Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Peut-on marquer un focus contrastif par le geste manuel en suppléance vocale ?

Delphine Charuau Nathalie Henrich Bernardoni Silvain Gerber Olivier Perrotin
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, F-38000, France
delphinecharuau1@gmail.com, olivier.perrotin@gipsa-lab.grenoble-inp.fr

RÉSUMÉ

Un paradigme expérimental élicitant la focalisation sur une syllabe a été élaboré dans une tâche de conversion chuchotement-parole avec contrôle manuel de l'intonation. Deux interfaces de contrôle intonatif ont été testées : contrôle isométrique par pression du doigt et isotonique par rotation du poignet. La réalisation de la focalisation par le geste a été observée, démontrant un transfert du contrôle naturel vers manuel de l'intonation. Les résultats sont également discutés en fonction de la position de la syllabe dans l'énoncé, et en fonction de l'interface de contrôle gestuel employée.

ABSTRACT

Can a contrastive focus be achieved using hand gestures in a voice substitution paradigm ?

An experimental paradigm to elicit a focus on a syllable was developed, in a whisper-to-speech conversion task with manual control of intonation. Two interfaces for controlling intonation were tested : an isometric control by finger pressure and an isotonic control by wrist rotation. A succesful realisation of focus with gesture was observed, demonstrating a transfer from natural to manual control of intonation. Results are also discussed in terms of syllable position in the utterance, and with regards to the gestural control interface employed.

MOTS-CLÉS : Focus contrastif, Contrôle chironomique, Conversion chuchotement-parole, Suppléance vocale.

KEYWORDS: Contrastive focus, Chironomic control, Whisper-to-speech conversion, Voice substitution.

1 Introduction

Notre étude porte sur le contrôle de l'intonation dans le cadre d'une dégradation ou d'une absence des capacités phonatoires chez des patients laryngectomisés. Les solutions médicales actuelles pour remplacer la source vocale défectueuse ou absente consistent à injecter une source sonore artificielle dans le conduit vocal, souvent à l'aide d'un électrolarynx (Liu & Ng, 2007; Fuchs *et al.*, 2016; Kaye *et al.*, 2017; Ahmadi *et al.*, 2018). Ce vibrateur génère une source vocale de substitution sur laquelle l'utilisateur peut articuler la parole normalement. Il est également possible d'utiliser un microphone pour capter la parole non vocalisée (par exemple un chuchotement), et d'y réintroduire une phonation en temps-réel par synthèse vocale à partir d'un modèle de source glottique filtré par la réponse du conduit vocal mesurée (Perrotin & McLoughlin, 2020). La voix reconstruite est ensuite diffusée en temps-réel sur un haut-parleur. La principale limite à la fois des électrolarynx et des systèmes de

Actes des 35èmes Journées d'Études sur la Parole, pages 142–152.

synthèse est le peu d'information disponible pour reconstruire l'intonation. Par défaut, ces systèmes génèrent une intonation relativement constante, privant ainsi l'expression parlée d'une partie de l'information prosodique nécessaire à la fois à la structuration du discours (Mertens, 2008; Di Cristo, 2016) et à l'expression d'attitudes et d'émotions (Ward, 2019).

La synthèse vocale performative consiste à proposer à l'utilisateur un contrôle temps-réel de certains paramètres de la voix à générer, à l'image d'un instrument de musique numérique. En particulier, un axe d'étude a exploré l'utilisation de la chironomie, c'est-à-dire la gestuelle de la main, pour le contrôle de l'intonation dans la synthèse vocale (d'Alessandro, 2022). Ainsi, un tel paradigme permet d'aller chercher l'information d'intonation auprès d'un geste non-vocal (la main), en faisant l'hypothèse forte que l'utilisateur est capable de transférer une production de l'intonation implicite lorsque produite par la vibration des plis vocaux vers une production explicite par le geste de la main. Cette hypothèse a été validée dans des *tâches d'imitation*, où des utilisateurs ont reproduit le contour intonatif de phrases données en contrôlant la fréquence fondamentale de synthétiseurs vocaux par la position d'un stylet sur une tablette graphique, à la fois sur des tâches de parole (d'Alessandro *et al.*, 2011) et de chant (d'Alessandro *et al.*, 2014). Le transfert inverse, du contrôle manuel de l'intonation vers le contrôle naturel, a aussi été observé et s'est montré efficace dans l'apprentissage de l'intonation du français (Xiao *et al.*, 2022) et de l'anglais (Xiao *et al.*, 2023) en tant que langues étrangères, toujours en tâches d'imitation.

