Pause position analysis in spontaneous speech for L2 English fluency assessment *

Sylvain COULANGE[1], Tsuneo KATO[2]
([1]LIDILEM/LIG, Université Grenoble Alpes; [1,2]SLPL, Doshisha University)

1 Introduction

Within the realm of language teaching and linguistics, fluency is often associated with the smoothness of speech flow and is considered a crucial aspect to ensure listener comprehension [1]. While fluency is mainly a perceived impression by the listener [2], it is possible to measure some acoustic phenomena in the speech that can potentially influence the listener's ease of understanding. In the context of assessment of spontaneous speech, automated tools often analyze disfluencies, phenomena that may disrupt the smoothness of speech, and mainly deal with the frequency and average duration of pauses [3]. Although pauses may become detrimental when they occur at unexpected positions, they are not inherently problematic and can help segment a speech flow or convey ideas when strategically used.

This paper presents an automated pipeline for transcribing and identifying pause positions at both the structural and the lexical level in non-native spontaneous speech. At structural level, we conducted constituency analysis on the transcribed text and categorized the pause positions into inter-clause, inter-phrase, and intra-phrase classes. In the lexical analysis, we classified the pause positions in accordance with the part-of-speech (POS) of preceding and following contextual words, and further conducted co-clustering of POS and speakers to identify how each student tends to pause in the most frequent syntactic contexts, regardless of their proficiency level.

The pipeline was used to analyze the spontaneous argumentative speech of 176 French learners of English at the CEFR B1 and B2 proficiency levels, in which the latter is widely recognized as a threshold for achieving a certain level of fluency. We hypothesize that B2 students make fewer disfluent pauses and more structural ones than B1 students, thanks to a greater ability to plan and structure their speech and retrieve vocabulary.

The rest of the paper is organized as follows. In Section 2, we review previous studies on pause position analysis in both native and non-native English speech. Section 3 presents the speech corpus used in our analysis and Section 4 provides a detailed explanation of the processing pipeline. The results are presented in section 5, followed by a discussion.

2 Related work

First, it is necessary to define pauses. Silent pauses are commonly described as interruptions of phonation [4]. The duration at which such an interruption is considered a pause varies significantly across studies, typically ranging from 100 to 400 milliseconds [5, 6]. Pauses can also be filled by phoneme lengthening, or filler words like "uh."

Furthermore, pauses can be categorized on the basis of their functions, such as respiratory, hesitation, grammatical, or stylistic [4]. Two major types of pauses are identified here: structuring and non-structuring pauses [7]. Structuring pauses aid in segmenting and structuring discourse, while non-structuring pauses are typically preceded by hesitation and serve the purpose of self-correction or finding the appropriate following word, and can add to the listener's cognitive load.

The relationship between pause position and syntax has been studied for several decades and appears to be significant. [8] utilized POS information and syntactic structures to predict intra-utterance pauses in spontaneous English speech of native speakers from the Switchboard corpus. He concluded that combining both types of information yielded better predictions than using solely word-level information. Most pauses tended to appear near conjunctions, hesitation fillers, before pronouns, or subjects. Conversely, pauses were unlikely to occur after subjects, between verbs and the particle "to," between verbs and prepositional phrases, or between prepositions and noun phrases. To exclusively examine priori structuring pauses, [9] analyzed the speech of "successful speakers," including both native and non-native English speakers delivering political speeches or short TED talk-style speeches. They found that, aside from emphasizing particular words, pauses primarily occurred between clauses, often around subordinate conjunctions such as "which," "that," and "when" with no discernible difference between native and non-native speakers.

Pauses therefore play an important role in structuring the flow of speech and, in addition to its duration and frequency, it is important to study its position in an utterance to determine whether its distribution reflects a better mastery of the L2 language.

