LREC-COLING 2024

Constructing Korean Learners' L2 Speech Corpus Of Seven Languages For Automatic Pronunciation Assessment

Seunghee Han, Minhwa Chung, Sunhee Kim

Seoul National University

Research Team

Seunghee Han

Learning Sciences Research Institute,
Seoul National University

Minhwa Chung

Department of Linguistics,
Seoul National University

Sunhee Kim

Department of French Language Education,
Seoul National University

TABLE OF CONTENT

Background

Transition to Hybrid Learning Environments Post-COVID-19

- Widespread adoption of online learning during the COVID-19 pandemic.
- Persistence of hybrid models blending online and offline methods in the post-pandemic landscape.
- Increased reliance on online platforms for foreign language learning.
- Escalated demand for AI-enhanced educational tools.

Enhancing AI Technologies for Non-native Speakers

- Critical need for robust AI solutions to support non-native speakers.
- Importance of improving Speech-to-Text (STT) recognition for varied accents.
- Development of automatic pronunciation and speaking assessment models.

Research Significance and Gap

Table 1. Comparison of L2 Corpora

Corpus Name	Details	Comment
GlobalPhone	• Recordings from native speakers across various languages.	• Small-sized corpus with 17.5 hours for each 21 languages on average.
Librispeech	• Recordings from native speakers across various languages. • Derived from read audiobooks from the LibriVox project.	• Not collected from ordinary speakers in a natural conversational interactions.
ACCENT	• Speakers read the same English paragraph. • Constructed to be used by linguists as well as other people who simply wish to listen to and compare the accents of different English speakers. • Demonstrates that accents are systematic rather than merely mistaken speech	• Largely centered around English speech from non-native speakers.
L2 Arctic	• Includes recordings from twenty-four (24) non-native speakers of English. • Each speaker recorded approximately one hour of read speech. • Includes speech recordings, word-level transcriptions, phoneme-level transcriptions, and manual annotations.	• Largely centered around English speech from non-native speakers.
CHILDES	• Comparable corpora made up from transcripts of child speech across 40 languages.	• Comprises transcripts of bilingual children, older school-aged children, adult second-language learners, children with various types of language disabilities and aphasics who are trying to recover from language loss.
Spechocean762	• A Benchmark in L2 Speech Corpora • Introduced in 2021, tailored for CALL. • Contains 5,000 English utterances from 250 non-native Mandarin speakers. • Includes detailed scoring for phonetic accuracy, fluency, prosody, completeness, and stress. • Evaluations from a panel of five experts provide both average and median scores.	• Limited focus on English with only 20 sentences per speaker. • Evaluation focused on within-sentence rather than across-sentences, limiting insight into broader proficiency.

- Focus on native speech across multiple languages
- L2-focused corpora – Primarily English from non-native speakers
- Limited multilingual speech data from adult L2 speakers, especially at intermediate and low proficiency levels
- Underrepresentation in speech datasets impacts research in language acquisition and speech processing

Addressing the Gaps in Existing L2 Corpora

Proposed Expansions

- Develop datasets with greater linguistic diversity and more extensive speaker contributions.
- Enhance evaluation criteria to include across-sentence analysis for comprehensive proficiency insights.

Challenges in Scale and Evaluation

- The complexity of the scoring matrix complicates uniform evaluations across diverse criteria.
- Logistical challenges in scaling up expert evaluations, highlighting the need for consistent and accurate assessments.

Future Directions

- Creation of a multilingual error classification system to streamline pronunciation error analysis.

Enhancing Assessment Model Accuracy

Overcoming Ambiguity in Scoring Criteria and Matrix

- Ambiguity in cumulative scoring approach within ETS's delivery rubrics, incorporating factors like clarity, fluency, and pronunciation completeness.
- Differentiated by scoring ranges leads to ambiguity in evaluation consistency.

