Combining manual control of intonation with whisper articulation in voice substitution: the case of contrastive focus

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

Combining manual control of intonation with whisper articulation in voice substitution: the case of contrastive focus

Delphine Charuau[1], Nathalie Henrich Bernardoni[1], Silvain Gerber[1], Olivier Perrotin[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab Grenoble, F-38000 France

delphinecharuau1@gmail.com, olivier.perrotin@grenoble-inp.fr

Introduction. The substitution of the glottal source with a synthetic one, for instance following a laryngectomy, requires a substitute control of intonation, such as pressing a button on an electrolarynx. This paradigm combines a manual control of intonation with natural articulation, which has poorly been evaluated in terms of efficiency in fulfilling linguistic functions. This speech task is made challenging by the diverse prosodic functions of the fundamental frequency (f_o) such as focus, boundary marking, attitudes, and emotions (Mertens 2008). In this study, we are particularly interested in the realisation of contrastive focus, which is characterised by an increase in the f_o curve (Jun & Fougeron 2000; Grice *et al.* 2017). Our aim is to analyse to which extent participant are able to produce the emphasising of syllables through variations in intonation controlled by manual gestures. Following the numerous developments in human-machine interfaces for speech control (d'Alessandro 2022), two complementary gestures will be compared.

Methods. Our experiment was conducted using a whisper-to-speech conversion (WSC) software (Ardaillon *et al.* 2022; Perrotin & McLoughlin 2020), which enables the real-time conversion of whispers acquired with a microphone into vocalised speech while providing control of intonation with hand gesture through human-machine interfaces. In this study, we compared two complementary control modalities: the first is isometric and allows modulation of intonation through finger pressure on a button as in the Trutone electrolarynx, while the second is isotonic and allows control through wrist rotation, similarly as beat gestures (Leonard & Cummins 2011). Both the degree of button depression and the angle of wrist rotation are linearly mapped to f_o, in semitones, in the range of an octave around the speaker's mean f_o value, measured in a calibration step.

To encourage speakers to produce a contrastive focus at a specific location but without giving any explicit instruction (Dohen 2005), they were recorded in simulated dyadic interactions guided by a scenario displayed on a screen. Speakers started with the production of an initial utterance (condition "*pre*"), followed by a pre-recorded question simulating the misunderstanding of a target word. The speaker had then to repeat the same utterance, potentially introducing a focus on the target word that was misunderstood (condition "*post*"). These interactions were based on a corpus of 6 sentences, each composed of 9 monosyllabic words (CV-type), evenly distributed among the subject, verb, and object constituents. The subject and object constituents of each sentence each contain the syllable [lu], of which only one is targeted at a time with the question that follows. The position of the [lu] syllable within the constituents (S1, S2, S3, O1, O2, O3) varied from one sentence to another, so that overall it is seen as targeted and non-targeted for all syllable positions. The interaction task was carried out in three production modes: with *Natural voice*, with *Finger pressure* control and with *Wrist movement* control. The two latter conditions use the WSC system and their order was randomly chosen across participants. Each production mode was preceded by a training phase to become familiar with the interaction scenario, and the interfaces.

Sixteen speakers were recorded (median age = 24.5 years old; Q1 = 22.5; Q3 = 27). They did not report any speech, hearing, arm, or hand motor disorders. The acoustic data was semi-automatically segmented and annotated using Astali (Loria 2016) and Praat (Boersma & Weenink 2021). Matlab was used to extract temporal (relative duration of syllables, utterance duration, articulation rate) and intonation data (height, position, and width of the centred f_o peak). Centred f_o (f_{oc}), expressed in semi-tones (st), corresponds to the subtraction of median f_o values computed for one speaker and one interface to the corresponding raw f_o. The relative duration (Dr) of syllables is expressed as a percentage of the sentence duration. Statistical analyses were conducted using R. The significance of the results was tested through a mixed-effects linear regression model to examine the effects of syllable position, interfaces, and syllable condition. Random factors such as *speaker* and *repetition* were also taken into account. The overall significance level was set to $p < 0.05$.

Results. Fig. 1 displays peak f_{oc} height per production mode and syllable position. In *Natural voice*, speakers tend to mark the focus on the target syllable by raising the f_o curve in the "*post*" condition (dark green) compared to the "*pre*" condition (dark red), regardless of the target syllable position in the sentence. However, this difference is only significant, when the target syllable is in the second position within the object constituent (O2). In contrast, in the *Finger pressure* task, this difference is significant, except when the target syllable is in the first position of a constituent. In the case of the *Wrist movement* task, the difference between the "*pre*" and "*post*" conditions of the target syllable is significant, regardless of the syllable position.

We also observed that the [lu] syllables are significantly longer when we expect a focus, both in *Natural voice* (Dr_{mean} = 16.1 ± 2.9%), *Finger pressure* (Dr_{mean} = 18.1 ± 4.5%) and *Wrist movement* (Dr_{mean} = 18.3 ± 4.4%), than when they do not

(*Natural voice:* Dr_{mean} = 12.3 ± 1.9%; *Finger pressure*: Dr_{mean} = 12.1 ± 3.1%; *Wrist movement:* Dr_{mean} = 12.3 ± 2.9%), regardless of the syllable position in the utterance.
Finally, we analysed the position of the f_o peak relatively to the target syllable boundaries (Pos) in the *"post*" condition. In *Natural voice*, the f_o peak tended to be located towards the end of the marked syllable (Pos_{mean} = 70.3 ± 27.5%). When using an interface for f_o control, the f_o peak was achieved slightly earlier during the production of the target syllable, i.e., the peak being more centred relatively to the boundaries of the syllable (*Finger pressure:* Pos_{mean} = 42.7 ± 35%; *Wrist movement:* Pos_{mean} = 51.8 ± 36.1%).