Dans la lignée de ces recherches, nous avons introduit la synthèse performative dans une solution de suppléance vocale basée sur la conversion chuchotement-parole en temps-réel. L'utilisateur doit alors à la fois articuler le message avec son conduit vocal et contrôler l'intonation de manière synchrone à l'aide du geste manuel (Perrotin & McLoughlin, 2020; Ardaillon *et al.*, 2022). Néanmoins, le paradigme de suppléance vocale diffère des études précédentes selon trois aspects. D'abord, si l'hypothèse de transfert du contrôle intonatif entre plis vocaux et geste de la main a été démontré sur des tâches d'imitations, ces dernières sont peu fréquentes en situation de communication orale. La question du transfert du contrôle intonatif sur des tâches que nous appellerons *tâches de production*, i.e. où l'utilisateur doit produire des contours intonatifs avec un but communicationnel mais sans références immédiates, reste donc ouverte. Par ailleurs, les synthétiseurs performatifs développés pour les études en tâche d'imitation demandent un contrôle de l'intonation uniquement, le contenu phonétique étant pré-défini (Feugère *et al.*, 2017; Locqueville *et al.*, 2020). En suppléance vocale, le contrôle simultané de l'articulation et de l'intonation doit être pris en compte. Enfin, ces mêmes instruments de synthèse proposent un contrôle de l'intonation par la position d'un objet (stylet ou doigt) sur une surface. Si cela introduit une modalité visuelle dans la représentation de l'intonation qui a été démontrée comme prépondérante dans le contrôle (Perrotin & D'alessandro, 2016), celle-ci n'est pas souhaitable dans des situations de communication orale, pour des raisons ergonomiques.

Dans cet article, nous cherchons donc à évaluer la capacité de locuteurs à externaliser le contrôle de l'intonation pour la réalisation de fonctions prosodiques dans un contexte communicationnel de suppléance vocale, c'est-à-dire dans des *tâches de production*, en synchronie avec l'articulation, et en utilisant des interfaces gestuelles ne sollicitant pas la modalité visuelle. Nous nous intéressons ici à la focalisation contrastive, qui se caractérise par la mise en évidence d'un ou plusieurs mots d'un énoncé, jugés comme les plus informatifs. Un focus contrastif est marqué par une augmentation notable de l'intensité et de la fréquence fondamentale f_o sur le ou les mots d'intérêt (Jun & Fougeron, 2000; Grice *et al.*, 2017), pouvant s'accompagner d'un allongement temporel de la syllabe portant le focus (Dahan & Bernard, 1996; Astésano *et al.*, 2004). La combinaison de ces paramètres contribue à la perception auditive du focus et permet une évaluation objective en termes de variations de durée syllabique et de fréquence fondamentale. Des variations fines de ces paramètres s'opèrent au niveau

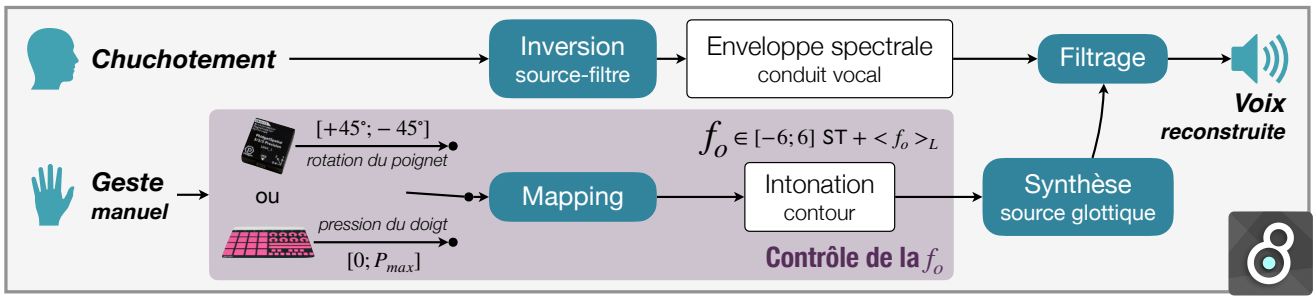

FIGURE 1 – Schéma du système de suppléance vocale avec contrôle manuel de l'intonation.

syllabique de manière complémentaire, lorsqu'il s'agit de marquer le focus sur l'ensemble du mot (Astésano *et al.*, 2004). En effet, les variations d'intensité et de f_o sont plutôt localisées sur la première syllabe ou au milieu du mot, tandis que les variations de durée arrivent à la fin du mot. Il s'agit de voir si les locuteurs reproduisent des dépendances à la position de la syllabe en suppléance vocale.

