Constructing Korean Learners' L2 Speech Corpus of Seven Languages for Automatic Pronunciation Assessment

Seunghee Han[1], **Minhwa Chung**[2], **Sunhee Kim**[3]
[1]Learning Sciences Research Institute, [2]Department of Linguistics,
[3]Department of French Language Education
Seoul National University, Republic of Korea
seunghee.hahn@gmail.com, {mchung, sunhkim}@snu.ac.kr

Abstract

Multilingual L2 speech corpora for developing automatic speech assessment are currently available, but they lack comprehensive annotations of L2 speech from non-native speakers of various languages. This study introduces the methodology of designing a Korean learners' L2 speech corpus of seven languages: English, Japanese, Chinese, French, German, Spanish, and Russian. We describe the development of reading scripts, reading tasks, scoring criteria, and expert evaluation methods in detail. Our corpus contains 1,200 hours of L2 speech data from Korean learners (400 hours for English, 200 hours each for Japanese and Chinese, and 100 hours each for French, German, Spanish, and Russian). The corpus is annotated with spelling and pronunciation transcription, expert pronunciation assessment scores (accuracy of pronunciation and fluency of prosody), and metadata such as gender, age, self-reported language proficiency, and pronunciation error types. We also propose a practical verification method and a reliability threshold to ensure the reliability and objectivity of large-scale subjective evaluation data.

Keywords: non-native L2 corpus, automatic pronunciation assessment, computer-assisted language learning, resources and evaluation, corpus construction and annotation

1. Introduction

Rapid technological advancements, particularly in storing and processing extensive, unstructured speech data, have ushered in a new era for ASR (automatic speech recognition) systems. These developments have significantly bolstered the capabilities of such systems, making them invaluable tools for various applications, including computer-assisted language learning (CALL). Of particular interest in CALL is the development of automatic evaluation systems for assessing non-native foreign language speech, leveraging the capabilities of speech recog nition technology (Shi et al., 2020).

As highlighted by Kannan and Munday (2018), research has unequivocally demonstrated the effectiveness of AI-based learning programs in delivering personalized education tailored to individual learners (Junying et al.; Liu et al., 2018). These programs have effectively transcended the limitations of traditional classroom settings by providing feedback rooted in specific learning outcomes. Substantiating this notion, studies conducted by Kim et al. (2019) and Lee (2019) have underscored the positive impact of AI chatbots on learners' proficiency in foreign language speaking. Meanwhile, Dodigovic (2007) has stressed their effectiveness in rectifying grammar errors.

In 2021, the South Korean Ministry of Education introduced AI PengTalk, an English-speaking practice system, to elementary schools nationwide, further highlighting the potential of AI-driven language learning tools. Extensive research findings have indicated that AI PengTalk has played a pivotal role in enhancing the English proficiency of elementary students and fostering self-directed learning (Seong and Lee, 2021). However, concerns have arisen among users, particularly regarding the system's limitations in the realm of speech recognition. This limitation stems from the observation that, while speech recognition technology exhibits high accuracy when processing native-speaker speech, its performance significantly declines when confronted with non-native-speaker speech due to pronunciation errors influenced by the speaker's native language (Markl and Lai, 2021).

Within the context of foreign language education, it becomes imperative for instructors to provide detailed feedback on learners' L2 speech, especially where incorrect pronunciations or deviations from native speaker speech patterns are evident. AI-based automatic pronunciation assessment systems have the potential to fill this need. However, these systems currently face challenges in accurately assessing such deviations. A primary reason for this shortfall is the lack of sufficiently labeled data to train models for pronunciation assessment. This data deficiency hinders the ability to distinguish various error types and evaluate them within speech samples.

The construction of extensive assessment data mandates the preservation of consistency and objectivity throughout the assessment process. This necessitates developing intricate rubric designs and implementing assessor training to minimize subjectivity discrepancies among assessors. Also,

LREC-COLING 2024, pages 3772–3781
20-25 May, 2024.

a novel approach is indispensable for validating the reliability of large-scale assessment data annotated by multiple experts, distinguishing it from the conventional validation methods employed in foreign language education, which typically rely on only a limited number of expert assessors.

Given the crucial role of large-scale assessment data in training artificial intelligence systems, this article delves into the intricate challenges involved in its generation, particularly in the context of Korean learners' L2 pronunciation assessment. Addressing these essential aspects, this research aims to pave the way for more accurate, reliable, and effective automatic pronunciation assessment systems, ultimately contributing to the fields of phonetics, foreign language education, speech recognition, and the advancement of AI-driven language learning tools.