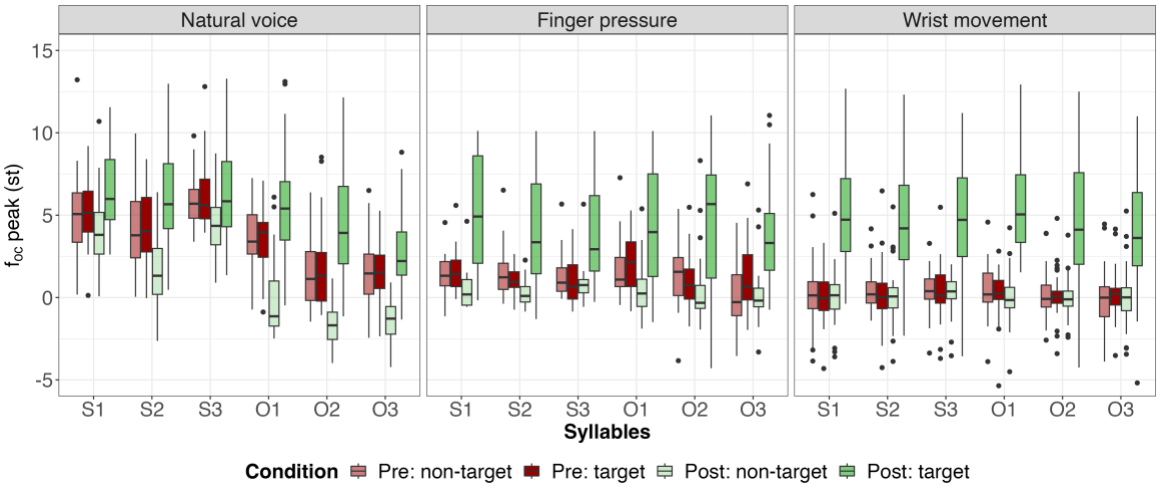

Figure 1: *f_{oc} peak height of the [lu] syllables according to their location on the utterance.*

Discussion. Explicit manual control of intonation requires to become aware of one's own intonation curve in speech, which is usually implicit in typical speech production. The question of the difficulty of external manual f_o control to realise a specific linguistic function was therefore not trivial. However, all speakers were able to successfully produce an elicited contrastive focus in a paradigm of external and explicit intonation control, within the relatively limited time of this experiment (one hour).
In *Finger pressure* and *Wrist movement* tasks, increased f_o on [lu] focused syllables demonstrated: i) the speakers' intention to distinguish this syllable from the others by modulating intonation, and ii) their awareness of the important role of its function for emphasising the target syllable. Focus realisation was also marked by a significant lengthening of the relative duration of the target syllable, both in *Natural voice* and in whispered speech with manual intonation control, regardless of articulation rate. Control of intonation with hand gesture was synchronised with syllable production: if the f_o peak was reached earlier than *Natural voice*, it was mostly realised within the boundaries of the [lu] target syllable. More specifically, in *Wrist movement* task, the f_o peak gravitated from the centre of syllable, regardless of the syllable position, while it was slightly anterior in *Finger pressure* task. The comparison of interface usage showed no significant differences between these two types of control, although we observed some specificities in their use, which will be investigated in future work. These encouraging results call for the exploration of other linguistic functions in a less controlled speech task, to fully validate such control paradigm in voice substitution applications.

References

Ardaillon, L., Henrich Bernardoni, N., & Perrotin, O. (2022). Voicing decision based on phonemes classification and spectral moments for whisper-to-speech conversion. In *Interspeech 2022*, 2253-2257.

Boersma, P., & Weenink, D. (2021). *Praat: Doing by computer* [Computer program], Version 6.1.42.

d'Alessandro C. (2022). Une nouvelle organologie de la voix : chironomie et prosodie de la parole et du chant. *Actes des Journées d'Etudes sur la Parole,* 625-636.

Dohen, M. (2005). Deixis prosodique multisensorielle : production et perception audiovisuelle de la focalisation contrastive en français. *Thèse de doctorat, Institut National Polytechnique de Grenoble*.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics, 64*, 90-107.

Jun, S.-A., & Fougeron, C. (2000). A phonological model of French intonation. In Botinis, A. (Ed.), Intonation: Analysis, Modeling and Technology, Dordrecht: Kluwer Academic. 209-242.

Laboratoire lorrain de recherche en informatique et ses applications – UMR 7503 (Loria). (2016). ASTALI [Outil]. *ORTOLANG (Open Resources and TOols for LANGuage)* – www.ortolang.fr, v2.

Leonard T. & Cummins F. (2011). The temporal relation between beat gestures and speech. *In Language and Cognitive Processes*, 26(10), 1457-1471.

Martens, P. (2008). Syntaxe, prosodie et structure informationnelle : une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de linguistique,* 56, 97-124.

Perrotin, O., & McLoughin, I. (2020). Glottal flow synthesis for whisper-to-speech conversion. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 889-900.