Nous posons une première hypothèse selon laquelle les locuteurs pourront combiner un contrôle implicite du focus sur le plan articulatoire, en allongeant la durée de la syllabe à mettre en relief, avec un contrôle explicite sur le plan gestuel, en élevant le contour intonatif sur la syllabe d'intérêt, par le biais de l'interface manuelle. Notre seconde hypothèse porte sur un transfert de l'effet de la position de la syllabe sur la f_o ainsi que sur la durée des syllabes. Le protocole expérimental pour l'élicitation d'un focus contrastif et l'évaluation de sa réalisation par le geste manuel est décrit en Section 2. Les résultats sont discutés en Section 3.

2 Matériel and méthodes

2.1 Système de suppléance vocale

Le système utilisé dans ces travaux et présenté en Fig. 1 est composé d'un module de conversion chuchotement-parole et d'interfaces gestuelles pour le contrôle de l'intonation, détaillés ci-dessous.

Conversion chuchotement-parole : Le système est une extension de la méthode proposée par Perrotin & McLoughlin (2020) qui consiste en : 1) la décomposition source-filtre du chuchotement par la méthode GFM-IAIF (Perrotin & McLoughlin, 2019), pour isoler l'enveloppe spectrale du conduit vocal du bruit coloré correspondant à la source sonore du chuchotement ; 2) la génération d'un signal de source glottique par le modèle LF (Fant *et al.*, 1994), de fréquence fondamentale f_o donnée ; 3) le filtrage du signal de source par l'enveloppe spectrale du conduit vocal. Ces trois étapes sont implémentées en temps-réel sur la plate-forme Max/MSP (Cycling74, 2024).

Interfaces gestuelles : La f_o utilisée en synthèse est contrôlée linéairement sur une échelle en demi-tons (ST), sur un intervalle d'une octave (± 6 ST) autour de la f_o moyenne de la voix du locuteur, notée $< f_o >_L$. Celle-ci est mesurée lors de la première phase d'entraînement du protocole sur de la parole naturelle (voir Section 2.2). Nous avons proposé deux types de geste pour le contrôle de l'intonation : un geste de *pression* isométrique et un geste de *rotation* isotonique. Un contrôle intonatif par geste de pression du pouce est déjà proposé dans la solution commercialisée et très usitée de l'électrolarynx Trutone (2024). Le geste de rotation du poignet s'inspire des gestes de battement qui peuvent accompagner la focalisation (Leonard & Cummins, 2011). Le geste de *pression* est réalisé sur

Scénario							Condition
– Participant :	Le	loup	doux	a suivi	le beau	loup .	*Pré*
– Expérimentatrice :	Le loup doux			a suivi	le beau chien ?		*Question*
– Participant :	Le	loup	doux	a suivi	le beau	loup .	*Post*

(a) Scénario pour l'expérience. La syllabe soulignée est celle ciblée par la question de l'expérimentatrice.

Syllabe		Énoncé			Contraste
cible	*non-cible*	*Sujet (S)*	*Verbe (V)*	*Objet (O)*	*Mot changé dans la 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

(b) Corpus d'énoncés.

TABLE 1 – Corpus (bas) et exemple de scénario sur un des énoncés (haut).

une tablette Sensel de la marque Morph (2024), qui mesure la pression de l'index de la main préférée de l'utilisateur, d'une pression nulle à une pression maximale P_{max}, respectivement associées à -6 et $+6$ ST autour de $< f_o >_L$. Le geste de *rotation* est réalisé à l'aide d'un accéléromètre 1044_1B de la marque Phidget (2024), tenu dans la main préférée de l'utilisateur, dont l'avant-bas est posé à l'horizontale sur un accoudoir. L'accéléromètre mesure le mouvement haut/bas du poignet de la main en degrés de rotation, $0°$ étant la position horizontale du poignet correspondant à $< f_o >_L$. Les rotations $+45°$ et $-45°$ sont respectivement associées à -6 et $+6$ ST autour de $< f_o >_L$. Ainsi, une descente du poignet est liée à une augmentation de f_o.

2.2 Protocole expérimental

Scénario : Nous avons construit un scénario permettant l'induction d'un focus contrastif sans donner d'instruction explicite, suivant les travaux de Dohen & Lœvenbruck (2009). Il s'agit d'une tâche de parole sous la forme d'interactions simulées entre le participant et l'expérimentatrice. L'interaction comporte trois tours de parole, résumés en Table 1a, dont le texte s'affiche au fur et à mesure sur l'écran disposé face au participant. Le participant a d'abord la consigne de lire un énoncé affiché. Ensuite, une question pré-enregistrée par l'expérimentatrice est affichée et jouée au participant. Cette question reprend l'énoncé du participant en y changeant un mot pour simuler une erreur de compréhension. Enfin, le participant a pour consigne de répéter l'énoncé initial qui s'affiche à l'écran. On appelle *Pré* et *Post* la première et deuxième répétition de la phrase par le participant.

