (plain line) and 2nd to 3rd quartiles (shaded area) of the syllable-aligned GAMM fits of f_{oc} contours per utterance (rows), production mode (columns) and participant turn (red: baseline; green: corrective) across participants and repetitions. The solid grey rectangles with thick and thin edges on each utterance indicate the target and non-target /lu/syllables, respectively. The thick orange thick lines on the x-axes indicate portions of fitted f_{oc} contours where baseline and corrective realisation of the same utterance are NOT significantly different according to the testing of the participant turn factor with GAMMs (similar utterances: U1 *vs.* U5 ; U2 *vs.* U6, and U3 *vs.* U4).

Data availability

Data will be published with acceptation of article.

References

Ahmadi, F., Noorian, F., Novakovic, D., van Schaik, A., 2018. A pneumatic Bionic Voice prosthesis—Pre-clinical trials of controlling the voice onset and offset. In: Baumert, M. (Ed.), PLoS One 13 (2), http://dx.doi.org/10.1371/journal.pone.0192257.

Ardaillon, L., Henrich Bernardoni, N., Perrotin, O., 2022. Voicing decision based on phonemes classification and spectral moments for whisper-to-speech conversion. In: Proc. of Interspeech. Incheon, Korea, pp. 2253–2257. http://dx.doi.org/10.21437/Interspeech.2022-10675.

Astésano, C., 2001. Rythme et accentuation en français : Invariane et variabilité statistique. L'Harmattan.

Astésano, C., Magne, C., Morel, M., Coquillon, A., Espesser, R., Besson, M.R., Lacheret-Dujour, A., 2004. Marquage acoustique du focus contrastif non codé syntaxiquement en français. In: Actes des Journées d'Etudes sur la Parole. JEP, Fès, Marocco, pp. 4–p.

Card, S.K., Mackinlay, J.D., Robertson, G.G., 1991. A morphological analysis of the design space of input devices. ACM Trans. Inf. Syst. 9 (2), 99–122. http://dx.doi.org/10.1145/123078.128726.

Carignan, C., Esteve-Gibert, N., Lœvenbruck, H., Dohen, M., D'Imperio, M., 2024. Co-speech head nods are used to enhance prosodic prominence at different levels of narrow focus in French. J. Acoust. Soc. Am. 156 (3), 1720–1733. http://dx.doi.org/10.1121/10.0028585.

Cave, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R., 1996. About the relationship between eyebrow movements and fo variations. ICSLP, In: International Conference on Spoken Language Processing, vol. 4, Philadelphia, PA, USA, pp. 2175–2178. http://dx.doi.org/10.1109/ICSLP.1996.607235.

Dahan, D., Bernard, J.M., 1996. Interspeaker variability in emphatic accent production in French. Lang. Speech 39 (Pt 4), 341–374. http://dx.doi.org/10.1177/002383099603900402.

d'Alessandro, C., 2022. Une nouvelle organologie de la voix : chironomie et prosodie de la parole et du chant. In: Actes des Journées d'Etudes sur la Parole. JEP, Noirmoutiers, France, pp. 625–636. http://dx.doi.org/10.21437/JEP.2022-66.

d'Alessandro, C., d'Alessandro, N., Le Beux, S., Simko, J., Cetin, F., Pirker, H., 2005. The Speech Conductor: Gestural Control of Speech Synthesis. Technical Report, eNTERFACE workshop.

d'Alessandro, N., Dutoit, T., 2009. Advanced techniques for vertical tablet playing a overview of two years of practicing the HandSketch 1.x. In: Proc. International Conference on New Interfaces for Musical Expression. NIME, Pittsburgh, PA, United States, pp. 173–174. http://dx.doi.org/10.5281/zenodo.1177465.

d'Alessandro, C., Feugère, L., Le Beux, S., Perrotin, O., Rilliard, A., 2014. Drawing melodies: Evaluation of chironomic singing synthesis. J. Acoust. Soc. Am. 135 (6), 3601–3612. http://dx.doi.org/10.1121/1.4875718.

d'Alessandro, C., Rilliard, A., Le Beux, S., 2011. Chironomic stylization of intonation. J. Acoust. Soc. Am. 129 (3), 1594–1604. http://dx.doi.org/10.1121/1.3531802.

Delais-Roussarie, E., Rialland, A., Doetjes, J., Marandin, J.M., 2002. The prosody of post-focus sequences in French. In: Proc. Speech Prosody. Aix-en-Provence, France, pp. 239–242. http://dx.doi.org/10.21437/SpeechProsody.2002-45.

