

Article

Mispronunciation Detection and Diagnosis for Non-Native Korean Learners Using Iterative Pseudo-Label Refinement Based on Self-Supervised Learning

Na Geng ¹, Hee-Jung Na ¹, Jiyu Won ², Dong-Gyu Kim ² and Jeong-Sik Park ^{3,*}

¹ Department of English Linguistics, Hankuk University of Foreign Studies, Seoul 02450, Republic of Korea; gengna0324@hufs.ac.kr (N.G.); 920313@naver.com (H.-J.N.)

² Department of AI Convergence, Hankuk University of Foreign Studies, Seoul 02450, Republic of Korea; wonjiyu@hufs.ac.kr (J.W.); dkkim2008@naver.com (D.-G.K.)

³ Department of English Linguistics and Language Technology, Hankuk University of Foreign Studies, Seoul 02450, Republic of Korea

* Correspondence: parkjs@hufs.ac.kr; Tel.: +82-02-2173-8814

Featured Application

This research can be applied to automated Mispronunciation Detection and Diagnosis for Korean language-learning applications, identifying specific error types to provide targeted correction guidance for non-native Korean learners. The proposed approach can be extended to mispronunciation detection in other languages.

Abstract

Accurate Mispronunciation Detection and Diagnosis (MDD) for non-native Korean learners is critical for effective pronunciation feedback, but it is hindered by the lack of training labels that reflect learners' actual pronunciations. This paper presents a pseudo-label generation framework that fine-tunes Whisper to output pronunciation-oriented sequences, supported by data-quality management and iterative label refinement. We convert orthographic transcripts into pronunciation targets using an existing Grapheme-to-Phoneme (G2P) tool to reduce reliance on standard written forms, and apply multi-stage refinement with cross-model agreement validation under progressively adjusted thresholds to filter unreliable pseudo-labels. We further improve robustness by incorporating larger and more diverse non-native speech corpora and by applying dataset-specific preprocessing, including noise removal, duration-based selection, and duplicate control. Evaluation on a manually annotated test set of actual learner pronunciations shows that models trained with refined pseudo-labels achieved a lower Phoneme Error Rate (PER) and performed better than the baseline model on MDD. Overall, the proposed framework enables practical MDD for non-native Korean speech without requiring large-scale manual phoneme annotation.

Keywords: MDD; non-native Korean speech; pseudo-label refinement; self-supervised learning; iterative label refinement

Academic Editor: Douglas O'Shaughnessy

Received: 14 January 2026

Revised: 26 February 2026

Accepted: 27 February 2026

Published: 2 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The growing global interest in learning Korean as a second language (L2 Korean) [1] demands effective technological tools for Mispronunciation Detection and Diagnosis

(MDD). Learners frequently produce phoneme substitution, deletion, and insertion errors due to interference from their native languages, hindering pronunciation accuracy and communication [2–4]. Phoneme-level recognition that identifies specific error types and provides targeted feedback is needed to address these challenges. An MDD system based on phoneme recognition that provides feedback on phoneme substitution, deletion, and insertion errors is therefore needed. For instance, if a learner pronounces “학교” [hak-gyo] (school) as “하교” (/ ha-gyo/), the system detects the consonant deletion and indicates the missing “ㄱ” (/k/) sound.

However, developing such systems faces critical obstacles. Phoneme recognition research has primarily focused on major languages such as English, which benefits from abundant labeled data enabling development from early HMM-GMM models to modern deep learning approaches [5–7]. Korean phoneme recognition research remains relatively limited. Since Korean employs “Hunminjeongeum”, a phonogram that combines properties of phoneme-level and syllable-level writing [8], research has primarily emphasized grapheme-based speech recognition rather than pronunciation-based phoneme recognition [9–13]. Existing models output standardized written forms (e.g., “학교”) that do not capture actual pronunciation features, rather than phonetically accurate transcriptions (e.g., [hak.gjo]) [14] needed for effective MDD. Nonetheless, MDD for non-native Korean speakers has been extensively studied by leveraging Korean speech recognition models [15,16].

The scarcity of labeled training data reflecting actual non-native pronunciations presents the most significant challenge for pronunciation error detection. Effective MDD requires models trained on actual learner productions rather than standard forms. Manual phonetic transcription by expert phoneticians is prohibitively expensive and time consuming, requiring specialized linguistic knowledge to accurately capture non-native pronunciation variations. While existing Korean speech datasets provide orthographic transcripts based on standard pronunciation, these fail to represent the acoustic reality of learner speech, creating a fundamental mismatch between training labels and actual pronunciations.

The development of self-supervised learning (SSL) offers potential solutions to these data scarcity challenges. Pre-trained models such as Wav2Vec 2.0 [17], HuBERT [18], and WavLM [19], trained on large unlabeled corpora, learn robust speech representations that transfer effectively to downstream tasks through fine-tuning [20]. The Whisper model demonstrates particularly strong transcription performance across diverse languages [13] and produces Korean outputs that largely correspond to actual pronunciation rather than standard orthographic forms based on our prior experiments.

Building on these advances, pseudo-labeling has become increasingly practical and accurate across a wide range of speech tasks [21–25]. Most prior studies primarily focus on automatic speech recognition (ASR) [26–28], where ground-truth transcriptions for the source (teaching) data are relatively easy to obtain. Moreover, even when the target domain exhibits substantial pronunciation variation under the same annotation, ASR outputs can still be evaluated against the corresponding text labels, since multiple realizations map to the same transcript.

Well-performing SSL and weakly supervised models have also advanced MDD and diagnosis across multiple languages [29–31]. However, in most of these studies, the target languages still provide ground-truth teaching labels for L2 speech. For Korean, there is currently no publicly available speech dataset with ground-truth, pronunciation-based labels for either native or non-native speakers. Moreover, few studies have explored

training models that output pronunciation-oriented symbol sequences that Korean learners can directly use to understand and correct their pronunciation.

Building on this observation, we develop a framework to generate pronunciation-based output sequences from non-native Korean speech for MDD, leveraging iterative label refinement as a practical training strategy. We first convert orthographic labels in existing datasets into pronunciation-based labels using a Grapheme-to-Phoneme (G2P) tool (available at: <https://github.com/Kyubyong/g2pk>), and then apply multi-stage iterative label refinement to fine-tune Whisper toward pronunciation-oriented outputs. To reduce pseudo-label noise, we apply cross-model agreement filtering by comparing the pronunciation-based sequences predicted by different models. Since pronunciation-based ground truth is not available during pseudo-labeling, we use Phoneme Difference Rate (PDR) as the length-normalized edit distance between two predicted sequences.

Data coverage and quality are also critical for robust pronunciation-oriented modeling. We therefore incorporate larger and more diverse non-native Korean corpora and apply dataset-specific preprocessing, including noise removal, duration-based filtering, and duplicate control.

Our methodology makes two key contributions: (1) we fine-tune Whisper to produce pronunciation-based sequences from non-native Korean speech, improving MDD performance; (2) we derive pronunciation targets by converting orthographic transcripts using the G2P tool, reducing reliance on standard orthography and guiding the model toward pronunciation-faithful outputs.

To evaluate models fine-tuned on refined pseudo-labels, we use Phoneme Error Rate (PER) as the primary metric on a manually annotated evaluation set, measuring substitutions, insertions, and deletions relative to the reference phoneme labels. Importantly, these manual annotations are used only for evaluation, while training relies on refined pseudo-labels, enabling MDD without large-scale manual phoneme annotation.

The paper is organized as follows. Section 2 reviews related works and limitations of previous studies. Section 3 describes the datasets and preprocessing strategies. Section 4 presents experimental configurations and results. Section 5 concludes with key findings and future directions.

2. Related Work

Recent studies have leveraged SSL models for MDD, demonstrating their effectiveness in capturing phonetic features without extensive labeled data. However, pronunciation-based MDD for Korean remains underexplored, largely due to the lack of pronunciation-faithful L2 annotations. In contrast, English has seen substantial progress, supported by publicly available corpora such as L2-ARCTIC [32], which provide phoneme-level references and have enabled extensive research on MDD.

