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

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble - France

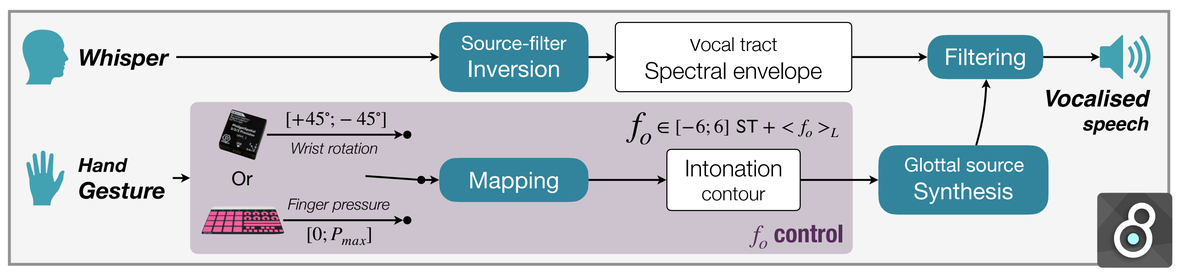
delphinecharuau1@gmail.com ; olivier.perrotin@grenoble-inp.fr

Introduction

- The substitution of glottal source with a synthetic one requires a substitute control of intonation → manual control of intonation + natural articulation
- Have been poorly evaluated in terms of efficiency in fulfilling linguistic functions [8]
- Contrastive focus → increase in the f_o curve [1] [2]

- **Our aim :** analyse to which extent participant are able to produce the emphasis of syllables through variations in intonation controlled by manual gestures

Whisper-to-speech conversion (WSC) [3, 4]



Experimental protocol

Speech task

- Simulated dyadic interactions guided by a scenario displayed on a screen
- A scenario enabling the induction of contrastive focus **without giving explicit instruction** [5]
- Speech task = 6 interactions x 3 repetitions

Scenario					Condition
– Participant :	Le	loup	doux	a suivi le beau loup .	*Pre*
– Experimenter :	Le loup doux			a suivi le beau chien ?	*Question*
– Participant :	Le	loup	doux	a suivi le beau loup .	*Post*

→ 4 syllable conditions

Example of scenario for the experiment. We make the hypothesis that only the Post-target [lu] *syllable will be accented by the participant.*

Syllable		Sentence			Contraste
Target	Non-target	*Subject (S)*	*Verb (V)*	*Object (O)*	*Word changed in the question*
S1	O2	Lou du Mans	a suivi	le loup doux.	Jean
S2	O3	Le loup doux	a suivi	le beau loup.	chat
S3	O1	Le beau loup	a suivi	Lou du Mans.	chien
O1	S3	Le beau loup	a suivi	Lou du Mans.	Jean
O2	S1	Lou du Mans	a suivi	le loup doux.	chat
O3	S2	Le loup doux	a suivi	le beau loup.	chien

Sentences corpus

- 9 monosyllabic words [CV]
- Target syllable : [lu]
- Randomly ordered

Experiment

16 speakers (median age = 24.5 y.o.)

NATURAL VOICE Training + Speech task → **WHISPER + FINGER PRESSURE** Training + Speech task → **WHISPER + WRIST MOVEMENT** Training + Speech task

Data processing

Data processing

- Text-speech alignment: Astali [6]
- Syllable annotation: Praat [7]
- Data extraction: Matlab

Measurements

Centred f_o (ST):

$$f_{oc} = f_o \text{ trajectory} - median(f_o) \; for \; 1 \; speaker$$

Statistical analyses

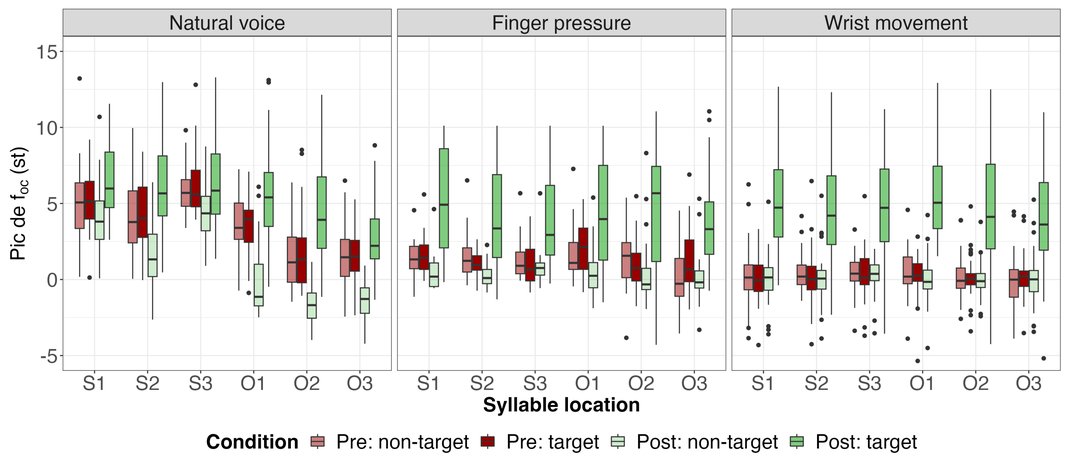
- Mixed linear regression model → syllable position, condition and interfaces
- Random factors : speaker and repetition
- Overall significance level : $p < 0.05$

