Exploring Impact of Pausing and Lexical Stress Patterns on L2 English Comprehensibility in Real Time

Sylvain Coulange[1,2,3], *Tsuneo Kato*[3], *Solange Rossato*[2], *Monica Masperi*[1]

[1] Laboratory of Linguistics and Didactics of Foreign and Mother Tongues (LIDILEM), University Grenoble Alpes, France
[2] CNRS, Institute of Engineering, Grenoble Computer Science Laboratory (LIG), University Grenoble Alpes, France
[3] Spoken Language Processing Laboratory (SLPL), Doshisha University, Japan

{sylvain.coulange, solange.rossato, monica.masperi}@univ-grenoble-alpes.fr,
tsukato@mail.doshisha.ac.jp

Abstract

A significant obstacle to effective L2 English speech lies in the inappropriate use of pauses and lexical stress.

We observed the impact of pauses within phrases (WP) and incorrectly stressed words on real-time perceived comprehensibility. Sixty native English listeners were asked, while listening to short recordings of L2 speakers, to click on a button whenever they were struggling to understand the speaker. Analysis showed that click frequency tends to increase after WP pauses and incorrectly stressed words, especially 2-3 s after onset, while it remains under the average click frequency after pauses between clauses and correctly stressed words. These results demonstrate that WP pauses and incorrect stress patterns directly impact the listener's perception, and that this impact can measured with a dynamic rating protocol. Moreover, such a protocol appears to be a promising tool to expand our knowledge about real-time comprehension of L2 speech.

Index Terms: L2 comprehensibility, pause positions, lexical stress, dynamic rating, prosody, fluency

1. Introduction

Pronunciation is key to success in spoken communication. The speakers most likely to be successful, however, are not necessarily those who have a native-like pronunciation, but rather those who adequately use certain linguistic features that help listeners easily understand and follow the speech [1]. In foreign-language acquisition, this concept of being easily understood is generally referred to as *speaker comprehensibility* [2]. Assessing speaker comprehensibility is a formidable challenge since it is a listener-based construct, influenced by various listener-side variables such as the listener's first language(s) (L1) [3], level of engagement in the speaking task [4], familiarity with the speaker, or the subject matter of the conversation [5].

There are, however, linguistic features that may help improve comprehensibility whatever the situation. Pauses are often correlated with L2 proficiency since lower level speakers tend to pause more and longer [6]; but their position is key to segment and structure speech so that it is more easily processed by the listener. It is therefore crucial to consider their location in speech. More specifically, pauses occurring at grammatical boundaries, i.e. between clauses (BC) and to a lesser extent between phrases (BP), tend to help the listener [7], while those occurring within phrases (WP) are often categorized as *hesitant* or *non-structuring* pauses [8, 9] and appear to be strongly correlated with poorer comprehensibility, fluency, and proficiency judgments [10, 11, 12].

Lexical stress, especially in English, also contributes to speech segmentation at a finer level and shapes discourse's rhythm [13, 14]. Stressed syllables in English generally present a higher fundamental frequency (F0), stronger intensity, and longer duration compared with neighboring syllables, and surrounding vowels tend to shrink to reduced forms [13]. Lexical stress is known to be correlated with comprehensibility judgments at all degrees of proficiency [15, 16] and often cited as a determinant feature when it comes to assessing speaking proficiency [17, 18]. Lexical stress is a particularly impactful factor with learners whose L1 do not have lexical stress such as French [15, 19].

Assessing comprehensibility is commonly done by asking listeners to holistically rate speakers' performance on a Likert scale, but some studies used more dynamic approaches. Previous studies [20, 21, 22] objectively predicted comprehensibility perceived by listeners by making them shadow learners' speech and measure the degree of smoothness of their shadowing. Another method involves asking listeners to rate comprehensibility while they are listening to the speaker. Nagle et al. [23] asked 24 native speakers of Spanish to dynamically rate 3 English-L1 speakers of Spanish using Idiodynamic Software [24] and stimulated recall interviews. The main causes reported by the participants for down-grading comprehensibility were lexical and grammar misuses, while up-grading was associated with fluency and discourse structure. However, participants reacted in various ways to the experiment, leading to difficulties when analysing click patterns.