Di Cristo, A., 2016. Les musiques du français parlé: essais sur l'accentuation, la métrique, le rythme, le phrasé prosodique et l'intonation du français contemporain. Études de linguistique française, (1), De Gruyter, http://dx.doi.org/10.1515/9783110479645.

Di Cristo, A., Jankowski, L., 1999. Prosodic organisation and phrasing after focus in French. Proc. Int. Congr. Phon. Sci. (ICPhS) 1565–1568.

Dohen, M., Loevenbruck, H., 2004. Pre-focal rephrasing, focal enhancement and postfocal deaccentuation in French. In: Proc. Interspeech. ISCA, Jeju, Korea, pp. 785–788. http://dx.doi.org/10.21437/Interspeech.2004-296.

Dohen, M., Lœvenbruck, H., 2009. Interaction of audition and vision for the perception of prosodic contrastive focus. Lang. Speech 52 (2–3), 177–206. http://dx.doi.org/10.1177/0023830909103166.

Dohen, M., Lœvenbruck, H., Cathiard, M.A., Schwartz, J.L., 2004. Visual perception of contrastive focus in reiterant French speech. Speech Commun. 44 (1), 155–172. http://dx.doi.org/10.1016/j.specom.2004.10.009, Special issue on audio visual speech processing.

Dudley, H., Riesz, R., Watkins, S., 1939. A synthetic speaker. J. Franklin Inst. 227 (6), 739–764. http://dx.doi.org/10.1016/S0016-0032(39)90816-1.

Evrard, M., Delalez, S., d'Alessandro, C., Rilliard, A., 2015. Comparison of chironomic stylization versus statistical modeling of prosody for expressive speech synthesis. In: Proc. Interspeech. Dresden, Germany, pp. 3370–3374. http://dx.doi.org/10.21437/Interspeech.2015-142.

Fant, G., Kruckenberg, A., Liljencrants, J., Bavegard, M., 1994. Voice source parameters in continuous speech. Transformation of LF-parameters. In: International Conference on Spoken Language Processing. ICSLP, Yokohama, Japan, pp. 1451–1454. http://dx.doi.org/10.21437/ICSLP.1994-377.

Fels, S., Hinton, G., 1998. Glove-talk II - A neural-network interface which maps gestures to parallel formant speech synthesizer controls. IEEE Trans. Neural Netw. 8 (5), 205–212. http://dx.doi.org/10.1109/72.623199.

Feugère, L., d'Alessandro, C., Doval, B., Perrotin, O., 2017. Cantor Digitalis: chironomic parametric synthesis of singing. EURASIP J. Audio, Speech, Music. Process. 2017 (1), http://dx.doi.org/10.1186/s13636-016-0098-5.

Fuchs, A.K., Hagmuller, M., Kubin, G., 2016. The new bionic electro-larynx speech system. IEEE J. Sel. Top. Signal Process. 10 (5), 952–961. http://dx.doi.org/10.1109/JSTSP.2016.2535970.

Grice, M., Ritter, S., Niemann, H., Roettger, T.B., 2017. Integrating the discreteness and continuity of intonational categories. J. Phon. 64, 90–107. http://dx.doi.org/10.1016/j.wocn.2017.03.003.

't Hart, J., 1981. Differential sensitivity to pitch distance, particularly in speech. J. Acoust. Soc. Am. 69 (3), 811–821. http://dx.doi.org/10.1121/1.385592.

Heeren, W., Van Heuven, V., 2014. The interaction of lexical and phrasal prosody in whispered speech. J. Acoust. Soc. Am. 136 (6), 3272–3289. http://dx.doi.org/10.1121/1.4901705.

Houle, N., Levi, S.V., 2020. Acoustic differences between voiced and whispered speech in gender diverse speakers. J. Acoust. Soc. Am. 148 (6), 4002–4013. http://dx.doi.org/10.1121/10.0002952.

Jun, S.A., Fougeron, C., 2000. A phonological model of French intonation. In: Botinis, A. (Ed.), Intonation: Analysis, Modelling and Technology. Kluwer Academic Publishers, Dordrecht, pp. 209–242. http://dx.doi.org/10.1007/978-94-011-4317-2_10.

Kaye, R., Tang, C.G., Sinclair, C.F., 2017. The electrolarynx: voice restoration after total laryngectomy. Med. Devices: Evid. Res. 10, 133–140. http://dx.doi.org/10.2147/MDER.S133225.

