



# A Bayesian Approach to L2 Fluency Ratings by Native and Nonnative Listeners

*Kakeru Yazawa<sup>1</sup>, Takayuki Konishi<sup>2</sup>*

<sup>1</sup>Institute of Humanities and Social Sciences, University of Tsukuba, Japan

<sup>2</sup>School of Languages and Communication, Kobe University, Japan

yazawa.kakeru.gb@u.tsukuba.ac.jp, tkonishi@port.kobe-u.ac.jp

## Abstract

This study investigates how native and nonnative listeners evaluate the fluency of Japanese speakers' English using a Bayesian modeling framework. Data were obtained from 16 listeners with diverse linguistic backgrounds (Cantonese, English, French, German, Japanese, Korean, Mandarin, Polish, Punjabi, and Spanish), who rated English read speech samples from 180 Japanese speakers, in the J-AESOP corpus. Utterance fluency measures included speed (syllable- or segment-based articulation rate), breakdown (pause frequency and duration), and repair (repetitions). Results revealed that nonnative listeners, particularly those with Asian language backgrounds, were generally more lenient and less reliant on speech rate than native listeners, highlighting inter-listener variability previously overlooked. Model comparisons also revealed that segment-based articulation rate better captures utterance speed fluency than the commonly adopted syllable-based articulation rate.

**Index Terms:** fluency ratings, (non)native listeners, Bayesian modeling, Japanese speakers' English, J-AESOP corpus

## 1. Introduction

Perceived fluency plays a crucial role in the assessment of second language (L2) speech, as it strongly influences listeners' judgments of a speaker's overall L2 proficiency and comprehensibility [1, 2]. Over the past four decades, substantial progress has been made in understanding how temporal aspects of L2 speech affect fluency ratings (see [3] for a historical perspective). However, most existing research focuses on native (L1) listeners' judgments, leaving unclear how nonnative listener evaluations may differ from these established norms.

The current study aims to address this gap by comparing fluency ratings of L1 Japanese speakers' L2 English speech, as evaluated by 16 phonetically trained listeners with various L1 backgrounds, using data from the J-AESOP corpus [4]. Unlike previous studies, it employs a Bayesian approach in place of a frequentist one, allowing inter-listener variability to be investigated in terms of probabilistic distributions rather than statistically significant differences. These methodological shifts are also useful for revisiting and refining commonly applied measures of utterance fluency, as discussed below.

### 1.1. Measures of L2 utterance fluency

Previous studies have explored and employed various measures of L2 utterance fluency, often grouped into three sub-categories: speed, breakdown, and repair [5].

Speed refers to the rate at which speech is produced. The most common measure is articulation rate (AR), usually employed as the number of syllables per second excluding pauses.

Breakdown refers to interruptions in the flow of speech. Often used measures are the frequency and duration of filled and unfilled pauses, while pause location may also be examined.

Repair refers to corrections speakers make to their speech. This includes instances of repetitions, self-corrections, false starts, and reformulations, typically measured by frequency.

### 1.2. Perceived fluency of Japanese speakers' English

Several studies have investigated L1 Japanese speakers' L2 English fluency using the aforementioned measures (see [6] for a meta-analysis). Overall, these studies suggest that perceived fluency as judged by native English listeners is strongly associated with AR and pause frequency, moderately with pause duration, and weakly with repair frequency. Each study also addressed different research questions, such as how learners' L2 fluency relates to their L2 proficiency [7], L1 fluency [8], and cognitive skills [9], how it develops over time [10], and how it is affected by interactions with other learners [11].

Magne et al. [12] represent one of very few studies that examined nonnative English listeners' judgments of Japanese speakers' English fluency. They asked 10 L2 users of English with various European language backgrounds (French = 2, Hungarian = 2, Russian = 1, Spanish = 4, Ukrainian = 1) to evaluate English speech samples from 90 Japanese speakers for perceived fluency. The results were largely consistent with native listeners' judgments, with AR and pause frequency as primary predictors of fluency ratings.