2. Background

2.1. L2 Pronunciation Corpora

While multilingual speech datasets such as GlobalPhone (Schultz and Schlippe, 2014) and Librispeech (Panayotov et al., 2015) exist, they predominantly feature recordings from native speakers across various languages. Numerous L2 corpora are available, particularly for speech recognition and pronunciation assessment; however, these are largely centered around English speech from non-native speakers, as seen in collections like ACCENT, the Speech Accent Archive, and L2 Arctic. Unique compilations like the MANY corpus, encompassing child speech across 40 languages, and the International Corpus of Learner English (ICLE), providing high-level L2 written samples, are available, but there is a glaring absence of multilingual speech corpora from adult L2 speakers at intermediate to low proficiency levels. This gap signifies an urgent need for comprehensive datasets that include this underrepresented demographic, pivotal for enhancing research and technological innovation in language acquisition and speech processing.

Introduced in 2021, Speechocean762 (Zhang et al., 2021) emerges as a significant stride in this domain. This globally accessible dataset, tailored for automatic pronunciation assessment in computer-assisted language learning (CALL), encompasses 5,000 English utterances from 250 non-native Mandarin speakers, balancing both adult and child speech samples. It stands out for its detailed scoring system, which ranges from phonetic accuracy within words (0 to 2) to comprehensive sentence-level assessments, including pronunciation accuracy, fluency, prosody (0 to 10), utterance completeness (0 to 1), and stress accuracy. Notably, Speechocean762 offers both average and median scores, consolidating evaluations from a panel of five experts.

2.2. Opportunities in L2 Pronunciation Studies

Despite its strengths, Speechocean762's scope is limited, primarily focusing on English with an average contribution of 20 sentences per speaker. A more expansive and diverse dataset, capturing a wider spectrum of speakers and linguistic variations, would be a game-changer, enabling models to derive richer insights from an extensive range of pronunciation patterns and errors. Furthermore, while Speechocean762's multi-tiered evaluation of speech quality is impressive, there lies an opportunity to broaden assessment parameters to "across-sentences" evaluations, yielding a more holistic understanding of pronunciation proficiency and diverse error manifestations.

The complexity of Speechocean762's scoring matrix, while comprehensive, introduces challenges in maintaining uniformity across different assessment criteria and score ranges, a demanding task even for experts. Its smaller scale permits a team of five experts to undertake in-depth evaluations, ensuring objectivity but also underscoring logistical hurdles when scaling pronunciation assessment datasets for AI training. Balancing transcription accuracy with evaluation consistency is critical when dealing with extensive pronunciation data. The operational difficulty in assembling a larger panel of expert phoneticians raises concerns about sustaining accuracy and uniformity in assessments. Thus, the development of a versatile, multilingual error classification system that simplifies pronunciation error categorization and aligns them with actual instances is imperative. Perfecting this system could significantly propel advancements in related AI technologies.

2.3. Key Considerations for Pronunciation Assessment

The TOEFL IBT Speaking Test, developed by ETS, reflects a rigorous effort to ensure consistent scoring among assessors—an essential aspect of language assessment. Their evaluation rubrics, documented extensively in research and reports, provide comprehensive guidelines for assessing various dimensions such as holistic scoring, delivery, language proficiency, and topic development. Each scoring criterion ranges from 0 to 5 points. The "delivery" category within these rubrics, which covers fluency, intonation, rhythm, and pronunciation, offers valuable insights for designing criteria for automatic pronunciation assessments.

However, delving deeper into ETS's "delivery" rubrics reveals that scores incorporate factors like clarity, intelligibility, fluency, intonation, pace, and completeness of pronunciation. This cumulative approach can introduce ambiguity in the scoring process.

In contrast, Speechocean762's pronunciation evaluation adopts a more structured approach. Their criteria focus on accuracy (emphasizing phonemic awareness), fluency (addressing coherence, pauses, and repetitions), and prosody (analyzing intonation, speed, and rhythm). Yet, the divisions into 2, 4, or 5 score ranges for each criterion and variations within these ranges can also render the evaluation ambiguous.

