Évaluation de la stylisation chironomique pour l'apprentissage de l'intonation du français L2

Xiao Xiao, Nicolas Audibert, Grégoire Locqueville, Christophe d'Alessandro, Barbara Kühnert, Rebecca Kleinberger, Claire Pillot-Loiseau

Évaluation de la stylisation chironomique pour l'apprentissage de l'intonation du français L2

Xiao Xiao[1, 3] Nicolas Audibert[1] Grégoire Locqueville[2] Christophe d'Alessandro[2] Barbara Kuhnert[1] Rébecca Kleinberger[3] Claire Pillot-Loiseau[1]

(1) Laboratoire de Phonologie et Phonétique, Sorbonne Nouvelle, UMR7018, Paris, France

(2) Institut Jean le Rond d'Alembert, Sorbonne Université, UMR7190 CNRS, Paris, France

(3) MIT Media Lab, Massachusetts Institute of Technology, Cambridge, USA

[xiao.xiao, nicolas.audibert, barbara.kuhnert, claire.pillot]@sorbonne-nouvelle.fr, [gregoire.locqueville, christophe.dalessandro]@sorbonne-universite.fr, rebklein@mit.edu

RÉSUMÉ

Cet article présente la nouvelle analyse d'une étude pilote sur l'imitation de la prononciation de phrases françaises dans laquelle des locuteurs natifs et non-natifs sont aidés d'une interface gestuelle de modulation de synthèse vocale. L'interface est envisagée pour aider l'apprentissage de l'intonation d'une langue étrangère et permet le contrôle en temps-réel de la prosodie d'une phrase cible par le tracé de contours mélodiques sur une tablette tactile. Le présent travail propose une analyse des données de l'étude pilote en utilisant des modèles mixtes additifs généraux (GAMM), afin de modéliser et comparer directement les trajectoires chironomiques et les courbes de f0. Cette approche confirme les résultats antérieurs : la chironomie guidée serait comparable à l'imitation vocale outre le timing dont le contrôle est difficile pour tous les sujets. Elle suggère en plus des différences notables entre locuteurs natifs et non-natifs pour certaines phrases sous certaines conditions.

ABSTRACT

Comparing Chironomic Stylization and Vocal Pronunciation of French Intonation with General Additive Mixed Models

This paper presents a follow-up analysis of a pilot study where native and non-native speakers imitated the pronunciation of French phrases using their natural voice and a gesture-controlled interface for the real-time modulation of vocal synthesis. A pilot study was conducted to better understand how this interface might be used for non-native speakers to practice the intonation of a foreign language. Previous analysis of the pilot data used Bayesian multilevel linear models on similarity scores of subjects' f0 curves compared to the reference curves. The present work reanalyzes the same data using general additive mixed models (GAMMs). This new analysis confirms the prior results—that guided chironomy without considering timing is comparable to vocal imitation and that timing control is difficult for all subjects. It also reveals previously undiscovered differences between conditions and between native and non-native speakers.

MOTS-CLÉS : intonation, acquisition de langue étrangère, geste, GAMM, synthèse vocale, interaction homme-machine.

KEYWORDS: intonation, second-language acquisition, gesture, GAMM, vocal synthesis, human-computer interaction..

1 Introduction

Notre recherche explore l'utilisation de la stylisation chironomique pour l'apprentissage de l'intonation d'une langue étrangère. "Stylisation chironomique" désigne ici la synthèse vocale avec l'intonation stylisée et contrôlée en temps réel par la gestuelle de la main. Une telle approche multimodale procure trois avantages pour l'apprentissage de l'intonation. Tout d'abord, elle peut permettre d'entraîner l'oreille à percevoir des caractéristiques vocales peu familières en les présentant à travers des modalités visuelles et kinesthésiques. Deuxièmement, le contrôle de la prononciation par des gestes de la main (chironomie) contourne les schémas enracinés dans la voix naturelle qui peuvent être difficiles à corriger (d'Alessandro *et al.*, 2011). Enfin, la synthèse vocale permet à un utilisateur de se concentrer sur le niveau suprasegmental sans se préoccuper des détails phonétiques fins au niveau segmental. L'hypothèse principale est que l'approche multimodale apportée par la chironomie (kinesthésique, visuelle et auditive) pourrait renforcer l'expérience sensorielle du sujet et l'aider dans le processus d'apprentissage (c'est-à-dire saisir et mémoriser les caractéristiques d'intonation).