Shahin et al. [33] move beyond phoneme-level MDD by predicting articulatory phonological features (e.g., place and manner) to provide more informative feedback and better handle non-categorical or distorted errors. They propose a multi-label CTC objective on top of a wav2vec2-based detector and report improvements over a strong phoneme-level baseline on English L2 speech (L2-ARCTIC), achieving lower FAR, FRR, and DER and stronger fine-grained diagnosis. Yang et al. [25] address the scarcity of phoneme-level labels by fine-tuning wav2vec 2.0 with pseudo-labeled L2 data, using an on-the-fly ensemble to generate dynamic pseudo-labels that are more robust to label noise and improve PER and MDD F1 over offline pseudo-labeling. Both works rely on English resources with pronunciation-faithful supervision (e.g., L2-ARCTIC), whereas comparable ground-truth, pronunciation-based annotations are largely unavailable for Korean.

El Kheir et al. [34] define Qur’anic MDD as a phoneme recognition task and provide standardized training/test resources with a dedicated 62-phoneme inventory and clear evaluation protocol for error localization and diagnosis.

With access to speech data paired with ground-truth labels, Mansour [35] achieves mispronunciation conduction through domain adaptation through multi-stage fine-tuning. Similarly, Qandos et al. [36] fine-tune the Whisper-large-v3 model to improve phoneme recognition performance.

To further expand the dataset, Al-Zaro et al. [37] use semi-supervised self-training to leverage the EqraTech corpus, where unlabeled data is about three times larger than labeled data. Pseudo-labels are generated by a model trained on labeled samples, then filtered by requiring agreement with an HMM recognizer (CMU Sphinx-4 [38]) supported by page-specific language models. Only utterances with a 100% match are kept and added to the training set.

For Korean, there is currently no publicly available L2 dataset with ground-truth pronunciation labels, which motivates us to explore pseudo-labeling in the absence of pronunciation-level supervision.

Xu et al. [21] have conducted iterative pseudo-labeling on ASR tasks by training the newly generated models through refined models. Lugosch et al. [39] extend pseudo-labeling from monolingual to large-scale multilingual ASR across 60 languages. It shows that fine-tuning on each target language before generating pseudo-labels improves low-resource performance and helps the model generalize better to new domains such as LibriSpeech.

Table 1 compares representative prior work with our approach. Most existing methods rely on fully or partially ground-truth pronunciation-based labels for L2 speech, whereas our method is developed without ground-truth pronunciation-based labels for Korean L2 speech (manual phoneme labels are used only for evaluation). In addition, prior studies mainly target other languages, while Korean L2 MDD remains underexplored due to the lack of ground-truth pronunciation-based datasets.

Table 1. Comparison of our approach with representative existing approaches.

Method	Language	Pron. GT*	Backbone	PL	Diff. vs Ours
Shahin et al. [33]	EN	Yes	wav2vec2	—	EN + phoneme GT
Yang et al. [25]	EN	Yes	wav2vec2.0	on-the-fly ensemble PL	EN + phoneme GT
Qandos et al. [36]	AR	Yes	Whisper-large-v3	—	AR + supervised PR
Our approach	KO	No (eval. only)	Whisper-medium	PDR/CER-based agreement	KO L2 without Pron. GT

* Abbreviations: Pronunciation (Pron.), Ground-Truth (GT), Pseudo-labeling (PL), Diff. (Differences), English (EN), Arabic (AR), Korean (KO), Evaluation (eval.) and Phoneme Recognition (PR).

For ASR, pronunciation variability under the same transcript is typically treated as noise to be reduced or ignored so that different realizations map to the same label. In our task, however, these pronunciation-driven deviations are the signal we want to capture. Therefore, we leverage unlabeled speech to generate pseudo-labels and apply a series of preprocessing and filtering steps to construct a training set that supports pronunciation-based sequence outputs.

3. Methodology

3.1. Model Selection and Phoneme Unit Definition

The primary challenge of this work lies in the absence of pronunciation-based labels for the L2 data. To address this, a robust ASR or pre-trained model is required to generate pronunciation-based labels from limited native speech data for fine-tuning. We experimented with Wav2Vec2 and HuBERT, and ultimately adopted Whisper, as it not only demonstrates relatively accurate ASR performance for Korean, but also tends to produce orthographic outputs that more closely reflect actual speech variations rather than canonical forms.

In this study, we fine-tune Whisper-medium to produce pronunciation-oriented outputs for MDD. Since Whisper is designed to generate orthographic word sequences, we convert its labels into pronunciation-based forms using a Korean G2P process that applies phonological rules (e.g., liaison, liquid assimilation, nasalization). For example, ‘ 좋겠다’ (/joh-get-da/, “would be good”) is converted to ‘조켄따’ (/jo-kyeot-tta/), and ‘막내’ (/mak-nae/, “the youngest”) becomes ‘망내’ (/mang-nae/).

The G2P conversion serves three key purposes. First, it aligns training targets and outputs in a pronunciation-based notation that reflects phonological alternations. Second, it reduces Whisper’s contextual bias toward standard spellings by training on pronunciation targets. Finally, it better accommodates accent and pronunciation variation for non-native speakers.

Recognition units are defined at the syllable level rather than the phoneme level. For Korean, syllable strings are more readable and practical for learner-facing feedback, while still capturing pronunciation deviations. For example, if a non-native speaker pronounces “학교” (school, standard: /hak-kkyo/ 학교) as /ha-kyo/, the label is recorded as ‘하교’ to match the actual pronunciation, rather than using decomposed phonemes (‘ㅎ’, ‘ㅏ’, ‘ㄱ’, ‘ㅛ’), which are difficult to read, or the standard form “학교”, which does not reflect actual pronunciation. For evaluation, syllable outputs are converted to phoneme sequences (e.g., ‘하교’ → ‘ㅎ’, ‘ㅏ’, ‘ㄱ’, ‘ㅛ’) to calculate PER.

This choice is also compatible with Whisper’s tokenizer because byte-level BPE (byte-pair encoding) covers the full Unicode Hangul Syllables block U+AC00–U+D7AF [40], allowing any realized L2 syllable form to be represented as valid tokens [41, 42]. This exhaustive coverage ensures that even non-standard L2 pronunciations can be represented as valid tokens.

Moreover, Korean coda neutralization restricts syllable-final realizations to a small set of consonants [25,26], so phoneme substitutions or deletions by L2 learners are still realized as well-formed syllables within the Hangul block. Overall, syllable-level modeling preserves readable feedback and avoids phoneme-to-syllable reconstruction while allowing phoneme-level evaluation.

Based on training data characteristics, different preprocessing strategies are applied to fine-tune the Whisper-medium model and refine pseudo-labels. Through iterative fine-tuning, pseudo-labels become more accurate in capturing actual pronunciations, leading to improved MDD performance.

3.2. Training Data Sources and Characteristics

We initially conducted label refinement using several non-native Korean speech datasets that are provided by the AI-Hub. (This study used datasets from “The Open AI Dataset Project (AI-Hub, S. Korea)”. All data information can be accessed through “AI-Hub (www.aihub.or.kr)”. We used “the spoken Korean voice of native Asian speakers” (<https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71479>), “the spoken Korean

voice of native Chinese and Japanese speakers” (<https://aihub.or.kr/aihub-data/data/view.do?dataSetSn=71490>), “the spoken Korean voice of native European speakers” (<https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71489>), and “the spoken Korean voice of native English speakers” (<https://aihub.or.kr/aihub-data/data/view.do?dataSetSn=71469>) as the non-native dataset (named AIHub-L2A below) and “News Transcript and Anchor Voice Data” (<https://aihub.or.kr/aihub-data/data/view.do?dataSetSn=71557>) as the native data (named AIHub-news below) provided by AI-Hub. This initial refinement produced a high-performance model with labels that reflect actual pronunciation. To enhance objectivity and data diversity, we subsequently expand the corpus by incorporating “Establishing Foreign Speakers’ Voice Data of Korean for Artificial Intelligence Learning” (<https://aihub.or.kr/aihub-data/data/view.do?dataSetSn=505>) (named AIHub-L2B below). The final training corpus consists of three distinct components: two non-native Korean speech datasets (AIHub-L2A and AIHub-L2B) and one native Korean speech dataset (AIHub-news). Although both datasets were obtained from AIHub, they differ in several key aspects. The AIHub-L2A and AIHub-L2B datasets were collected by different organizations, and speech data was recorded on different equipment under varying environmental conditions. These differences in data collection protocols introduce natural acoustic variability between the two datasets, which contributes to a more diverse and comprehensive evaluation setting. The AIHub-news data is used to accurately obtain all phonetic information that appears in Korean pronunciation. Since the news dataset contains standard pronunciations from professional Korean news anchors, its labels are directly generated by converting transcripts to pronunciation forms using the G2P tool, without requiring model-based pseudo-labeling.