### 1.3. Current study

The bias toward native listeners' judgments, which is pervasive in the L2 fluency literature [6], is problematic at both conceptual and practical levels, as the diverse perspectives and evaluation criteria of nonnative listeners are virtually overlooked. The current study attempts to bridge this gap by comparing fluency ratings by (non)native listeners with various language backgrounds against Japanese speakers' English speech to extend prior research.

A further limitation of previous studies is their reliance on frequentist statistical models, such as ANOVA [10], regression [7, 8, 11, 12], and structural equation modeling [9]. These models provide binary decisions about whether a factor (e.g., listeners' L1) has a significant effect or not, although the absence of statistical significance does not imply the factor is irrelevant. Bayesian modeling, as adopted in the current study, is more suitable for capturing the variability and uncertainty in fluency ratings within and across listeners. By applying such analysis to (non)native listeners' data, the study also re-assesses the validity of utterance fluency measures that have been identified as significant predictors of native listener judgments.

## 2. Methods

### 2.1. Speakers

The speech samples came from 180 native Japanese speakers (114 female and 66 male) in the J-AESOP corpus [4]. They were undergraduate or graduate students at universities in Tokyo and surrounding areas, aged between 18 and 38 (mean = 20.3, standard deviation = 2.4). Most began learning English at age 13 as part of Japan’s compulsory junior high school curriculum, while others were first exposed to English at an earlier age (mean = 10.6, standard deviation = 3.7). Approximately two-thirds had resided in Japan throughout their lives, while the remaining third had lived abroad for varying durations.

### 2.2. Materials

While the corpus features various types of recording tasks, this study focuses on Task 6\_01, where participants read aloud the English version of “The North Wind and the Sun” [13]. The full text is provided in Table 1.

Table 1: Full text of “The North Wind and the Sun”

Section	Material
1	<i>The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.</i>
2	<i>Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt.</i>
3	<i>Then the Sun shone out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.</i>

### 2.3. Measurement

The standard word- and segment-level annotation files of the corpus were used to calculate the following measures of utterance fluency for each speaker and each section.

- *Speed*
  - The number of syllables (as identified by vowel labels, including epenthetic ones) divided by speech duration excluding pauses (*syllable-based AR*)
  - The number of segments (as identified by vowel and consonant labels, including epenthetic ones) divided by speech duration excluding pauses (*segment-based AR*)
- *Breakdown*
  - The number of (un)filled pauses (*pause frequency*)
  - The mean duration of (un)filled pauses (*pause duration*)
- *Repair*
  - The number of word repetitions, as identified by the “(RPT)” tag used in the corpus (*repetitions*)

### 2.4. Fluency ratings

The perceptual rating data from the corpus [14] were used. For Task 6.01, 16 phonetically trained listeners with various L1 backgrounds (Table 2) rated each speaker’s recording of each section for perceived fluency on a scale of 1 to 10. Note that the group labels in Table 2 (“American,” “Japanese,” “Asian,” and “European”) are unique to this study and do not reflect the classifications used in the original corpus; the possibility of alternative groupings is discussed in §4. The intraclass correlation coefficient (ICC) using a two-way model for consistency was 0.882 for “American,” 0.813 for “Japanese,” 0.844 for “Asian,” and 0.837 for “European” listeners, respectively, with an overall ICC of 0.844.

Table 2: Listeners examined in the study

ID	L1	Group
eng1	American English	“American”
eng2	American English	
eng3	American English	
eng4	American English	
jpn1	Japanese	“Japanese”
jpn2	Japanese	
jpn3	Japanese	
jpn4	Japanese	
kor1	Korean	“Asian”
pan1	Punjabi	
yue1	Cantonese	
zho1	Mandarin	
deu1	German	“European”
fra1	French	
pol1	Polish	
spa1	Spanish	

### 2.5. Statistical analysis

All statistical analyses were performed in R<sup>1</sup>. Bayesian mixed-effects models (“models” hereafter) were fitted using the *brms* package [15], and posterior parameter distributions were visualized using the *bayesplot* package [16]. Each model had the following structure:

```
model <- brm(
  Fluency ~ Speed + Breakdown + Repair +
  (1 + Speed + Breakdown + Repair | Listener),
  data = df, family = gaussian(),
  iter = 2000, warmup = 1000, chains = 4
)
```

where perceived fluency was predicted by a given combination of speed, breakdown, and repair measures of utterance fluency, with random intercepts and slopes for each listener. The default (i.e., weakly informative) priors were used to minimize potential biases from prior knowledge, especially because judgments by the “Japanese” and “Asian” listeners may differ from those by native English and European language listeners tested previously.