Designing an effective hierarchical classification system is crucial. However, the challenge lies in overcoming inherent subjectivity in pronunciation assessment, as highlighted by numerous research studies. Notably, while many automatic pronunciation assessments utilize accuracy, fluency, and speech completeness as evaluation criteria, the intertwining of fluency and speech completeness can complicate subjective assessments, affecting consistency.

A thorough examination of past research underscores the need for well-defined evaluation criteria and rubrics. The development of these tools should prioritize ensuring reliability and consistency in subjective assessments.

3. Corpus Methodology

3.1. Task Design

3.1.1. Script Composition

To capture the nuanced errors of non-native L2 speech, which can vary significantly from single words to complex interactions within and between sentences, our study expands beyond the conventional focus on individual words and short phrases. This comprehensive approach is particularly pertinent given the diverse accuracy and fluency levels exhibited by foreign language learners in extended readings.

Reading Tasks Design: We instituted two distinct reading tasks: single sentence readings and paragraph readings. The scripts, meticulously prepared by our team, integrated words identified as frequent pronunciation challenges for Korean L2 learners across seven languages.

Difficulty Categorization: Employing the CEFR standards as a guide, we classified vocabulary into three tiers of difficulty — High (B2 - C2), Medium (A2 - B1), and Low (Pre-A1 - A1) — to diversify the learning curve and closely mimic real-life language use scenarios.

Balanced Distribution: In the interest of fairness and consistency, each participant encountered an equal mix of these difficulty levels. We crafted our reading tasks, which included 10 to 15-word sentences and 5 to 6-sentence paragraphs, to secure a minimum engagement time of 25 minutes per participant, aiming to elicit a wide range of pronunciation errors.

Expert-Driven Script Development: Our scripts were crafted through a collaborative effort between Korean professors and native-speaking professors from abroad, all of whom specialize in the relevant foreign languages and currently hold teaching positions at universities in Korea. Each contributor brought their unique expertise to the project. Initial drafts, created by Korean professors, were subsequently refined by their foreign native-speaking counterparts to ensure the linguistic authenticity of the scripts and the inclusion of words or phrases that are commonly mispronounced.

Pilot Testing and Feedback Integration: Before full-scale script production, we created a prototype containing 50 questions per language, which underwent a rigorous review by phonetics and foreign language instruction experts. Feedback from these experts was instrumental in identifying key areas of focus for script development. Subsequently, assessors conducted trial recordings with these scripts, extracting critical insights that were relayed back to the scriptwriters. This feedback loop ensured that the final scripts were not only academically sound but also aligned with real-world language applications and pedagogical standards.

3.1.2. Topic Selection

This corpus is intricately designed to bolster artificial intelligence capabilities in linguistic applications, acknowledging the indispensability of exhaustive data. Our pivotal goal was harvesting eclectic spoken samples, reflecting an extensive array of topics and practical scenarios.

Our strategy amalgamated globally accepted standards with domestic acumen. We harmonized the CEFR (Common European Framework of Reference for Languages) guidelines with the taxonomy advocated by the Korea Institute of Curriculum and Evaluation, thereby devising a robust framework for our thematic matrix.

This matrix included broad categories: Personal, Public, Occupational, and Educational realms. For nuanced insights and depth, we segmented these primary domains into detailed subcategories, such as Location, Institution, Persons, Objects, Events, Operations, and Texts. Our corpus stands out due to its dependency on pre-existing linguistic repos-

	Pronunciation Accuracy	Prosodic Fluency
5	No errors or awkwardness of segmental phonemes in the speech. Easy to understand.	Natural stress, rhythm and intonation. The speaking rate is moderate, and the number and duration of pauses are natural. There are few speech mistakes, and the pauses are appropriately used to separate units of speech.
4	A few errors or awkwardness of segmental phonemes in the speech. But intelligibility is not significantly affected.	Slightly awkward stress, rhythm and intonation. The speaking rate is mostly consistent, with some hesitations and breaks. The pauses are appropriately used to separate units of speech, but their number and duration are slightly awkward.
3	Some errors or awkwardness of segmental phonemes in the speech. Intelligibility is somewhat affected due to certain consistent errors.	Somewhat awkward stress, rhythm and intonation. The speaking rate is inconsistent and a bit slow, with frequent breaks. The pauses are not appropriately used to separate units of speech.
2	Frequent errors or awkwardness of segmental phonemes in the speech. Intelligibility is only achieved when the listener pays attention to the speaker's intonation due to some persistent pronunciation errors.	Considerably awkward stress, rhythm and intonation. The speaking rate is slow, with many breaks. The pauses last long and do not appropriately separate units of speech.
1	The speech lacks clarity of segmental phonemes, with too many errors and awkwardness. Hard to understand.	Terrible stress, rhythm and intonation. The speaking rate is too slow, with too many breaks. The pauses last too long and do not serve to separate units of speech at all.

