Mutual Shadowing among Speakers of World Englishes and Its Application to Investigating the Influence of Accents on Listening Fluency

Akari Fujiwara[1], *Nobuaki Minematsu*[1], *Noriko Nakanishi*[2], *Daisuke Saito*[1]

[1]Graduate School of Engineering, The University of Tokyo, Japan
[2]Faculty of Global Communication, Kobe Gakuin University, Japan

{fujiwara, mine}@gavo.t.u-tokyo.ac.jp, nakanisi@gc.kobegakuin.ac.jp, dsk_saito@gavo.t.u-tokyo.ac.jp

Abstract

English is used as a lingua franca with diverse pronunciations, known as World Englishes (WE), whose intelligibility varies depending on the listener. This study investigates how the accents of WE speakers influence listening fluency when heard by other WE speakers. We used data from 28 WE speakers (in three groups), who each read one of 28 passages, shadowed group recordings a week later, and then reread the same passage. On the collected data, we measured shadowing disfluency, interpreted as listening disfluency (LD). We also measured the phonetic gap (PhG) and prosodic gaps (PrGs) from each speaker's accent to three reference accents: General American, Received Pronunciation, and the listener's own English. Correlation and regression analyses revealed a general trend: while PhG has a greater influence on LD than PrGs, rhythmic deviations still increase LD. Further, the influences of the gaps to the three reference accents were found to depend on the listener's learning background.

Index Terms: World Englishes, listening disfluency, phonetic gap, prosodic gaps, correlation and regression analyses

1. Introduction

English has come to function as a lingua franca, leading to diversification in grammar, vocabulary, and pronunciation, and such diverse forms of English are known as World Englishes (WE) [1]. English pronunciation diversity is mainly due to language background diversity, and it is easy to assume that listening behaviors are also diverse due to language learning diversity. Then, how can we assess this listening diversity?

Previous research on WE listening diversity has often relied on subjective evaluations (e.g., surveys) to assess how listeners perceive different varieties of English [2, 3]. However, these studies have not employed objective and quantitative methods to measure listening diversity. To address this, a previous study [4] proposed a quantitative method using shadowing (a task in which listeners almost simultaneously repeat a presented speech while listening) [5]. In [4], each of 28 WE speakers shadowed all the others' recordings, and disfluencies measured in the shadowing speech were interpreted to reflect listening disfluency (LD). The study found LD to be correlated to some extent with the phonetic gap between the presented speech and one of the three reference accents: General American (GA), Received Pronunciation (RP), and the listener's own English (OE).

In addition to phonetic deviations, however, prosodic deviations such as pitch, intensity, and duration are also expected to contribute to LD[1]. For example, research on Japanese language education [6] demonstrates that lexical and phrasal prosody training using tools such as Online Japanese Accent Dictionary (OJAD) [7] improves learners' intelligibility effectively, underscoring the importance of prosody training. Given the relatively small inventory of vowels and consonants in Japanese, prosodic deviations are expected to have a larger impact on intelligibility than phonetic deviations[2]. While English has a richer phonemic inventory than Japanese, we hypothesize that phonetic deviations may contribute significantly to LD in WE with prosodic deviations still causing LD to some extent. In educational context, teachers and learners want to know which of phonetic training and prosodic training should be prioritized to reduce LD when the learners talk to others in English. In this study, we analyze the factors underlying LD by measuring phonetic gap (PhG) and prosodic gaps (PrGs) between the accent in the presented speech and the three reference accents of GA, RP, and OE.

2. Related research

2.1. Measuring intelligibility based on shadowing

In applied linguistics, several measures have been proposed to evaluate learner speech, including intelligibility, comprehensibility, and accentedness [8]. Among these, intelligibility is often defined as the percentage of correctly transcribed words by a listener (or rater) in a dictation task. However, several limitations of this approach have been pointed out:

1. Since writing or typing is generally slower than speaking, listeners must retain the content in memory while transcribing. Therefore, the presented speech has to be short enough.
2. Listeners may rephrase rather unconsciously what they heard during transcription, introducing subjectivity.
3. Transcription imposes additional cognitive demands, such as recalling orthography, that are not related at all to the actual listening process [9].

