



Language Proficiency and F0 Entrainment: A Study of L2 English Imitation in Italian, French, and Slovak Speakers

Zheng Yuan^{1,2}, Štefan Beňuš^{3,4}, Alessandro D'Ausilio^{1,2}

¹Italian Institute of Technology, Italy

²University of Ferrara, Italy

³Constantine the Philosopher University in Nitra, Slovakia

⁴Institute of Informatics, SAS, Slovakia

{zheng.yuan, alessandro.dausilio}@iit.it, sbenus@ukf.sk

Abstract

This study explores F0 entrainment in second language (L2) English speech imitation during an Alternating Reading Task (ART). Participants with Italian, French, and Slovak native languages imitated English utterances, and their F0 entrainment was quantified using the Dynamic Time Warping (DTW) distance between the parameterized F0 contours of the imitated utterances and those of the model utterances. Results indicate a nuanced relationship between L2 English proficiency and entrainment: speakers with higher proficiency generally exhibit less entrainment in pitch variation and declination. However, within dyads, the more proficient speakers demonstrate a greater ability to mimic pitch range, leading to increased entrainment. This suggests that proficiency influences entrainment differently at individual and dyadic levels, highlighting the complex interplay between language skill and prosodic adaptation.

Index Terms: speech entrainment, phonetic convergence, second language proficiency, speech imitation, pitch contour analysis

1. Introduction

Speech entrainment [1], often termed alignment [2], convergence [3], or accommodation [4], is observed as individuals unconsciously or deliberately modify their speaking style during interactions, leading to heightened similarity in acoustic-prosodic features.

Entrainment, functioning as implicit imitation, is theorised to share neural-cognitive mechanisms with speech imitation [5], involving automatic priming [2], motor control [6] and the perception-production loop [7]. Dialect formation and language change stem from multiple imitative speech interactions [8].

Studies have shown that entrainment plays a positive role in second language (L2) acquisition. Learners can effectively adapt to the target language by aligning their pronunciation with either native (L1) speakers [9] or proficient L2 speakers [10]. Gnevshva et al. [11] suggest second language learners exhibit greater flexibility in adapting their pronunciation compared to native speakers. Jiang et al. [9] found that L2 learners' belief in an interlocutor's proficiency can influence L2 phonetic entrainment, ultimately bringing them closer to native speakers with an improved vowel pronunciation. Additionally, the entrainment effect varies based on individual differences, such as language talent [12] and context (e.g., task type). Imitative tasks exhibit a higher degree of entrainment than interactive tasks [13, 14], particularly for non-native (L2) sounds [15].

Despite extensive research focusing on phonetic entrainment in L1-L2 interactions, the prosodic entrainment in L2-L2 scenarios remains largely unexplored, particularly considering the potential influence of L2 proficiency. Utilising the Alternating Reading Task (ART) corpus [16], an L2 English speech dataset featuring recordings from Italian, French, and Slovak speakers, we conducted subjective evaluations to assess spoken English proficiency and subsequently delve into its connection with the fundamental frequency (F0) entrainment in L2-L2 speech imitation. Our investigation focuses on global static proximity and synchrony in entrainment as framed by [1, 17].

Past research on global static entrainment predominantly involves statistical analyses, including paired t-tests applied to "partner distance" and "other distance" [1], as well as techniques like Time-aligned Moving Average (TAMA) [18], Cross-Recurrence Quantification Analysis [19], and Windowed Lagged Cross-Correlation [20]. While effective, these methods have limitations in controlling the unit of analysis, such as phonemes or words.

We employed a novel hybrid method that involves aligning timestamps at the word level using the WhisperX [21] ASR tool, adopting F0 contour parameterization [22, 23], and computing F0 entrainment as the distance between two parameterized F0 contours using Dynamic Time Warping (DTW). Three experiments were performed to validate the robustness of the F0 entrainment measurement algorithm and examine the correlation between F0 entrainment and L2 proficiency at both the individual speaker and dyadic levels.