Ensuring Objectivity and Consistency in Scoring

- Need for overcoming subjectivity inherent in pronunciation assessments.
- Importance of well-defined criteria and rubrics to ensure reliability and consistency.
- Comprehensive assessor training to minimize subjective discrepancies among assessors.

Methodological Advances and Contributions

Innovative Validation Approaches

- Introduction of novel methods to validate large-scale data annotated by multiple experts.
- Differentiation from traditional validation methods, which rely on fewer assessors.

Impact on AI and Language Education

- Application of consistent, reliable assessment data in training AI systems.
- Focus on Korean learners' L2 pronunciation—advancing phonetics, foreign language education, and speech recognition technology.

Advancements in AI-Driven Language Tools

- Aiming for more accurate and practical automatic pronunciation assessments.
- Potential to enhance AI-driven language learning applications.

TABLE OF CONTENT

Script Design Considerations

Task Design

- Incorporates single-sentence and paragraph readings to cover a spectrum from simple to complex linguistic interactions.
- Scripts integrate challenging vocabulary for Korean L2 learners across seven languages.

Difficulty Categorization

- Vocabulary classified by CEFR levels: High (B2-C2), Medium (A2-B1), Low (Pre-A1-A1).
- Ensures diverse learning scenarios, mimicking real-life language usage.

Balanced Distribution

- Equal exposure to different difficulty levels for all participants.
- Tasks designed for minimum engagement of 25 minutes to elicit varied pronunciation errors.

Collaborative Script Development

- Developed jointly by Korean and native-speaking professors specializing in the target languages.
- Ensures linguistic authenticity and relevance to common pronunciation challenges.

Pilot Testing and Feedback Integration

- Initial scripts tested with 50 questions per language, reviewed by phonetics and language instruction experts.
- Continuous feedback loop from trial recordings to refine scripts for final use.

Addressing the Limitations in Existing Resources

Objective

- Enhance AI capabilities in linguistic applications through a robust corpus of diverse spoken samples.

Framework Integration

- Combines CEFR guidelines with the Korea Institute of Curriculum and Evaluation taxonomy.
- Creates a solid framework for a comprehensive thematic matrix.

Thematic Matrix

- Broad categories: Personal, Public, Occupational, and Educational.
- Further segmented into subcategories: Location, Institution, Persons, Objects, Events, Operations, Texts.

Script Creation and Data Enrichment

- Utilizes pre-existing linguistic repositories and thematic keywords from Korea's National Information Society Agency (NIA).
- Ensures wide thematic scope and high representational accuracy in script creation.

Development of a New Scoring Rubric

Table 2. Scoring Rubrics

	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.

Revised Assessment Criteria

Objective

- Enhances uniformity and objectivity in evaluations of L2 speech pronunciation.
- Combines evaluative elements from ETS and Speechocean762 with insights from phonetic research and automated pronunciation evaluation studies.

Primary Categories (1~5 scale each)

- Pronunciation Accuracy: Focuses on the clarity of individual speech segments.
- Prosodic Fluency: Evaluates stress, intonation, rhythm, and speech rate comprehensively.

Speech Completeness

- Initially considered for inclusion; measures non-native speakers' speech-to-text recognition rate.
- Excluded from primary benchmarks to maintain scoring consistency due to its high dependency on accuracy and fluency.

TABLE OF CONTENT

Corpus Composition and Data Allocation

Table 3. Distribution of the Corpus

			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	35.53%	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	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

Corpus Composition and Data Allocation

Multilingual Focus

- Includes speech samples in seven languages: English, Japanese, Chinese, French, German, Spanish, and Russian.
- Addresses the need for broader linguistic research and the complexities of multilingual studies.

Speech Data Volume

- Total of 1,000 hours of L2 adult speech data distributed as follows:
- 200 hours each for English, Chinese, and Japanese.
- 100 hours each for French, German, Spanish, and Russian.