To address these issues, an alternative has been proposed using shadowing where listeners speak to replicate rather than write to replicate [5]. This method offers several advantages:

1. Since shadowing can be done at the same pace of listening, longer utterances can be presented to listeners or raters.
2. For the same reason above, we can measure objectively what is happening in the listeners' mind while they are listening.
3. Knowledge of orthography is not needed at all. Shadowing can be applied even to languages with no writing system.

When listeners fail to identify certain words in a given

[1]Naturally, the content of the presented speech can also affect intelligibility, but as described later, the semantic difficulty of the presented speeches was controlled in the experiments.

[2]The phonemes of Japanese are often a subset of the phonemes of learners' L1, and their phonetic deviations are generally less frequent and less salient to native listeners than their prosodic deviations.

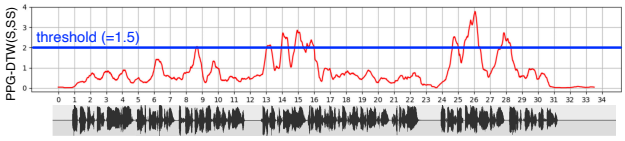

Figure 1: *An example of the listening disfluency (LD) curve drawn for a presented audio*

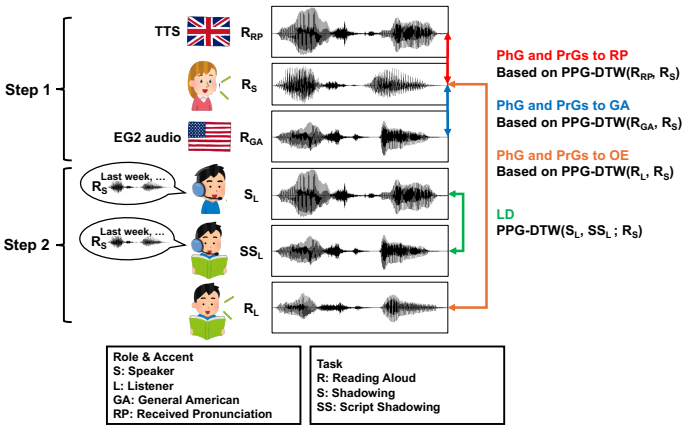

Figure 2: *Measuring LD and PhG using S, SS, and R*

speech, their shadowing speech typically becomes disfluent or disrupted on those words. Furthermore, when the same utterance is shadowed again but with its content (i.e., the script) visually presented, the resulting shadowing speech is generally fluent and coherent with no listening difficulty. This guided shadowing is called script-shadowing. By comparing these two types of shadowing speech, it is possible to find out where listening was disrupted in the presented speech and to draw the listening disfluency (LD) curve [5], as illustrated in Figure 1. The process for constructing this curve is described in the following section.

2.2. Quantifying LD and PhG in sequence

To measure LD in sequence between shadowing (S) and script-shadowing (SS), we convert both into Phonetic PosteriorGrams (PPGs) and align them using Dynamic Time Warping (DTW). PPG-DTW(S, SS) quantifies LD, which can also draw the LD curve for the presented audio [10]. Prior work [11] showed that, if word-based mean LD is larger than 1.5, we can judge that listening difficulty was large enough for that word (Figure 1). A part of Figure 2 illustrates this process, where speaker S's read-aloud input ($\mathrm{R_S}$) is shadowed by listener L, and his shadowing speeches are $\mathrm{S_L}$ and $\mathrm{SS_L}$. PPG-DTW($\mathrm{S_L}$, $\mathrm{SS_L}$) quantifies LD.

After SS, L reads the script aloud, producing $\mathrm{R_L}$ in the figure. PPG-DTW($\mathrm{R_S}$, $\mathrm{R_L}$) quantifies PhG in sequence between S and L for that script.

2.3. Comparison of prosodic control between two speakers

[12] investigated the PrGs between speech produced by a model speaker (model utterance) and that produced by learners (learner utterance), focusing specifically on prosodic control. As a preprocessing step, PPG-DTW was applied to align the temporal axes of the two utterances. Then, forced alignment (FA) was performed on the model utterance to detect its phoneme boundaries. Using the detected boundaries and the DTW alignment, the boundaries were projected onto the learner utterance as well. Following this, the degree of prosodic similarity was quantified between the two utterances by calculating correlation coefficients between the two in terms of pitch, intensity, and duration control, separately.