Table 1: Scoring Rubrics

itories. We enriched our script creation process by incorporating a diverse array of thematic keywords extracted from multiple corpora, courtesy of Korea's National Information Society Agency (NIA). This methodology not only widened the thematic scope but also amplified the representational accuracy of our samples.

3.2. Assessment Criteria Development

With a focus on enhancing uniformity and objectivity in subjective evaluations, we streamlined our approach by synthesizing evaluation metrics from existing research. This synthesis birthed two primary categories: "pronunciation accuracy" and "prosodic fluency." The former concentrates on the meticulous assessment of individual speech segments, while the latter provides an all-encompassing evaluation of various speech attributes, including stress, intonation, rhythm, and speech rate.

Table 1 delineates the refined rubrics, an amalgamation of evaluative elements from sources like ETS and Speechocean762, supplemented with insights from phonetic research and studies centered on automated pronunciation evaluations. Our enhanced scoring rubrics define "accuracy" in terms of speech segment clarity, and "prosodic fluency" consolidates traditional fluency parameters with prosodic components. We also recognized the opportunity to evaluate speech "completeness" based on non-native speakers' speech-to-text (STT) recognition rate in L2 utterances. However, to reinforce consistency in evaluations, "completeness" was purposefully excluded from the primary evaluative benchmarks, following evidence suggesting its scoring is substantially influenced by factors like accuracy and fluency, necessitating its separate analysis.

4. Data Collection

4.1. Participant Recruitment

Our corpus diverges from previous studies deliberately encompassing a broad demographic spectrum, including various age groups, genders, and language proficiency levels, thereby enhancing the data's robustness for training speech recognition systems.

Adult Participants and Language Learning: Predominantly, our participants are adults, reflecting the trend in Korea where intensive study of foreign languages, particularly languages other than English, often commences at the university level. This focus is based on observable educational trends, underscoring the importance of foreign language acquisition in higher education.

Inclusion of Younger Demographics: We also recognize the critical role of English proficiency from an early age in educational trajectories. Thus, our corpus incorporates speech samples from middle and high school students, maintaining a 1:1

			English	Japanese	Chinese	German	Spanish	French	Russian
Demographics	Speakers		882	677	489	264	287	213	229
	Age	10s	43.20%	N/A	N/A	N/A	N/A	N/A	N/A
		20s	35.83%	69.42%	71.37%	81.06%	84.67%	79.34%	82.97%
		30s	14.51%	21.42%	20.65%	15.15%	12.54%	16.90%	14.41%
		+40s	6.46%	9.16%	7.98%	3.79%	2.79%	3.76%	2.62%
	Gender	M	36.05%	22.01%	18.40%	17.42%	21.25%	15.49%	26.20%
		F	63.95%	77.99%	81.60%	82.58%	78.75%	84.51%	73.80%
		Overlap Rate	75.14%	54.07%	49.04%	47.69%	53.00%	45.07%	60.08%
	Proficiency	H	20%	15%	15%	15%	15%	15%	15%
		M	30%	20%	20%	20%	20%	20%	20%
		L	50%	65%	65%	65%	65%	65%	65%
Speech Characteristics	Duration (h)		400	200	200	100	100	100	100
	Samples		114,494	63,678	88,712	34,596	37,437	28,003	30,223
	Avg. Tokens/Characters*		26.87	63.06*	35.59*	17.41	20.31	26.37	17.83
	Duration/Speaker (h)		0.48	0.30	0.42	0.38	0.43	0.51	0.46
Assessment Panel	Assessors		28	10	10	11	9	8	8
	Groups of Two		14	5	6	5	5	5	4
	Samples/Group		8,178	17,742	10,613	6,919	7,487	5,601	7,556
Recording Standards	Format (WAV), Bit rate (256kb/s), Channels (1 channel), Sampling rate (16kHz), Bit depth (16 bits)								

* Peak and RMS levels are included as metadata in the corpus.

Table 2: Demographics and Speech Data Overview Across Languages

ratio with adult samples. This ensures diverse representation across educational stages, enriching the corpus with variations in linguistic development.