3.2.1. AIHub-L2A Dataset Characteristics and Limitations

The AIHub-L2A dataset contains 1.39 million utterances with durations ranging from 2 to 9 s. The data comprises 70% isolated words and 30% short sentences, covering 225 unique scripts read by different speakers. Speakers are categorized into four regional groups (Chinese/Japanese, European, Asian except Chinese and Japanese, English) representing 70 native languages including English, Chinese, Japanese, Vietnamese, and others.

The dataset exhibits severe label imbalance. For example, in an initial filtering of 680,000 samples, 31,086 unique pseudo-labels were generated for the 225 unique scripts. The script “사과” (apple) generated 27 different pronunciation variants. Among these, the correct pronunciation /사과/ (sa-gwa) appeared 5,727 times, accounting for 92.6%, while other variants included /사가/ (sa-ga), /사와/ (sa-wa), and other mispronunciations. In contrast, the script “창문이 열렸네요” (The window is open) generated 789 different predictions, with the most frequent prediction appearing only 550 times and the top 10 predictions covering merely 49.6% of instances.

This imbalance stems from pronunciation difficulty variation. Simple words like “사과” showed relatively consistent pronunciations across speakers, resulting in high-frequency labels. Complex sentences produced diverse actual pronunciations, leading to lower repetition rates per label. While pronunciation diversity benefits learning, high-frequency labels risk biasing the model toward specific pronunciation patterns.

Additionally, the dataset suffers from cross-contamination due to the recording environment. Since data was collected during speaking exams with multiple speakers in adjacent seats, recordings often captured speech from neighboring speakers. For example, for the script “참 잘 했구나” (Well done), Speaker A’s intended pronunciation /참잘했구나/ (/cham-jal-haet-kku-na/) was followed by Speaker B’s speech

/얼구미안조아네/ (/eol-gu-mi-an-jo-a-ne/) captured in the silence period, resulting in the combined audio /참잘했꾸나얼구미안조아네/ (/cham-jal-haet-kku-na-eol-gu-mi-an-jo-a-ne/). When models repeatedly trained on the same script, they overfitted to produce only /참잘했꾸나/ (/cham-jal-haet-kku-na/), generating incorrect labels that failed to reflect the actual audio content. This contamination issue was widespread and required systematic removal of affected samples.

To address label imbalance and audio contamination, duplicate labels were removed based on native language groups rather than regional categories, and noise removal was applied to eliminate contaminated recordings.

3.2.2 AIHub-L2B dataset characteristics and limitations

The AIHub-L2B dataset contains 485,000 utterances with durations ranging from 5 to over 40 s, consisting entirely of sentence-level recordings from speakers representing 62 native languages including Vietnamese, English, Japanese, Chinese, Thai and others.

A key characteristic of this dataset is the uneven distribution of speaker proficiency levels. Recording scripts were randomly assigned without considering speakers' Korean language abilities, resulting in low-proficiency speakers reading complex sentences with high pronunciation difficulty. This led to extreme duration variation for identical scripts. For example, the script “한국의 짬 문화는 여전히 우리나라의 자부심이며, 한국인은 전통을 지켜 내기 위해 현재도 꾸준히 노력하고 있다” (Korean ssam culture remains our nation's pride, and Koreans continue to make steady efforts to preserve this tradition) showed a mean duration of 12.90 s and median of 12.93 s across 674 speakers, but the range spanned from 7.90 to 36.72 s. Recordings exceeding 35 s likely contained disfluencies such as repetitions, pauses, and restarts that negatively impact phoneme recognition accuracy.

This study focuses on data from a specific language, Korean, but proposes a method for efficiently refining training data, which is expected to be applicable to building MDD and speech recognition systems for minority languages with limited training data.

3.3. Label Refinement and Data Preprocessing Strategy

3.3.1. Multi-Stage Iterative Label Refinement

Our approach uses an offline multi-stage refinement strategy to progressively improve label quality. Figure 1 illustrates our multi-stage iterative refinement procedure using Whisper-medium as the base model. In each fine-tuning round n , we fine-tune the base model on different datasets or combinations (Data1- n .0, Data2- n .0, ..., Data d - n .0) to obtain multiple models (Model1- n .0, Model2- n .0, ..., Model d - n .0).

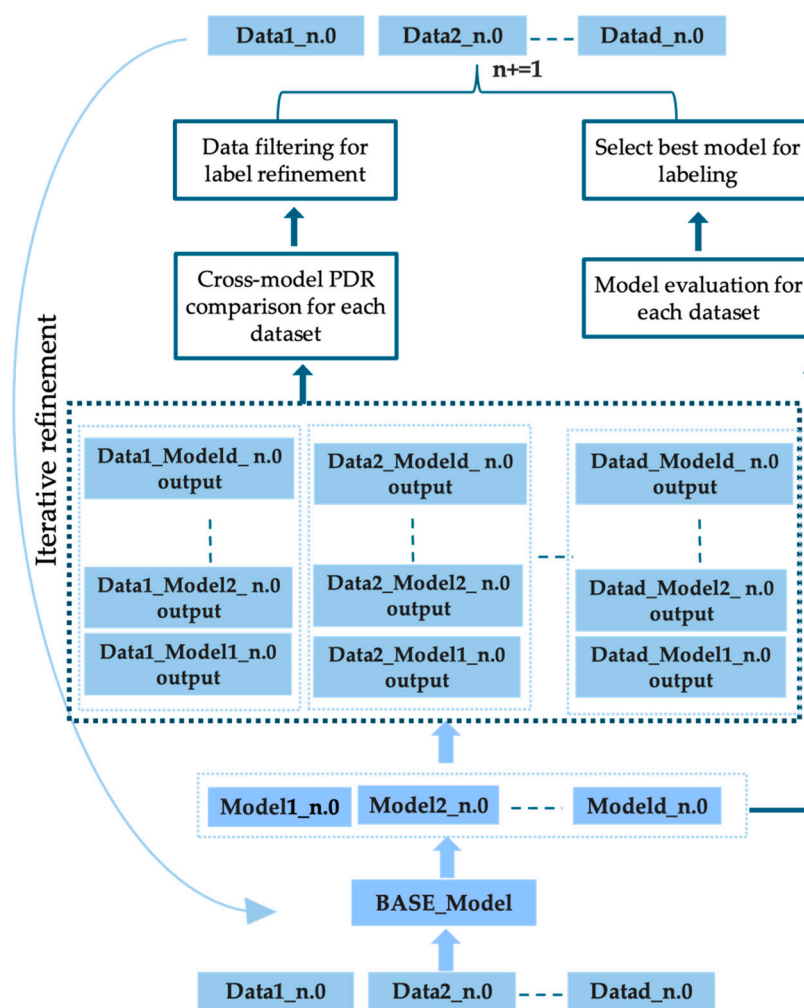


Figure 1. Multi-stage iterative label refinement framework (n: stage number starting from 0; d: number of datasets).