3 Data

Our dataset consists of the L2 English speech of 176 French learners recorded during the oral interaction speaking task of the CLES (See https://www.certification-cles.fr/english/), a national, government-certified test of language proficiency in France. This task involved a 10-minute role play where two or three candidates engaged

*L2 英語流暢性評価のための自然発話におけるポーズ位置分析. クランジュ・シルヴァン [1], 加藤恒夫 [2] ([1] グルノーブルアルプ大学院, [1,2] 同志社大学院)

in an argumentative discussion on a controversial topic, such as e-cigarettes, security cameras, or the use of technology in the classroom. Each participant was evaluated by two experts, who assessed them on various dimensions and assigned a final speaking proficiency level of either B1 or B2, in accordance with the CEFR [10]. The overall proficiency distribution of the students was 56% at level B2 and 44% at level B1 (based on the global score obtained from the CLES exam), while speaking proficiency was at B2 for 66% and at B1 for 34% of the participants.

4 Methodology

The automated processing pipeline involved several steps: neural speaker diarization using Pyannote [11], speech recognition and force alignment using WhisperX [12], morphosyntactic analysis using SpaCy [13] and constituency analysis using the Berkeley Neural Parser [14]. The recordings were segmented into mono-speaker continuous speech segments using Pyannote's voice activity detection, with a threshold set at 1 second. Segments with a duration of 8 seconds or less were excluded to avoid short utterances. This resulted in a corpus of 11 hours of continuous speech. The average duration of speech per speaker was 3'44" (min 0'32", max 6'51", SD 1'20"). The transcribed text was annotated in POS tag and aligned to the corresponding audio signal, with an empty interval tagged as "<p:>" separating the left and right words. From this data, all <p:> segments, along with their left and right POS tags, as well as the largest ending and starting constituents identified through constituency analysis, were extracted. Pauses were defined as <p:> segments with a duration equal or greater than 180 milliseconds. <p:> segments could either be silent or filled with phoneme lengthening, hesitation, laughter, etc., which explains why several segments exceeded 1 second in duration. <p:> segments longer than 2 seconds, often resulting from inaccurate word alignment, were excluded. This paper will focus on the analysis of the 21,942 pauses extracted from the 72,594 <p:> segments of the corpus.

Our approach encompasses conducting a comparative analysis of the distribution of pauses within the syntactic structure of each utterance for both B1 and B2 proficiency groups, then looking at pausing patterns in the most common lexical contexts. We posit that B1 students are more likely to exhibit pauses in unexpected contexts, specifically within phrases, as opposed to at clause junctures where pauses are typically anticipated. In terms of lexical patterns, we anticipate a higher occurrence of pauses between word categories that normally do not expect pauses, such as between prepositions and determiners, determiners and nouns, or pronouns and verbs. Conversely, we expect fewer pauses before or after conjunctions. At the syntactic level, we expect to observe a greater frequency of pauses within phrases, and a lesser frequency between clauses.

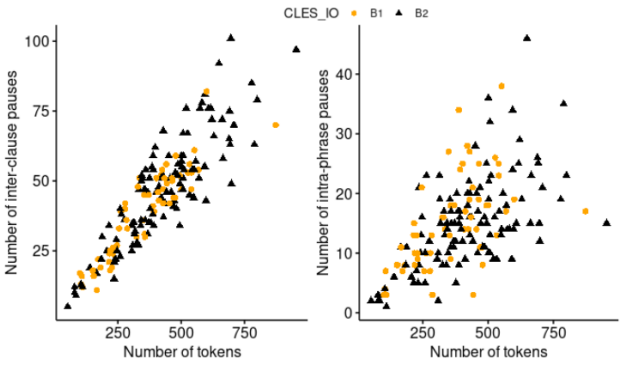

Fig. 1 Number of inter-clause (left) and intra-phrase (right) pauses per speaker.

5 Results

This section compares students from the B1 and B2 speaking proficiency levels. The duration of speech per speaker is similar for both groups, as indicated by the non-parametric rank test (Wilcoxon Mann Whitney) that shows no significant difference. However, the speech rate of B2 students is faster (median at 110 tokens/minute) compared with B1 students (97 tokens/minute), with a significant difference at $p < .0001$. This indicates that B2 students convey more content within the same duration of speech time. Additionally, B2 students make more pauses (median at 34.3 pauses/minutes/speaker) compared with B1 students (30.7), with a significant difference at $p < .01$. However, the mean duration of their pauses is shorter (592ms) compared with B1 students (615ms) at a significance level of $p < .005$. Note that the proportion of silence per speaker is similar between the two groups, with no significant difference (median at 33% for both groups).