2. Data

2.1. The ART corpus

The Alternating Reading Task (ART) corpus [16] is a collection of recordings from a collaborative L2 English speech production experiment designed to investigate entrainment and imitation. The corpus encompasses data gathered from 58 subjects, organised into same-sex dyads, with 18 native Italian speakers (6 males), 20 native French speakers (all female), and 20 native Slovak speakers (10 males). The experiment includes three distinct conditions: solo, interactive, and imitative utterance reading. Participants engaged in these tasks by reading utterances individually, taking turns in interactive sessions, and mimicking their dyadic partner's delivery of the target utterance during the imitative condition.

The textual material utilised in the experiment comprises a simplified adaptation of a Wikipedia article, totalling 801 words and segmented into 80 speaking turns. These speaking turns range from 6 to 13 words, and turn boundaries were strategi-

cally positioned within sentences to enhance prosodic continuity. During the experiment, participants were seated side by side, facing two screens, and separated by a curtain to mitigate the potential influence of mutual visual contact on speech entrainment. Notably, the focus of our analysis primarily revolves around the recordings from the imitation condition.

2.2. L2 proficiency evaluation

To discern the potential influence of interlocutors' L2 speaking proficiency on speech entrainment, six language experts were enlisted to evaluate the spoken English skills of subjects in the ART corpus. The assessment focused on the initial 10 utterances from the solo recordings for each speaker, based on four key criteria: pronunciation, intonation, fluency, and overall impression. For each criterion, evaluators assigned scores on a scale ranging from 1 to 5, and a final score was derived as the average across the four criteria.

Table 1: Intraclass Correlation Coefficients for Spoken English Proficiency Assessments

Indicator	ICC	p-value	CI95%
pronunciation	0.828	< 0.001	[0.71, 0.90]
intonation	0.767	< 0.001	[0.65, 0.85]
fluency	0.796	< 0.001	[0.68, 0.87]
overall	0.800	< 0.001	[0.67, 0.88]
final	0.840	< 0.001	[0.73, 0.91]

Table 1 delineates the degree of agreement among experts for each criterion, quantified through Intraclass Correlation Coefficients (ICC) values and their corresponding 95% confidence intervals. The ICC values were computed using a two-way mixed-effects model with the mean of raters and 57 degrees of freedom. All ICC values indicate a level of "good reliability" (ICC between 0.75 and 0.9) with statistical significance (p-value < 0.001), aligning with the criteria stipulated by [24].

3. Method

3.1. Word segmentation and alignment

Our approach to speech imitation analysis relied on utterance-level F0 contour comparison, comprising 4,640 (58 × 80) audio segments of speech imitation data. All instances of spoken words, inclusive of stutters, repetitions, and self-corrections, were retained.

To enhance the precision of time-series comparison, we executed word segmentation for each utterance. Audio files underwent transcription and force-alignment using the WhisperX ASR tool [21], providing precise starting and ending timestamps for each word. The models employed for this task were the `base.en` Whisper model and the `Base_960h` phoneme model.

The selected models exhibited a commendable precision score of 93.1 [21] on word segmentation for the Switchboard-1 Telephone Corpus (SWB), instilling confidence in their suitability for our analysis on the ART corpus. Through visual inspection by the authors, a comparable precision rate was observed for the ART corpus, with minor inaccuracies primarily associated with the cutting of the initial vowel in certain words. It's worth mentioning that manual calibration of the alignment was not conducted.

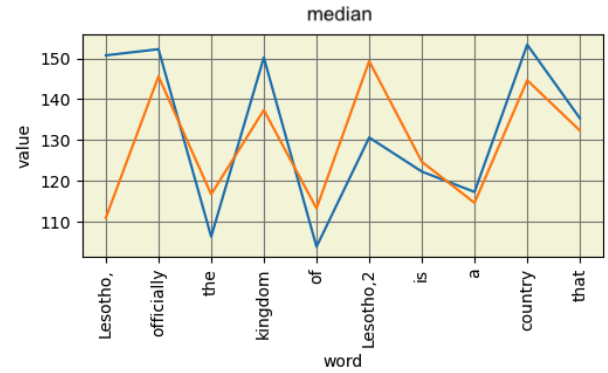


Figure 1: F0 contour parameterization (median). On the x-axis is the utterance transcription separated into words and y-axis represents the parameter value. Yellow and blue lines are the plots of F0 parameters of the imitator and model utterance.