As shown in Algorithm 1, each refinement stage, we first filter candidate utterances using cross-model agreement. Multiple Whisper models fine-tuned on different datasets or dataset combinations generate pronunciation-based sequences, and we retain only samples with high agreement measured by PDR, which is analyzed in detail in Section 3.3.3. We then select the best-performing model based on PER on a manually annotated evaluation set and use its outputs as pseudo-labels. Finally, to reduce model-specific artifacts from relying on a single architecture, we compute PDR between the generated labels and canonical phoneme references, which are used only as a reference rather than ground truth, and filter labels with excessive deviations. The resulting pseudo-labeled data is used for the next stage of fine-tuning.

Algorithm 1. Iterative label refinement process.

Algorithm 1: Label refinement process

Input: Preprocessed audio dataset D , fine-tuned models $M_1, M_2, \dots, M_n$, cross-model PDR threshold θ_1 , output-canonical label PDR threshold θ_2 , CER threshold θ_3 , canonical labels $L_{canonical}$, content labels $L_{content}$

Initial Filtering: CER-based filtering (model output vs. $L_{content}$)

- For each sample $x_i \in D$, obtain transcriptions from baseline Whisper-medium model
- Compute CER between model output and $L_{content}$
- If $CER > \theta_3$, discard sample x_i
- Retain samples where $0 < CER \leq \theta_3 \rightarrow$ produce D_0

Iterative Label Refinement (repeat for each stage $t = 1, 2, \dots, T$):

Step 1: PDR-based filtering (cross-model outputs)

- For each sample $x_i \in D_0$, obtain outputs from all models $M_1, \dots, M_n$
- Compute PDR between each pair of model outputs
- If $PDR > \theta_1$, discard sample x_i
- Retain samples where $PDR \leq \theta_1 \rightarrow$ produce D_t^1

Step 2: PER-based best model selection

- Evaluate all models $M_1, \dots, M_n$ on evaluation set
- Select best-performing model M_t^*

Step 3: PDR-based filtering (model output vs. $L_{canonical}$)

- For each sample $x_i \in D_t^1$, compute PDR between M_t^{*} 's output and $L_{canonical,i}$
- If $PDR > \theta_2$, discard sample x_i
- Retain samples where $0 < PDR \leq \theta_2 \rightarrow$ produce D_t

Step 4: Fine-tuning

- Fine-tune the baseline Whisper-medium model using D_t
- Update models $M_1, \dots, M_n$ for the next stage

End Iterative Loop

Output: $D^* = D_T$ with labels from M_T^*

The implementation details, including specific dataset configurations, threshold settings for each iteration, and quantitative results of the refinement process, are presented in Section 4.1.1.

3.3.2. Data Preprocessing Based on Different Dataset Characters

The distinct characteristics of the AIHub-L2A and AIHub-L2B datasets require tailored preprocessing strategies. For AIHub-L2A, preprocessing focuses on addressing label imbalance and cross-contamination from the recording environment. For AIHub-L2B, strategies target duration variation caused by proficiency differences and label diversity from pronunciation inconsistencies. This section presents the dataset-specific quality control methods applied to optimize training data.

First, quantitative acoustic metrics were introduced to systematically filter low-quality recordings [43,44]. Signal-to-Noise Ratio (SNR) is used as the primary quality indicator, with samples below 10 dB excluded as noise data. To address cross-contamination

from neighboring speakers captured during recording, Root Mean Square (RMS) energy filtering is applied, removing samples with RMS below 0.01, where background speech from adjacent speakers reduces RMS values. These acoustic thresholds effectively eliminate recordings that could introduce training instability and recognition errors.

Second, we revise duplicate management from a region-based scheme to a mother tongue-based scheme, denoted as MT. The region-based approach removed too many samples, reducing data utilization and potentially limiting generalization. MT-level control retains speakers with different native languages within the same region, preserving acoustic diversity while mitigating label imbalance.

Duplicate criteria are dataset-specific. For AIHub-L2A, which contains short utterances, we control duplicates by pseudo-labels and allow up to four occurrences of each pseudo-label per MT group to balance label diversity and repetition. For AIHub-L2B, where label diversity is high and context-dependent overfitting is a concern, we restrict each utterance to one occurrence per MT group.

Third, duration-based filtering for AIHub-L2B is conducted. Due to the extreme duration variation for identical scripts in the AIHub-L2B dataset, statistical analysis is conducted on recording durations for all scripts, calculating mean and median values. Based on this analysis, only recordings under 15 s are selected for training. This filtering removes excessively slow or error-prone samples with frequent disfluencies, retaining fluent and accurate pronunciations for model training. The filtering reduces the dataset from 485,615 to 410,299 samples, with script types decreasing from 2639 to 2624. Despite this reduction, the dataset maintains sufficient scale and script diversity while ensuring training data quality.

3.3.3. Metrics Definition and Formulation

We employ three evaluation metrics at different pipeline stages, all based on Levenshtein distance [45]: Character Error Rate (CER), PDR, and PER. Each metric calculates the minimum number of substitutions (S), deletions (D), and insertions (I) required to transform a hypothesis into a reference, normalized by the reference length (N):

$$\text{Error Rate} = (S + D + I)/N. \quad (1)$$

CER is used in the initial data selection stage to measure the character-level mismatch between the model prediction and the canonical text before G2P conversion, serving as a proxy for pronunciation-driven deviation. PDR is used for cross-model validation to quantify sequence-level discrepancies between two models' outputs, as well as between a model's pronunciation outputs and canonical pronunciation labels. PER is reserved for final evaluation on manually annotated evaluation datasets, measuring absolute phonetic accuracy.

These metrics enable systematic evaluation across the pipeline: CER guides initial data filtering, PDR facilitates unsupervised quality control via cross-model consensus, and PER provides the ultimate performance benchmark against human annotations. Further preprocessing is conducted on the AIHub-L2A and AIHub-L2B datasets, with strategies tailored to each dataset's characteristics. The final dataset configurations and the performance of models fine-tuned on each dataset variant are presented in Section 4.

4. Experiments and Results

This section presents the datasets constructed through the preprocessing strategies described in Section 3 and evaluates their impact on phoneme recognition performance.

4.1. Fine-Tuning Dataset Construction Through Label Refinement and Data Preprocessing Strategies

The label refinement and data preprocessing strategies described in Section 3 generated multiple fine-tuning dataset variants from the AIHub-L2A and AIHub-L2B corpora. This section details the composition of these variants, including sample sizes, label distributions, and the preprocessing configurations applied to create each dataset.

4.1.1. Datasets from Multi-Stage Iterative Refinement

We applied the multi-stage iterative refinement framework (Section 3.3.1) to AIHub-L2A. Because this dataset provides only canonical text labels rather than pronunciation-faithful transcripts, we first generated pseudo-labels with Whisper and computed CER between Whisper’s hypothesis and the canonical label.

To select data that is informative for L2 acoustic modeling, we conducted a small qualitative check by listening to representative samples and inspecting Whisper outputs. We found that CER = 0% typically corresponds to native-like productions that closely match the canonical text and thus provides little L2-specific variation for accent modeling. At the other extreme, samples with very high CER (above 40%) were usually recognition failures rather than meaningful pronunciation deviations, often caused by poor audio quality, background noise, or severe disfluencies. Based on this empirical observation, we retained utterances within $0\% < \text{CER} \leq 40\%$ to initialize pseudo-labels with a controlled noise level. In the subsequent fine-tuning stages, canonical label PDR filtering was additionally applied as described in Section 3.3.1, with a threshold of $0\% < \text{PDR} \leq 40\%$.

The labels of the selected 68,000 utterances were converted into pronunciation-based labels using G2P for the first refinement stage. We evaluated several fine-tuning configurations. L2-68k means models using only the 68k AIHUB-L2A dataset. News-L2A-68k means sequential fine-tuning on 12k AIHub-news data followed by the 68k AIHUB-L2A dataset. L2A-68k-News refers to sequential fine-tuning on 68k AIHub-L2A data followed by 12k AIHub-news data. And AIHub-L2A-68k refers to joint fine-tuning on the merged set of 68k AIHUB-L2A data and 12k AIHub-news data. Because the model fine-tuned on the merged set of AIHub-L2A and the 12k AIHub-news data performed best, but our initial experiments used only AIHub-L2A as the L2 data, we did not use this best-performing model for cross-model validation.