Des travaux antérieurs sur l'intonation chironomique de phrases françaises avec des locuteurs natifs ont conclu à des résultats comparables entre imitation vocale et gestuelle (d'Alessandro *et al.*, 2011). L'acquisition de l'intonation d'une langue étrangère pourrait ainsi être facilitée par le recours à la chironomie. Dans une étude précédente, les participants ont utilisé une tablette graphique et un stylet pour contrôler la mélodie de phrases synthétisées, mais sans permettre la modification des paramètres rythmiques. Le présent travail vise à évaluer l'utilisation de la chironomie comme substitution vocale par les locuteurs non natifs, et nous nous intéressons au contrôle simultané du rythme et de la mélodie de la parole. Pour cela, une interface mobile, Gepeto, a été développée pour permettre la modulation en temps réel de la mélodie et la synchronisation de la prononciation synthétisée.

Une étude pilote, utilisant un paradigme d'imitation pour la désambiguïsation prosodique, a été menée pour évaluer la capacité des locuteurs natifs et non natifs à percevoir, contrôler et modifier des modèles d'intonation correspondant à différentes réalisations possibles de phrases ambiguës, affichées dans le tableau 1 (Xiao *et al.*, 2021).

L'étude pilote visait également à recueillir des informations pour de futures itérations de l'interface Gepeto et pour la conception d'études additionnelles. Notre première analyse des données pilotes consistait à quantifier l'écart entre les courbes de F0 des sujets et le contour de référence, les scores étant analysés à l'aide de modèles linéaires bayésiens à plusieurs niveaux. Les résultats ont suggéré que pour les locuteurs natifs et non natifs, l'imitation chironomique à l'aide d'un guide visuel est comparable en précision à l'imitation vocale, et que le contrôle de la synchronisation était une source de difficulté. Étonnamment, aucune différence significative n'a été trouvée entre apprenants étrangers et natifs avec l'analyse précédente.

Cet article réexamine les données pilotes à l'aide de modèles mixtes additifs généraux (GAMM), qui permettent l'analyse statistique de données dynamiques non-linéaires (Wood, 2017; Wieling, 2018; Sóskuthy, 2017). Au lieu d'être réduites à un seul score de similarité par rapport à la référence, les courbes F0 sont directement modélisées indépendamment de leur courbe de référence.

Dans cet article, nous présentons tout d'abord l'interface de contrôle gestuel (Section 2) et un bref résumé de l'étude pilote (Section 3). Nous décrivons ensuite les modèles construits pour déterminer (1) si les courbes de chaque paire de phrases présentent une différence significative et (2) s'il y a des différences significatives entre les deux groupes de sujets.

2 Interface de contrôle

L'architecture est basée sur Voks (Locqueville *et al.*, 2020), un synthétiseur vocal performatif de haute qualité qui permet, en temps réel, le contrôle mélodique et rythmique d'échantillons de parole précédemment enregistrés ou de synthèse texte-parole, grâce à l'utilisation de gestes de la main. Voks est une application Max/MSP basée sur le vocodeur WORLD (Cycling74, 2011; Morise *et al.*, 2016), initialement développée pour des applications de synthèse chantée, par l'utilisation d'une tablette graphique contrôlée par un stylet ou d'un thérémine (Xiao *et al.*, 2019; D'Alessandro *et al.*, 2019).

Compte tenu de la diffusion massive des appareils mobiles, nous avons développé une interface mobile personnalisée pour Voks, contrôlée par les mouvements du doigt (au lieu d'un stylet). Le logiciel fonctionne sur un ordinateur Mac OS X, connecté via la bibliothèque logicielle Websockets à un appareil mobile (ici une tablette Samsung Galaxy S2 de 9,7 pouces) exécutant l'interface Gepeto, avec une latence moyenne d'environ 10 ms dans chaque direction. Voir (Xiao *et al.*, 2021) pour plus de détails sur l'architecture et l'interface.