Relative duration (D_r) of syllable:

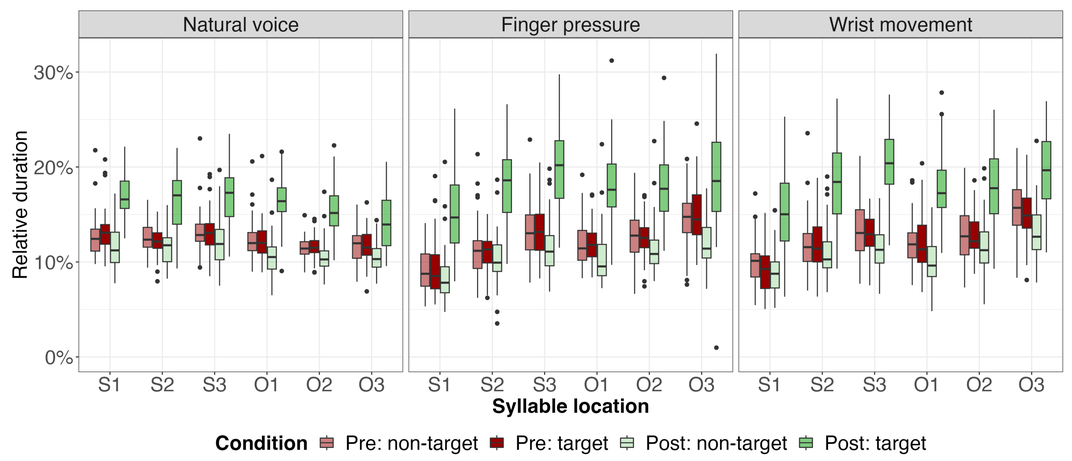
$$D_r = \frac{syllable \; duration}{sentence \; duration}$$

Results

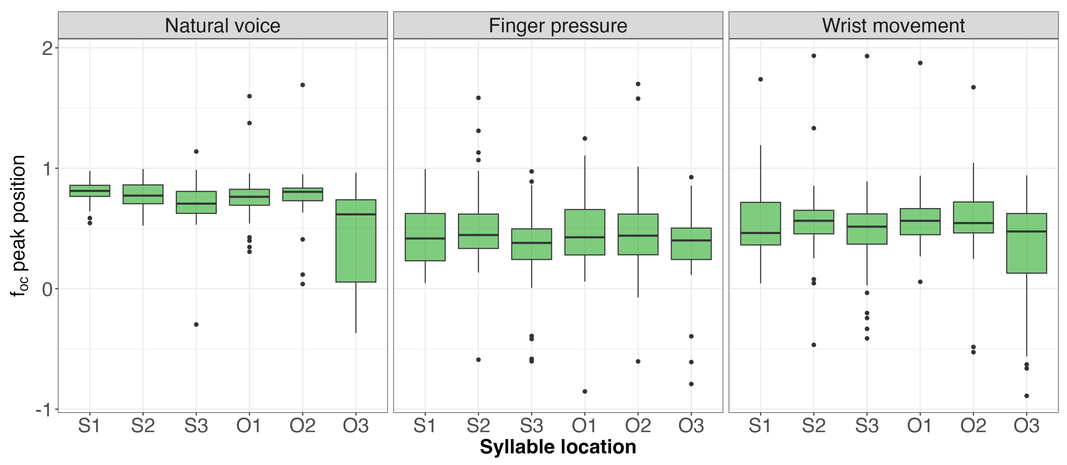
1. f_{oc} peak height of the [lu] syllables



2. Relative duration of the [lu] syllables



3. f_{oc} peak relative position on each 'post: target' [lu] syllable



Discussion

Focus production. 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 (1h)

- Articulatory level (D_r) and gestural level (f_{oc})
- Control of intonation with hand gestures was synchronised with syllable production → f_{oc} peak within the boundaries of the [lu] target syllable

Interfaces uses. Increased f_o on [lu] syllables demonstrated :

1. The speakers' intention to distinguish syllable from others by modulating intonation
2. Their awareness of the important role of its function for emphasising the target syllable

Comparison of interface usage. No significant differences between these two types of control, although we observed some specificities in their use

Perspective. These encouraging results call for the exploration of other linguistic functions in less controlled speech tasks, to fully validate such control paradigm in voice substitution applications

References

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

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

Grenoble Image Parole Signal Automatique
UMR CNRS 5216 - Grenoble Campus
38400 Saint Martin d'Hères - FRANCE