INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Prosodic disambiguation using chironomic stylization of intonation with native and non-native speakers

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

[1]LPP, Sorbonne Nouvelle, France
[2]LAM, Sorbonne Université, France

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

Abstract

This paper introduces an interface that enables the real-time gestural control of intonation in phrases produced by a vocal synthesizer. The melody and timing of a target phrase can be modified by tracing melodic contours on the touch-screen of a mobile tablet. Envisioning this interface as a means for non-native speakers to practice the intonation of a foreign language, we present a pilot study where native and non-native speakers imitated the pronunciation of French phrases using their voice and the interface, with a visual guide and without. Comparison of resulting F0 curves against the reference contour and a preliminary perceptual assessment of synthesized utterances suggest that for both non-native and native speakers, imitation with the help of a visual guide is comparable in accuracy to vocal imitation, and that timing control was a source of difficulty.

Index Terms: human-computer interaction, intonation, second-language acquisition

1. Introduction

Our research explores how chironomic stylization can be used by non-native speakers for intonation practice of a foreign language. "Chironomic stylization" means here vocal synthesis using real-time hand gesture control of stylized intonation patterns. Such multimodal practice addresses three sources of difficulty for intonation learning. First, it can train the ear to perceive unfamiliar features in speech by presenting them through visual and kinesthetic modalities. Second, control of pronunciation with hand gestures (chironomy) bypasses ingrained patterns in the natural voice that are difficult to correct [1]. Finally, vocal synthesis enables a learner to focus on the suprasegmental level without being preoccupied by fine-phonetic detail on the segmental level. Prior research has shown the benefits of pitch visualization [2] and pitch gestures [3] for L2 intonation learning. We hypothesize that the multimodal approach provided by chironomy (kinesthetic, visual and auditory) can similarly reinforce the sensory experience of the learner and help in the learning process (i.e. grasping and memorizing intonation features).

An imitation paradigm for prosodic disambiguation has been chosen for assessing the ability of both native and non-native speakers to perceive, control and modify linguistically meaningful intonation patterns. This paradigm proved useful for studying intonation in language acquisition tasks [4, 5, 6]. Prior work on chironomic intonation of French phrases with native speakers yielded similar results for vocal and gestural imitation [1]. In other words, chironomy can be a substitute for the human voice. Our work seeks to assess the feasibility of chironomy as vocal substitution for non-native speakers. The previous study, using a graphic tablet and stylus, allowed only the modulation of melody, with no change of rhythmic parameters.

Our work examines the simultaneous control of speech rhythm and melody. To this end, a mobile interface was developed that enables the control of both melody and timing of the synthesized pronunciation (Section 2). A performance and perception pilot (with native and non-native speakers of French) was conducted using a corpus of ambiguous French phrases (with identical phonemic content that change meaning according to intonation) (Section 3). The pilot addresses the following questions: 1/ How does chironomic intonation compare with vocal intonation for phrase disambiguation? 2/ How much do visually guided and non-guided chironomic intonations differ? 3/ To what extent does the additional control of timing add to the difficulty of the task? 4/ Do non-native and native speakers differ in their performance in different modalities?

2. Chironomic Control Interface

2.1. Performative Synthesis Architecture

The architecture is based on Voks [7], a high-quality performative vocal synthesizer that enables the real-time melodic and rhythmic control of previously recorded or Text-to-Speech speech samples, through the use of hand gestures. It is a MaxMSP application based on the WORLD vocoder [8, 9] initially developed with singing synthesis applications in mind, using a stylus on a graphic tablet or a Theremin [10, 11]. Given the widespread availability and popularity of mobile devices and their apps, we built a custom mobile interface for Voks, controlled by finger motions (instead of a stylus). It runs on a Mac OS X computer, which also runs a Node.js server that allows external devices to control Voks wirelessly. This server-client architecture was created to simulate the user experience of a mobile app, enabling proof-of-concept testing without the need to implement the synthesis software on a new platform.