La phrase à contrôler apparaît graphiquement en haut de l'interface Gepeto. Sous la phrase cible, la zone de contrôle utilise le tracé du doigt pour contrôler la resynthèse de la phrase de Voks. L'axe horizontal détermine la position temporelle dans l'échantillon d'origine à resynthétiser, qui est segmenté en fonction de la subdivision des syllabes spécifiée par un fichier Praat TextGrid (Boersma & Weenink, 2019). Les syllabes sont indiquées sur l'écran en alphabet phonétique international (API). La langue française étant souvent considérée comme "isochronique" (Ramus, 2002), toutes les syllabes sont affichées avec une largeur égale pour cette étude. Différents rythmes peuvent être réalisés en changeant la vitesse du mouvement du doigt sur la surface de contrôle. L'axe vertical détermine la hauteur mélodique du signal de sortie sur une échelle en demi-tons (DT) espacée de manière régulière, avec une plage de 24DT (2 octaves) calibrée autour du corpus d'étude (116-466Hz). Une position verticale plus élevée donne un son de sortie plus aigu.

Chaque phrase cible est accompagnée d'un guide visuel montrant la courbe d'intonation stylisée de l'enregistrement de référence généré par Prosogram (Mertens, 2004). Basé sur un modèle perceptif de (d'Alessandro & Mertens, 1995), Prosogram simplifie la courbe de hauteur d'un enregistrement en segments de droite, la parole resynthétisée avec ces courbes de hauteur stylisées ayant été évaluée perceptuellement identique aux stimuli d'origine. Un guide visuel stylisé a été choisi pour faciliter le traçage et éviter que les artefacts vocaux micro-prosodiques ne perturbent les utilisateurs.

Par défaut, un geste dans la zone de contrôle laisse une trace colorée dont la persistance dépend du mode choisi. L'utilisateur a le choix entre le "mode en fondu" où les traces s'estompent après 1,5 secondes et le "mode maintenu" où les gestes restent à l'écran jusqu'à ce qu'ils soient effacés (Figure 1). En mode maintenu, les trois boutons suivants sont activés. Un bouton rejoue le geste. Le bouton suivant efface le geste et le dernier enregistre le geste sur le serveur. Le bouton dans le coin inférieur gauche déclenche la lecture de l'audio de référence.

3 Corpus, Sujets, Tâche

Six paires de phrases lexicalement ambiguës ont été sélectionnées à partir d'un corpus plus large (Table 1). Chaque paire est composée de la même séquence de phonèmes, mais les deux phrases ont des sens différents dus à leur intonation qui induit un découpage prosodique différent dans les deux

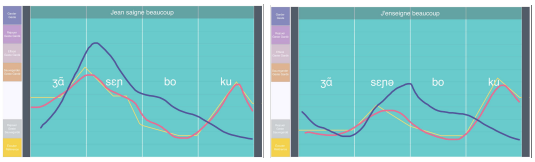

FIGURE 1 – Captures d'écran de l'interface Gepeto montrant une paire de phrases du corpus. Les lignes rose vif et violet foncé sont des traces gestuelles de l'utilisateur. Le rose s'estompe tandis que le violet reste affiché jusqu'à ce qu'il soit effacé et puisse être réécouté. La ligne jaune est le guide visuel.

cas. Les deux significations ont des probabilités d'occurrence à peu près égales. Les enregistrements de référence sont des énoncés déclaratifs lus par une locutrice francophone native.

Dix sujets ont participé à l'étude (2 hommes, 8 femmes, âgés de 20 à 48 ans, âge moyen 32,7 ans) [1]. Cinq d'entre eux sont des locuteurs non natifs issus de différentes langues maternelles : cantonais, portugais (deux sujets), mandarin, slovène. Tous ont suivi un cours de prononciation française d'un semestre et ont un niveau officiel avancé. Les 5 participants natifs sont des étudiants de premier cycle en orthophonie. Un non-natif et trois natifs ont déclaré avoir une expérience musicale (6-16 ans).