Additionally, the *loo* package was used to perform leave-one-out cross-validation (LOO-CV) for comparing the predictive performance of different models, as described below.

<sup>1</sup><https://www.r-project.org>

### 3. Results

#### 3.1. Evaluating fluency measures

Before proceeding to inter-group comparisons, it is important to identify which combination of speed, breakdown, and repair measures is most suitable as the basis of analysis. To this end, a series of models with different fluency measures as independent variables (Table 3) were compared, using all data (180 speakers  $\times$  3 sections  $\times$  16 listeners = 8640 observations).

Model A serves as the base model, incorporating speed and breakdown measures that have demonstrated particular effectiveness in previous studies. Model B investigates whether replacing syllable-based AR with segment-based AR improves predictive performance; the rationale for this substitution will be discussed later in detail (§4). Models C and D assess the effects of including pause duration and repetitions—moderate and weak predictors of native listeners’ fluency judgments, according to previous research—on the model’s predictive power. Finally, Model E is a maximally different model from the base model.

Table 3: *Models compared*

Model	Speed	Breakdown	Repair
A	AR (syllable)	Pause freq.	-
B	AR (segment)	Pause freq.	-
C	AR (syllable)	Pause freq. & dur.	-
D	AR (syllable)	Pause freq.	Repetitions
E	AR (segment)	Pause freq. & dur.	Repetitions

The results of the model comparison are shown in Table 4. Here,  $\Delta$ ELPD (the difference in expected log predictive density (ELPD)) is a measure of predictive performance of the models, with higher (less negative) values indicating better performance. It is important to interpret  $\Delta$ ELPD in conjunction with its standard error ( $\Delta$ SE), which quantifies the uncertainty associated with the estimated predictive performance.

Table 4: *Results of model comparison*

Model	$\Delta$ ELPD	$\Delta$ SE
E	0.0	0.0
B	-122.4	16.3
C	-765.3	30.5
D	-774.7	27.8
A	-827.7	29.1

Model A (the base model) shows the lowest  $\Delta$ ELPD, suggesting that relying solely on syllable-based AR and pause frequency is insufficient for optimal prediction. Substituting syllable-based AR with segment-based AR in Model B substantially improves performance, indicating that segment-based AR captures additional information not effectively represented by syllable-based AR. Models C and D, which incorporate pause duration and repetitions, also show performance gains compared to the base model, though these improvements are less pronounced than those with Model B. Model E, which integrates the substituted and added predictors from Models B to D, achieves the highest performance, demonstrating the synergistic benefits of combining these measures.

Based on the above results, Model E was selected for detailed analysis of inter-group differences, presented next.

#### 3.2. Comparing native and nonnative listener groups

Figures 1 to 4 present the posterior distributions of parameters in Model E for data from each of the four listener groups.