- English: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71463
- Japanese & Chinese: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71464
- French, German, Spanish & Russian: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71466

Additional English Data

- Additional 200 hours of English speech from Korean middle and high school students, reflecting compulsory English education.
- Ensures 400 total hours of English speech, with all data held to consistent curation standards.

Standardization and Methodology

- Uniform curation criteria applied across all data subsets to maintain consistency and comparability.
- Strategic allocation ensures broad linguistic representation and robust comparative analyses.

Demographic Diversity and Proficiency Classification

Broad Demographic Spectrum

- Includes various age groups, genders, and language proficiency levels to enhance data robustness.

Focus on Adult Learners

- Predominantly adults, reflecting language study trends in higher education in Korea.
- Emphasis on languages beyond English, started typically at university level.

Proficiency Stratification

- Participants primarily from domestic universities, majoring in foreign languages.
- Proficiency categorized into beginner, intermediate, and advanced based on CEFR standards and academic progression.

Gender Distribution

- Aim for at least 50% representation of each gender in every language category.
- Gender balance assessed via a formula to compare intended vs. actual participation (Table 3).

Prioritization of Beginner and Intermediate Levels

- Focus on capturing the variable data from non-native speech, crucial for enhancing STT systems and developing pronunciation models

Technical Consistency in Data Collection

Challenges and Solutions

- Addressed the issue of initial response truncation observed in preliminary trials.
- Adjusted recording onset to ensure complete capture of participant responses from the very beginning.

Preprocessing Techniques

- Applied noise-cancellation processing to enhance audio clarity.
- Added a 0.5-second silent period at both the beginning and end of each recording to facilitate precise labeling.

TABLE OF CONTENT

Assessment Process

Objective

- Focuses on professional evaluations concerning pronunciation precision and linguistic fluency.
- Data volumes for each evaluative criterion and error category designed to mirror real-world linguistic diversity.
- Strategic curation aligns with foundational aims of authenticity in linguistic representation.

Expert Scoring System

- Assessors use a custom interface to score pronunciation accuracy and prosodic fluency on a scale of 1 to 5, as per criteria in Table 2.
- Each item is concurrently scored by two assessors for comprehensive evaluation.
- Random pairing of assessors for each item, detailed in Table 2 showing team configurations and workload.

Phoneme Error Analysis

Table 4. Error Types

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

Automatic Error Tagging

- Due to the challenge of securing phonetics specialists for seven languages, an automated strategy for transcription and error tagging was implemented.
- Correct phonemes for reading passages were initially transcribed.
- A phoneme recognition tool transcribed uttered phonemes from STT results.
- Discrepancies between correct and uttered phonemes were automatically tagged through force-alignment, categorized into substitution, omission, insertion, and other errors (Table 4).

Error Type Distribution

Table 5. Pronunciation Error Types Across Seven Languages

	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%)	448,432 (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%)

Comprehensive Data Records

Demographic and Technical Data

- Includes age, gender, and language proficiency of speakers.
- Records date, duration, location, and device used for each recording.

Audio Quality Indicators

- Documents peak and RMS levels to standardize audio loudness.
- Supports consistent automatic gain control thresholds across recordings.

TABLE OF CONTENT

Assessor Selection Criteria

Assessor Expertise and Panel Composition

- 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.
- Selected based on theoretical knowledge and practical expertise in language proficiency.

Qualification Criteria

- University faculty with at least a PhD in a relevant field and a minimum of three years teaching the target language to Korean students.
- Professional simultaneous interpreters with at least three years of field experience and one year of language instruction at the tertiary level.

Research Justification

- Selection exclusively of native assessors aligns with research indicating evaluation variance between native and non-native assessors.
- Supports the authenticity and consistency of linguistic assessments.

Minimizing Subjective Bias

Comprehensive Assessor Training

- Conducted a detailed training session for assessors in each language category.
- Covered diagnostic queries and grading rubrics, supplemented by relevant case studies.