When connected to the same WiFi network as the host computer, a mobile tablet or phone (herein a 9.7in Samsung Galaxy S2 tablet) can open the Gepeto interface in a Chrome browser. Communication between the interface and server uses the Websockets protocol. Messages for Voks are first sent to the server, where they can be saved, and then routed to Max via Open Sound Control (OSC). One way latency between the mobile device and the computer averages around 10ms. Subjects can control intonation for recorded sentences using the tip of their fingers on the touch-screen of the mobile device.

http://dx.doi.org/10.21437/Interspeech.2021-182

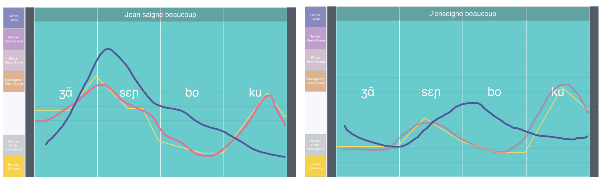

Figure 1: *Screenshots of the Gepeto interface showing a pair of phrases from the corpus. The hot pink and dark purple lines are gesture traces from the user. The pink fades away after 1.5 second while the purple remains until it is erased and can be played back. The yellow line is the visual guide, a stylized version of the reference phrase's F0 curve.*

2.2. Finger Intonation Control

The phrase to be controlled is shown at the top of the Gepeto interface. Below the target phrase is the control region, where tracing one's finger outputs a resynthesis of the phrase from Voks. The horizontal axis determines the temporal position in the original sample to resynthesize. It is divided based on syllable segmentation specified by a Praat TextGrid file [12]. Syllables are indicated with the International Phonetic Alphabet (IPA). For the present study, all syllables appeared with equal width in the interface, as French is often said to be ”isochronic” [13]. Different rhythms can be realized by changing the speed of the finger's movement across the control surface (the ”scrub” mode in Voks [7]). Although only syllable-level segmentation is displayed, intra-syllabic temporal modifications are possible based on the speed of finger movement within each syllabic region. The vertical axis determines the output frequency regularly spaced on a semitones (ST) scale, with a range of 24ST (2 octaves) calibrated around the study corpus (116-466Hz). A higher vertical position results in a higher output sound.

A visual guide was made for each target phrase, which shows the stylized intonation curve of the reference recording generated by Prosogram [14]. Based on a perceptual model by [15], Prosogram simplifies a recording's pitch curve into straight line segments. Speech resynthesized with these stylized pitch curves are perceptually identical to the original stimuli. A stylized visual guide was chosen for ease of tracing and to avoid distraction from micro-prosodic vocal artifacts.

By default, a gesture in the control region leaves a hot pink trace that fades away after 1.5 seconds. The top button in the buttons panel toggles between the default ”fade mode” and “held mode” where gestures stay on the screen until erased. In held mode, gesture traces appear in dark purple. When in held mode, the next three buttons are activated. One replays the gesture, with each point highlighted as it is resent to Voks. The next erases the gesture, and the last saves the gesture to the server. The button on the bottom left corner triggers the reference audio to play. When the visual guide is on, a cursor moves along the curve to indicate playback position.

3. Prosodic Disambiguation Pilot

3.1. Subjects, Corpus and Task

Ten subjects took part in the pilot study (2 male, 8 female, aged 20-48, mean age 32.7). Five were non-native speakers with different L1 backgrounds: Cantonese (S1), Portuguese (S2, S3), Mandarin (S6), Slovenian (S10). Subjects all reside in France and have lived in France for 2.5-7.5 years. All have completed a semester-long course on French pronunciation and have DELF-DALF level between B1 and C2 [16]. The 5 native speakers were undergraduate students in speech therapy. One non-native and three native subjects have musical experience (6-16 years).

Table 1: *Corpus of phrases*

Id #	Phrase #	Phrase #bis
2	Tu parais très soucieux.	Tu paraîtrais soucieux.
	You seem very worried.	*You would seem worried.*
7	Jean lève son verre.	J'enlève son verre.
	Jean lifts his glass.	*I lift his glass.*
8	Jean porte un journal.	J'emporte un journal.
	Jean carries a newspaper.	*I carry a newspaper.*
10	Jean saigne beaucoup.	J'enseigne beaucoup.
	Jean is bleeding a lot.	*I teach a lot.*
11	Jean cadre la photo.	J'encadre la photo.
	Jean frames a photo.	*I frame a photo.*
21	C'est la morsure.	C'est la mort sûre.
	It's the bite.	*It's death for sure.*

Six pairs of lexically ambiguous French phrases (Table 1) were selected from a larger corpus used in the above-mentioned French pronunciation course, with phrases from [17, 18] and the instructor. Each pair shares the same sequence of phonemes and syllabic segmentation, but can take on two different meanings with roughly equal plausibility, based on the location of the first prosodic word boundary, marked by an intonational rise and the elongation of the last syllable of the group. Reference recordings featured a female native francophone speaker reading each as a declarative utterance.