Un premier enregistrement est effectué, durant lequel il est demandé aux participants de lire les phrases en fonction de leur propre interprétation. Deuxièmement, les sujets sont invités à utiliser l'interface Gepeto pour imiter de leur mieux l'enregistrement de référence de chaque phrase. Lors de cette phase, les sujets doivent d'abord trouver un geste pour la phrase sans aide visuelle. Une fois le premier geste soumis, la courbe de hauteur stylisée de la phrase de référence apparaît et les sujets ont alors une autre chance de tracer un geste. Deux essais de familiarisation sont donnés pour la tâche d'imitation gestuelle. Enfin, les sujets enregistrent leur imitation vocale des phrases de référence.

Les paires de phrases appariées sont présentées en ordre aléatoire dans toutes les parties de l'étude. Pour les tâches d'imitation, aucune limite n'est imposée sur le nombre de fois où les références sont réécoutées, ni sur le temps pris pour imiter chaque phrase. L'ensemble de l'étude, y compris les instructions verbales et l'enquête d'information sur le sujet, dure entre 1h à 1h30. L'étude se déroule dans un studio insonorisé. Le son est transmis via un casque filaire et les enregistrements vocaux sont effectués à l'aide d'un microphone AKG C414 XLS connecté via une interface audio à un ordinateur portable Macbook Air. Un moniteur externe affiche la phrase en cours et offre une interface pour les sections d'enregistrement audio.

№de phrase	A	B (bis)
2	Tu parais très soucieux.	Tu paraîtrais soucieux.
7	Jean lève son verre.	J'enlève son verre.
8	Jean porte un journal.	J'emporte un journal.
10	Jean saigne beaucoup.	J'enseigne beaucoup.
11	Jean cadre la photo.	J'encadre la photo.
21	C'est la morsure.	C'est la mort sûre.

TABLE 1 – Corpus des phrases

1. L'expérience a été réalisée pendant la pandémie de COVID 19 et la disponibilité des sujets était limitée

4 Analyse

4.1 Préparation des données

Pour chaque participant, 4 prononciations par phrase avec différentes modalités sont recueillies : lecture vocale, imitation gestuelle non guidée, imitation gestuelle guidée et imitation vocale. Une extraction automatique de la fréquence fondamentale (F0) des enregistrements vocaux des sujets est effectuée à l'aide de Praat puis validée manuellement. La segmentation des énoncés dans chaque enregistrement est étiquetée au format TextGrid. Chaque enregistrement est normalisé pour égaliser la F0 moyenne et normalisé dans le temps en rééchantillonnant les valeurs de F0 à des intervalles de 10 millisecondes dans l'enregistrement de référence.

Les énoncés d'imitation gestuelle sont encodés sous la forme d'une série de points dans un espace tridimensionnel. Chaque point est basé sur une position 2D touchée par le sujet dans la région de contrôle de l'interface Gepeto et comprend les informations suivantes :

- `f` : la fréquence en demi-tons relative à la fréquence la plus basse dans l'interface Gepeto
- `scrub` : l'instant de référence dans l'enregistrement original, normalisé sur une échelle de 0 à 1
- `t` : le temps d'apparition du point courant mis à l'échelle de 0 à 1 en fonction du temps de l'ensemble du geste.

Dans le cas d'une parfaite synchronisation temporelle entre le geste effectué par le sujet et le modèle de phrase à imiter, les valeurs de `scrub` et `t` sont donc égales.

Deux types de données gestuelles ont été analysés : l'un avec les données t pour la dimension temporelle, qui représente les courbes de fréquence des sujets avec leur contrôle de synchronisation, un autre avec les données de "scrub" pour la dimension temporelle, qui représente à quel point les tracés F0 des sujets ressemblaient au guide courbes indépendamment du temps. Une modélisation statistique a ainsi été réalisée sur 4 conditions : lecture, imitation vocale, ainsi que deux versions de données chironomiques, les unes non guidées et les autres guidées.