Calibration Process

- Assessors evaluated a set of 50 test items specific to their language group.
- Aimed to standardize assessment criteria and reduce individual bias.
- Statistical analysis conducted post-evaluation to measure agreement levels.
- Focused on identifying assessors' tendencies towards stringency or leniency.

Feedback and Recalibration

- Provided tailored feedback when scoring deviations exceeded predefined thresholds.
- Enabled assessors to self-regulate and adjust their scoring methods to maintain consistency.

Assessment Protocol Details

Dual Assessor Evaluation

- Each item is concurrently rated by two assessors on a scale from 1 to 5.
- Ensures initial mitigation of personal bias and promotes consistent evaluation standards.

Adjudication Process

- In cases of score divergences exceeding two points, a third adjudicator with senior professorial expertise intervenes.
- This adjudicator scrutinizes and recalibrates the final scores to resolve significant discrepancies.

Final Scoring Protocol

- If discrepancies are confined to a one-point margin or non-existent, the mean score is ratified as the final outcome.
- Facilitates fairness and maintains objectivity in final score determination.

Ensuring Consistency on a Large Scale

Challenges in Traditional Validation

- Traditional use of FACETS analysis in language research to identify assessor biases.
- Limited by the scale of evaluations, typically involving a small group of assessors.

Adaptation to Large-Scale Data

- •Employed Krippendorff's alpha to measure agreement among assessors.
- Suitable for large datasets with many assessors and diverse items.

Methodology for Ensuring Reliability

- Calculated inter-assessor reliability for each of the 44 pairs, then determined an overall average.
- Aimed to capture directional concordance among pairs assessing the same items.

Results of Krippendorff's Alpha

- Achieved an average reliability score of .65, indicating good agreement.
- High significance due to the volume of samples assessed (average of 9,157 per assessor) and the subjectivity involved in linguistic and phonetic evaluations.

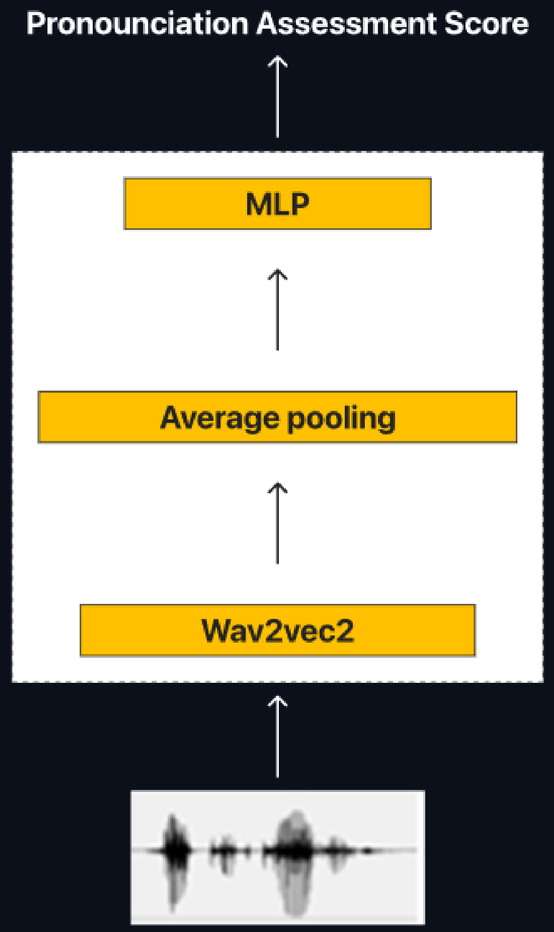
Ensuring Consistency on a Large Scale

Table 6. Inter-Rater Reliabilty

	Language	Samples	Mean	SD
Pronunciation Accuracy	EN	114,494	0.6731	0.1020
	JP	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
	JP	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	28,003	0.7133	0.1213