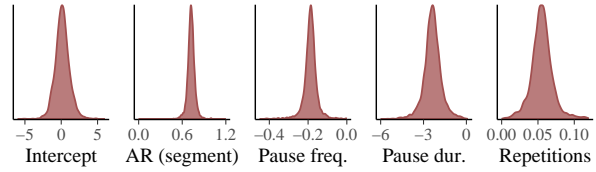


Figure 1: *Posterior distributions for American listeners’ results.*

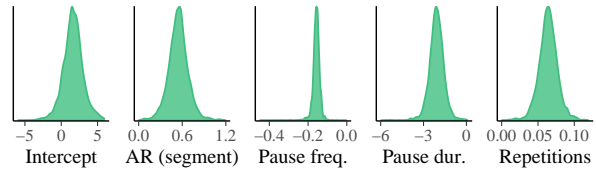


Figure 2: *Posterior distributions for Japanese listeners’ results.*

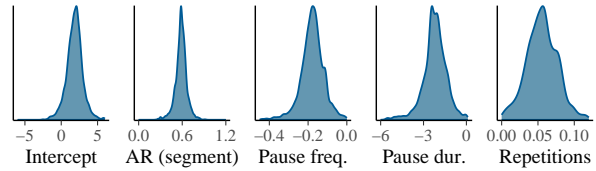


Figure 3: *Posterior distributions for Asian listeners’ results.*

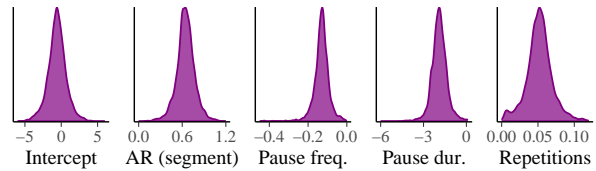


Figure 4: *Posterior distributions for European listeners’ results.*

The intercepts reflect baseline fluency ratings across groups. “Japanese” and “Asian” listeners showed generally higher values than “American” and “European” listeners, suggesting more lenient judgments by the former groups than the latter groups.

The effect of segment-based AR was positive across all groups, indicating that faster speech is consistently associated with higher fluency ratings. Notably, native (i.e., “American”) listeners appear to rely more heavily on AR than nonnative listeners, particularly “Japanese” and “Asian” listeners.

Pause frequency and duration exhibited similarly negative effects on perceived fluency across all groups, with more frequent and longer pauses leading to lower ratings. However, the degree of uncertainty associated with these effects varied by group. For example, “Japanese” listeners showed less uncertainty in the effect of pause frequency compared to other groups.

Repetitions showed weak effects across all groups, with confidence intervals close to zero. This finding aligns with prior research suggesting a relatively minor role of repair fluency.

## 4. Discussion

### 4.1. Summary

This study investigated how native and nonnative listeners with diverse L1 backgrounds evaluate the fluency of Japanese speakers' English, using a Bayesian modeling framework to analyze various utterance fluency measures. Among the tested models, the one incorporating segment-based AR, pause frequency, pause duration, and repetitions (Model E) demonstrated the best predictive performance. While the overall patterns of fluency judgments were generally consistent across the four listener groups ("American," "European," "Japanese," and "Asian"), some notable differences emerged. First, "Japanese" and "Asian" listeners tended to assign more lenient fluency ratings than "American" and "European" listeners. Second, "American" listeners were more reliant on AR in fluency judgments than their "Japanese" and "Asian" counterparts. Finally, the degree of uncertainty associated with the effects of pause frequency and duration varied considerably across groups.

### 4.2. Beyond native listener norms

A central contribution of the current study is its demonstration of the variability and uncertainty in L2 fluency ratings across native and nonnative listeners. While all four groups generally agreed that faster speech with fewer and shorter pauses indicates higher fluency, systematic differences emerged in their overall leniency and reliance on specific temporal measures, as highlighted above. The findings add to prior research that focused on native English listeners [7, 8, 9, 10, 11] (or European language listeners [12]), extending the understanding of how fluency judgments may qualitatively differ among diverse listener populations. Importantly, the adoption of Bayesian modeling proved useful in revealing these inter-group differences. Unlike frequentist approaches, which often reduce findings to binary significance tests, Bayesian methods allowed for a nuanced visualization and interpretation of variability and uncertainty, both within and across listeners.

The observed inter-group differences may reflect listeners' varying levels of familiarity with, or expectations for, Japanese speakers' English. Other factors, such as nonnative listeners' proficiency in English, may also play a role. Regardless of the underlying causes, these findings illustrate the need to move beyond native listener norms in fluency assessments (similar to how the field has largely shifted away from native speaker norms), especially in light of English's status as a lingua franca [17]. By incorporating nonnative listener perspectives, we are one step closer to fairer and more inclusive evaluations that better reflect the diverse community of English users worldwide.