Corpus : Le corpus est constitué de 3 phrases de 9 syllabes de type Sujet-Verbe-Objet (SVO), avec 3 syllabes par constituant. Le constituant verbal est fixe pour toutes les phrases et les constituants sujet et objet sont chacun composés de 3 mots monosyllabiques entièrement voisés (pour s'affranchir d'une décision de voisement dans la conversion chuchotement-parole). Chacun de ces deux constituants contient une syllabe /lu/ et chaque phrase apparaît deux fois dans le corpus, où soit le premier /lu/ soit le deuxième est la syllabe *cible*, c'est-à-dire changée par l'expérimentatrice. On appellera l'autre syllabe /lu/ *non-cible*. Au total, le corpus est composé de 6 énoncés, possédant chacun la syllabe *cible* à une ***position dans l'énoncé*** différente, comme indiqué en Table 1b.

Au final, lors de la production d'un scénario (Table 1a), la syllabe /lu/ est prononcée 4 fois par le

participant, selon 4 ***statuts de la syllabe*** : *Pré non-cible* (rouge clair), *Pré cible* (rouge foncé), *Post non-cible* (vert clair) et *Post cible* (vert foncé). Alors qu'aucune autre consigne que de lire chaque énoncé est donnée, nous faisons l'hypothèse que le participant produira naturellement un focus contrastif dans la condition *Post cible* uniquement. Dans la suite, on appellera *tâche de production* la réalisation des 6 scénarios correspondant à chacun des énoncés, répétés 3 fois chacun. L'ordre de présentation des scénarios et répétitions est aléatoire.

Productions vocales : L'expérience est divisée en trois phases associées à trois ***modes de production***, chacune précédée d'une ou plusieurs activités de familiarisation au protocole et/ou au contrôle. Dans la première, le participant utilise sa *voix* naturelle et commence par 3 répétitions de 6 scénarios sur des énoncés autres que ceux présentés en Table 1a pour se familiariser avec la tâche. C'est dans cette phase que $< f_o >_L$ est mesuré. Ensuite, le participant réalise la *tâche de production*. En deuxième et troisième phase, il utilise le système de conversion chuchotement-parole en contrôle par *rotation* puis *pression* ou inversement. L'ordre de passage de ces deux dernières phases est attribué aléatoirement selon le participant. Pour chacune de ces phases, le participant commence par la lecture du texte MonPage 2 (Pommée, 2021, p. 114) qui lui permet de se familiariser à l'interface par un contrôle libre de l'intonation. Le deuxième entraînement est une tâche d'imitation, où il s'agit de reproduire 6 phrases enregistrées par d'Alessandro *et al.* (2011) en imitant leur intonation, avec 3 répétitions chacune. Cela permet d'apprendre au participant à contrôler l'intonation. Enfin, la phase se termine par la réalisation de la *tâche de production*.

Conditions expérimentales : L'expérience s'est déroulée en chambre anéchoïque au laboratoire. Chaque participant est assis face à un écran sur lequel s'affiche les consignes et les supports des tâches de parole. Un casque audio fermé Beyerdynamic DT797 équipé d'un micro est utilisé pour capter la voix de l'utilisateur et lui restituer la voix de synthèse en temps-réel, ainsi que les stimuli sonores. Le participant devait passer manuellement au scénario suivant et pouvait faire des pauses librement à ces moments-là. L'expérience complète durait environ 1h15 et était rémunérée 15€ en bon d'achat dans un grand magasin.

Participants : Nous avons enregistré 16 locuteurs (âge médian = 24.5 ans ; Q1 = 22.5 ; Q3 = 27), de langue maternelle française et sans trouble rapporté de la parole, de l'audition et de la motricité du bras et de la main. Le protocole expérimental a été approuvé par le comité d'éthique de l'université Grenoble Alpes (CERGA-Avis-2023-21) et respecte le Règlement Général sur la Protection des Données.