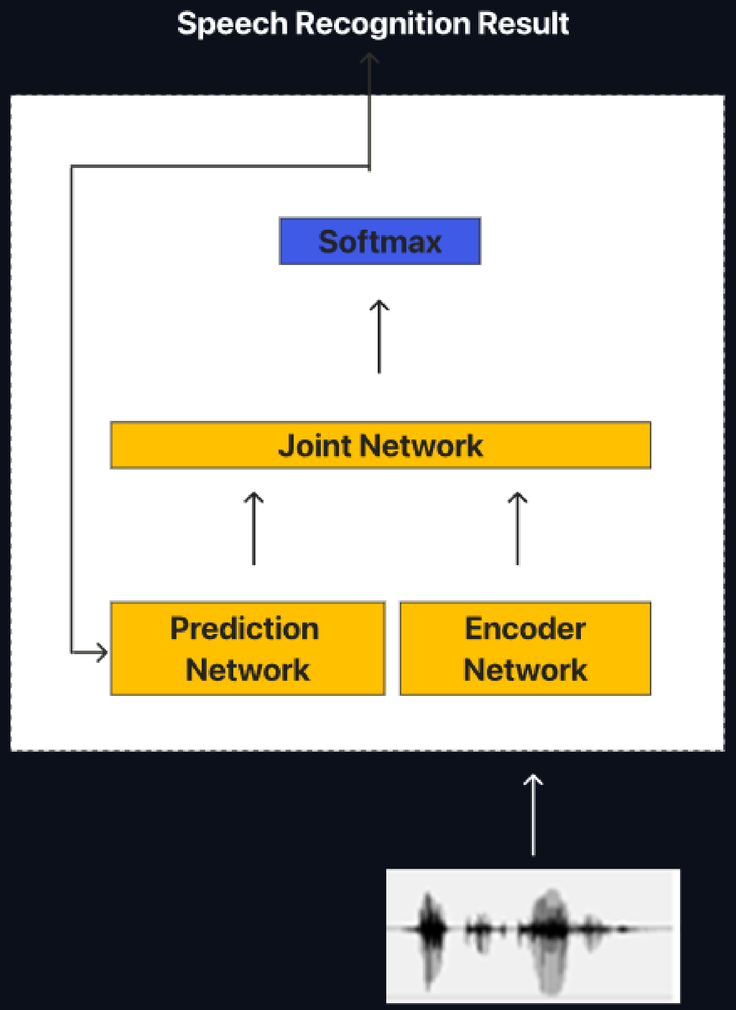
Performance Testing of STT and Pronunciation Models

Table 7. Automatic Pronunciation Assessment Model



Language	PCC
EN	0.72
ZH	0.73
JP	0.62
DE	0.67
ES	0.61
FR	0.68
RU	0.72

Table 8. Speech Recognition Model



Language	WER/CER
EN	5.4%
ZH	6.0%
JP	4.2%
DE	8.7%
ES	8.1%
FR	14.2%
RU	14.0%

TABLE OF CONTENT

Limitations and Future Works

Unique Contributions

- First dataset of its kind with multilingual speech from speakers of a single language origin (Korean) annotated with speech qualities and expert assessments.
- Targets the gap in ASR models trained primarily on native speaker speech, which often misinterprets non-native pronunciations.

Existing Limitations

- Gender imbalance remains despite efforts to adjust overlap rates during corpus design.
- Discrepancies in the volume of speech data available for each language, potentially affecting model training.

Strategies for Mitigation

- Consider securing additional male speech samples or adjusting sampling strategies to achieve a balanced gender ratio.
- Propose acquiring more data for languages with lesser content to ensure balanced training exposure.

Future Research Directions

- Further studies to explore and address gender and volume disparities in the corpus.
- Expand the scope to assess and enhance speech recognition and automatic pronunciation assessment uniformly across languages.

Thank You

seunghee.han@snu.ac.kr

Our corpus is available at:

- English: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71463
- Japanese & Chinese: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71464
- French, German, Spanish & Russian: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71466