Subjects first recorded themselves reading the phrases based on their own interpretation of the phrase. Next, subjects used the Gepeto interface to imitate the reference recording of each phrase to the best of their ability. Initially, subjects were asked to find a gesture for the phrase without any visual guidance. After the first gesture is submitted, the stylized pitch curve for the reference phrase appeared, and subjects were given another chance to find a gesture. Two familiarization trials were given for the gestural imitation task. Finally, subjects recorded vocal imitations of the reference phrases.

Phrases appeared in random order in all parts of the study, with paired phrases next to each other. For the imitation sections, no limits were imposed on the number of times references are played and the amount of time subjects spent finding the pronunciation of each phrase. The entire study, including verbal instructions and the subject information survey took between 1 and 1.5 hours. The study took place in a sound isolated studio. All audio was heard through monitor headphones, and voice recordings were made with an AKG C414 XLS microphone connected via an audio interface to a Macbook Air laptop computer. An external monitor displayed the current phrase and a graphical user interface for the audio recording sections.

3.2. Intonation Contours comparison

The study collected 4 pronunciations per phrase with different modalities: vocal reading, non-guided gestural imitation, guided gestural imitation, and vocal imitation. Intonation contour distances were computed for each utterance.

3.2.1. F0 analysis and Intonation Contour Determination

To extract the F0 of subjects' vocal recordings, pitch analysis was performed with Praat and then manually verified. The start-

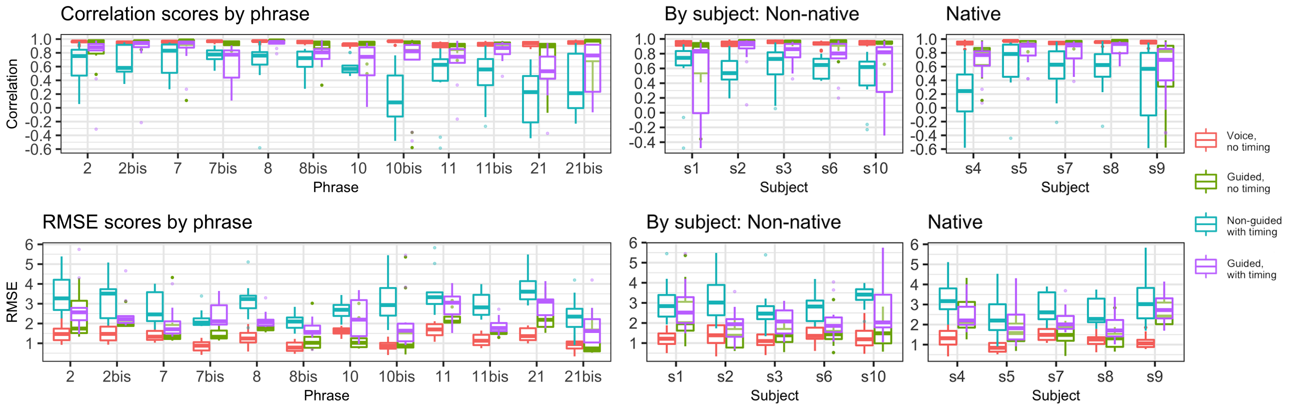

Figure 2: *Boxplots of scores used in statistical analysis. Horizontal line indicates median, hinges indicate 25th and 75th percentiles.*

ing and ending timestamp of the utterance within each recording were labeled in a Praat TextGrid. To compare with the results of [1], vocal imitations were fist aligned with the reference recordings using Dynamic Time Warping, which removes differences in timing. F0 values are sampled and compared at 10-millisecond intervals in the reference.

Gestural imitation utterances were encoded as a series of data points. Each point is based on a 2D position touched by the subject in the control region of the Gepeto interface and includes the following information:

- *f*: the frequency in semitones (ST) relatives to the lowest frequency in the Gepeto interface
- *scrub*: the point in the original recording where phonemic information is taken, specified from 0 to 1
- *t-start*, *t-end*: the start and end time of the current point, relative to the start time of the gesture

3.2.2. Intonation Contours Distances

F0 contours were compared using two distance measures [19, 1]: correlation between contours and weighted Root Mean Square Error (RMSE), i.e. differences between contours. For both measures, the mean of each curve was subtracted from the F0 contours to normalize global register differences between individual voices.