We propose to adapt Nagle et al.'s experiment to observe the real-time impact of pause positions and lexical stress patterns on perceived comprehensibility. We asked 60 English-L1 naive crowd-sourced participants to rate the comprehensibility of 15 French-L1 speakers of English. While listening, raters had to click on a button whenever they were struggling to understand the speaker. Recordings where automatically annotated in pause positions and lexical stress quality with a tool developed in a previous study, and click frequency following each category of pauses and stressed word were analyzed.

Our research questions are as follows: Q1) Do listeners exhibit consistent behavior in dynamically rating comprehensibility of L2 speakers, despite inter-rater variations? Q2) Do we observe a decrease in comprehensibility after WP pauses and incorrectly stressed words?

More specifically, we are expecting to observe:

- H1: a higher click frequency after WP pauses compared with pauses between clauses or phrases,
- H2: a higher click frequency after words with inappropriate lexical stress compared with those with correct stress,
- H3: that speech segments with more WP pauses and inappropriate stress should receive more clicks from listeners and lower global ratings of fluency, pronunciation accuracy, and comprehensibility.

Section 2 provides an overview of the automated annotation in pauses and stress, the proposed dynamic rating protocol, and data-analysis methodology we used. Section 3 presents the speech data and recruited raters. Section 4 presents the initial results of our experiment, further discussed in Section 5.

2. Methodology

2.1. Objective analysis of pauses and stresses in L2 speech

Pause positions and lexical stress in L2 speech data are first automatically annotated by a pipeline that combines the latest speech processing and natural language processing tools [25]. This pipeline is based on WhisperX speech recognition and forced alignment [26] and it annotates pauses with its corresponding largest syntactic boundary–either clause-boundary, phrase-boundary, or word-boundary–on the basis of the constituency analysis of the transcribed text using the Berkeley Neural Parser [27].

The pipeline extracts stress-related acoustic parameters from syllable nuclei points of polysyllabic content words (nouns, verbs, adjectives, and adverbs), henceforth called *target words*. Each target word is given a ”stress score” indicating whether the most prominent syllable corresponds to the prescriptive (primary) stress position and how contrasted it is compared with the other syllables of the word. This score is calculated using the following equation:

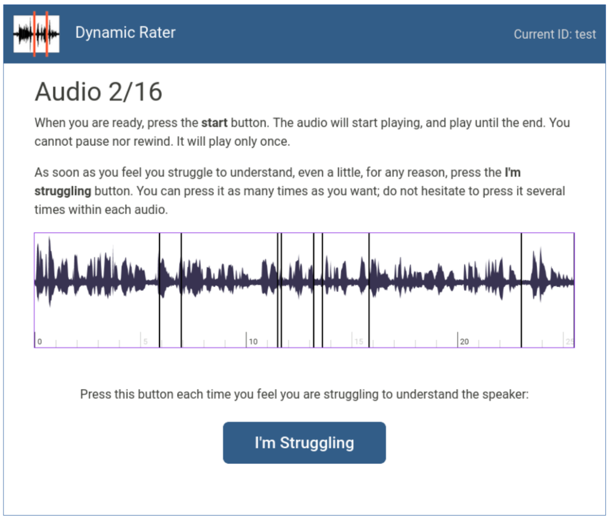
$$S_w = \frac{P_{s,w} - \overline{P_{u,w}}}{P_{s,w} + \overline{P_{u,w}}} \tag{1}$$

with w being the current word, $P_{s,w}$ being the prominence value of the prescriptive stressed syllable (mean of normalized F0, intensity, and duration), and $\overline{P_{u,w}}$ the mean prominence value of other syllables of the word. A score of 0 means that no prosodic contrast is measured between the prescriptive primary stress and other syllables, high positive scores reflect correct stress position and high contrast, and low negative scores reflect incorrect position and high contrast. Based on this score, target words are categorized as ”StressO” (appropriate stress position and strong prosodic contrast, score $> .2$), ”StressΔ” (low prosodic contrast and unclear stress pattern, score between $.2$ and $-.2$), and ”StressX” (wrong stress position and strong prosodic contrast, score $< -.2$).