4.2 Modélisation statistique

À l'aide de la commande `bam` du package `mgcv` du language de programmation R, deux types de modèles mixtes additifs généraux (GAMM) ont été construits pour répondre à chacune des deux questions de recherche (Wood, 2017, 2011). Le premier type examine la différence de F0 entre les versions A et B de chaque phrase pour chaque condition et chaque type de sujet. Le deuxième type examine la différence entre les sujets natifs et non natifs pour chaque phrase et chaque condition distincte. Des modèles distincts ont été créés pour chaque sous-ensembles de données pour faciliter leur interprétation. Le type de spline de lissage par défaut ("tp") est utilisé pour tous les modèles, et une correction pour l'autocorrélation est incluse pour tous les modèles.

Question 1 : Différence significative entre les types de phrases

Le premier type de modèles s'exprime par la formule suivante :

```
f ~ typePhrase + s(t) + s(t, by=typePhrase) # effets fixes
    + s(t, subject, by="fs", m=1)   # effets aléatoires
    + s(t, subject, by=typePhrase, bs="fs", m=1)
```

`typePhrase` est une variable binaire définie sur `TRUE` pour les versions B de la phrase et `FALSE` sinon. Elle est incluse dans la spécification du modèle en tant qu'effet fixe, pour prendre en compte les différences constantes, et en tant que spline de lissage, pour tenir compte des différences non linéaires. Pour tenir compte de la variabilité entre sujets, deux lissages aléatoires sont ajoutés à la spécification du modèle, représentant les différences globales non linéaires pour chaque sujet et les différences liées aux versions A et B de la phrase.

Étant donnée la localisation variable de la première séparation prosodique entre les paires de phrases, des instances distinctes de ce modèle ont été créées pour chaque paire de phrases.

Question 2 : Différence significative entre les sujets natifs et non natifs

Le second type de modèles s'exprime par la formule :

```
f ~ isNative + s(t) + s(t, by=isNative)   # effets fixes
    + s(t, subject, bs="fs", m=1)   # effet aléatoire
```

La variable binaire `isNative` indique pour chaque sujet s'il est francophone natif ou non. Elle est incluse dans le modèle à la fois en tant qu'effet fixe et en tant que spline de lissage. Un seul lissage aléatoire basé sur le sujet capture les différences individuelles entre les locuteurs. Un modèle séparé a été construit pour chaque phrase individuelle, dans chaque condition.

4.3 Test de signification

Sur la base des recommandations de (Sóskuthy, 2017), deux méthodes sont utilisées pour tester la significativité de chaque modèle. Pour la première méthode, chaque modèle est comparé à une version simplifiée excluant les termes paramétriques et courbes lissées à effet fixe tout en gardant la même courbe lissée de base et les mêmes effets aléatoires. La deuxième méthode repose sur l'affichage des différences entre conditions intégrées dans le package `itsadug` via la fonction `plotdiff`, qui indique visuellement les régions significativement différentes (van Rij *et al.*, 2020). Une différence est considérée comme significative si le modèle à effets fixes est évalué significativement meilleur que le modèle de base ($p < 0{,}05$) et si l'inspection visuelle de la courbe de différence indique que la zone dans laquelle les conditions comparées diffèrent significativement correspond bien à celle attendue, en l'occurrence en excluant les parties initiales et finales des énoncés comme illustré par la figure 2. Un résumé de la significativité pour tous les modèles construits est présenté dans les tableaux 2 et 3.

5 Résultats et discussion

Les résultats des deux analyses sont résumés dans les tableaux 2 et 3. Pour les sujets natifs et non natifs, les versions A et B de toutes les phrases présentent des différences significatives dans l'imitation vocale. Pour l'imitation chironomique avec les données temporelles des sujets, seulement la moitié des paires de phrases ont des courbes lissées significativement différentes pour les deux versions. Les phrases dont les courbes lissées sont significativement différentes varient selon les sujets natifs et non natifs. Lorsque le temps n'est pas pris en compte, les deux groupes de sujets distinguent les phrases A et B pour 5 des 6 paires de phrases à l'aide du guide. Lorsqu'ils ne sont pas guidés, les locuteurs non natifs tracent des courbes significativement différentes pour 4 paires de phrases sur 6,

tandis que les locuteurs natifs n'ont des courbes significativement différentes que pour 2 paires de phrases. Ces résultats sont cohérents avec l'analyse précédente, la chironomie guidée sans timing est comparable à l'imitation vocale et le timing apparaît comme une source de difficulté.