In the present study, we adopt a similar approach to analyze the factors that influence LD measured as PPG-DTW($\mathrm{S_L}$, $\mathrm{SS_L}$). For the four types of speech, $\mathrm{R_S}$, $\mathrm{S_L}$, $\mathrm{SS_L}$, and $\mathrm{R_L}$ as well as two other samples of read-aloud of the script with GA and that with RP, indicated as $\mathrm{R_{GA}}$ and $\mathrm{R_{RP}}$ in Figure 2, we examine both PhG and PrGs as potential contributors to LD. Unlike [12], where analysis focused on vowel intervals, our analysis of pitch and intensity control focuses on voiced regions instead.

Table 1: *L1 distribution for each participant group*

Group A	Group B	Group C
CHN1	CHN4	CHN7
CHN3	CHN6	CHN8
JPN1	JPN2	JPN5
KOR1	KOR3	KOR5
FRA2	FRA3	FRA5
ITA1	ITA2	SPN1
HIN1	HIN2	HIN5
SRB1	UKR1	UKR2
HUN1	MAL1	HUN2
KOR4		

ID	languages
CHN	Chinese
JPN	Japanese
KOR	Korean
FRA	French
ITA	Italian
SPN	Spanish
HIN	Hindi
SRB	Serbian
UKR	Ukrainian
HUN	Hungarian
MAL	Malayalam

3. Experiments and results

3.1. Collection of reading-alouds and shadowings

All speech data used in this study were collected in [4]. Below is a brief summary of the data collection procedure. The participants were 28 university students. They were non-native English speakers who demonstrated sufficient ability to shadow GA or RP speech smoothly. Taking the diversity of their L1 into account, the 28 participants were divided into three groups (A to C) to enlarge L1 diversity within each group, as shown in Table 1 along with their L1[3].

The data collection consisted of two steps (see Figure 2). In Step 1, $\mathrm{R_S}$ audio files were prepared. Since LD is affected partially by the linguistic content of the presented audio, to control for semantic difficulty, 28 passages were extracted from the listening sections of Eiken Grade-2 Tests (EKG2) [13]. EKG2 is a standardized English proficiency test designed for high school students in Japan. Each participant read one of the 28 passages. The Automated Readability Index (ARI) [14] of all the passages varied from 6.2 to 7.0, indicating that they were easy enough for university students to read. We used the EKG2 GA recordings for those passages as $\mathrm{R_{GA}}$ in Figure 2, which were converted to $\mathrm{R_{RP}}$. The conversion was performed by a high-performance commercial TTS converter [15]. As a result, for each script, we prepared three versions: the participants' own English (OE), GA, and RP. Each audio was approximately 30 seconds long.

In Step 2, each participant listened to all $\mathrm{R_S}$ recordings but his/her own within the group, and performed shadowing ($\mathrm{S_L}$ and $\mathrm{SS_L}$) as well as reading-aloud ($\mathrm{R_L}$) for each of $\mathrm{R_S}$. Namely, each participant served as both a "reader" who read aloud every script ($\mathrm{R_S}$ and $\mathrm{R_L}$) and a "listener" who shadowed others' speech ($\mathrm{S_L}$ and $\mathrm{SS_L}$).

Hereafter, we refer to the four measured indices, PhG and

[3]For the classification of languages into language families and subfamilies, see [4].

Table 2: *Correlation between LD and acoustic gaps*

Reference accent = GA				
Group	PhG	Pitch	Intensity	Duration
A	0.50	−0.36	0.14	−0.21
B	0.32	−0.20	−0.03	−0.23
C	0.38	0.14	−0.41	−0.17
Reference accent = RP				
Group	PhG	Pitch	Intensity	Duration
A	0.44	−0.19	0.07	−0.07
B	0.22	0.18	−0.42	−0.09
C	0.36	0.04	−0.34	−0.14
Reference accent = OE				
Group	PhG	Pitch	Intensity	Duration
A	0.55	−0.17	−0.17	−0.24
B	0.38	−0.01	−0.12	−0.32
C	0.35	0.01	0.07	−0.22

PrGs (pitch, intensity, and duration), collectively as "acoustic gaps". While the previous sections described methods for quantifying the acoustic gaps between the presented speech and OE, similar methods were also used to quantify the acoustic gaps between the presented speech and GA and RP.