2.2. Dynamic perceptive test of L2 speech

We developed a rating web application called Dynamic Rater[1]. Inspired by McIntyre's Idiodynamic Software [24], the rating procedure was slightly simplified and adapted for crowd-sourced rating. Raters are presented with successive audio recordings and asked to signal by clicking a button while they are listening whenever they are experiencing difficulty understanding the speaker (cf. Figure 1). Unlike McIntyre's software, only negative judgments are solicited, since raters tend to comment negatively rather than positively in relation to comprehensibility [28, 29], and no incrementing of judgment is done when clicking several times consecutively. Raters are required to listen to each recording in its entirety without the ability to rewind or pause, although breaks between recordings are allowed.

Following each recording, a short form appears to globally rate the speaker's pronunciation accuracy, fluency, and easiness to understand using a 100-step cursor button. An optional text zone is also provided to specify pronunciation features hindering comprehension or suggest improvements for better comprehensibility.



Audio 2/16

When you are ready, press the **start** button. The audio will start playing, and play until the end. You cannot pause nor rewind. It will play only once.

As soon as you feel you struggle to understand, even a little, for any reason, press the **I'm struggling** button. You can press it as many times as you want; do not hesitate to press it several times within each audio.

Press this button each time you feel you are struggling to understand the speaker:

Figure 1: *View of rating screen*

2.3. Normalization & click-pattern analysis

Raters should click with a range of delay after a speech event that harms comprehensibility occurs. Thus, we count clicks in a time window with a delay. Furthermore, some raters may click more frequently than others, it is therefore important to consider mean click frequencies depending on raters. We propose to mean-normalize clicks by subtracting rater's mean click rate across the all recordings. A sliding 1-second window is used to compute the mean-normalized sum of clicks, henceforth called *m-clicks*. Equation 1 details the normalization process:

$$M_w = \sum_{r=1}^{R} \left(C_{r,w} - \overline{C_r} \right) \tag{2}$$

where M_w is the number of m-clicks in window w, R is the number of raters, $C_{r,w}$ is the number of clicks by rater r within w, and $\overline{C_r}$ is the mean click rate of r. Mean-normalization centers all data points around 0, interpreting positive values as abnormally high clicking activity and negative values as less clicks than the mean clicking activity.

The frequency of m-clicks within the five consecutive windows following the onset of pauses or stressed words are then compared to see if the quantity of clicks increases, stagnates, or decreases.

At the recording level, the total number of m-clicks per rater is compared with the WP and BC ratio (number of pauses / file duration) and mean stress score to verify H3. Recordings are divided in two groups for each dimension: those above and those below the median ratio of WP pauses, BC pauses, and mean stress score. Both distributions are then compared with a Wilcoxon-Mann-Whitney rank test.

Consistency among raters is tested with the intraclass correlation coefficient [30] on raw global ratings. These ratings are then z-standardized for each rater across all the rated recordings to ensure comparability among ratings, accounting for individual raters' tendencies with a cursor-type rating system.

[1] https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/dynamic-rater

Table 1: *Number of pauses and target words per category*

	LOW		HIGH			LOW		HIGH	
	nb	%	nb	%		nb	%	nb	%
BC	59	29,9	73	39,5	**StressO**	1	1,4	22	31,9
BP	99	50,3	98	53	**Stress∆**	35	50	44	63,8
WP	39	19,8	14	7,6	**StressX**	34	48,6	3	4,3
total	**197**		**185**		**total**	**70**		**69**	

3. Experiment

3.1. L2 speech data

The speech data used in this study comprises 16 brief audio excerpts derived from dyadic argumentative discussions among French-L1 learners of English at CEFR B1 and B2 levels. The excerpts were extracted from the CLES Corpus of Spontaneous L2 English [31], totaling 11 hours of speech.

The whole corpus was annotated with the pipeline described in Section 2.1. The pause-duration threshold was set at 180 ms, as shorter pauses can be mistaken for stop closures [32], with an upper limit of 2 s as suggested in a previous study [25]. Given that the processing pipeline does not take into account secondary stress, only 2- and 3-syllable words were considered. Pauses were categorized as between clauses (BC), between phrases (BP), and within phrases (WP) for consistency with prior studies [10, 11].

The number of selected recordings was limited to 16 to enable all raters to rate all recordings. This selection was based on specific criteria: a) eight segments with a high ratio of WP pauses and a low overall lexical stress score, and b) eight segments with a low ratio of WP pauses and a high lexical stress score. Additional criteria included a minimum length of 60 tokens and an equal distribution of speaker proficiency and gender in each category. Table 1 summarizes the number of pauses per category and target words.