5.1 Structural analysis

To further analyze the structural aspects, the number of pauses between clauses and within phrases was examined. The total number of clause and word boundaries with pauses was counted for each speaker. The results reveal that B2 students make on average more pauses between clauses (47 pauses) compared with B1 students (42), with a significant difference at $p < .05$. Regarding the distribution of pauses within phrases, B1 students demonstrate a wider range, although no significant difference is observed between the two proficiency levels. It is important to keep in mind that the absolute number of pauses strongly correlates with the quantity of speech. Nevertheless, Figure 1 shows that at an equal number of tokens, students can make a very different number of intra-phrase pauses (such as 10 and 36 pauses at 500 tokens for two B2 students, respectively). However, the variation for inter-clause pauses is much narrower.

Comparing the proportion of pauses to mitigate the effect of speech quantity, the difference between B1 and B2 disappeared for clause boundaries (median at 10.9% for B1 and 10.6% for B2, no signifi-

cant difference), but is significant for pauses within phrases (4.2% for B1 and 3.4% for B2 at $p < .005$). No correlation is seen between the proportion of pauses between clauses and within phrases for both groups as shown in Figure 2.

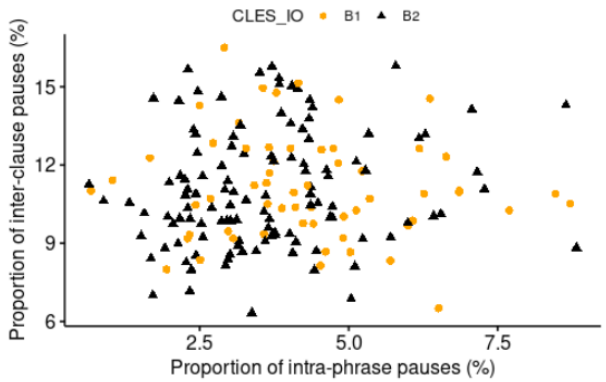

Fig. 2 Proportion of inter-clause and intra-phrase pauses per speaker B1 and B2, both correlations are not significant.

5.2 Lexical analysis

Furthermore, the pausing patterns at the lexical level between B1 and B2 were analyzed. This subsection focuses on the immediate syntactic context of pauses within the top 15 most frequent consecutive POS pairs observed in the corpus. The proportion of occurrences with a pause was computed for each pair in both the B1 and B2 subcorpura. This analysis enabled for a comparison of pausing tendencies between B1 and B2 students in each context. Despite a very thin difference, the results show that B2 students generally make fewer pauses than B1 students in these 15 contexts, with the largest gaps observed between nouns and pronouns (-4 points), nouns and coordination conjunctions (-3.5 points), and subordinate conjunctions (SCONJ) and pronouns (-3.4 points). These contexts are likely to be clause boundaries, which contradicts the hypothesis that B2 students make more pauses between clauses to enhance speech structure. However, B2 students noticeably make more pauses than B1 students in two contexts: between nouns and prepositions (ADP, +4.2 points) and between verbs and determinants (DET, +2.7), which are likely to be phrase boundaries.