For subsequent refinement stages, we performed cross-model validation using three model variants. By calculating pairwise PDR across model outputs and applying stage-specific thresholds, we selected high-confidence labels for each iteration. Table 2 presents the progressively refined datasets and models.

Table 2. Cross-model validation for iterative refinement on AIHub-L2A corpus.

Stage	Models for Data Filtering			PDR (sent)	Data Size (k)	Best Model for Pseudo-labels
	Model 1	Model 2	Model 3			
1st		Whisper-baseline		40%	68	AIHub-L2A-68k
2nd	L2-68k	L2A-68k-News	News-L2A-68k	15%	22	AIHub-L2A-22k
3rd	L2-22k	L2A-22k-News	News-L2A-22k	10%	17	AIHub-L2A-17k
4th	L2-17k	L2A-22k-News	News-L2A-17k	0%	15	AIHub-L2A-15k
5th	L2-15k	L2A-22k-News	News-L2A-15k	0%	16	AIHub-L2A-16k

The initial experiment, fine-tuned on the merged 16k dataset and AIHub-news data (AIHub-L2A-16k), achieved the best performance on AIHub-L2A. To improve objectivity and data diversity beyond the single AIHub-L2A source, we incorporated the AIHub-L2B

dataset, applying the same iterative refinement framework with adaptations for this new corpus.

Since AIHub-L2B lacks pronunciation-based labels, we generated initial pseudo-labels using the AIHub-L2A-16k model. For the first refinement stage, cross-model validation employed two architecturally diverse models—Whisper-medium and HuBERT-large, both fine-tuned on the combined AIHub-L2A and AIHub-news dataset—to ensure label reliability through architectural consensus.

All experiments were conducted using the HuggingFace Transformers library. We employed Whisper-medium as the base architecture. For the optimization, the AdamW optimizer was used with a weight decay of 0.01 and a linear learning rate warm-up ratio of 0.1. Based on empirical experiments, the learning rate was set to 1e-6 for AIHub-L2A and 5e-6 for AIHub-L2B. The models were trained for 10 epochs, with a per-device batch size of 4. A fixed seed of 42 was maintained for reproducibility. Decoding was performed using greedy search (num_beams=1) with a maximum generation length of 225 tokens, following the default HuggingFace Whisper generation configuration.

Subsequent stages used progressively refined models trained on updated AIHub-L2A and AIHub-L2B data (each combined with AIHub-news), with PDR thresholds tightened at each iteration. Table 3 summarizes the cross-validation configuration across refinement stages. Following the preprocessing strategy in Section 3.3.2, only recordings under 15 s were retained to filter disfluent speech.

Table 3. Cross-model validation for iterative refinement on AIHub-L2B corpus.

Stage	PDR	Models for Data Filtering	
		Model 1	Model 2
1st	30%	Hubert-AIHub-L2A-16k	AIHub-L2A-16k (label)
2nd	5%	AIHub-L2B-1st-pdr-30% (label)	AIHub-L2A-16k
3rd	5%	AIHub-L2B-2nd-pdr-5% (label)	AIHub-L2A-47k
4th	5%	AIHub-L2B-3rd-pdr-2% (label)	AIHub-L2A-47k-two-stage

4.1.2. Datasets from Data Preprocessing

For the AIHub-L2A dataset, preprocessing was conducted based on the AIHub-L2A-16k dataset using region-based duplicate management with a four-occurrence limit. Table 4 summarizes the resulting preprocessing variants and their characteristics.

Table 4. AIHub-L2A dataset variants.

AIHub-L2A	Files	Balance	Max Duplication	Noise Removal
AIHub-L2A-16k	16k	region (4)	4	X
AIHub-L2A-680k	680k	X	5727	X
AIHub-L2A-93k	93k	MT (4)	61	X
AIHub-L2A-47k	47k	MT (4)	34	O
AIHub-L2A-19k	19k	MT (4)	1	O

The initial AIHub-L2A-680k dataset was constructed by reprocessing the AIHub-L2A corpus using the optimal model trained in the initial study with revised output processing. Cross-model validation employed two independently trained models: one fine-tuned on AIHub-L2A data and another on the newly acquired AIHub-L2B data. Since these models were trained on different datasets, this approach provided improved

objectivity in label validation compared to the previous method where all models shared the same training data.

The filtering procedure selected samples where both models produced identical predictions. These selected samples were then compared against standard pronunciations, retaining only those with PDR below 50%. This process yielded 680,000 samples with pseudo-labels reflecting cross-model agreement.

To address the severe label imbalance in the AIHub-L2A-680k dataset, MT-based instead of region-based duplicate control was implemented. Each unique predicted label was limited to four occurrences per native language group, reducing the dataset from 680,000 to 93,000 samples (AIHub-L2A-93k). Noise removal was then applied following the procedures described above. This generated the AIHub-L2A-47k dataset. To examine the impact of duplicate removal stringency, an additional variant (AIHub-L2A-19k) was created with complete deduplication, limiting each label to one occurrence per native language group.

Analysis of generated pseudo-labels for “사과” /sa-gwa/ (apple) revealed distinct patterns across datasets. The 680k dataset showed extreme imbalance, with maximum duplication reaching 5,727 occurrences, while the maximum duplication of 93k and 47k datasets were reduced to 41 and 22, respectively. The 19k dataset achieved nearly single-label distribution with minimal redundancy. Performance comparisons across these variants are presented in Section 4.2.

The training datasets are presented in Tables 4 and 5. While pseudo-labeled training data is suitable for fine-tuning, objective evaluation requires manually transcribed data. The evaluation dataset construction is described in detail in Section 4.2.

Table 5. AIHub-L2B dataset variants.

AIHub-L2B Dataset	Stage	PDR Threshold	Sample Count
AIHub-L2B-15s	Initial	N/A	410,299
AIHub-L2B-1st-pdr-30%	1st	≤ 30%	71,569
AIHub-L2B-2nd-pdr-5%	2nd	≤ 5%	26,193
AIHub-L2B-3rd-pdr-5%	3rd	≤ 5%	55,616
AIHub-L2B-4th-pdr-5%	4th	≤ 5%	108,795

4.2. Evaluation Data Construction

To ensure a reliable benchmark for performance assessment, the non-native evaluation datasets were constructed through a strict dual-annotator verification protocol. The annotation work was carried out by two native Korean researchers specializing in linguistics.

A unanimous consensus policy was enforced for sample inclusion: each utterance was independently transcribed through direct auditory inspection by both annotators. In cases of discrepancy, a collaborative review was conducted through repeated listening until a common phonetic representation was agreed upon. If a persistent disagreement or phonetic ambiguity occurred, the sample was strictly excluded and replaced with a new candidate to maintain the integrity of the reference labels. While this exhaustive verification process prioritized label precision over dataset volume, it ensures that the resulting evaluation set serves as a high-confidence ‘gold standard’ for PER analysis.

The AIHub-L2A-test dataset comprises 800 samples. The dataset is structurally balanced across four geographical regions (200 samples each), with equal distribution between short utterances (≤4 syllables) and longer sentences (≥5 syllables) in each region.

Importantly, the dataset retains real-world acoustic challenges including background noise, cross-contamination, and disfluencies, enabling evaluation under realistic conditions.

The AIHub-L2B-test dataset contains 1200 samples balanced across six native language groups: Chinese, Japanese, Vietnamese, Thai, English, and other languages (200 samples each). Labels were generated through the same cross-verification procedure. The news-test dataset comprises 400 samples with direct G2P conversion.

These three evaluation datasets collectively assess the performance of Whisper-medium models fine-tuned on the refined datasets described in Section 4.1.1 and 4.1.2, validating both pseudo-label quality and the effectiveness of preprocessing strategies.