2.3 Traitement des données

Extraction des données : L'alignement texte-parole des fichiers audio a été réalisé à l'aide de l'application Astali (Loria, 2016). La segmentation en syllabes ainsi que leurs annotations (*cible*/ *non-cible*) ont été réalisées manuellement sur Praat. Pour chaque syllabe, nous reportons sa durée relative par rapport à la durée de l'énoncé pour lequel les pauses ont été exclues, appelée D_r. Chaque énoncé ayant 9 syllabes, la durée relative moyenne d'une syllabe est de 11%. La f_o de la voix naturelle a été mesurée automatiquement par la fonction `To pitch` de Praat et la f_o contrôlée par le geste est fournie directement par le système. Ces valeurs sont exprimées en ST. Pour s'affranchir de l'effet du locuteur et de la production vocale, nous soustrayons à chaque trajectoire de f_o mesurée la médiane de f_o calculée pour l'ensemble des productions du locuteur avec la production vocale correspondante. Nous appelons f_{o_c} la fréquence fondamentale centrée résultante, exprimée aussi en ST. Dans cette étude, nous reportons le pic de f_{o_c} sur les syllabes d'intérêt (*cible* et *non-cible*).

Analyses statistiques : Nous avons étudié l'impact du ***statut de la syllabe***, de sa ***position dans l'énoncé***, et de son ***mode de production*** sur la durée relative et sur le pic de f_{oc}. Pour la durée relative, compte tenu du fait que ses valeurs sont bornées dans l'intervalle (0;1), nous avons appliqué une régression beta avec effet aléatoire et utilisé la fonction `glmmTMB` du package `glmmTMB` du logiciel R. Pour le pic de f_{oc}, nous avons appliqué un modèle linéaire mixte et utilisé la fonction `lme` de la librairie `nlme` du logiciel R. Dans les deux cas, le participant et le numéro de répétitions ont été ajoutés comme effets aléatoires du modèle. Nous avons utilisé ensuite la fonction `glht` de la librairie `multcomp` du logiciel R pour réaliser des comparaisons multiples d'où sont issues les p-values données ci-après. Les résultats sont considérés comme significatifs si $p < 0.05$.

3 Résultats

3.1 Réalisation de la focalisation

La figure 2 rend compte des valeurs du pic du f_{oc} sur les syllabe /lu/ et de leur durée relative, selon leur ***position*** au sein des constituants et selon le ***mode de production***. Il convient de préciser que pour une même position, les syllabes *cible* et *non-cible* appartiennent à des énoncés différents.

Validité du protocole : Aucune différence significative n'est observée entre les syllabes *Pré non-cible* (rouge clair) et *Pré cible* (rouge foncé), tant en termes de variation de f_o qu'en termes de durée, et ce, quel que soit le ***mode de production***. Ces résultats indiquent qu'il n'y a pas d'anticipation du focus et que nous avons bien induit une causalité entre question et focus observés en condition *Post* ci-après.

Réalisation de la focalisation : La figure 2 montre que le focus est réalisé d'une part, par une augmentation de l'intonation sur la syllabe *Post cible*, et d'autre part, par un allongement sa durée relative, quel que soit le ***mode de production***. En condition *voix*, les locuteurs ont tendance à marquer le focus en augmentant, en médiane sur l'ensemble des syllabes, le pic de f_{oc} de 1.64 ST sur la syllabe *Post cible* (vert foncé), par rapport à la syllabe *Pré cible*. Toutefois, sur chaque syllabe, cette différence n'est significative que lorsque le focus tombe sur la deuxième syllabe du constituant objet (O2). En revanche, durant les tâches de production avec interface, une augmentation significative du pic de f_{oc} entre l'ensemble des syllabes *Pré cible* et *Post cible*, de 3.28 ST pour le contrôle par *pression* et de 4.59 ST pour le contrôle par *rotation*, rend compte de la volonté des locuteurs de marquer le focus sur la syllabe cible. En outre, l'ensemble des syllabes *Post cible* sont significativement allongées de 3.8% en parole naturelle, et de 6% lors des tâches de parole impliquant un contrôle manuel de l'intonation. Au regard de ces résultats, il apparaît que la focalisation est réalisée sur le plan articulatoire (durée), ainsi que par le geste manuel (f_{oc}). Par ailleurs, la hausse du pic de f_{oc} est particulièrement marquée sur les interfaces.

3.2 Effet de la position de la syllabe

Afin de vérifier les effets de la ***position dans l'énoncé*** de la syllabe sur la réalisation du focus, nous comparons la durée relative et le pic de f_{oc} sur les syllabes *Pré cible* et *Post cible* en *voix* selon leur position au sein des constituants, puis en production *pression* et *rotation*.