3.2. Correlation analysis between LD and acoustic gaps

In the perception of WE, the ease or difficulty of listening depends on each listener's linguistic background and learning history [16]. Therefore, this study focuses on how PhG and PrGs between the presented speech and GA, RP, and OE influence LD. When the phonetic or prosodic features of the presented speech differs significantly from that of GA, RP, and OE, it is expected that the speech will be less familiar and maybe more difficult to understand, thus leading to larger LD. In other words, a positive correlation is hypothesized between LD and PhG, while negative correlations are expected between LD and PrGs.

When a participant in a group listened to 10 $\mathrm{R_S}$ samples (See Table 1), it provides 10 mean LDs and their corresponding mean PhGs and PrGs. Since we have about 10 participants in each group, correlations between the LDs and the PhGs and those between the LDs and the PrGs were calculated out of about 100 data points, and the results are shown in Table 2 for each case of the three reference accents.

For PhG, moderate positive correlations are always found irrespective of the reference accents and the groups. On the other hand for PrGs, only weak negative correlations are found, depending on the accents, the groups, and the prosodic features. In GA, weak correlations are found in A and B for pitch and duration. In RP, they are found in B and C only for intensity and in OE, they are found in all the groups but only for duration. Although it was difficult to discuss why these dependencies were observed, we focused on the group-independent correlations found only in the case of OE for duration control.

Direct comparison of the three prosodic features between any two WE speech samples of the same content ($\mathrm{R_S}$ and $\mathrm{R_L}$) was performed. Mean correlations were 0.37, 0.59 and 0.86 for pitch, intensity, and duration, respectively. This indicates that duration control tends to be more consistent across speakers, and when duration control in the presented speech differs from the listener's own control, it tends to increase LD.

Table 3: *Results for participants with salient characteristics*

Listener	accent	R^2	Normalized Weights			
			PhG	Pitch	Intensity	Duration
CHN3	GA	0.44	**0.88**	0.00	0.12	0.00
SRB1	OE	0.71	**0.42**	0.27	0.11	0.20
ITA2	RP	0.95	0.42	0.01	**0.48**	0.09
SPN1	RP	0.66	0.25	0.33	**0.36**	0.07
HIN1	GA	0.90	0.36	0.14	0.12	**0.37**
CHN6	OE	0.91	0.033	0.32	0.16	**0.50**
HUN2	$*$	0.00	0.0	0.0	0.0	0.0
JPN2	$*$	0.00	0.0	0.0	0.0	0.0

$*$: any accent

3.3. Analysis of acoustic gaps contributing to LD

The above section examined the overall trends across the listener groups in how the acoustic gaps between the presented speech and GA, PR, and OE contribute to LD. In Table 1, however, the participants speak various native languages, and it is reasonable to assume that the dependency of the LD on the PhG, the PrGs, and the reference accents varies across the participants.

For example, in [4], listening behaviors were found to be diverse across the participants. It is interesting that the authors of [4] found three "super listeners" who are non-native speakers of English but can understand WE speech samples regardless of the accents actually observed in the presented speech. These results highlight the importance of participant-based analysis.

To address this, we formulated a regression problem to predict the LD of each participant using the acoustic gaps. By selecting an adequate model for regression, we examine for each participant which features and accent are more influential to predict his/her LD. For this aim, we employed Elastic Net regression [17]. When both dependent and independent variables are standardized, Elastic Net naturally performs feature selection by assigning zero weights to irrelevant independent variables, allowing us to identify more influential acoustic features for each participant. We will run Elastic Net regression separately for each participant and each reference accent. Comparison among the three accents will be done by comparing the coefficient of determination (R^2) among the three accents.

Some learners may have limited experience in speaking English, resulting in less exposure to their own accents. In this case, they may be more accustomed to textbook-like accents such as GA and RP, which will show larger R^2.

3.3.1. Participants with salient characteristics

In examining overall trends based on the participant-based analysis, we identified several learners with distinctive listening abilities. We therefore highlight several ones exhibiting salient characteristics in this section. Table 3 shows the selected accent, its R^2, and the normalized weights assigned to the four independent variables of PhG, pitch, intensity and duration. Here, the largest ratios are shown in bold. If R^2 is zero, meaning that none of the four features are relevant to the dependent variables, we instead show the absolute value of the weights (e.g., for HUN2 and JPN2).

• *Listeners highly affected by PhG*

As suggested by Table 2, many listeners show higher dependency on PhG. Among them, SRB1 stands out, with a high R^2 and a notably greater influence of PhG compared to PrGs. This indicates that SRB1's LD is primarily affected by PhG between the accent of the presented speech and his/her own accent.