Participant Proficiency Levels: We primarily recruited participants from domestic universities, notably those majoring in relevant foreign languages, ensuring a homogeneous sample group. We stratified language proficiency into three categories: beginner, intermediate, and advanced. Whenever possible, we aligned proficiency designations with the Common European Framework of Reference for Languages (CEFR) standards. For university students, we correlated linguistic competencies with their academic progression, assigning years 1-2 as beginners, 3-4 as intermediates, and postgraduates or professional interpreters as advanced. Non-university participants were assessed based on their study duration in accredited language institutions or length of overseas residencies.

Addressing Gender Disparity: Acknowledging the prevalent gender disparity among language learners in Korea, we strived for a practical gender distribution in our corpus, deviating from a rigid 1:1 ratio. Instead, we aimed for a minimum of 50% representation of each gender in every language category, ensuring a balanced dataset while reflecting the actual demographics of the language learning community. To quantitatively assess our success in achieving this gender balance, we employed a measurement formula that compares the intended representation (goal value) with the actual participation (result value) for each gender. This formula is applied as follows:

- Male: MIN(goal value, actual value)/MAX(goal value, actual value)
- Female: MIN(goal value, actual value)/MAX(goal value, actual value)

By computing the average of the male and female scores, we obtain a comprehensive view of our dataset's gender distribution. (Table 2)

Prioritizing Beginner and Intermediate Speakers: Our corpus intentionally emphasizes intermediate and beginner speakers over advanced speakers. This strategy stems from the understanding that non-native speech, often characterized by a diverse range of accents and common mispronunciations, presents substantial data variability, making it especially valuable for enhancing Speech-to-Text (STT) systems and developing pronunciation assessment models (Table 2).

4.2. Recording Protocols

We implemented recording protocols tailored for digital environments, specifically optimizing for the Chrome browser on personal computers. Recognizing the detrimental impact of ambient noise on speech recordings, we provided participants with explicit instructions to conduct their sessions in acoustically controlled environments — these included private spaces or noise-free classrooms, thereby mitigating the intrusion of extraneous sounds. The use of earphones was recommended to further isolate their speech from potential environmental interference.

A notable challenge identified in preliminary trials was the inadvertent truncation of the initial part of participants' responses. This issue was predominantly observed at the commencement of recordings, attributable to the latency between the activation of the recording interface and the participant's consequent verbal response. To counteract this, participants were instructed to observe a brief pause, precisely 0.5 to 1.0 seconds, post-initiation of the recording process before commencing their speech. This buffer ensures the capture of the entirety of the response from its onset, thereby preserving the integrity of the data.

4.3. Data Overview

In developing our corpus, we move beyond the traditional emphasis on English, integrating speech samples from seven diverse languages: English, Japanese, Chinese, French, German, Spanish, and Russian. This expansion responds to a critical need to broaden the scope of linguistic research and tackle the complexities inherent in multilingual studies.

Our corpus amasses roughly 1,000 hours of adult L2 speech data, strategically allocated to ensure a broad linguistic representation: 200 hours each for English, Chinese, and Japanese, and 100 hours each for French, German, Spanish, and Russian. Moreover, recognizing English as a compulsory subject in the Korean education system, we included an additional 200 hours of English speech from middle and high school students, totaling 400 hours of English speech. This subset was subjected to the same stringent curation criteria as adult data, ensuring uniformity in methodology and a standard of comparison throughout our corpus.

5. Data Processing and Annotation

5.1. Preprocessing

The audio recorded on the authoring platform underwent noise-cancellation processing, followed by the introduction of a 0.5-second silent period at the beginning and end of the speech segments before labeling commenced. Pronunciation assessments were labeled based on expert scoring results at the task unit level, specifically at the sentence and paragraph levels. However, spelling and speech transcription, as well as tagging of pronunciation error types, were conducted at the sentence level.

5.2. Annotation Strategy

5.2.1. Scoring System

Assessors meticulously reviewed audio recordings through a custom evaluation interface, assigning scores from 1 to 5 for both pronunciation accuracy and prosodic fluency, following the established criteria outlined in Table 1. The evaluation protocol required concurrent scoring by two assessors for each item, enabling a comprehensive assessment approach. This detailed process was implemented within the framework of the authoring tool, where assessors were randomly paired for each item. Table 2 shows the number of two-person teams that evaluated the same item and the average number of items scored per assessor.