Kessous, L., 2004. Gestural control of singing voice, a musical instrument. In: Sound and Music Computing Conference. SMC, Paris, France.

Krivokapić, J., Tiede, M.K., Tyrone, M.E., 2017. A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. Lab. Phonol. 8 (3), 1–26. http://dx.doi.org/10.5334/labphon.75.

Leonard, T., Cummins, F., 2011. The temporal relation between beat gestures and speech. Lang. Cogn. Process. 26 (10), 1457–1471. http://dx.doi.org/10.1080/01690965.2010.500218.

Liu, H., Ng, M.L., 2007. Electrolarynx in voice rehabilitation. Auris. Nasus. Larynx. 34 (3), 327–332. http://dx.doi.org/10.1016/j.anl.2006.11.010.

Locqueville, G., d'Alessandro, C., Delalez, S., Doval, B., Xiao, X., 2020. Voks: Digital instruments for chironomic control of voice samples. Speech Commun. 125, 97–113. http://dx.doi.org/10.1016/j.specom.2020.10.002.

Loria, 2016. ASTALI. ORTOLANG (open resources and tools for language) –www.ortolang.fr. URL https://hdl.handle.net/11403/astali/v2.

Marshall, M.T., Wanderley, M.M., 2005. Evaluation of sensors as input devices for computer music interfaces. In: International Symposium on Computer Music Multidisciplinary Research. CMMR, Springer-Verlag, Pisa, Italy, pp. 130–139. http://dx.doi.org/10.1007/11751069_12.

Matsui, K., Kimura, K., Nakatoh, Y., Kato, Y.O., 2013. Development of electrolarynx with hands-free prosody control. In: ISCA Speech Synthesis Workshop. Barcelona, Spain, pp. 273–277.

Mertens, P., 2008. Syntaxe, prosodie et structure informationnelle : une approche prédictive pour l'analyse de l'intonation dans le discours. Trav. Linguist. 56 (1), 97–124. http://dx.doi.org/10.3917/tl.056.0097.

Pagel, L., Roessig, S., Mücke, D., 2024. The encoding of prominence relations in supra-laryngeal articulation across speaking styles. Lab. Phonol. 15 (1), 1–55. http://dx.doi.org/10.16995/labphon.10900.

Parrell, B., Goldstein, L., Lee, S., Byrd, D., 2011. Temporal coupling between speech and manual motor actions. In: International Seminar on Speech Production. ISSP, Montreal, Canada, http://dx.doi.org/10.1016/j.wocn.2013.11.002.

Pérez Zarazaga, P., Malisz, Z., 2023. Recovering implicit pitch contours from formants in whispered speech. In: Proc. International Congress of Phonetic Sciences (ICPhS). Prague, Czech Republic, pp. 3146–3150.

Perrotin, O., 2015. Chanter avec les mains: Interfaces chironomiques pour les instruments de musique numériques (Ph.D. thesis). Université Paris-Sud, Orsay, France.

Perrotin, O., d'Alessandro, C., 2016. Seeing, listening, drawing: Interferences between sensorimotor modalities in the use of a tablet musical interface. ACM Trans. Appl. Percept. 14 (2), 10:1–10:19. http://dx.doi.org/10.1145/2990501.

Perrotin, O., McLoughlin, I.V., 2019. A spectral glottal flow model for source-filter separation of speech. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP, IEEE, Brighton, UK, pp. 7160–7164. http://dx.doi.org/10.1109/ICASSP.2019.8682625.

Perrotin, O., McLoughlin, I.V., 2020. Glottal flow synthesis for whisper-to-speech conversion. IEEE/ACM Trans. Audio, Speech, Lang. Process. 28, 889–900. http://dx.doi.org/10.1109/TASLP.2020.2971417.

Pirker, J., Pojer, M., Holzinger, A., Gütl, C., 2017. Gesture-based interactions in video games with the leap motion controller. In: Kurosu, M. (Ed.), Human-Computer Interaction. User Interface Design, Development and Multimodality. Springer International Publishing, pp. 620–633.

Pommée, T., 2021. Les mesures d'intelligibilité : État de l'art, considérations pratiques pour l'applicabilité clinique et explorations acoustiques (Ph.D. thesis). Université Toulouse III Paul Sabatier.