3.2. F0 Preprocessing

The extraction of F0 was performed utilising autocorrelation in PRAAT software through its Python interface Parselmouth (version 0.4.3) [25], with all parameters set to their default values. For precise and smooth F0 contours, voiceless utterance segments underwent linear interpolation, and a two-pass method [26] was implemented to address F0 outliers. Subsequently, a Savitzky-Golay filter was employed to enhance the smoothness of the F0 contour [22, 23], adopting third-order polynomials in 7-sample windows.

3.3. Parameterization

To explore the static global entrainment, specifically proximity and synchrony, as defined by [1, 17], we adopted a simplified CoPaSul [22, 23] style F0 parameterization. Five widely recognised and easily interpretable F0 features were selected for this purpose.

- **Mean:** the average of pitch values y .
- **Median:** the median of pitch values y .
- **Slope:** the slope of the first-order linear fit of pitch values y against normalised time t_n .
- **Range:** the difference between the 95th and 5th percentiles of the fitted values y_n .
- **Drop:** the difference between the last and first fitted values y_n , normalised by time t .

These selected features serve distinct roles in the analysis: mean and median pertain to a more stable indicator of the physiological quality of speakers' articulators, while slope, range, and drop contribute to the dynamics of prosody. This parameterization framework provides a nuanced perspective on the entrainment dynamics captured in the F0 data.

3.4. F0 Entrainment

The quantification of F0 entrainment for each speaker involved assessing the average distance across 40 utterances where the speaker (imitator) mimicked their partner (model speaker). The degree of entrainment, denoted by E , is represented by the distance between the parameterized F0 contours of the model and imitated utterances utilising the Dynamic Time Warping (DTW)

algorithm with the Euclidean distance metric. A larger E value indicates less entrainment.

The raw value of F0 entrainment E_{raw} is expressed by the distance equation:

$$E_{raw} = \frac{1}{N} \sum_{i=1}^N DTW(\mathbf{s}_i^{imit}, \mathbf{s}_i^{model}) \quad (1)$$

Here, N denotes the number of utterances imitated by the speaker, \mathbf{s}_i^{imit} and \mathbf{s}_i^{model} represent the F0 parameter vectors for the utterances uttered by the imitator and the model speaker, respectively. The DTW function gauges the degree of entrainment between these parameterized F0 contours.

In the experiments detailed in Sections 4.2 and 4.3, we employed an optimised entrainment function denoted as:

$$E_{opt} = \frac{1}{N} \sum_{i=1}^N \text{Norm} \left(DTW(\mathbf{s}_i^{imit}, \mathbf{s}_i^{model}) \right) \quad (2)$$

Here, the DTW distance is normalised by subtracting the mean of all utterance-level DTW samples and dividing by the sample standard error, as defined by the Norm function. The sample statistics and N , representing the number of utterances per speaker, are computed with outliers removed using the quantile method.

To assess the robustness of the distance algorithms, two additional measures were considered: partner distance and other distance, i.e., the distances of real dyads and surrogate dyads. The partner distance is defined as:

$$\text{partner distance} = E_{raw}^{(t,p)} \quad (3)$$

where $E_{raw}^{(t,p)}$ signifies the average raw distance between the target speaker and their partner speaker.

Similarly, the other distance is expressed as:

$$\text{other distance} = \frac{1}{C} \sum_i^C E_{raw}^{(t,i)} \quad (4)$$

Here, $E_{raw}^{(t,i)}$ represents the average distance between the target speaker and non-partner speakers, and C denotes the number of combinations of the target speaker with other speakers (i.e., non-partners). Our hypothesis posits that the partner distance will be smaller than the other distance, indicative of the entrainment effect in interactions with the partner.

In Section 4.3, we investigated the correlation between the inner-dyad difference in English proficiency and the disparity in the degree of entrainment, as represented by Equation 5:

$$\text{inner-dyad distance} = \text{Norm}(E_{raw}^A - E_{raw}^B) \quad (5)$$

Here, A and B denote the speaking partners. The Norm function transforms the raw distance by dividing it by the mean of E_{raw}^A and E_{raw}^B . A positive inner-dyad distance indicates that speaker A is less entrained to speaker B in the imitation task, and vice versa.