Variations prosodiques en voix naturelle : La comparaison des valeurs de f_{oc} pour les 6 syllabes *Pré cible* selon leur position fait ressortir 11 combinaisons significatives sur les 15 testées. L'analyse

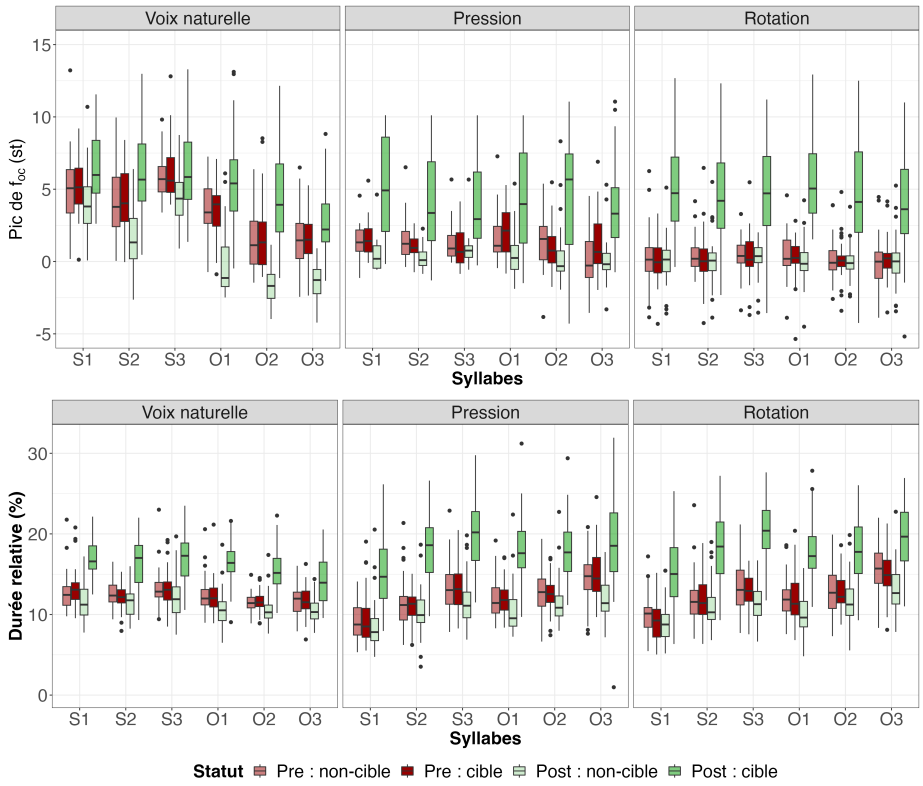

FIGURE 2 – Pic de f_{oc} (haut) et durée relative (bas) des syllabes /lu/ selon le ***statut de la syllabe*** et le ***mode de production***.

statistique révèle une variabilité significative du f_{oc} entre l'ensemble des syllabes du constituant objet. Cette variabilité s'atténue au sein du constituant sujet pour lequel une différence significative n'est observée qu'entre S2 et S3. Le nombre de combinaisons significatives rend compte d'une forte dépendance de la variation de f_{oc} à la position de la syllabe en condition *Pré*. Nous constatons donc une variation de f_{oc} des syllabes *Pré cible* selon leur position dans les constituants et dans la phrase, notamment une baisse significative de f_{oc} à la fin du constituant objet, coïncidant avec l'intonation descendante en fin de phrase. En revanche, la position de la syllabe ne présente que peu d'effet sur la variation du f_{oc} des syllabes *Post cible* : sur 15 combinaisons testées, seules 6 s'avèrent significatives. Nous n'observons aucune différence significative entre les syllabes d'un même constituant, sujet et objet, excepté entre O2 et O3. La réduction du nombre de différences significatives entre les syllabes *Pré cible* et *Post cible* suggère donc un plafonnement de f_{oc} lors de la réalisation du focus.

En parole naturelle, la durée des syllabes *Pré cible* reste relativement constante tout au long des constituants. Nous relevons une constance similaire dans la durée des syllabes *Post cible* au sein du constituant sujet. Une légère baisse de la durée de la syllabe initiale du constituant objet est relevée, tandis que la durée des syllabes en deuxièmes et troisièmes positions de ce constituant décroît significativement. De manière générale, les syllabes sont plutôt isochrones, exceptées pour les deux dernières syllabes de la phrase. Nous ne relevons pas de dépendance à la position de la syllabe.

Variations prosodiques en contrôle de l'intonation : Lors des tâches de parole impliquant un contrôle manuel de l'intonation, nous n'observons aucun effet de la position de la syllabe sur le contour de f_{oc}. En effet, aucune différence significative n'est observée selon la position syllabique, aussi bien pour les syllabes *Pré cible* que pour les syllabes *Post cible*, bien que nous retrouvons un plafonnement de la focalisation pour les syllabes *Post cible*, à l'instar de ce qui a été constaté en *voix*. Dans le contexte d'un contrôle externe de l'intonation, en dehors de la réalisation du focus, il y a

donc peu de mouvements intonatifs durant la production de la phrase. Les participants se concentrent exclusivement sur la tâche de focalisation, au détriment des autres fonctions de la prosodie.