Using the wCorr R package, Pearson's correlation was calculated between the reference and an imitation F0 contour, weighed by the intensity of the original contour to give more importance to phonemes with a higher sound level [20, 21]. Correlation measures similarity between two curves, and is 1 for identical curves. The mltools package [22] was used to compute the RMSE, which is 0 for identical contours and increases for divergent curves. Vocal contours were compared with the original F0 of the reference while gestural contours were compared with Prosogram stylizations of the reference curves. Between the stylized and original contours, mean correlation is 0.94 (SD: 0.04) and mean RMSE is 1.11 semitones (SD: 0.32 ST). For stylized contours across phrase pairs, mean correlation is 0.51 (SD: 0.38) and mean RMSE is 4.47 ST (SD: 0.91 ST).

Two sets of correlation and RMSE scores were computed for each gestural imitation, with F0 values sampled and compared at 10-millisecond intervals in the reference. For both, the region of interest in the reference file is determined by the first and last scrub values in the gesture. One set retains the original timing of the gesture and linearly scales the t-start values to match the length of the reference, interpolating when necessary. A second set of scores aligns the gesture by linearly scaling the gesture's scrub values in the region of interest, reflecting only how closely a subject followed the stylized reference F0 curve and do not take into account distortions in timing (e.g. if a subject moved too slowly when tracing one section of the gesture).

As neither correlation nor RMSE follow a Gaussian distribution, the Fisher Z and log transforms are used respectively for correlation and RMSE scores to obtain Gaussian distributions for statistical analysis.

3.3. Statistical Modeling of Comparison Scores

Multilevel models were fitted for further analysis, focusing on two types of comparisons. A first set of models were fitted for vocal imitation and guided gestural imitation scores without timing. Only guided scores were used because they represent the "best attempts" of gestural imitation. A second set of models were fitted for guided and non-guided gestural imitation scores with subjects' timing included, to assess the effect of the visual guide. Difference in intercepts for guided imitation scores between the no-timing and with-timing models represents the effect of timing control on the gestural imitation task. Figure 2 shows aggregate plots of scores used in the models.

All models were given random intercepts for phrases and random slopes for subjects based on condition. Frequentist multilevel models failed to converge when random slopes were included [23]. To prevent type 1 errors from removing the random slopes [24], we turned to Bayesian multilevel models using BRMS and Stan [25, 26, 27]. Weakly informative, "regularizing," priors based on [28, 29, 30] were used (intercepts (α) & slopes (β)- Normal(0, 10); σ_e & σ_{group}- HalfCauchy(0,10), Correlation Parameter - LKJ(2)) All models used two chains with 10000 iterations (2000 warm-up).

For each data subset and score type, four models were made with different fixed effects: on condition only, on subject nativeness only, on both without correlation, and on both with correlation. Condition and nativeness were contrast coded in each model (0.5 and -0.5). Models with the same dataset and score type were compared with LOO (Leave One Out).

3.4. Analysis of Results

All LOO comparisons yielded standard error (SE) differences of less than 3, indicating that no best model stands out. Table 2 shows the results of the full model, given by the following for-

Table 2: *Posterior estimates for Fisher Z transformed correlation* (z(r)) *and log transformed RMSE* (log(r)) *values. The mean, lower 95%Credible Interval and upper 95%CrI estimates are shown for four Bayesian multi-level models. The left two models were fitted on vocal and guided imitation scores with neutralized timing. The right two models were fitted on guided and non-guided imitation scores with subjects' timing input. Rhat was 1.00 for all parameters. Parameters with entirely positive or entirely negative 95% credible intervals are in bold. The bottom row gives intercept values reverse transformed (f^{-1}) into correlation and RMSE scores.*

	Condition: 0.5 vocal, -0.5 gestural						Condition: 0.5 guided, -0.5 non-guided					
	$f = z(correlation))$			$f = log(RMSE)$			$f = z(correlation)$			$f = log(RMSE)$		
	Mean	Lower	Upper	Mean	Lower	Upper	Mean	Lower	Upper	Mean	Lower	Upper
Condition, β_1	0.18	-0.16	0.52	**-0.29**	**-0.57**	**-0.02**	**0.45**	**0.24**	**0.65**	**-0.28**	**-0.39**	**-0.17**
Native, β_2	-0.09	-0.44	0.25	0.06	-0.17	0.29	-0.03	-0.38	0.32	-0.03	-0.24	0.19
Cond.:Native, β_3	0.33	-0.35	1.01	-0.21	-0.78	0.36	0.10	-0.30	0.50	0.06	-0.16	0.28
$f^{-1}(\alpha)$	0.95	0.92	0.97	1.35	1.09	1.68	0.73	0.60	0.83	2.43	2.07	2.85

mula in lme4 notation [23]: $Score \sim Condition * Native + (1 + Condition|Subject) + (1|Phrase)$, applied to each data subset and score type. This formula models score as a function of condition, nativeness, and their interaction. Random intercepts account for different baseline results across phrases and subjects. Random slopes were modeled for subjects based on condition.