In the following experiments, we adopt a significance level of $p < 0.05$. Following [1, 27] results with $p < 0.1$ are considered to trend towards significance.

4. Results

4.1. Partner vs non-partner experiment

The comparison between partner and non-partner distances, assessed through a paired t-test across multiple F0 features, reveals statistically significant differences. The negative t-statistics in all cases (See Table 2) indicate that the mean differences for partner pairs are consistently smaller than those for non-partner pairs, suggesting that the chosen F0 features and the distance algorithms' capability to capture the intricate entrainment effect during speech imitation.

Table 2: *T-Test Results Comparing Partner Differences to Other Differences*

Feature	t	df	p-value	Sig.
Median	-4.44	57	2.1e-5	*
Mean	-4.43	57	2.2e-5	*
Range	-3.67	57	0.0003	*
Slope	-4.53	57	1.5e-5	*
Drop	-3.42	57	0.0006	*

4.2. Individual speaker experiment

The correlation analysis between individual speakers' degree of F0 entrainment and their spoken English proficiency reveals a nuanced relationship between how closely speakers mimic their partners' speech patterns and their L2 English competence. The F0 entrainment here is quantified as a distance, with higher values indicating less entrainment or mimicry of speech patterns.

As depicted in Fig. 2, "range" and "drop" parameters exhibit the most notable correlations with all language proficiency scores except "pronunciation". "Range" displays moderate associations ($0.26 < r < 0.31$), while "drop" demonstrates stronger correlations ($0.3 < r < 0.34$). "Intonation", among the score parameters, shows statistically significant correlations with "slope" ($r = 0.26$), "range" ($r = 0.31$), and "drop" ($r = 0.34$). Conversely, "median" and "mean" F0 parameters exhibit weaker, less consistent associations with language proficiency, lacking statistical significance, except for a minor connection with "pronunciation".

4.3. Inner-dyad experiment

Instead of focusing on absolute English scores and F0 distance values, this experiment explores the intricate relationship between inner-dyad differences in L2 English proficiency scores and corresponding variations in parameterized F0 measures to ascertain whether a more proficient L2 English speaker within a dyad exhibits more or less entrainment with their partner. The findings, presented in Fig.3, reveal a complex interplay of the two aspects, particularly when considering the collaborative effects within a dyad.

For the "range" parameter and "fluency" score, a significant negative correlation was found (-0.424 , $p = 0.0219$), suggesting that within a dyad, the speaker with higher English fluency tends to be more entrained in terms of pitch variation with their partner. While many other F0 parameters and English score aspects display negative correlations, they do not reach statistical significance, with "final" and "intonation" showing close-to-significant results.

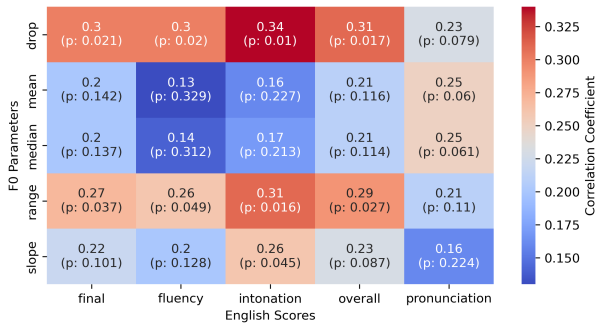


Figure 2: Heatmap of Pearson correlation coefficients between F0 entrainment measures and English proficiency scores. Cells indicate correlations of F0 distance (median, mean, range, slope, drop) with language scores (final, fluency, intonation, overall, pronunciation). Colour intensity reflects correlation strength (red: positive, blue: negative), with values in parentheses denoting p-values.

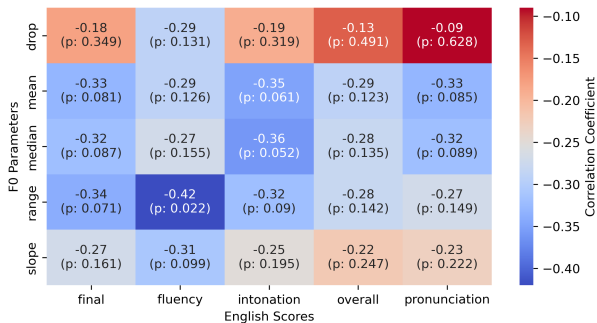


Figure 3: Heatmap of Pearson correlation coefficients between inner-dyad differences in English proficiency scores and F0 entrainment measures. Negative correlations are illustrated, with darker shades indicating stronger relationships. The correlation coefficients are accompanied by p-values in parentheses.