5.2.2. Error Type Tagging

Acknowledging the challenges of securing phonetics specialists for each of the seven foreign languages in Korea and mindful of the high error rate in human annotation of speech transcriptions and phonemic errors, we employed an automated strategy for transcribing speech and tagging error types. Initially, we transcribed the correct phonemes for the prepared reading passages. Subsequently, a phoneme recognition tool was deployed to transcribe the uttered phonemes from the speech-to-text (STT) results. By force-aligning both sets of data, we mapped the correct phonemes to the uttered ones, facilitating the automatic tagging of discrepancies. These were classified into four distinct categories: substitution, omission, insertion, and other types of errors. (Table 3)

Type	Description
Substitution	A phoneme is pronounced as other phonemes than the correct one.
Deletion	A phoneme is not pronounced where it is supposed to be pronounced.
Insertion	A phoneme is pronounced where it is not supposed to be pronounced.
Others	When applying g2p (grapheme-to-phoneme), the data with a warning tag are set to "null," while the tagging field is labeled as "O" for "Other." 1) no sentence, 2) no speech, 3) g2p error, 4) sentence with numbers, 5) decoding error

Table 3: Error Types

5.2.3. Metadata Analysis

Our corpus is enriched with detailed metadata for each recording, covering demographics, technical specifics, and audio quality indicators. This metadata includes the speaker's age, gender, and language proficiency level, which is critical for diverse linguistic analyses. We also document the recording date, duration, location, and device used, supporting in-depth studies into language trends and variations over time. Uniquely, our corpus incorporates precise audio quality metrics, notably peak and RMS levels. These crucial parameters aid in establishing consistent automatic gain control

	English	**Japanese**	**Chinese**	**German**	**Spanish**	**French**	**Russian**
Substitution	1,754,015 (65.27%)	779,480 (77.09%)	3,102,608 (94.83%)	2,294,674 (90.38%)	3,353,366 (95.54%)	2,942,016 (90.21%)	2,031,092 (81.38%)
Deletion	365,186 (13.59%)	203,307 (20.11%)	95,843 (2.93%)	199,119 (7.84%)	48,432 (1.38%)	194,879 (5.98%)	384,620 (15.41%)
Insertion	552,677 (20.56%)	3,341 (0.33%)	5,148 (0.16%)	42,732 (1.68%)	107,008 (3.05%)	121,401 (3.72%)	72,094 (2.89%)
Others	15,618 (0.58%)	24,970 (2.47%)	68,116 (2.08%)	2,290 (0.09%)	1,170 (0.03%)	2,933 (0.09%)	7,936 (0.32%)
Total	2,687,496 (100.00%)	1,011,098 (100.00%)	3,271,715 (100.00%)	2,538,815 (100.00%)	3,509,976 (100.00%)	3,261,229 (100.00%)	2,495,742 (100.00%)

Table 4: Pronunciation Error Types Across Seven Languages

thresholds, ensuring uniform loudness perception across recordings. Consequently, this enhances the reliability of subsequent analyses or applications in machine learning tasks.

5.2.4. Label Distribution

This dataset was meticulously curated with the objective of annotating professional evaluations concerning the precision of pronunciation and linguistic fluency, following the accumulation of non-native speech samples across diverse language competencies — encompassing 400 hours of English, 200 hours each of Chinese and Japanese, and 100 hours of assorted European languages. In adherence to our foundational aim of authenticity, the volume of data corresponding to each evaluative criterion and error category has been crafted to mirror real-world linguistic diversity. The specific quantity of data per attribute is detailed in the Table 4.

5.3. Quality Assurance

5.3.1. Assessor Selection

In this study, assessors were meticulously selected based on their theoretical knowledge and practical expertise in language proficiency, adhering to stringent criteria. The panel included 28 experts for English, 10 each for Chinese and Japanese, 8 each for French and Russian, 11 for German, and 9 for Spanish. Assessors were required to meet one of the following qualifications: they must either be university faculty possessing at least a PhD degree in a field pertinent to foreign language education, with no less than three years of experience teaching the relevant language to Korean students; or they must be professional simultaneous interpreters with a minimum of three years of field experience, having specialized in the target language at an advanced school of interpretation and translation and possessing at least one year of language instruction experience at the tertiary level. These stringent qualifications ensured the selection of evaluators with an in-depth understanding of language nuances, which is critical for accurate assessment. Furthermore, acknowledging previous research indicating variance in L2 evaluation patterns between non-native and native assessors, our panels were composed exclusively of native speakers for each target language, ensuring evaluation authenticity and consistency.