Table 4: *The accent with the highest R^2*

GA	RP	OE
CHN3, CHN8	CHN7, JPN1	CHN1, CHN4
KOR1, HIN1	MAL1, ITA1	CHN6, JPN5
HIN2, UKR2	ITA2, UKR1	HIN5, HUN1
FRA2, FRA5	SPN1	SRB1, FRA3

Table 5: *Model accents adopted in schools*

GA	RP	GA+RP
CHN1, CHN4	CHN3, HIN5	CHN8
CHN6, CHN7	MAL1, HUN2	HIN2
JPN1, JPN2	SRB1, UKR2	UKR1
JPN5, KOR1	ITA2, FRA2	HUN1
KOR3, KOR4	FRA3, SPN1	ITA1
KOR5, HIN1		
FRA5		

SRB1 was identified as one of the three "super listeners" in [4], meaning that s/he showed small enough LD for almost all the other participants. Despite this, the results in Table 3 suggest that even super listeners exhibit dependency of LD on acoustic gaps between the presented speech and his/her own accent.

• *Listeners highly affected by PrGs*

Some participants showed greater influence of PrGs on their LD. For example, ITA2 and SPN1 were most affected by the acoustic gaps in intensity control, while HIN1 and CHN6 were most affected by the gap in duration control.

• *Listeners unaffected by any acoustic gaps*

For some participants, their regression models assigned zero weights to all the independent variables, irrespective of the reference accents. This means that all the four features were completely useless for regression, and R^2 has to be 0.0. These participants fell into one of two distinct groups.

One group represents what we might call "true" super listeners, who appear to understand a given speech regardless of PhG and PrGs between the speech and GA, RP, and OE. HUN2 is one of the three super listeners in [4], who showed his/her R^2 to be 0.0. This implies that HUN2 may be a true super listener.

The other group consisted of "listeners with low proficiency" who showed high LD across all input speeches, regardless of their PhG and PrGs. JPN2, for instance, consistently exceeded the LD threshold of 1.5, even when listening to his/her own English. For this participant, we did an additional analysis where his/her repeated recordings of the same passage showed a large PhG between them, indicating difficulty of producing stable pronunciations. Similarly, ITA1 and UKR2 had their R^2 of 0.0 when using their OE as reference accent, indicating that the acoustic gaps between the presented audio and their OE were not informative for modeling their LD.

3.3.2. The accent with the highest R^2

For each participant, we examined which regression model among the three reference accents yielded the highest R^2. Table 4 shows the selected accent for each participant in different colors. The colors indicate which accent was adopted as model accent in the participants' school days, shown in Table 5. With this table, we can find in Table 4 that, in the GA group, 5 out of 8 heard GA as model accent in their school days, and in the RP group, 5 out of 7 heard RP as model accent. It is reasonable to consider the accent adopted in schools as familiar accent, and we can say that LD of the participants tend to be characterized better with the acoustic gaps to their familiar accent.

As for the OE group in Table 4, 2 out of 3 super listeners and 3 out of 6 Chinese participants were found in this group. HUN1 and SRB1 were identified as super listeners in [4], implying that they have frequent chances of speaking in English and listening to WE. Therefore, we can say that their pronunciations have been well established, and their LD can be characterized better with the acoustic gaps to their own accent.

As for the three Chinese participants (CHN1, CHN4, and CHN6), we cannot claim any strong reason, but we can point out unique speech training conventions in China, where reading aloud practices were conducted intensively and extensively [18]. Indeed, [19] reports that Chinese learners engage in oral reading practice significantly more frequently than Japanese learners. These conventions may help Chinese learners establish their own pronunciations well.

In summary, the English accent that can characterize a listener's LD well depends on the learning profile of that listener.

4. Conclusions and future work

In this study, we quantitatively measured LD, PhG, and PrGs between each speaker's accent and three reference accents: GA, RP, OE. We analyzed the relationship between LD and these acoustic gaps, finding that PhG had a greater impact on LD than PrGs, although duration control also contributed. These findings suggest that phonetic training is more likely to be effective than prosodic training in reducing LD for many learners. However, the contribution of duration control also indicates that prosodic features should not be entirely neglected, depending on each learner's specific needs. Listeners' reliance on each reference accent varied depending on listener's learning profile.