### 4.3. Revisiting and refining fluency measures

Another notable finding of this study is that segment-based articulation rate (AR) clearly outperformed syllable-based AR as a measure of speed fluency. In L2 fluency research, the de facto standard practice for calculating AR is to divide the number of syllables (often approximated by the number of vowels or other highly sonorant segments) by the total duration of speech segments. However, syllable-based AR may not accurately capture 'articulation rate' per se, as the total duration of speech segments is dependent on the complexity of the syllables [18, 19]. For example, a one-syllable word like "strength" /stɹɛŋkθ/ (CCVCVCC) naturally takes longer to produce than a simpler one-syllable word like "ten" /tɛn/ (CVC) due to the greater number of consonants.

This limitation with syllable-based AR becomes especially problematic when measuring speed fluency of L2 speech, as less proficient L2 learners often simplify complex syllable structures through vowel epenthesis (e.g., pronouncing "strength" as [sʊtɹɛŋgʊsu] (CVCVCVCCVCV)) or consonant deletion (e.g., pronouncing "strength" as [stɹɛŋ] (CCVC)) [20, 21]. Such simplifications reduce the duration of each syllable and artificially inflate AR values (although, in the case of consonant deletion, the magnitude of syllable shortening can be mitigated by compensatory lengthening of the vowel). As a result, syllable-based AR fails to reliably capture speed fluency, since slower speech in less proficient speakers (which should otherwise yield lower AR values) is obscured by the confounding effects of syllable simplification. Counting the number of segments provides a more direct and accurate measure of 'articulation rate.'

### 4.4. Limitations and future directions

Although the present findings offer valuable insights into how various listener groups perceive L2 fluency and its quantification, some limitations must be acknowledged.

First, the speech data in the current study were obtained from a reading task, which differs from spontaneous speech in several respects. For example, previous research has suggested that pause locations reflect distinct stages of speech production (where between-clause pauses correspond to conceptualization and within-clause pauses to formation), but this distinction was not addressed in the current study, as reading pre-written sentences likely involves minimal conceptualization. Furthermore, filled and unfilled pauses were analyzed together, since filled pauses (e.g., fillers) are infrequent in reading tasks. Similarly, repair strategies were restricted to word repetitions following stumbling or misreading, with no possibilities of false starts or reformulations typical of spontaneous speech. Future research should compare fluency ratings across read-aloud and spontaneous tasks (e.g., Tasks 6 and 8 of the J-AESOP corpus) to explore whether listeners rely on similar fluency measures under varying speaking conditions.

Second, the sample size of listeners and the distribution of their language backgrounds were limited. Further research with more listeners, and with greater diversity in L1 groups—particularly of less-represented languages in L2 English contexts such as African languages—is needed to extend these findings. Additionally, the validity of grouping listeners by geographic regions, as was done in this study, warrants further scrutiny. Alternative groupings could consider typological differences among listeners' L1s in terms of rhythm (e.g., mora-timed Japanese, syllable-timed Korean, or stress-timed German) and syllable complexity (e.g., simple Japanese, moderately complex French, or complex Polish), which may provide a more accurate explanation for inter-listener variability.

Finally, while Bayesian mixed-effects modeling proved effective for analyzing listener variability, further refinements in modeling techniques could yield deeper insights. Promising directions include: (1) incorporating more informative priors (e.g., drawn from previous studies on native listeners' judgments), (2) using the cumulative rather than the Gaussian family (since responses were discrete rather than continuous), and (3) examining the main effect of listener group (to better capture inter-group differences in overall leniency or strictness) as well as the interactions between listener groups and fluency measures. Also, the flexibility of Bayesian modeling allows for the inclusion of qualitative data, such as listener interviews, to shed light on the cognitive processes underlying fluency judgments.

## 5. Acknowledgements

This study was supported by JSPS JSPS Grant-in-Aid for Scientific Research (B) [23K20468] and JSPS Grant-in-Aid for Early-Career Scientists [22K13159].