A manual verification of these annotations was conducted on the first 8 recordings. This involved 193 pauses and 89 stressed words. Pauses with both correct time-alignment and constituent tags totaled 82.4%, and words with correct recognition, time-alignment and syllable-alignment totaled 82.0%. Consequently, no annotations from the pipeline were modified.

3.2. Ratings on crowd-sourcing

Sixty raters were recruited through the crowd-sourcing platform Prolific. Raters were selected on the basis of specific criteria: a) native English speakers, b) declaring no proficiency in languages other than English, c) residing in the United Kingdom during the experiment, and d) ensuring an equal gender distribution. Raters' ages ranged from 25 to 72 (M: 44, SD:12).

After a training on a supplementary unanalyzed recording, the 16 recordings followed one another in a randomized order. The whole experiment was designed to take approximately 35 minutes. Only participants who had clicked at least once and had not concentrated more than 50% of their clicks on a single recording were included in this study.

4. Results

4.1. Raters behavior

The task of rating the 16 recordings consumed an average of 26 min 59 s for the 60 participants, ranging from 12 min 42 s to 1 h 3 min 2 s (with four raters exceeding 45 min). As anticipated, there was considerable variation in clicking activity among raters, with a total number of clicks ranging from 12 to 272 across the 16 recordings (mean: 76.7, standard deviation (SD): 48.65). Five raters exhibited notably high click frequencies, surpassing 120 clicks. Though frequencies of clicks differ between raters, clear peaks of clicks appeared as shown in Figures 2 and 3.

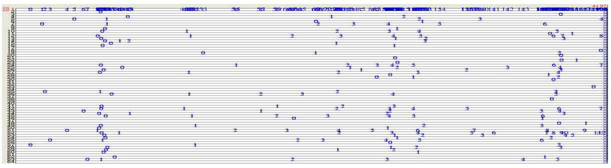


Figure 2: *TextGrid view of click activity of recording n°5, one point per click, one rater per tier with first tier containing sum of all clicks*

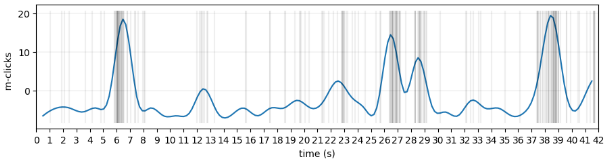


Figure 3: *M-clicks for recording n°5 (with raw clicks shown as vertical bars)*

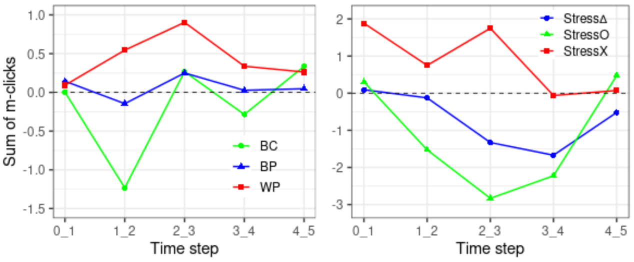


Figure 4: *Mean m-clicks over* 5 s *following pause (left) or target word (right)*

The intraclass correlation coefficient revealed an average absolute agreement among raters of .97, with an average consistency of raters' scores reaching .98 across the three global rating categories (respectively ICC1k and ICC3k from psych package 2.4.1).

4.2. Impact of pauses and lexical stress on click patterns

Click-pattern analysis involved observing the click activity following each target event, namely, each pause and stressed word. A sliding-window approach tallied m-clicks for each second from the beginning of the event to the fifth second. Figure 4 left illustrates the mean m-clicks over the 5 s following pauses BC, BP, or WP. Notably, normalized clicking activity remained above 0 for WP pauses, below 0 for BC pauses, and close to 0 after BP pauses. Clicking activity also increased during the subsequent 3 s after a WP pause (+.8) but significantly decreased during the following 2 s after a BC pause (−1.23) before returning to 0. A Wilcoxon-Mann-Whitney rank test showed significance only during the second timeframe (between 1 and 2 s following the pause, cf. Table 2).

A similar approach was applied to the lexical stress quality. Among the 139 target words, StressO constituted 17% (n=23), Stress∆ 57% (n=79), and StressX 27% (n=37) of the words.