The unsupervised co-clustering [15] of students and their pausing patterns in the 15 analyzed contexts, resulted in the identification of three distinct student clusters shown in Figure 4. These clusters exhibit two predominant profiles that are primarily differentiated by the overall frequency of pauses (clusters 1 and 2). Additionally, there is an additional cluster (cluster 0, left) consisting of students with extreme values, likely due to insufficient observations in certain contexts, leading to a less structured grouping. Cluster 2 demonstrated a higher frequency of pauses across all 15 contexts. It encompassed 53% of B2 students and 42% of B1 students, while cluster 1 included 28% of B2 students

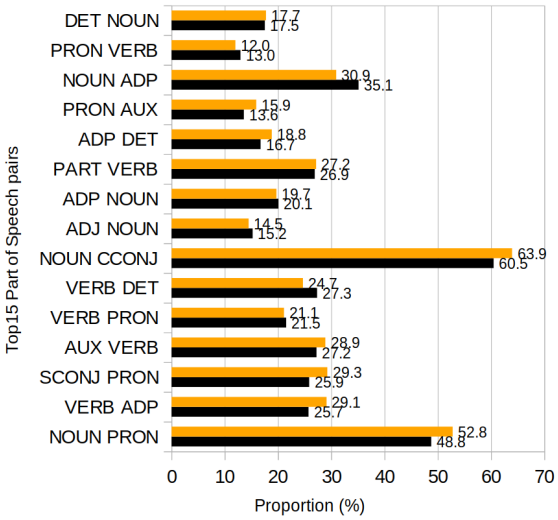

Fig. 3 Proportion of POS pairs containing a pause for B1 (yellow) and B2 (black) speakers.

and 29% of B1 students, and cluster 0 consisted of 19% of B2 students and 29% of B1 students.

The disparity in pause frequency between clusters 1 and 2 within each context was significantly larger than the differences observed between the B1 and B2 proficiency levels (*cf.* Figure 5). However, while cluster 2 has almost half the number of students of cluster 1, distributions of pause frequencies per context showed wider ranges of values.

6 Discussion

We analyzed the position of pauses of students from the B1 and B2 proficiency levels. At a structural level, B2 students showed as expected a significantly lower proportion of intra-phrasal pauses, that are more likely to hinder the speech, than B1 students ($p < .005$), but both groups made the same proportion of inter-clausal pauses, which are more likely to help structuring it. At a lexical level, B1 and B2 students generally made their pauses in the same proportion in each of the 15 most frequent POS contexts, with slightly fewer pauses for B2 even in contexts where pauses should have a positive effect. This could be explained by a better performance in discourse planning for B2 students, but it rather seems due to a wide diversity of pausing profiles as shown in Figure 2.

The unsupervised clustering of students suggested groups mixing B1 and B2 students, on the basis of the overall frequency of pauses.

By plotting proportions of inter-clausal and intra-phrasal pauses for each speaker from clusters 1 and 2 (cf. Figure 6), it appears that there is no significant correlation between both types of pauses among students from cluster 1. However, there is one among those of cluster 2, in which students who make more inter-clausal pauses tend to do fewer intra-phrasal ones ($R = -.3$, $p < .05$).

In summary, B2 learners from our corpus made in average fewer potentially disfluent pauses, but

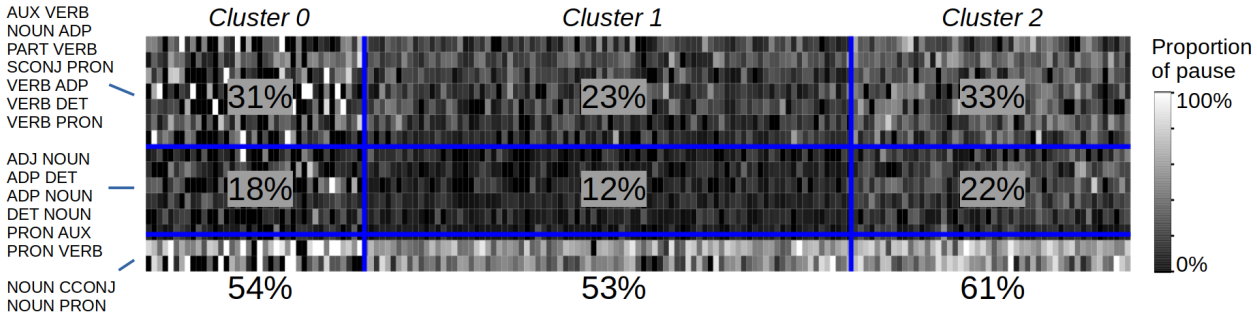

Fig. 4 Clustering output of pausing patterns in top 15 POS contexts, speakers in columns, POS pairs in rows, with the mean value of each block. Darker areas mean fewer pauses.