Lors d'un contrôle par *rotation*, la durée relative des syllabes *Pré cible* augmente significativement au fur et à mesure que l'on s'approche de la fin du constituant, aussi bien sujet qu'objet. Si nous observons un schéma similaire lors d'un contrôle par *pression*, seule la hausse de durée entre les syllabes *Pré cible* initiale et finale du constituant objet est significative. Les syllabes *Post cible* montrent également une augmentation significative de leur durée au fur et à mesure de leur avancée au sein du constituant sujet, quelle que soit l'interface employée. Néanmoins, si une hausse similaire est également observée pour les syllabes du constituant objet, elle n'est pas significative. Nous observons un schéma d'allongement des syllabes au sein de chaque constituant, qui se confirme à la fois dans les contextes avec et sans focalisation, quel que soit le contrôle manuel exercé (*pression* et *rotation*). Ces allongements spécifiques aux interfaces pourraient être imputables au chuchotement, qui ralentit naturellement le débit de parole (Schwartz, 1967; Houle & Levi, 2020), et/ou à la hausse de la charge cognitive, induite non seulement par l'externalisation de l'intonation, mais également par la coordination entre l'articulation et le geste manuel.

4 Conclusion

Au regard des résultats présentés, nous pouvons conclure au succès du transfert de la production du focus à travers la variation de f_{oc} et de la durée sur la syllabe cible. En effet, lors des tâches de production impliquant un contrôle manuel de l'intonation, tous les locuteurs ont clairement explicité le focus en augmentant les valeurs de f_{oc} à l'endroit attendu. Ceci indique que les locuteurs ont non seulement perçu l'importance de f_{oc} dans la réalisation du focus, mais aussi qu'ils ont su utiliser les interfaces pour mettre en relief la syllabe cible. Ce comportement a été observé chez tous nos locuteurs. Par ailleurs, une augmentation significative de la durée des syllabes *Post cible* a également été constatée, à l'instar des données en parole naturelle attestées dans la littérature.

Contrairement aux observations faites en *voix*, lors des tâches de production impliquant un contrôle gestuel, nous ne remarquons aucune variation significative de f_o en dehors de la focalisation. Hormis lors de la réalisation du focus, la voix de synthèse est relativement monotone. Nous ne relevons que peu ou pas de variation de f_o, y compris à la fin des énoncés, où l'on pourrait s'attendre à une intonation descendante. Si les locuteurs ont montré une aptitude à reproduire précisément les contours intonatifs en tâche d'imitation (d'Alessandro *et al.*, 2011), dans notre tâche de production, les participants semblent s'être concentrés exclusivement sur la tâche de focalisation, au détriment des autres fonctions de la prosodie. La méthode employée se révèle néanmoins encourageante. En effet, cette étude met en évidence la capacité des participants à 1) comprendre qu'il fallait produire un focus ; 2) intégrer que le focus s'exprime en partie par une augmentation de f_o ; 3) planifier et mettre en œuvre ce contrôle par le biais des interfaces. Cependant, il apparaît que l'utilisation des interfaces ait été limitée à une seule fonction prosodique. Ceci suggère qu'un apprentissage complet de toutes les fonctions de la communication pourrait nécessiter un entraînement prolongé.

Au-delà de l'étude d'autres fonctions prosodiques, il conviendrait d'examiner le phénomène de ralentissement de la production des syllabes, observé en conditions *pression* et *rotation*, possiblement attribuable à la charge cognitive engendrée par la coordination entre l'articulation et le geste manuel, et/ou au chuchotement. Cette question suscite un certain intérêt dans le cadre de travaux futurs.

Références

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. *PLOS ONE*, **13**(2), e0192257. DOI : 10.1371/journal.pone.0192257.

ARDAILLON L., HENRICH N. & PERROTIN O. (2022). Voicing decision based on phonemes classification and spectral moments for whisper-to-speech conversion. In *Interspeech 2022*, p. 2253–2257 : ISCA. DOI : 10.21437/Interspeech.2022-10675.

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 *25èmes Journées d'Études sur la Parole*, p.ã, Fès, Maroc : AFCP.

CYCLING74 (2024). Max 8, http://cycling74.com.

DAHAN D. & BERNARD J.-M. (1996). Interspeaker variability in emphatic accent production in french. *Language and Speech*, **39**(4), 341–374. DOI : 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)*, p. 625–636, Noirmoutiers, France : ISCA. DOI : 10.21437/JEP.2022-66.

D'ALESSANDRO C., FEUGÈRE L., LE BEUX S., PERROTIN O. & RILLIARD A. (2014). Drawing melodies : Evaluation of chironomic singing synthesis. *The Journal of the Acoustical Society of America*, **135**(6), 3601–3612. DOI : 10.1121/1.4875718.