Table 2: *Rank tests comparing number of m-clicks after BC and WP pauses (left) and after StressO and StressX (right), along with correlation coefficients between number of m-clicks in each window and stress score of target word. (−: not significant, *:$p < .05$, **:$p < .01$)*

	Rank tests		*Pearson correlations*	
	BC vs. WP	**StressO vs. StressX**	**Stress score**	
window	p-value	p-value	R	p-value
0-1s	–	–	-0.13	–
1-2s	*	*	-0.1	–
2-3s	–	**	-0.25	**
3-4s	–	*	-0.062	–
4-5s	–	–	-0.027	–

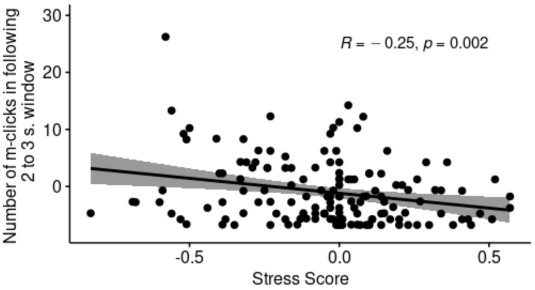


Figure 5: *Scatter plot depicting relationship between number of m-clicks occurring* 2 *and* 3 s *after target words and their corresponding stress scores.*

Figure 4 right illustrates the mean m-click sums over the 5 s following each word category. The clicking activity after incorrectly stressed words (StressX) averaged higher than after words with correct stress patterns (StressO), with a local increase between 2 and 3 s for StressX (reaching 1.75), while it continuously decreased after StressO until the third second (from .30 to -2.83). Click activity following unclear stress words stayed below 0 and progressively decreased to -1.67 between 3 and 4 s before increasing again, like StressO, on the fifth second for obscure reasons. The difference in click activity after correct and incorrect stress was significant from the second to the fourth timeframes (cf. Table 2).

As stress score is continuous, its linear correlation with the number of clicks was also analyzed without considering categories. It emerged that the lower the stress score, the higher the number of clicks, albeit with a relatively weak correlation (maximum between 2 and 3 s, $R = -.25$, $p < .01$, cf. Table 2 and Figure 5).

4.3. Impact of pauses and lexical stress on click frequency and global ratings

An analysis was conducted to determine if recordings containing more WP pauses and incorrect stress would garner more clicks than those without. As expected, recordings with a higher ratio of WP pauses received more clicks than those with less WP pauses ($p < .01$). The distribution of m-clicks was nearly equal for recordings with lower and higher ratios of BC pauses (no significant difference). Interestingly, recordings with fewer pauses overall (ratio of all pauses, regardless of category) received significantly more clicks ($p < .001$). Recordings with a lower mean stress score received significantly more clicks than those with a higher mean stress score ($p < .001$).

The same analysis was applied to global ratings. The fluency rating exhibited a negative correlation with the WP pause ratio ($p < .001$) and, surprisingly, with the BC pause ratio ($p < .05$). More pauses overall in a recording were associated with higher fluency ratings ($p < .001$). Lexical stress was compared with the global pronunciation rating, revealing that higher mean stress scores were associated with higher pronunciation ratings ($p < .001$). Similar patterns for both pauses and stress were observed with global comprehensibility ratings ($p < .001$).

5. Discussion

We conducted a real-time comprehensibility analysis on L2 English spontaneous speech. Our experiment demonstrated that listeners experienced difficulty understanding similar regions, and that dynamic perceptive tests enabled the identification of these zones. We focused on the actual impact of WP pauses and incorrect stress patterns on listeners' perception. Analyses revealed a general increase in click frequency within the 3 s following WP pauses, whereas it remained under the average click frequency following BC and BP pauses. Recordings exhibiting more pauses in general were rated with higher fluency and comprehensibility, which can be counter-intuitive, but those with more WP pauses had lower ratings. We also observed an increase in click frequency from 2 to 3 s after an incorrectly stressed word, while it decreased during the 3 s following a correctly stressed word. These results were expected, as WP pauses and incorrect lexical stress are known to be correlated with comprehensibility ratings, but this study proposed a method to observe their real-time impact on the listeners' perception.