L'analyse révèle un résultat nouveau et inattendu : les courbes des locuteurs natifs et non natifs ne sont pas significativement différentes pour la plupart des paires lors de la lecture initiale en fonction de leur propre interprétation des phrases. Cela suggère que malgré le contrôle inhérent à la tâche de lecture proposée, la distinction entre les paires de phrases est facilement cachée par la variabilité de prononciation dans chaque version de la phrase.

Des différences significatives entre les sujets non-natifs et natifs n'ont pu être trouvées que dans certaines des phrases. Il est intéressant de noter qu'aucune différence significative n'a été trouvée pour aucune des phrases en condition d'imitation vocale et de chironomie guidée, ni avec le timing propre aux sujets, ni en considérant le timing relatif (scrub). Ce résultat suggère que la chironomie guidée permet aux non-natifs d'atteindre le même degré de précision que les natifs.

	Condition					
№ phrase	**Lecture**	**Imitation Vocale**	**Chironomie Non Guidée**	**Chironomie Guidée**	**Chiro. Non Guidée Scrub**	**Chiro. Guidée Scrub**
	Sujets Natifs					
2	Non	Oui	Non	Non	Non	Non
7	Non	Oui	Oui	Oui	Non	Oui
8	Non	Oui	Non	Non	Oui	Oui
10	Non	Oui	Oui	Oui	Oui	Oui
11	Non	Oui	Oui	Oui	Non	Oui
21	Oui	Oui	Non	Non	Non	Oui
	Sujets Non Natifs					
2	Oui	Oui	Oui	Non	Oui	Oui
7	Non	Oui	Oui	Oui	Non	Oui
8	Non	Oui	Non	Oui	Oui	Oui
10	Non	Oui	Non	Non	Non	Oui
11	Non	Oui	Non	Non	Oui	Non
21	Oui	Oui	Oui	Oui	Oui	Oui

TABLE 2 – Résumé des situation où une différence significative peut être trouvée entre les courbes A et B pour chaque phrase

	Condition		
№ phrase	**Lecture**	**Chironomie Non Guidée**	**Chiro. Non Guidée Scrub**
2A Tu parais très soucieux	Non	Oui	Oui
2B Tu paraîtrais soucieux	Oui	Oui	Oui
11A Jean cadre la photo	Non	Oui	Non
11B J'encadre la photo	Non	Oui	Non
21A C'est la mort sûre	Non	Non	Oui

TABLE 3 – Phrases et conditions dans lesquelles les courbes des locuteurs natifs diffèrent significativement de celles des non natifs. Aucune différence significative n'a été trouvée pour toutes les combinaisons d'expressions et de conditions ne figurant pas dans ce tableau.

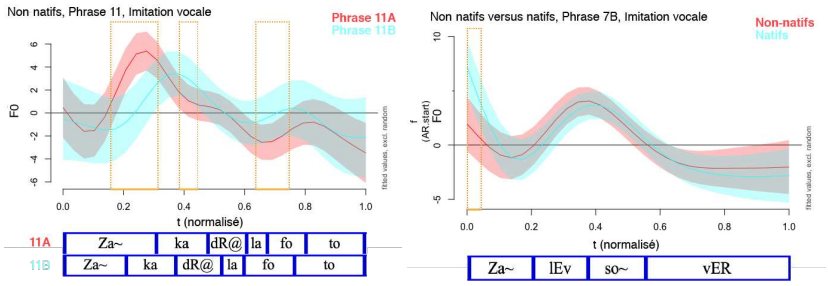

FIGURE 2 – Trajectoires modélisées et zones dans lesquelles la différence entre conditions est significative. Les différences au début des énoncés (graphique à droite) sont éxclues.

6 Conclusions

Les résultats de cette analyse montrent que les modèles de type GAMM apparaissent comme un outil approprié pour l'analyse des trajectoires gestuelles et leur comparaison avec les courbes de fréquence fondamentale. A ce titre, l'identification visuelle des zones dans lesquelles les divergences entre conditions sont les plus importantes pourrait se révéler particulièrement utile dans une optique didactique. Confirmant les résultats de (Xiao *et al.*, 2021), le contrôle du temps s'est avéré plus complexe tant pour les sujets natifs que non-natifs et sera donc un objectif prioritaire des travaux visant à l'amélioration de l'interface Gepeto. Il serait également utile de collecter des données sur chaque tentative du sujet plutôt que sur l'essai final uniquement.