## 6. References

- [1] S. Suzuki and J. Kormos, “Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech,” *Studies in Second Language Acquisition*, vol. 42, no. 1, pp. 143–167, 2020.
- [2] K. Saito, K. Macmillan, M. Kachlicka, T. Kunihara, and N. Mine-matsu, “Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies,” *Studies in Second Language Acquisition*, vol. 45, no. 1, pp. 234–263, 2023.
- [3] P. Foster, “Pauline Foster’s essential bookshelf: Oral fluency in a second language,” *Language Teaching*, pp. 1–9, 2023.
- [4] K. Yazawa, T. Konishi, and M. Kondo, “The J-AESOP corpus: Design, application, and future directions of a japanese-english bilingual speech corpus,” *Acoustical Science and Technology*, pp. 1–12, 2025.
- [5] P. Skehan, “Task-based instruction,” *Language Teaching*, vol. 36, no. 1, pp. 1–14, 2003.
- [6] S. Suzuki, J. Kormos, and T. Uchihara, “The relationship between utterance and perceived fluency: A meta-analysis of correlational studies,” *The Modern Language Journal*, vol. 105, no. 2, pp. 435–463, 2021.
- [7] K. Saito, M. Ilkan, V. Magne, M. N. Tran, and S. Suzuki, “Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency,” *Applied Psycholinguistics*, vol. 39, no. 3, pp. 593–617, 2018.
- [8] S. Suzuki and J. Kormos, “The moderating role of L2 proficiency in the predictive power of L1 fluency on L2 utterance fluency,” *Language Testing*, vol. 42, no. 1, pp. 73–99, 2024.
- [9] —, “The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency,” *Studies in Second Language Acquisition*, vol. 45, no. 1, pp. 38–64, Mar. 2023.
- [10] K. Hanzawa, “Development of second language speech fluency in foreign language classrooms: A longitudinal study,” *Language Teaching Research*, vol. 28, no. 3, pp. 816–838, May 2024.
- [11] M. Sato, “Exploring the construct of interactional oral fluency: Second Language Acquisition and Language Testing approaches,” *System*, vol. 45, pp. 79–91, 2014.
- [12] V. Magne, S. Suzuki, Y. Suzukida, M. Ilkan, M. Tran, and K. Saito, “Exploring the dynamic nature of second language listeners’ perceived fluency: A mixed-methods approach,” *TESOL Quarterly*, vol. 53, no. 4, pp. 1139–1150, 2019.
- [13] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to Use the International Phonetic Alphabet*. Cambridge: Cambridge University Press, 1999.
- [14] T. Konishi, “A corpus-based study on Japanese English rhythm,” Ph.D. dissertation, Waseda University, 2022.
- [15] P.-C. Bürkner, “brms: An R package for Bayesian multilevel models using Stan,” *Journal of Statistical Software*, vol. 80, pp. 1–28, 2017.
- [16] —, “Advanced Bayesian multilevel modeling with the R package brms,” *The R Journal*, vol. 10, no. 1, pp. 395–411, 2018.
- [17] J. Jenkins, *The Phonology of English as an International Language*. Cambridge: Cambridge University Press, 2000.
- [18] Y. Ozaki, K. Yazawa, and M. Kondo, “L2 English speech rhythm of Japanese speakers: An alternative implementation of the Varco metrics,” in *Proceedings of the Phonetics Teaching and Learning Conference 2017*, London, 2017, pp. 84–88.
- [19] K. Yazawa and M. Kondo, “A comparison of rhythm metrics for L2 speech,” in *Proceedings of the 11th International Conference on Speech Prosody*. International Speech Communication Association, 2022, pp. 332–336.
- [20] K. Yazawa, T. Konishi, K. Hanzawa, G. Short, and M. Kondo, “Vowel epenthesis in Japanese speakers’ L2 English,” in *Proceedings of the 18th International Congress of Phonetic Sciences*, The Scottish Consortium for ICPHS 2015, Ed. Glasgow: The University of Glasgow, 2015, pp. 969:1–4.
- [21] K. Yazawa, “Annotating second language learner speech: Reflecting on the development of the J-AESOP corpus,” *Journal of the Phonetic Society of Japan*, vol. 26, no. 3, pp. 111–123, 2022.