4.3. Experiment Results and Analysis

This section presents experimental results evaluating the impact of dataset refinement and preprocessing strategies on phoneme recognition performance measured by PER. All experiments employed the Whisper-medium architecture as the base model.

4.3.1. AIHub-L2A Dataset Experiment Results and Analysis

As described in Section 3, only the AIHub-L2A dataset was used for iterative label refinement in the initial stage. The learning rate was set to $1e-5$, and the results are shown in Table 6. As the evaluation dataset contains only AIHub-L2A data, results show progressive accuracy improvements on AIHub-L2A with each refinement stage. However, AIHub-news performance remained inconsistent, without clear improvement trends. This indicates limited generalization, leading us to add the AIHub-L2B dataset.

Table 6. Performance (PER) of AIHub-L2A models across iterative refinement stages.

Model	AIHub-L2A	AIHub-News
Whisper-medium (baseline)	9.03%	3.69%
AIHub-L2A-68k	7.47%	1.55%
AIHub-L2A-22k	6.51%	1.04%
AIHub-L2A-17k	5.60%	1.36%
AIHub-L2A-15k	5.47%	0.55%
AIHub-L2A-16k	5.18%	1.50%

To compare with the baseline, we also evaluated the Whisper outputs by converting the recognized text into pronunciation-based labels using the G2P tool, following the same procedure used to generate our initial pronunciation-based labels. Table 7 presents results after incorporating the AIHub-L2B dataset and applying further preprocessing to the best-performing model trained on AIHub-L2A data. Reducing the learning rate from $1e-5$ to $1e-6$ improved the AIHub-L2A-16k model from 8.88% to 7.48% PER on AIHub-L2A evaluation data, with 8.72% PER on the unseen AIHub-L2B dataset (17.2% relative improvement). Based on these results, a learning rate of $1e-6$ was adopted for all subsequent experiments. To assess whether the performance differences between systems are statistically significant, we employ the Matched Pairs Sentence-Segment Word Error (MAPSSWE) test [46]. MAPSSWE is a non-parametric significance test that compares the word error rates of two systems by examining error counts over matched sentence segments, where segments are defined to approximately satisfy the independence assumption. The significance levels are denoted as follows, where “*” indicates $p \leq 0.05$ (significant), “**” indicates $p \leq 0.01$ (highly significant), and “***” indicates $p \leq 0.001$ (extremely significant), while entries without any marker indicate no statistically significant difference.

Table 7. Performance (PER, p -value) by data preprocessing method and learning rate.

Model	All	AIHub-L2A	AIHub-L2B	AIHub-news
Whisper-medium (baseline)	10.66%	9.03%	11.23%	3.69%
AIHub-L2A-16k	8.88% (***)	5.18% (***)	10.53% (***)	1.22% (***)
AIHub-L2A-16k-lr-1e-6	7.48% ($p = 0.015$ *)	5.22% ($p = 0.2$)	8.72% ($p = 0.008$ **)	1.34% ($p = 0.2$)
AIHub-L2A-93k	8.78%	5.16%	12.70%	2.05%
AIHub-L2A-47k	7.35% (***)	4.87% ($p = 0.029$ *)	8.59% (***)	1.34% ($p = 0.84$)
AIHub-L2A-47k_two_stage	7.44% ($p = 0.09$)	4.97% ($p = 0.082$)	8.62% ($p = 0.71$)	1.44% ($p = 0.21$)
AIHub-L2A-19k	7.54%	5.12%	8.79%	1.45%

“*” indicates $p \leq 0.05$ (significant), “***” indicates $p \leq 0.01$ (highly significant), and “****” indicates $p \leq 0.001$ (extremely significant).

As shown in Table 7, the baseline model performs poorly on all datasets. Changing the learning rate improves performance on the AIHub-L2B evaluation data without affecting the other two datasets. The p -values between the AIHub-L2A-47k model and the AIHub-L2A-16k model indicate that further refinement of the dataset through noise removal and modified deduplication strategies leads to significant improvements on both L2 evaluation datasets, while leaving performance on the L1 dataset unchanged. Since the AIHub-L2A data includes word-level utterances (with fewer than four syllables), we divided the data into words and sentences and conducted two-stage fine-tuning by first fine-tuning the baseline model on words and then on sentences. However, the results show that this approach does not yield any improvement.

4.3.2. AIHub-L2B Dataset Experiment Results and Analysis

This section presents experimental results evaluating models trained on AIHub-L2B dataset variants. The analysis examines two aspects: the impact of iterative label refinement across stages and the effect of varying PDR thresholds on data selection. Model performance is assessed across all three test datasets to examine performance consistency across different evaluation conditions.

As shown in Table 8, iterative label refinement progressively improves overall performance. At the first stage, the model achieves the best performance on AIHub-L2A, but fails on the other two evaluation datasets. We hypothesize that this is because the labels used for fine-tuning at the first stage are derived from the AIHub-L2A-16k model trained on the AIHub-L2A dataset. In subsequent stages, the model progressively improves on AIHub-L2B without affecting performance on the AIHub-L2A evaluation data. This demonstrates that iterative label refinement leads to improvement in model performance, confirming that the labels have been successfully refined.

Table 8. Performance (PER, p -value) of AIHub-L2B models across iterative refinement stages with PL-based duplicate control.

Model	All	AIHub-L2A	AIHub-L2B	AIHub-news
AIHub-L2B-1st-pdr-30%	8.53%	5.17%	9.95%	2.21%
AIHub-L2B-2nd-pdr-5%	8.15%	5.99%	9.52%	1.04%

	(***)	(***)	(***)	(***)
AIHub-L2B-3rd-pdr-5%	7.60%	5.78%	8.78%	1.40%
	(***)	($p = 0.33$)	(***)	(***)
AIHub-L2B-4th-pdr-5%	6.99%	5.86%	7.98%	1.50%
	(***)	($p = 0.97$)	(***)	($p = 0.30$)

“***” indicates $p \leq 0.001$ (extremely significant).

These results confirm that iterative refinement progressively improves label quality and model performance, though initial-stage validation model selection can introduce temporary dataset-specific biases that subsequent refinement stages successfully mitigate.

From stage 3 onward, as the dataset size increased, model performance was compared across tightened PDR thresholds. Table 9 presents model performance across PDR. As shown in Table 9, at the third stage, reducing the PDR value from 0.05 to 0.03 leads to significant improvements across all evaluation datasets. However, further reduction to 0.02 does not yield significant gains on the L2 evaluation data, though it does not negatively affect overall performance either. At the fourth stage, reducing the PDR value from 0.05 to 0.02 also results in meaningful improvement. However, possibly due to the larger data size, the model improves on AIHub-L2B but shows no effect on the other two datasets or overall performance, suggesting that a moderately sized dataset with well-refined labels may cause overfitting to the same dataset.

Table 9. Impact of PDR threshold on model performance.

Model	Size	All	AIHub-L2A	AIHub-L2B	AIHub-news
AIHub-L2B-3rd-pdr-5%	56k	7.60%	5.78%	8.78%	1.40%
AIHub-L2B-3rd-pdr-3%	36k	7.29%	5.60%	8.43%	1.25%
		(***)	($p = 0.045 *$)	(***)	($p = 0.0016 **$)
AIHub-L2B-3rd-pdr-2%	10k	7.26%	5.43%	8.44%	1.08%
		($p = 0.812$)	($p = 0.201$)	($p = 0.397$)	($p = 0.025 *$)
AIHub-L2B-4th-pdr-5%	109k	6.99%	5.86%	7.98%	1.50%
AIHub-L2B-4th-pdr-3%	56k	6.87%	5.53%	7.90%	1.29%
		(***)	(***)	($p = 0.01 **$)	($p = 0.01 **$)
AIHub-L2B-4th-pdr-2%	32k	6.87%	5.70%	7.83%	1.32%
		($p = 0.149$)	($p = 0.068$)	($p = 0.027 *$)	($p = 0.789$)

“*” indicates $p \leq 0.05$ (significant), “**” indicates $p \leq 0.01$ (highly significant), and “***” indicates $p \leq 0.001$ (extremely significant).