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*. Volume 1 de Études de linguistique française. De Gruyter. DOI : 10.1515/9783110479645.

DOHEN M. & LŒVENBRUCK H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech*, **52**(2-3), 177–206. PMID : 19624029, DOI : 10.1177/0023830909103166.

D'ALESSANDRO C., RILLIARD A. & LE BEUX S. (2011). Chironomic stylization of intonation. *The Journal of the Acoustical Society of America*, **129**(3), 1594–1604. DOI : 10.1121/1.3531802.

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)*, p. 1451–1454, Yokohama, Japan : ISCA.

FEUGÈRE L., D'ALESSANDRO C., DOVAL B. & PERROTIN O. (2017). Cantor digitalis : Chironomic parametric synthesis of singing. *EURASIP Journal on Audio, Speech, and Music Processing*, **2**. DOI : 10.1186/s13636-016-0098-5.

FUCHS A. K., HAGMULLER M. & KUBIN G. (2016). The New Bionic Electro-Larynx Speech System. *IEEE Journal of Selected Topics in Signal Processing*, **10**(5), 952–961. DOI : 10.1109/JSTSP.2016.2535970.

GRICE M., RITTER S., NIEMANN H. & ROETTGER T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics*, **64**, 90–107. Mechanisms of regulation in speech, DOI : https ://doi.org/10.1016/j.wocn.2017.03.003.

HOULE N. & LEVI S. V. (2020). Acoustic differences between voiced and whispered speech in gender diverse speakers. *The Journal of the Acoustical Society of America*, **148**(6), 4002. DOI : 10.1121/10.0002952.

JUN S.-A. & FOUGERON C. (2000). A phonological model of french intonation. *Intonation : Analysis, Modelling and Technology*, p. 209–242. DOI : 10.1007/978-94-011-4317-2_10.

KAYE R., TANG C. G. & SINCLAIR C. F. (2017). The electrolarynx : voice restoration after total laryngectomy. *Medical Devices : Evidence and Research*, **Volume 10**, 133–140. DOI : 10.2147/MDER.S133225.

LEONARD T. & CUMMINS F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, **26**(10), 1457–1471. DOI : 10.1080/01690965.2010.500218.

LIU H. & NG M. L. (2007). Electrolarynx in voice rehabilitation. *Auris Nasus Larynx*, **34**(3), 327–332. DOI : 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 Communication*, **125**, 97–113. DOI : 10.1016/j.specom.2020.10.002.

LORIA (2016). Astali. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

MERTENS P. (2008). Syntaxe, prosodie et structure informationnelle : une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de linguistique*, **56**(1), 97–124. DOI : 10.3917/tl.056.0097.

MORPH (2024). Surface tactile, https://morph.sensel.com.

PERROTIN O. & D'ALESSANDRO C. (2016). Seeing, Listening, Drawing : Interferences between Sensorimotor Modalities in the Use of a Tablet Musical Interface. *ACM Transactions on Applied Perception*, **14**(2), 1–19. DOI : 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)*, ICASSP '19, p. 7160–7164, Brighton, UK : IEEE. DOI : 10.1109/ICASSP.2019.8682625.

PERROTIN O. & MCLOUGHLIN I. V. (2020). Glottal Flow Synthesis for Whisper-to-Speech Conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**, 889–900. DOI : 10.1109/TASLP.2020.2971417.

PHIDGET (2024). Accéléromètre, https://www.phidgets.com/docs/Accelerometer_Guide.

POMMÉE T. (2021). *Les mesures d'intelligibilité : État de l'art, considérations pratiques pour l'applicabilité clinique et explorations acoustiques*. Thèse de doctorat, Université Toulouse III Paul Sabatier.

SCHWARTZ M. F. (1967). Syllable duration in oral and whispered reading. *The Journal of the Acoustical Society of America*, **41**(5), 1367–1369. DOI : 10.1121/1.1910487.

TRUTONE (2024). Electrolarynx, https://www.atosmedical.com/products/provox-trutone-emote-2.

WARD N. G. (2019). *Prosodic Patterns in English Conversation*. Cambridge University Press. DOI : 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)*, Journées d'Études sur la Parole " Parole, Geste, Musique : des unités à leur organisation ", p. 465–473, Noirmoutier, France : AFCP. HAL : hal-03838095.

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 *CHI '23 : CHI Conference on Human Factors in Computing Systems*, p. 1–9, Hamburg, Germany : ACM. DOI : 10.1145/3544548.3581210, HAL : hal-04113924.