A limitation of this study lies in the brevity of the recordings (ranging from 26 to 66 s each). Their lack of context challenged raters, potentially impacting their comprehensibility judgments. Additionally, the small number of occurrences of WP pauses and correctly stressed words may have contributed to the weakly significant results in rank tests. Furthermore, pause categories could be refined, as some WP pauses can be very natural (e.g., emphasis or enumeration), while certain BP pauses can impact comprehensibility, as they are rarely observed in L1 contexts (e.g., between a pronoun and a verb) [33]. Pause duration thresholds could also be further discussed. A previous study demonstrated that a cutoff point of 250 ms shows a higher correlation with fluency ratings [34]. Although we obtained similarly weakly significant results with this threshold, click frequencies showed more contrast with a cutoff point of 180 ms.

6. Conclusion

We propose a large-scale dynamic rating protocol for monitoring perceived comprehensibility of L2 spontaneous speech. We demonstrated the feasibility of investigating real-time comprehensibility perception through a crowd-sourced approach using our protocol and adequate number of raters. Our experiment showed that WP pauses and incorrect stress patterns directly impact comprehensibility. This protocol opens up promising perspectives for exploring linguistic phenomena impacting comprehensibility among speakers with different L1 backgrounds and proficiency levels. It may also aid in a better understanding of how listeners process L2 speech and how they react when confronted with it.

7. Acknowledgements

This work was supported by IDEX and JSPS Kakenhi n°23K25345. The authors thank the LIDILEM for funding the perception experiment and all the participants.

Scripts and data are available here: https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/dynamic-rater.

8. References

[1] R. Walker, E.-L. Low, and J. Setter, *English pronunciation for a global world*. Oxford University Press, 2021.

[2] T. M. Derwing and M. J. Munro, *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins, 2015.

[3] D. Crowther, P. Trofimovich, K. Saito, and T. Isaacs, "Second language comprehensibility revisited: Investigating the effects of learner background," *TESOL Quarterly*, vol. 49, no. 4, pp. 814–837, 2015, doi:10.1002/tesq.203.

[4] C. L. Nagle, P. Trofimovich, M. G. O'Brien, and S. Kennedy, "Comprehensible to whom? examining rater, speaker, and interlocutor perspectives on comprehensibility in an interactive context," *The Modern Language Journal*, Nov. 2022, doi:10.1111/modl.12809.

[5] D. Crowther, P. Trofimovich, K. Saito, and T. Isaacs, "Linguistic dimensions of l2 accentedness and comprehensibility vary across speaking tasks," *Studies in Second Language Acquisition*, vol. 40, no. 2, p. 443–457, 2018, doi:10.1017/S027226311700016X.

[6] S. Bhat, M. Hasegawa-Johnson, and R. Sproat, "Automatic fluency assessment by signal-level measurement of spontaneous speech," *Second Language Studies: Acquisition, Learning, Education and Technology*, 01 2010.

[7] J. Kahng, "The effect of pause location on perceived fluency," *Applied Psycholinguistics*, vol. 39, no. 3, p. 569–591, 2018, doi:10.1017/S0142716417000534.

[8] M. Candéa, "Contribution à l'étude des pauses silencieuses et des phénomènes dits d'hésitation en français oral spontané : étude sur un corpus de textes en classe de français," Ph.D. dissertation, Paris 3, 2000.

[9] F. Ferreira, "Prosody and performance in language production," *Lang. Cogn. Process.*, vol. 22, no. 8, pp. 1151–1177, 2007.

[10] S. Suzuki and J. Kormos, "The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency," *Studies in Second Language Acquisition*, vol. 45, no. 1, p. 38–64, 2023, doi:10.1017/S0272263121000899.

[11] H. Kallio, M. Kuronen, and L. Koivusalo, "The role of pause location in perceived fluency and proficiency in L2 Finnish," in *Proc. ISAPh 2022, 4th International Symposium on Applied Phonetics*, 2022, pp. 22–27, doi:10.21437/ISAPh.2022-5.

[12] Y. Cao and H. Chen, "World englishes and prosody: Evidence from the successful public speakers," *APSIPA ASC*, pp. 2048–2052, 2019.

[13] A. Cutler, "Lexical stress in english pronunciation," in *The Handbook of English Pronunciation*. Hoboken, NJ: John Wiley & Sons, Inc, 2015, pp. 106–124.