For the models using scores without timing, the posterior mean correlation of vocal and guided gestural imitation is 0.95 (95% CrI=[0.92, 0.97]), and the mean RMSE score is 1.35 ST (95% CrI=[1.09, 1.68]). For RMSE, the slope for condition is negative for the entire credible interval, indicating slightly better imitation results for the vocal modality. In other words, when timing is not taken into account, vocal and gestural imitation perform similarly, with vocal imitation slightly better. For correlation, the effect of condition is uncertain because its credible interval spans both positive and negative values.

When timing was taken into consideration, gestures had lower correlation and higher RMSE, with a reverse transformed posterior means of 0.73 (95%CrI=[0.60,0.83]) for correlation and 2.43 ST (95%CrI=[2.07, 2.85]) for RMSE. The slope for condition is positive for correlation and negative for RMSE across the entire credible interval, indicating the effect of the guide in improving gestural imitations. Nativeness and the interaction parameter had credible intervals that span both positive and negative values in all four models, so their effects are uncertain.

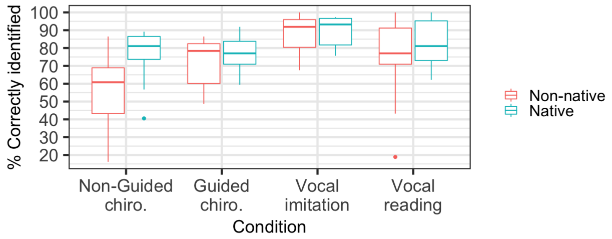

Figure 3: *Preliminary results from an online perceptual pilot using stimuli from one pair of phrases (10 and 10bis, 80 total stimuli). Gestural data with subjects' timing was resynthesized using WORLD [9]. 37 francophone natives listened to stimuli in random order and selected between the two meanings in a forced choice. Aggregate scores of "correctness" are presented by condition and nativeness. Horizontal line indicates median, hinges indicate 25th and 75th percentiles.*

4. Discussion and Conclusions

This paper introduced an interface that enables the real-time gestural control of intonation in phrases produced by a vocal synthesizer. A study using an imitation paradigm for prosodic disambiguation has been conducted as a means to explore chironomic stylisation of intonation for native and non-native speakers. The 4 questions raised in the introduction are revisited here.

For quantitative measures, vocal and guided gestural imitation performed comparably when timing was not taken into account, a finding which corroborates prior results. Guided chironomy performed significantly better than non-guided chironomy in our results. Another finding is the additional difficulty from the added dimension of timing control, confirmed by lower comparison scores when timing was taken into account. Surprisingly, no statistically significant difference in quantitatives scores was found between native and non-native subjects.

However, results of a preliminary perceptual test suggests further examination of this question (see Figure 3). While vocal imitation performed the best overall, guided chironomy was still able to produce intonation that is correctly identifiable by native listeners. Using Fisher's exact test ($p<0.05$) to compare subject-phrase pairings across conditions, guided chironomy performed significantly better than non-guided for 4 subject-phrases, where 3 were from non-native subjects. It also performed significantly better than reading for the 2 subject-phrase combinations with the lowest reading scores (0.19 and 0.43). The non-native subjects available for this study are already fairly advanced, and have already come across the stimuli in their studies. Errors in reading only occurred for a small percentage of stimuli. Nevertheless, the fact that subject-phrase combinations with mistakes in reading had high correctness scores in guided gestural imitation suggests an opportunity for chironomic practice.

In conclusion native and non-native speakers succeeded in prosodic disambiguation using chironomic stylization of intonation. This allows us to envision this interface as a means for non-native speakers to practice the intonation of a foreign language. Future work should explore easier ways to control timing, which should be tested with non-native subjects with more room for improvement in their intonation. Further perceptive testing should also be conducted.

5. Acknowledgements

This work has been supported by ANR GEPETO (ANR-19-CE28-0018) and ANR-10-LABX-0083. It contributes to the IdEx Université de Paris (ANR-18-IDEX-0001).

6. References

[1] C. d'Alessandro, A. Rillard, and S. LeBeux, "Chironomic stylization of intonation," *Journal of the Acoustical Society of America*, vol. 3, no. 129, pp. 1594–1604, Mar. 2011.

