Measuring speech rhythm through automated analysis of syllabic prominences

Noriko Nakanishi[1], Sylvain Coulange[2]

[1.] Kobe Gakuin University (Japan), [2.] Univ. Grenoble Alpes (France)

Perceived fluency refers to the smoothness of speech delivery (Lickley, 2015). A key element contributing to fluency is speech rhythm, defined as the recurrence of perceivable temporal patterns of strong and weak elements over time (Gibbon & Gut, 2001). English is commonly categorized as a stress-timed language, where stressed syllables contrast with unstressed ones, and content words tend to be stressed while function words are reduced. This contrasted pattern at syllabic and lexical levels helps listeners in segmenting speech and focusing on essential information (Cutler, 2015).

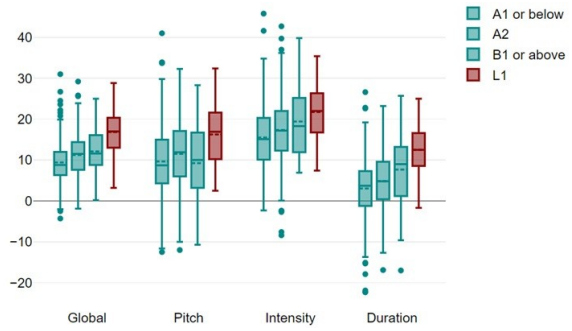
Learning English as a foreign language thus implies correctly stressing words for ease of understanding (Isaacs et al., 2017). This can be particularly challenging when the learner's L1 has a different rhythmic system. For instance, Japanese is characterized by a mora-timed rhythm, where each mora has a regular duration (Mihara & Takami, 2013). Thus, Japanese-English speech (JE) may exhibit different rhythmic patterns compared to Native English speech (NE). Specifically, we hypothesize that JE, compared to NE, demonstrates 1) lower prosodic contrast between syllables within polysyllabic words, and 2) a less pronounced difference between content and function words.

We tested these hypotheses using a 34-hour read-aloud corpus containing 877 JE and 91 NE samples. The JE samples were recordings of 42 Japanese university students with English proficiency levels ranging from CEFR A1 or below to B2, while the NE samples were recordings of 7 professional narrators. We aligned the reference texts using MFA3.0 (McAuliffe et al., 2017) and analyzed syllabic prominence with an adapted version of PLSPP (Pauses and Lexical Stress Processing Pipeline, Coulange et al., 2024). This pipeline uses syntactic analysis and speaker-normalized measures of pitch, intensity, and duration of each vowel interval to characterize the accuracy and degree of prominence of syllables in polysyllabic words. We extended the measures to monosyllabic words and compared how content and function words were pronounced. A manual evaluation of word-level alignment precision across the entire corpus showed 92.24% accuracy for NE and 79.67% for JE. Excluding words with incorrect syllable count (29%) had minimal impact on the results. Thus, we report findings based on the full corpus.

Two types of prosodic scores were calculated: *Syllabic Contrast Scores* between stressed and unstressed syllables within polysyllabic words (Figure 1), and *Lexical Contrast Scores* between content and function words, including monosyllabic words (Figure 2). These scores were compared according to the speakers' English proficiency levels. Repeated measures ANOVAs with Holm's Sequentially Rejective Bonferroni correction revealed significant tendencies: higher CEFR levels in JE corresponded with higher syllabic and lexical contrast scores, gradually approaching those of NE. Notably, duration was the strongest indicator of lexical contrast, suggesting that lower-level learners are more influenced by the rhythm of their L1.

This study proposed a method to automatically measure the degree of prosodic contrast between syllables to characterize the influence of L1 rhythm on L2 English. The method proved promising for assessing English speech rhythm across various proficiency levels. Our next step is to conduct stress analysis at the sentence level to investigate how learners emphasize essential information within their utterances.

Figure 1. Syllabic Contrast Scores within Polysyllabic Words by CEFR Level.



ANOVA between groups (***$p < .001$)

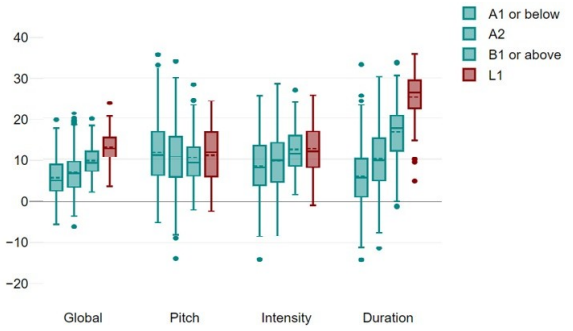
Global scores:
A1 <*** A2, B1, L1
A2, B1 < *** L1

Pitch scores:
A1 <*** A2, L1
A2, B1 < *** L1

Intensity scores:
A1 <*** A2, B1, L1
A2 < *** L1

Duration scores:
A1 <*** A2 <*** B1 < *** L1

Figure 2. Lexical Contrast Scores between Content and Function Words by CEFR Level.



ANOVA between groups (***$p < .001$)

Global scores:
A1 <*** A2 <*** B1 < *** L1

Pitch scores:
n.s.

Intensity scores:
A1 <*** A2, B1, L1
A2 <*** L1

Duration scores:
A1 <*** A2 <*** B1 < *** L1

References:

Coulange, S., Kato, T., Rossato, S., & Masperi, M. (2024). Enhancing language learners' comprehensibility through automated analysis of pause positions and syllable prominence. *Languages*, *9*(3), 78. doi:10.3390/languages9030078

Cutler, A. (2015). Lexical Stress in English Pronunciation. In *The Handbook of English Pronunciation*, 106–124. Hoboken, NJ: John Wiley & Sons, Inc.

Gibbon, D., & Gut, U. (2001, September 3). Measuring speech rhythm. *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*. Presented at the 7th European Conference on Speech Communication and Technology (Eurospeech 2001). doi:10.21437/eurospeech.2001-36

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*(2), 193–216. doi:10.1177/0265532217703433

Lickley, R. J. (2015). Fluency and Disfluency. In *The Handbook of Speech Production*, 445–474. doi:10.1002/9781118584156.ch20

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech 2017*, 498–502. doi:10.21437/Interspeech.2017-1386

Mihara, K., & Takami, K. (Eds.), Kubozono, H., Namiki, Y., Ono, N., Sugimoto, T., & Yoshimura, A. (2013). *The Basics of English Linguistics: A Contrastive Study of English and Japanese*. Kuroshio.