4.3.3. Verification Based on Actual Pronunciation Labels

To validate the effectiveness of data refinement strategies, model outputs were evaluated against two reference types: manually transcribed actual pronunciation labels and standard pronunciation labels. The results are shown in Table 10.

Table 10. Performance comparison based on actual vs. standard pronunciation labels.

Model	AIHub-L2A Actual	AIHub-L2A Standard	AIHub-L2B Actual	AIHub-L2B Standard
AIHub-L2A-93k	5.16%	8.78%	12.70%	15.88%
AIHub-L2A-47k	4.87%	8.74%	8.59%	11.80%

AIHub-L2B-3rd-pdr-2%	5.43%	9.35%	8.44%	11.76%
AIHub-L2B-4th-pdr-3%	5.53%	9.46%	7.90%	11.14%

All models demonstrated substantially lower PER on actual pronunciation labels compared to standard pronunciation labels, confirming that models trained through iterative refinement predict actual non-native acoustic realizations rather than standard forms. This alignment is critical for non-native MDD, as models can accurately recognize learner-specific phonetic variability.

4.3.4. Mispronunciation Detection and Diagnosis Performance

Based on the label refinements, models fine-tuned on pseudo-labels have shown significant improvement and are considered suitable for MDD. In this section, the Whisper-Baseline model, the AIHub-L2A-47k model with the best performance on the AIHub-L2A evaluation data, and the AIHub-L2B-4th-pdr-2% model with the best performance on the AIHub-L2B evaluation data are selected for MDD evaluation. The evaluation metrics are calculated based on the criteria defined in [34].

The evaluation framework classifies each phone into four categories, based on the agreement between the annotation and prediction: True Accept (TA) and True Reject (TR) represent correct detection of well-pronounced and mispronounced phones, respectively, while False Reject (FR) and False Accept (FA) capture the two types of detection errors. TR phones are further analyzed to distinguish between Correct Diagnosis (CD) and Error Diagnosis (ED), depending on whether the specific type of mispronunciation is correctly identified. And the evaluation metrics are computed as follows.

$$\text{FRR} = \text{FR}/(\text{TA}+\text{FR}), \quad (2)$$

$$\text{FAR} = \text{FA}/(\text{FA}+\text{TR}), \quad (3)$$

$$\text{DER} = \text{ED}/(\text{CD}+\text{ED}), \quad (4)$$

According to Table 11, both models fine-tuned on refined pseudo-labels tend to make more conservative acceptance decisions, but still developed compared with the Whisper-Baseline. The relatively low DER values indicate that our models perform well in diagnosing the specific type of mispronunciation on the AIHub-L2A evaluation data, with AIHub-L2A-47k achieving a DER below 10%. The differences between all compared models are statistically significant, with p -values below 0.001.

Table 11. MDD performance on AIHUB-L2A data.

Models	FRR	FAR	DER	Detection Accuracy	PER
Whisper-Baseline	2.62%	48.20%	23.21%	96.10%	9.03%
AIHub-L2B-4th-pdr-2%	1.59%	32.13%	12.09%	96.33%	5.70%
AIHub-L2A-47k	1.37%	29.33%	8.59%	97.85%	4.87%

Table 12 shows a similar trend to Table 10, though with overall lower performance, suggesting that the AIHub-L2B evaluation data presents a greater challenge for MDD compared to the AIHub-L2A data.

Table 12. MDD performance on AIHUB-L2B data.

Models	FRR	FAR	DER	Detection Accuracy	PER
Whisper-Baseline	4.38%	53.63%	26.95%	94.68%	11.23%
AIHub-L2A-47k	3.18%	57.82%	25.01%	95.88%	8.59%
AIHub-L2B-4th-pdr-2%	2.85%	57.77%	23.61%	95.60%	7.83%

Overall, the results demonstrate that models fine-tuned on refined pseudo-labels achieve consistent improvements in both detection and diagnosis performance across both L2 evaluation datasets, with all differences being statistically significant.

5. Conclusions

This study fine-tuned Whisper to produce pronunciation-based sequences from non-native Korean speech, improving MDD performance. Iterative label refinement improved data quality, which in turn enhanced both pronunciation recognition and mispronunciation detection for non-native Korean learners. The use of two distinct L2 datasets provides additional evidence for the consistency of our evaluation across different data conditions.

Application of G2P-based pronunciation rules converted orthographic text to actual pronunciations, reducing Whisper’s contextual dependency and enabling effective pronunciation-based sequence output.

Dataset-specific preprocessing, including noise removal, duration-based filtering, and deduplication, was applied to improve data quality. Noise removal contributed to improved performance consistency across the evaluation datasets, while deduplication ensured a balanced label distribution and contributed to overall performance gains.

This study has several limitations requiring further investigation. First, the experiments were limited to the Whisper-medium architecture and should be validated on other models. Second, incorporating more fine-grained statistical analyses across speaker characteristics, utterance lengths, L1 backgrounds, and acoustic conditions would provide deeper insight into whether the observed improvements reflect robust methodological gains or dataset-specific effects. Third, further analysis of FRR, FAR, and DER patterns is needed to guide future improvements.

Future work should address these limitations by incorporating speaker characteristics and acoustic conditions into curriculum learning, exploring alternative pseudo-labeling methods, and investigating other architectures suited for non-native phoneme recognition.

Author Contributions: Conceptualization, N.G. and J.-S.P.; methodology, N.G. and J.-S.P.; software, N.G.; validation, N.G., H.-J.N. and J.-S.P.; formal analysis, N.G. and J.-S.P.; resources, N.G., J.W. and D.-G.K.; data curation, J.W. and D.-G.K.; writing—original draft preparation, N.G.; writing—review and editing, J.-S.P.; supervision, J.-S.P.; project administration, J.-S.P.; funding acquisition, J.-S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Hankuk University of Foreign Studies Research Fund, and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2025-24523068).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This research paper used datasets from ‘The Open AI Dataset Project (AI-Hub, S. Korea)’. All data information can be accessed through ‘AI-Hub (www.aihub.or.kr)’.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDD	Mispronunciation Detection and Diagnosis
PER	Phoneme Error Rate
L2	Second Language
SSL	Self-supervised Learning
ASR	Automatic Speech Recognition
G2P	Grapheme-to-Phoneme
PDR	Phoneme Difference Rate
PL	Pseudo-labeling
PR	Phoneme Recognition