5. Discussion

The observation that higher proficiency speakers tend to exhibit less entrainment in terms of pitch variation and declination is a notable finding. It suggests that these individuals have developed a more stable and independent prosodic system in their L2 speech, allowing them to maintain their prosodic identity rather than conforming to the model’s patterns. Less proficient speakers have demonstrated relative entrainment flexibility which aligns with [11]’s comparison between L2 language speakers and native speakers.

Contrasting with the individual entrainment trends, within dyads, proficient speakers show a tendency for stronger mimicry of pitch range, indicating enhanced entrainment. This could be interpreted as proficient speakers’ adaptive linguistic behaviour, showcasing their capability to align more closely with their partner when necessary, perhaps as evidence of stronger sensorimotor adaptation [5] competence trained through L2 learning.

Our finding also shows that in imitation tasks speakers prioritise dynamic pitch features like range and slope. During imitation, speakers might pay more attention to these prominent and salient prosodic cues as they hold significant emotional and

communicative weight, making them more amenable to conscious replication. In contrast, mean and median F0 values, reflecting baseline pitch influenced by physiological characteristics, voice quality, and less linguistic context, might be less subject to conscious manipulation.

The use of DTW distance and F0 parameterization as a measure of entrainment offers a methodological contribution, demonstrating its effectiveness in capturing subtle prosodic alignment phenomena. This quantitative approach could be complemented by qualitative analyses to further understand the subjective aspects of prosodic adaptation including task involvement, partner attractiveness, and motivation [12].

The interpretation of the study’s findings is subject to certain limitations. Firstly, the relatively small sample size of speakers and language evaluators might exert a disproportionate influence on the results. Moreover, the accuracy of the DTW distance may be compromised by errors in force-alignment. Furthermore, the prosodic parameterization employed was relatively simplistic, relying on a first-order linear fitting that captures only rising or falling intonation patterns, thus overlooking the more complex nuances of prosody.

Future work could explore a wider range of F0 features and apply a Linear Mixed Effects model incorporating variables like sentence length and interaction time. It should also include subjective evaluations of the perceived similarity of the imitated utterances, more evaluators for proficiency ratings, and focus on local entrainment within intra-pausal units to deepen our understanding of L2 speech dynamics.

6. Conclusions

The study reveals that higher L2 English proficiency correlates with less F0 entrainment on an individual level but greater mimicry of pitch range within dyadic interaction. This suggests that advanced speakers maintain distinct prosodic patterns while also displaying adaptability in interactive settings. This plasticity sheds light on prosodic dynamics and language acquisition mechanism in L2-L2 interactions.

7. Acknowledgements

This work was supported by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 859588. We are grateful to Uwe D. Reichel for generously sharing the pitch parameterization code that served as the foundation for our analysis. We also extend our sincere thanks to Štefan Beňuš, Jana Beňušová, Lucia Mareková, Changyong Min, Qiuwen Zhang, and Fang Liu for their meticulous evaluation of the English language data.

8. References

- [1] R. Levitan and J. Hirschberg, “Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions,” in *Proc. Interspeech 2011*, 2011, pp. 3081–3084.
- [2] M. J. Pickering and S. Garrod, “Toward a mechanistic psychology of dialogue,” *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.
- [3] J. S. Pardo, “On phonetic convergence during conversational interaction,” *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.
- [4] H. Giles, N. Coupland, and J. Coupland, “Accommodation theory: Communication, context, and consequence,” in *Contexts of Accommodation*. Cambridge University Press, Sep. 1991, pp. 1–68.