5.3.2. Bias Mitigation

Before initiating the evaluation, we orchestrated a comprehensive training session for assessors across each language category to establish uniformity and impartiality in the scoring process. This session encompassed an in-depth overview of diagnostic queries and grading rubrics, supplemented by case studies that assisted assessors in calibrating their psychological benchmarks for assessment.

After the preparatory briefing, assessors were presented with a set of 50 test items specific to their language group, which they were instructed to evaluate concurrently. This was followed by a statistical analysis of the inter-assessor agreement, offering insights into individual assessor's tendencies towards stringency or leniency and their predilection for specific scoring metrics. When the standard deviation in scoring exceeded an established threshold, tailored feedback was provided, empowering assessors to self-regulate and recalibrate their grading inclinations accordingly.

5.3.3. Cross-Validation

Pronunciation evaluations, primarily, are known to be substantially influenced by the subjective discrepancies amongst assessors, a phenomenon documented in preceding research that can have consequential impacts on the efficacy of pronunciation assessment models. To bolster the objectivity within these subjective appraisals, our evaluation protocol mandated that two assessors concurrently

	Language	Samples	Mean	SD
Pronunciation Accuracy	EN	114,494	0.6731	0.1020
	JA	88,712	0.6114	0.0626
	ZH	63,678	0.7104	0.0531
	DE	34,596	0.6358	0.1367
	ES	37,437	0.6004	0.1016
	FR	28,003	0.6170	0.2032
	RU	30,223	0.7093	0.1270
Prosodic Fluency	EN	114,494	0.6704	0.0872
	JA	88,712	0.6008	0.0819
	ZH	63,678	0.7284	0.0663
	DE	34,596	0.6067	0.1581
	ES	37,437	0.6196	0.1156
	FR	28,003	0.5630	0.1941
	RU	30,223	0.7133	0.1213

Table 5: Inter-Rater Reliability

rate each item on a scale spanning from 1 to 5. In instances where divergences in scores between assessors exceeded two points, an intervention by a third adjudicator, holding senior professorial expertise, was solicited to scrutinize the results and enact final score recalibrations. The mean score was ratified as the conclusive outcome in scenarios where the discrepancy was non-existent or confined to a one-point margin.

5.3.4. Qualitative Assessment Validation

Traditionally, FACETS analysis is integral to foreign language pedagogy research, often utilized to pinpoint assessors' biases, such as severity or leniency, and to highlight various influences on their evaluations. This method typically involves a select group of expert assessors concurrently scoring a limited sample set. However, our corpus relies on extensive qualitative assessments, making it impractical for all assessors to evaluate the same items simultaneously. This challenge necessitated an innovative validation method to ensure reliability within our statistical framework, as we dealt with data gathered on an unusually large scale.

To address this, we employed Krippendorff's alpha—a statistical measure known for gauging the consistency or agreement among data coders—in our validation process. This metric allowed us to measure the level of agreement, or more specifically, the directional concordance among pairs of assessors evaluating the same items. Due to the large number of assessor pairs and the variability in the items they examined, our methodology focused on calculating inter-assessor reliability for each pair, then determining an overall average value (refer to Table 5).

A Krippendorff's alpha value of .60 or higher is generally considered reliable, with values above .70 signaling high reliability. Our study, factoring in the complex nature of linguistic and phonetic evaluations, the high volume of samples assessed (an average of 9,157 per assessor), and the configuration of 44 assessor duos, achieved an average reliability score of approximately .65. This score represents a good agreement, especially significant given the highly subjective and intricate aspects of the assessments involved.

6. Conclusion

This research delved into the intricacies of creating an annotated corpus for L2 pronunciation assessment for Korean learners across seven foreign languages and examined its validity as actual AI training material. To our knowledge, no existing dataset features multilingual speech from speakers of a single language origin, annotated with both speech qualities and expert assessment results. Considering that most ASR models are trained predominantly on native speakers' speech, rendering them vulnerable to inaccuracies with non-native pronunciations, and the prevailing focus on English-native speakers in existing corpora, we anticipate that our corpus of Korean speakers' multilingual speech will offer substantial research and technological merits.