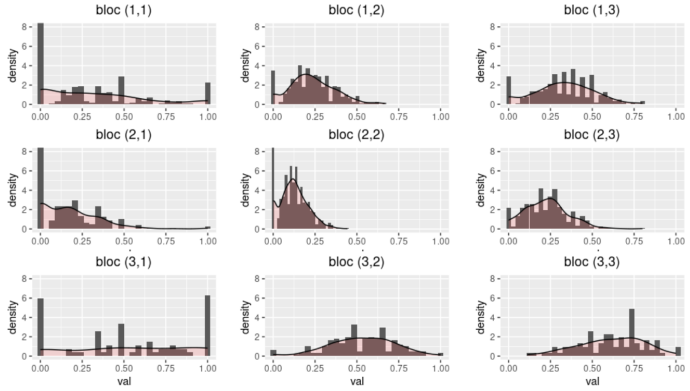

Fig. 5 Distributions for each block of the clustering shown in Figure 4. In columns from left to right: student clusters 0, 1, and 2.

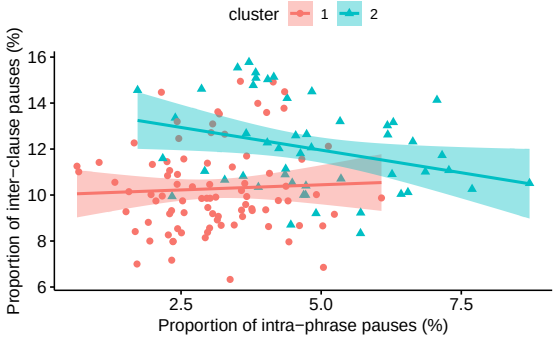

Fig. 6 Proportion of inter-clause and intra-phrase pauses per speaker from clusters 1 (red) and 2 (blue), correlation for cluster 1 is not significant, that for cluster 2 is $R = -.3$ $p < .05$.

not more structural ones than B1 learners. There is a great variety of speakers in the way they are using pauses, and this should impact the ease of understanding on the listener side. We assumed that the absence of pauses between clauses and the presence of pauses within phrases are likely to hinder the speech, but the pausing pattern is much more complex, especially speaking situations involving multiple speakers, where pauses can be used for regulating speech turns, as well as emphasizing ideas or touches of sarcasms. We plan to conduct a perceptive test to try to identify when pauses are more likely to hinder the speech, and when it helps understanding it.

Acknowledgment This work was supported by JSPS KAKENHI n° 23H00648. The authors thank the IDEX for funding S.C. mobility grant.

References

[1] 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.

[2] R. Lickley, *Fluency and Disfluency*, pp. 445–469. Chichester: Wiley Online Library, 2015.

[3] K. Evanini and K. Zechner, *Overview of automated speech scoring*, pp. 3–20. Innovations in Language Learning and Assessment at ETS, London, England: Routledge, 2019.

[4] I. Grosman, A. C. Simon, and L. Degand, "Variation de la durée des pauses silencieuses : impact de la syntaxe, du style de parole et des disfluences," *Langages*, vol. 211, no. 3, pp. 13–40, 2018.

[5] J. Trouvain, *Tempo Variation in Speech Production: Implications for Speech Synthesis*. PhD thesis, 2004.

[6] P. Tavakoli, "Pausing patterns: differences between L2 learners and native speakers," *ELT Journal*, vol. 65, no. 1, pp. 71–79, 2010.

[7] M. Candea, *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*. PhD thesis, 2000.

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

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

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

[11] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Interspeech*, 2021.

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

[13] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spacy: Industrial-strength natural language processing in python," 2020.

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

[15] P. Singh Bhatia, S. Iovleff, and G. Govaert, "blockcluster: An R package for model-based co-clustering," *Journal of Statistical Software*, vol. 76, no. 9, pp. 1–24, 2017.