- [5] C. Gambi and M. J. Pickering, "Prediction and imitation in speech," *Frontiers in psychology*, vol. 4, p. 340, 2013.
- [6] M. Sato, K. Grabski, M. Garnier, L. Granjon, J.-L. Schwartz, and N. Nguyen, "Converging toward a common speech code: imitative and perceptuo-motor recalibration processes in speech production," *Frontiers in psychology*, vol. 4, p. 422, 2013.
- [7] S. D. Goldinger, "Echoes of echoes? an episodic theory of lexical access." *Psychological review*, vol. 105, no. 2, p. 251, 1998.
- [8] V. Delvaux and A. Soquet, "The influence of ambient speech on adult speech productions through unintentional imitation," *Phonetica*, vol. 64, no. 2-3, pp. 145–173, 2007.
- [9] F. Jiang and S. Kennison, "The impact of L2 english learners' belief about an interlocutor's english proficiency on L2 phonetic accommodation," *Journal of Psycholinguistic Research*, vol. 51, no. 1, pp. 217–234, 2022.
- [10] P. Trofimovich and S. Kennedy, "Interactive alignment between bilingual interlocutors: Evidence from two information-exchange tasks," *Bilingualism: Language and Cognition*, vol. 17, no. 4, pp. 822–836, 2014.
- [11] K. Gnevsheva, A. Szakay, and S. Jansen, "Phonetic convergence across dialect boundaries in first and second language speakers," *Journal of Phonetics*, vol. 89, p. 101110, 2021.
- [12] N. Lewandowski and M. Jilka, "Phonetic convergence, language talent, personality and attention," *Frontiers in Communication*, vol. 4, p. 18, 2019.
- [13] D. de Jong, A. Pastore, N. Nguyen, and A. D'Ausilio, "Speech imitation skills predict automatic phonetic convergence: a GMM-UBM study on L2," in *Proc. Interspeech 2022*, 2022, pp. 769–773.
- [14] Z. Yuan, A. Pastore, D. de Jong, H. Xu, L. Fadiga, and A. D'Ausilio, "The ART of Conversation: Measuring Phonetic Convergence and Deliberate Imitation in L2-Speech with a Siamese RNN," in *Proc. Interspeech 2023*, 2023, pp. 132–136.
- [15] H. Wilt, Y. Wu, A. Trotter, and P. Adank, "Automatic imitation of human and computer-generated vocal stimuli," *Psychonomic Bulletin & Review*, vol. 30, no. 3, pp. 1093–1102, 2023.
- [16] Z. Yuan, D. de Jong, Š. Beňuš, N. Nguyen, R. Feng, R. Sabo, L. Fadiga, and A. D'Ausilio, "Art: The alternating reading task corpus for speech entrainment and imitation," *arXiv preprint arXiv:2404.02710*, 2024.
- [17] C. J. Wynn and S. A. Borrie, "Classifying conversational entrainment of speech behavior: An expanded framework and review," *Journal of Phonetics*, vol. 94, p. 101173, 2022.
- [18] S. Kousidis, D. Dorran, Y. Wang, B. Vaughan, C. Cullen, D. Campbell, C. McDonnell, and E. Coyle, "Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues," in *Proc. Interspeech 2008*, 2008, pp. 1692–1695.
- [19] R. Fusaroli, I. Konvalinka, and S. Wallot, "Analyzing social interactions: the promises and challenges of using cross recurrence quantification analysis," in *Translational recurrences: From mathematical theory to real-world applications*. Springer, 2014, pp. 137–155.
- [20] S. M. Boker, J. L. Rotondo, M. Xu, and K. King, "Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series," *Psychological methods*, vol. 7, no. 3, p. 338, 2002.
- [21] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," in *Proc. Interspeech 2023*, 2023, pp. 4489–4493.
- [22] U. D. Reichel, Š. Beňuš, and K. Mády, "Entrainment profiles: Comparison by gender, role, and feature set," *Speech Communication*, vol. 100, pp. 46–57, 2018.
- [23] U. D. Reichel, "Copasul manual – contour-based parametric and superpositional intonation stylization," *ArXiv*, 2023.
- [24] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.
- [25] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.
- [26] C. D. Looze and S. Rauzy, "Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration," in *Proc. Interspeech 2009*, 2009, pp. 2919–2922.
- [27] A. Weise, S. I. Levitan, J. Hirschberg, and R. Levitan, "Individual differences in acoustic-prosodic entrainment in spoken dialogue," *Speech Communication*, vol. 115, pp. 78–87, 2019.