[14] A. Cutler and A. Jesse, *Word Stress in Speech Perception*. John Wiley & Sons, Ltd, 2021, ch. 9, pp. 239–265, doi:10.1002/9781119184096.ch9.

[15] A. Tortel, "Le rythme en anglais oral : considérations théoriques et illustrations sur corpus," *Recherche et pratiques pédagogiques en langues - Cahiers de l'APLIUT*, no. Vol. 40 N°1, 2021, doi:10.4000/apliut.8857.

[16] T. Isaacs and P. Trofimovich, "Deconstructing comprehensibility: Identifying the linguistic influences on listeners' l2 comprehensibility ratings," *Studies in Second Language Acquisition*, vol. 34, no. 3, pp. 475–505, 2012, doi:10.2307/26328952.

[17] Council of Europe, *Common European framework of reference for languages*. Strasbourg, France: Council of Europe, 2020.

[18] T. Isaacs, P. Trofimovich, and J. A. Foote, "Developing a user-oriented second language comprehensibility scale for english-medium universities," *Language Testing*, vol. 35, no. 2, pp. 193–216, 2018, doi:10.1177/0265532217703433.

[19] A. Tortel and D. Hirst, "Rhythm metrics and the production of English L1/L2," in *Speech Prosody 2010*, 2010, p. paper 959.

[20] Y. Inoue, S. Kabashima, D. Saito, N. Minematsu, K. Kanamura, and Y. Yamauchi, "A Study of Objective Measurement of Comprehensibility through Native Speakers' Shadowing of Learners' Utterances," in *Proc. Interspeech 2018*, 2018, pp. 1651–1655, doi:10.21437/Interspeech.2018-1860.

[21] S. Kabashima, Y. Inoue, D. Saito, and N. Minematsu, "Dnn-based scoring of language learners' proficiency using learners' shadowings and native listeners' responsive shadowings," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 971–978, doi:10.1109/SLT.2018.8639645.

[22] Z. Lin, Y. Inoue, T. Trisitichoke, S. Ando, D. Saito, and N. Minematsu, "Native Listeners' Shadowing of Non-native Utterances as Spoken Annotation Representing Comprehensibility of the Utterances," in *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, 2019, pp. 43–47, doi:10.21437/SLaTE.2019-8.

[23] C. Nagle, P. Trofimovich, and A. Bergeron, "Toward a dynamic view of second language comprehensibility," *Studies in Second Language Acquisition*, vol. 41, no. 4, p. 647–672, 2019, doi:10.1017/S0272263119000044.

[24] P. D. MacIntyre, "The idiodynamic method: A closer look at the dynamics of communication traits," *Communication Research Reports*, vol. 29, no. 4, pp. 361–367, 2012, doi:10.1080/08824096.2012.723274.

[25] S. Coulange, T. Kato, S. Rossato, and M. Masperi, "Enhancing language learners' comprehensibility through automated analysis of pause positions and syllable prominence," *Languages*, vol. 9, no. 3, 2024, doi:10.3390/languages9030078.

[26] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *Interspeech*, 2023.

[27] N. Kitaev, S. Cao, and D. Klein, "Multilingual constituency parsing with self-attention and pre-training," in *ACL*, Florence, Italy, 2019, pp. 3499–3505.

[28] J. A. Foote and P. Trofimovich, "Is it because of my language background? a study of language background influence on comprehensibility judgments," *Can. Mod. Lang. Rev.*, vol. 74, no. 2, pp. 253–278, May 2018, doi:10.3138/cmlr.2017-0011.

[29] S. Kennedy, J. A. Foote, and L. K. D. Santos Buss, "Second language speakers at university: Longitudinal development and rater behaviour," *TESOL Q.*, vol. 49, no. 1, pp. 199–209, Mar. 2015, doi:10.1002/tesq.212.

[30] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979.

[31] S. Coulange, T. Kato, S. Rossato, and M. Masperi, "A corpus of spontaneous l2 english speech for real-situation speaking assessment," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, May 2024.

[32] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447010000628

[33] J. Tauberer, "Predicting intrasentential pauses: is syntactic structure useful?" in *Speech Prosody 2008*, 2008, pp. 405–408.

[34] N. H. De Jong and H. R. Bosker, "Choosing a threshold for silent pauses to measure second language fluency," in *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech, DiSS*, R. Eklund, Ed., 2013, pp. 17–20.