Plus de soin peut également être apporté au choix des sujets, d'abord en effectuant à nouveau l'expérience auprès de sujets non-natifs de niveau moins avancé que ceux de cette expérimentation. En outre, au lieu de traiter les sujets non natifs comme un groupe uniforme, nous pouvons imaginer un diagnostic pour identifier des problèmes spécifiques dans l'intonation d'un sujet, comme détecter la différence entre montée et descente d'intonation ou l'exécution d'un schéma rythmique particulier. L'interface Gepeto pourrait être adaptée en un outil pour déterminer si le problème est enraciné dans la capacité à écouter, la capacité à contrôler la voix naturelle ou une erreur dans la représentation interne d'un son.

Enfin, le fait que les sujets natifs aient eu des difficultés à différencier les paires de phrases dans la condition de lecture suggère que les études futures devraient accorder plus de soin à l'étude de la variabilité de production de ce type de phrase par plusieurs francophones natifs. Néanmoins, cette analyse renforce la constatation que la chironomie a un potentiel en tant qu'outil de pratique de l'intonation.

Remerciements

Ce travail est soutenu par le projet ANR GEPETO (ANR-19-CE28-0018), ainsi que par le programme « Investissements d'Avenir » Labex EFL (ANR-10-LABX-0083). Il contribue à l'IdEx Université de Paris (ANR-18-IDEX-0001).

Références

BOERSMA P. & WEENINK D. (2019). Praat : doing phonetics by computer [computer program]. http://www.praat.org/ Version 6.1.08. Accessed : 2021-03-21.

CYCLING74 (2011). Max 6. https://cycling74.com/. Accessed : 2021-03-21.

D'ALESSANDRO C. & MERTENS P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language*, **9**(3), 257–288.

D'ALESSANDRO C., RILLARD A. & LEBEUX S. (2011). Chironomic stylization of intonation. *Journal of the Acoustical Society of America*, **3**(129), 1594–1604.

D'ALESSANDRO C., XIAO X., LOCQUEVILLE G. & DOVAL B. (2019). Borrowed voices. In *International Conference on New Interfaces for Musical Expression NIME'19*, p. 2.2–2.4, Porto Alegre, Brazil.

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.

MERTENS P. (2004). The prosogram : Semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of Speech Prosody 2004*, Nara, Japan.

MORISE M., YOKOMORI F. & OZAWA K. (2016). World : A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, **E99.D**, 1877–1884.

RAMUS F. (2002). Acoustic correlates of linguistic rhythm : Perspectives. *Proceedings of Speech Prosody 2002*.

SÓSKUTHY M. (2017). Generalised additive mixed models for dynamic analysis in linguistics : a practical introduction.

VAN RIJ J., WIELING M., BAAYEN R. H. & VAN RIJN H. (2020). itsadug : Interpreting time series and autocorrelated data using gamms. R package version 2.4.

WIELING M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling : A tutorial focusing on articulatory differences between l1 and l2 speakers of english. *Journal of Phonetics*, **70**, 86–116.

WOOD S. (2017). *Generalized Additive Models : An Introduction with R*. Chapman and Hall/CRC, 2 edition.

WOOD S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, **73**(1), 3–36.

XIAO X., AUDIBERT N., LOCQUEVILLE G., D'ALESSANDRO C., KUHNERT B. & PILLOT-LOISEAU C. (2021). Prosodic Disambiguation Using Chironomic Stylization of Intonation with Native and Non-Native Speakers. In *Proc. Interspeech 2021*, p. 516–520.

XIAO X., LOCQUEVILLE G., D'ALESSANDRO C. & DOVAL B. (2019). T-Voks : the singing and speaking theremin. In *NIME 2019 International Conference on New Interfaces for Musical Expression*, p. 110–115, Porto Alegre, Brazil.