[2] M. Taniguchi and E. Abberton, "Effect of interactive visual feedback on the improvement of english intonation of japanese efl learners," *Speach, Hearing, and Language: work in progress*, vol. 11, pp. 77–89, 01 1999.

[3] C. Yuan, S. González-Fuente, F. Baills, and P. Prieto, "Observing pitch gestures favors the learning of spanish intonation by mandarin speakers," *Studies in Second Language Acquisition*, vol. 41, pp. 5–32, 03 2019.

[4] P. J. Price, M. Ostendorf, and M. Park, "The use of prosody in syntactic disambiguation," *Journal of the Acoustical Society of America*, no. 90, pp. 2956–2970, Dec. 1991.

[5] Y. Zhang, H. Ding, P. Zelchenko, X. Cui, Y. Lin, Y. Zhan, and H. Zhang, "Prosodic disambiguation by chinese efl learners in a cooperative game task," *Proceedings of Speech Prosody 2018*, pp. 979–983, 06 2018.

[6] A. Fultz, "The use of prosody for disambiguation in english-french interlanguage," *Proc. 9th Gener. Approaches to Second Lang. Acquis. Conf.*, pp. 130–139, 2007.

[7] G. Locqueville, C. d'Alessandro, S. Delalez, B. Doval, and X. Xiao, "Voks: Digital instruments for chironomic control of voice samples," *Speech Communication*, vol. 125, pp. 97–113, 12 2020.

[8] Cycling74, "Max 6," https://cycling74.com/, 2011, accessed: 2021-03-21.

[9] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, pp. 1877–1884, 07 2016.

[10] X. Xiao, G. Locqueville, C. D'Alessandro, and B. Doval, "T-Voks: the singing and speaking theremin," in *NIME 2019 International Conference on New Interfaces for Musical Expression*, Porto Alegre, Brazil, Jun. 2019, pp. 110–115.

[11] C. D'Alessandro, X. Xiao, G. Locqueville, and B. Doval, "Borrowed voices," in *International Conference on New Interfaces for Musical Expression NIME'19*, Porto Alegre, Brazil, Jun. 2019, pp. 2.2–2.4.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," http://www.praat.org/ Version 6.1.08, 2019, accessed: 2021-03-21.

[13] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," *Proceedings of Speech Prosody 2002*, 04 2002.

[14] P. Mertens, "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Proceedings of Speech Prosody 2004*, Nara, Japan, 2004.

[15] C. d'Alessandro and P. Mertens, "Automatic pitch contour stylization using a model of tonal perception," *Computer Speech and Language*, vol. 9, no. 3, pp. 257–288, 1995.

[16] "DELF-DALF," http://www.delfdalf.fr/, accessed:2021-06-13.

[17] J. Vaissière, *La Phonétique*. Que sais-je, 2020.

[18] L. Charliac and A. C. Motron, *Phonétique progressive du français avec 400 exercices: niveau avancé*. Clé international, 2006.

[19] D. Hermes, "Measuring the perceptual similarity of pitch contours," *Journal of speech, language, and hearing research : JSLHR*, vol. 41, pp. 73–82, 03 1998.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[21] A. E. . P. Bailey, *wCorr: Weighted Correlations*, 2017, r package version 1.9.1. [Online]. Available: https://CRAN.R-project.org/package=wCorr

[22] B. Gorman, *mltools: Machine Learning Tools*, 2018, r package version 0.3.5. [Online]. Available: https://CRAN.R-project.org/package=mltools

[23] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[24] D. Barr, R. Levy, C. Scheepers, and H. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of memory and language*, vol. 68, pp. 255–278, 01 2013.

[25] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[26] ——, "Advanced Bayesian multilevel modeling with the R package brms," *The R Journal*, vol. 10, no. 1, pp. 395–411, 2018.

[27] Stan Development Team, "Stan modeling language users guide and reference manual," https://mc-stan.org, 2021, accessed: 2021-03-21.

[28] R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd Edition*, 2nd ed. CRC Press, 2020. [Online]. Available: http://xcelab.net/rm/statistical-rethinking/

[29] S. Vasishth, B. Nicenboim, M. Beckman, F. Li, and E. J. Kong, "Bayesian data analysis in the phonetic sciences: A tutorial introduction," *Journal of Phonetics*, vol. 71, pp. 147–161, 08 2018.

[30] L. Nalborczyk, C. Batailler, H. Loevenbruck, A. Vilain, and P.-C. Bürkner, "An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard indonesian," *Journal of Speech, Language, and Hearing Research*, vol. 62, 05 2019.