Future work will proceed in two directions. First, since the Elastic Net regression used in this study assumes feature independence, it cannot capture interactions between pronunciation and prosody. Given their interdependence, we plan to apply factor analysis to better model these interactions. Second, PhG and PrGs were quantified using different methods: PhG was measured using symbolized PPG sequences, while PrGs relied on raw acoustic correlations. This inconsistency may have affected the comparison. To enable more consistent and interpretable metrics, we plan to explore symbolized representations for PrGs based on prosodic expectations.

5. References

[1] D. M. Eberhard, G. F. Simons, and C. D. Fennig, "Ethnologue: Languages of the World," Dallas, Texas, 2025, online version. [Online]. Available: http://www.ethnologue.com

[2] H. Jeong, A. Elgemark, and B. Thorén, "Swedish Youths as Listeners of Global Englishes Speakers With Diverse Accents: Listener Intelligibility, Listener Comprehensibility, Accentedness Perception, and Accentedness Acceptance," *Frontiers in Education*, vol. 6, 06 2021.

[3] G. Verbeke and E. Simon, "Listening to accents: Comprehensibility, accentedness and intelligibility of native and non-native English speech," *Lingua*, vol. 292, p. 103572, 2023.

[4] Y. Tomita, Y. Gao, N. Minematsu, N. Nakanishi, and D. Saito, "Analysis and Visualization of Directional Diversity in Listening Fluency of World Englishes Speakers in the Framework of Mutual Shadowing," in *Interspeech 2024*, 2024, pp. 4024–4028.

[5] C. Zhu, R. Hakoda, D. Saito, N. Minematsu, N. Nakanishi, and T. Nishimura, "Multi-Granularity Annotation of Instantaneous Intelligibility of Learners' Utterances Based on Shadowing Techniques," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1071–1078.

[6] H. Date, N. Nakamura, and N. Minematsu, "Evaluation of the Prosodic Naturalness of Japanese Learners' Utterances after Practicing with OJAD Suzuki-kun," *Journal of the Phonetic Society of Japan*, vol. 23, pp. 6–21, 2019.

[7] "OJAD." [Online]. Available: https://www.gavo.t.u-tokyo.ac.jp/ojad/

[8] M. J. Munro and T. M. Derwing, "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners," *Language Learning*, vol. 49, no. s1, pp. 285–310, 01 1999.

[9] C. Zhu, T. Kunihara, D. Saito, N. Minematsu, and N. Nakanishi, "Automatic Prediction of Intelligibility of Words and Phonemes Produced Orally by Japanese Learners of English," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1029–1036.

[10] T. Kunihara, C. Zhu, D. Saito, N. Minematsu, and N. Nakanishi, "Detection of Learners' Listening Breakdown with Oral Dictation and Its Use to Model Listening Skill Improvement Exclusively Through Shadowing," in *Interspeech 2022*, 2022, pp. 4461–4465.

[11] J. Choi, L. Zhand, Y. Gao, N. Minematsu, D. Saito, and N. Nakanishi, "Shadowing-based subjective annotation of semantic listening disfluency measured while listening to L2 speech," *Proc. Speech Research Meeting, The Acoustical Society of Japan*, vol. 4, no. 2, pp. SC–2024–18, 2024.

[12] C. Shoda, Y. Gao, Y. He, N. Minematsu, N. Nakanishi, and D. Saito, "Learners' Prosodic Control in the Task of Expressive Storytelling and Predicted Native Listeners' Impressions of the Learners' Speech," in *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 46–50.

[13] "EIKEN English Proficiency Test." [Online]. Available: https://www.eiken.or.jp/eiken/

[14] R. Senter and E. Smith, "Automated Readability Index," *AMRLTR*, vol. 106, no. 2, p. 1–14, 1967.

[15] "ElevenLabs." [Online]. Available: https://elevenlabs.io/

[16] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[17] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 03 2005.

[18] M. J. Chen, G. J. Yin, H. S. Goh, R. S. Soo, R. N. S. R. Harun, C. K. S. Singh, and W. L. Wong, "Theoretical Review of Phonics Instruction among EFL Beginner-level Readers in China," *International Journal of Academic Research in Progressive Education and Development*, vol. 11, no. 2, pp. 449–466, 2022.

[19] X. Teng, "The Lack of Japanese Students' Oral Reading Practice in Studying English," *Journal of Language and Culture of Hokkaido*, vol. 12, pp. 73–83, 2014.