Pritchard, B., Fels, S., 2006. GRASSP: Gesturally-realized audio, speech and song performance. In: Proc. International Conference on New Interfaces for Musical Expression. NIME, Paris, France, pp. 272–276. http://dx.doi.org/10.5281/zenodo.1176987.

Rochet-Capellan, A., Laboissière, R., Galván, A., Schwartz, J.L., 2008. The speech focus position effect on jaw-finger coordination in a pointing task. J. Speech, Lang. Hear. Res. 51 (6), 1507–1521. http://dx.doi.org/10.1044/1092-4388(2008/07-0173).

Roessig, S.B., 2023. The inverse relation of pre-nuclear and nuclear prominences in German. Lab. Phonol.: J. Assoc. Lab. Phonol. 15 (1), 1–43. http://dx.doi.org/10.16995/labphon.9993.

Roustan, B., Dohen, M., 2010. Co-production of contrastive prosodic focus and manual gestures: temporal coordination and effects on the acoustic and articulatory correlates of focus. In: Proc. Speech Prosody. (110), Chicago, IL, USA, pp. 1–4. http://dx.doi.org/10.21437/SpeechProsody.2010-246.

Ryan, R.M., 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. J. Pers. Soc. Psychol. 43 (3), 450—-461. http://dx.doi.org/10.1037/0022-3514.43.3.450.

Ryan, R.M., Mims, V., Koestner, R., 1983. Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. J. Pers. Soc. Psychol. 45 (4), 736–750. http://dx.doi.org/10.1037/0022-3514.45.4.736.

Scherer, K.R., 2003. Vocal communication of emotion: A review of research paradigms. Speech Commun. 40 (1–2), 227–256. http://dx.doi.org/10.1016/S0167-6393(02)00084-5.

Schwartz, M.F., 1967. Syllable duration in oral and whispered reading. J. Acoust. Soc. Am. 41 (5), 1367–1369. http://dx.doi.org/10.1121/1.1910487.

Sharf, D.J., 1964. Vowel duration in whispered and in normal speech. Lang. Speech 7 (2), 89–97. http://dx.doi.org/10.1177/002383096400700204.

Sóskuthy, M., 2021. Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. J. Phon. 84, 101017. http://dx.doi.org/10.1016/j.wocn.2020.101017.

Toda, T., Nakagiri, M., Shikano, K., 2012. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. 20 (9), 2505–2517. http://dx.doi.org/10.1109/TASL.2012.2205241.

Vercherand, G., 2011. Perceptual level of intonation in whispered voice. In: Proc. ISCA Workshop on Experimental Linguistics. Paris, France.

Vertegaal, R., Ungvary, T., Kieslinger, M., 1996. Towards a musician's cockpit: Transducers, feedback and musical function. In: Proc. International Computer Music Conference. ICMC, Hong Kong, China, pp. 308–311.

Wagner, P., Malisz, Z., Kopp, S., 2014. Gesture and speech in interaction: An overview. Speech Commun. 57 (Supplement C), 209–232. http://dx.doi.org/10.1016/j.specom.2013.09.008.

Wanderley, M.M., Viollet, J.P., Isart, F., Rodet, X., 2000. On the choice of transducer technologies for specific musical functions. In: Proc. International Computer Music Conference. ICMC, Berlin, Germany.

Ward, N.G., 2019. Prosodic Patterns in English Conversation. Cambridge University Press, http://dx.doi.org/10.1017/9781316848265.

Xiao, X., Audibert, N., Locqueville, G., d'Alessandro, C., Kühnert, B., Kleinberger, R., Pillot-Loiseau, C., 2022. Évaluation de la stylisation chironomique pour l'apprentissage de l'intonation du français L2. In: Actes des Journées d'Etudes sur la Parole. JEP, Noirmoutier, France, pp. 465–473. http://dx.doi.org/10.21437/JEP.2022-46.

Xiao, X., Kuhnert, B., Audibert, N., Locqueville, G., Pillot-Loiseau, C., Zhang, H., d'Alessandro, C., 2023. Performative vocal synthesis for foreign language intonation practice. In: Proc. of the CHI Conference on Human Factors in Computing Systems. (697), Hamburg, Germany, pp. 1–9. http://dx.doi.org/10.1145/3544548.3581210.

Zieliński, K., Rączaszek-Leonardi, J., 2022. A complex human-machine coordination problem: Essential constraints on interaction control in bionic communication systems. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. New York, NY, USA, pp. 1–8. http://dx.doi.org/10.1145/3491101.3519672.