Notwithstanding its contributions, the corpus introduces certain constraints. Although adjustments (i.e., overlap rate) were made during the design phase to counteract the imbalance in the number of learners and the gender ratio for each foreign language in Korea, thereby preventing distortion in model training performance, the remaining gender imbalance is a potential concern during training. It may be necessary either to secure additional male speech samples or to utilize the current data by sampling to achieve an equal gender ratio for training purposes. Similarly, given the varying volumes of speech data across languages, acquiring more speech data for languages with lesser content may be essential or ensuring that they are trained at a specific ratio. We have confirmed that the existing data alone exhibit similar levels of speech recognition and automatic pronunciation assessment performance across languages. However, these aspects have not been addressed due to the scope limitations of this research and are earmarked for exploration in subsequent studies.

7. Acknowledgments

The study was supported by the AI Data project from National Information Society Agency (NIA), Republic of Korea.

8. Ethical considerations

In the recruitment of participants, we engaged a specialized survey company to ensure professionalism and adherence to ethical standards. All participants were involved in the speech collection process only after giving informed consent through signing two critical documents: the "Copyright Usage Agreement for Data" and the "Consent Form for Personal Information Collection, Use, and Third-Party Provision," which are standard contracts released by relevant government agencies. Special attention was given to the recruitment of minors for English speech recordings. This was conducted through official requests for cooperation sent to various middle and high schools, and participation was limited to students who had obtained formal consent from their legal guardians. To ethically compensate for their participation, all participants were remunerated for contributing their speech data, upholding the ethical standards of research involving human subjects.

9. Bibliographical References

Marina Dodigovic. 2007. Artificial intelligence and second language learning: An efficient approach to error remediation. *Language Awareness - LANG AWARE*, 16:99–113.

Shin Dong-Kwang. 2019. Exploring the feasibility of ai chatbots as a tool for improving learners' writing competence of english. *Korean Journal of Teacher Education*, 35:41–55.

Fan Junying, Yang Lei, Mai Yinghong, and Xu Yujun. Effectiveness of automatic speech evaluation system in improving students' self-efficacy in oral english learning.

Jaya Kannan and Pilar Munday. 2018. New trends in second language learning and teaching through the lens of ict, networked learning, and artificial intelligence. *Círculo de Lingüística Aplicada a la Comunicación*, 76.

Heyoung Kim, Dongkwang Shin, Hyejin Yang, and Jang Ho Lee. 2019. A study of ai chatbot as an assistant tool for school english curriculum. *Korean Association For Learner-Centered Curriculum And Instruction*, 19:89–110.

Sak Lee. 2019. The effects of gamification-based artificial intelligence chatbot activities on elementary english learners' speaking performance and affective domains. *The Korea Association of Primary English Education*, 25:75–98.

Xiaobin Liu, Chunxiao Zhu, Jianli Jiao, and Manfei Xu. 2018. Promoting english pronunciation via mobile devices-based automatic speech evaluation (ase) technology. In *Blended Learning. Enhancing Learning Success: 11th International Conference, ICBL 2018, Osaka, Japan, July 31-August 2, 2018, Proceedings 11*, pages 333–343. Springer.

Nina Markl and Catherine Lai. 2021. Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 34–40.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Tanja Schultz and Tim Schlippe. 2014. GlobalPhone: Pronunciation dictionaries in 20 languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 337–341, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sujin Seong and Sak Lee. 2021. Analyzing learners' and teachers' perceptions of ai pengtalk for english learning and the suggestions for its use. *Korean Association For Learner-Centered Curriculum And Instruction*, 21:915–935.

Ke Shi, Kye Min Tan, Richeng Duan, Siti Umairah Md Salleh, Nur Farah Ain Suhaimi, Rajan Vellu, Ngoc Thuy Huong Helen Thai, and Nancy F Chen. 2020. Computer-assisted language learning system: Automatic speech evaluation for children learning malay and tamil. In *INTERSPEECH*, pages 1019–1020.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. pages 3710–3714.

10. Language Resource References

Korean Learners' L2 Speech Corpus (Chinese and Japanese). distributed via AI-Hub by National Information Society Agency (NIA), Korea. https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71464

Korean Learners' L2 Speech Corpus (English). distributed via AI-Hub by National Information Society Agency (NIA), Korea. https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71463

Korean Learners' L2 Speech Corpus (German, French, Spanish, Russian). distributed via AI-Hub by National Information Society Agency (NIA), Korea. https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71466