References

1. Shin, S.J.; Lee, J.S. Expanding Capacity, Opportunity, and Desire to Learn Korean as a Heritage Language. *Herit. Lang. J.* **2013**, *10*, 357–366. <https://doi.org/10.46538/hlj.10.3.7>.
2. Flege, J.E.; Frieda, E.M.; Nozawa, T. Amount of Native-Language (L1) Use Affects the Pronunciation of an L2. *J. Phon.* **1997**, *25*, 169–186. <https://doi.org/10.1006/jpho.1996.0040>.
3. Cho, T.; Keating, P.A. Articulatory and Acoustic Studies on Domain-Initial Strengthening in Korean. *J. Phon.* **2001**, *29*, 155–190. <https://doi.org/10.1006/jpho.2001.0131>.
4. Tu, M.; Grabek, A.; Liss, J.; Berisha, V. Investigating the Role of L1 in Automatic Pronunciation Evaluation of L2 Speech. *arXiv* **2018**, arXiv:1807.01738. <https://doi.org/10.48550/arXiv.1807.01738>.
5. Kermanshahi, M.A.; Homayounpour, M.M. Improving Phoneme Sequence Recognition Using Phoneme Duration Information in DNN-HSMM. *J. AI Data Min.* **2019**, *7*, 137–147. <https://doi.org/10.22044/jadm.2018.6136.1725>.
6. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Pallett, D.S.; Dahlgren, N.L.; Zue, V.; Fiscus, J.G. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.
7. Fernández, S.; Graves, A.; Schmidhuber, J. Phoneme Recognition in TIMIT with BLSTM-CTC. *arXiv* **2008**, arXiv:0804.3269.
8. Byun, J.; Oh, S. A Representation of Korean Syllable Characters Defined in Hunminjeongeum. In *Advances in Multimedia Modeling, Proceedings of the 13th International Multimedia Modeling Conference, MMM 2007, Singapore, 9–12 January 2007*; Cham, T.-J., Cai, J., Dorai, C., Rajan, D., Chua, T.-S., Chia, L.-T., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 662–670.
9. Koo, M.-W. A Korean speech recognition based on conformer. *J. Acoust. Soc. Korea* **2021**, *40*, 488–495. <https://doi.org/10.7776/ASK.2021.40.5.488>.
10. Kim, S.; Bae, S.; Won, C. Open-Source Toolkit for End-to-End Korean Speech Recognition. *Softw. Impacts* **2021**, *7*, 100054. <https://doi.org/10.1016/j.simpa.2021.100054>.
11. Lee, D.; Lim, M.; Park, H.; Kim, J.-H. LSTM RNN-based Korean Speech Recognition System Using CTC. *J. Digit. Contents Soc.* **2017**, *18*, 93–99. <https://doi.org/10.9728/dcs.2017.18.1.93>.
12. Na, J.; Park, Y.; Lee, B. A Comparative Study on the Biases of Age, Gender, Dialects, and L2 Speakers of Automatic Speech Recognition for Korean Language. In *Proceedings of the 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Macau, China, 3–6 December 2024*; IEEE: New York, NY, USA; pp. 1–6. <https://doi.org/10.1109/APSIPAASC63619.2025.10848815>.
13. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023*; pp. 28492–28518.
14. Choo, M.; O’Grady, W. *The Sounds of Korean: A Pronunciation Guide*; University of Hawaii Press: Honolulu, HI, USA, 2003.
15. Oh, Y.R.; Park, K.; Jeon, H.-B.; Park, J.G. Automatic Proficiency Assessment of Korean Speech Read Aloud by Non-Natives Using Bidirectional LSTM-Based Speech Recognition. *ETRI J.* **2020**, *42*, 761–772. <https://doi.org/10.4218/etrij.2019-0400>.

16. Yang, S.H.; Ryu, H.; Chung, M. A Corpus-Based Analysis of Korean Segments Produced by Chinese Learners. In *Proceedings of the 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Hong Kong, China, 16–19 December 2015; IEEE: New York, NY, USA; pp. 583–586. <https://doi.org/10.1109/APSIPA.2015.7415336>.
17. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the Advances in Neural Information Processing Systems, Virtual*, 6–12 December 2020; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 12449–12460.
18. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. <https://doi.org/10.1109/TASLP.2021.3122291>.
19. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. <https://doi.org/10.1109/JSTSP.2022.3188113>.
20. Yang, S.; Chi, P.-H.; Chuang, Y.-S.; Lai, C.-I.J.; Lakhotia, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.-T.; et al. SUPERB: Speech Processing Universal PERFORMANCE Benchmark. *arXiv* **2021**, arXiv:2105.01051.
21. Xu, Q.; Likhomanenko, T.; Kahn, J.; Hannun, A.; Synnaeve, G.; Collobert, R. Iterative Pseudo-Labeling for Speech Recognition. *arXiv* **2020**, arXiv:2005.09267.
22. Lin, Y.-C.; Liang, Y.-H.L.; Su, H.; Lin, T.-Q.; Chen, S.-T.; Chen, Y.-N.; Lee, H. Pseudo2Real: Task Arithmetic for Pseudo-Label Correction in Automatic Speech Recognition. *arXiv* **2025**, arXiv:2510.08047.
23. Prakash, J.; Kumar, B.; Hacioglu, K.; Sharma, B.; Gopalan, S.; Chetlur, M.; Venkatesan, S.; Stolcke, A. Better Pseudo-Labeling with Multi-ASR Fusion and Error Correction by SpeechLLM. In *Proceedings of the Interspeech 2025*, Rotterdam, The Netherlands, 17–21 August 2025; pp. 579–583. <https://doi.org/10.48550/arXiv.2506.11089>.
24. Hwang, D.; Sim, K.C.; Huo, Z.; Strohman, T. Pseudo Label Is Better Than Human Label. *arXiv* **2022**, arXiv:2203.12668.
25. Yang, M.; Hirschi, K.; Looney, S.D.; Kang, O.; Hansen, J.H.L. Improving Mispronunciation Detection with Wav2vec2-Based Momentum Pseudo-Labeling for Accentedness and Intelligibility Assessment. *arXiv* **2022**, arXiv:2203.15937.
26. Xiao, A.; Fuegen, C.; Mohamed, A. Contrastive Semi-Supervised Learning for ASR. In *Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA; pp. 3870–3874.
27. Higuchi, Y.; Moritz, N.; Le Roux, J.; Hori, T. Momentum Pseudo-Labeling: Semi-Supervised ASR With Continuously Improving Pseudo-Labels. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1424–1438. <https://doi.org/10.1109/JSTSP.2022.3195367>.
28. Zhou, J.; Zhao, S.; Liu, Y.; Zeng, W.; Chen, Y.; Qin, Y. KNN-CTC: Enhancing ASR via Retrieval of CTC Pseudo Labels. In *Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA; pp. 11006–11010. <https://doi.org/10.1109/ICASSP48485.2024.10447075>.
29. Daoud, M.N.; Ben Messaoud, M.A. Phoneme-Level Mispronunciation Detection in Quranic Recitation Using ShallowTransformer. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, Suzhou, China, 8–9 November 2025; Darwish, K., Ali, A., Abu Farha, I., Touileb, S., Zitouni, I., Abdelali, A., Al-Ghamdi, S., Alkhereyf, S., Zaghouani, W., Khalifa, S., et al., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 457–463. <https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.63>.
30. Jin, L.; Min, D.; Shin, J.E.; Kim, E.Y. Contrastive Learning-Based Syllable-Level Mispronunciation Detection and Diagnosis for Speech Audiometry. In *Proceedings of the Interspeech 2025*, Rotterdam, The Netherlands, 17–21 August 2025; ISCA; pp. 828–832. <https://doi.org/10.21437/Interspeech.2025-1285>.
31. Kim, J.I.; Kim, S.; Chung, M. Developing a Robust Mispronunciation Detection by Data Augmentation Based on Automatic Phone Annotation. In *Proceedings of the 2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Hsinchu City, Taiwan, 17–19 October 2024; IEEE: New York, NY, USA; pp. 1–5. <https://doi.org/10.1109/O-COCOSDA64382.2024.10800119>.
32. Zhao, G.; Sonsaat, S.; Silpachai, A.; Lucic, I.; Chukharev-Hudilainen, E.; Levis, J.; Gutierrez-Osuna, R. L2-ARCTIC: A Non-Native English Speech Corpus. In *Proceedings of the Interspeech 2018*, Hyderabad, India, 2–6 September 2018; ISCA; pp. 2783–2787. <https://doi.org/10.21437/Interspeech.2018-1110>.
33. Shahin, M.; Epps, J.; Ahmed, B. Phonological Level Wav2vec2-Based Mispronunciation Detection and Diagnosis Method. *Speech Commun.* **2025**, *173*, 103249. <https://doi.org/10.1016/j.specom.2025.103249>.

34. El Kheir, Y.; Meghanani, A.; Toyin, H.O.; Almarwani, N.; Ibrahim, O.; Elshahawy, Y.A.; Shahin, M.; Ali, A. Iqra'Eval: A Shared Task on Qur'anic Pronunciation Assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks, Suzhou, China, 8–9 November 2025*; Darwish, K., Ali, A., Abu Farha, I., Touileb, S., Zitouni, I., Abdelali, A., Al-Ghamdi, S., Alkhereyf, S., Zaghouani, W., Khalifa, S., et al., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 443–452. <https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.61>.
35. Mansour, A. Metapseud at Iqra'Eval: Domain Adaptation with Multi-Stage Fine-Tuning for Phoneme-Level Qur'anic Mispronunciation Detection. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks, Suzhou, China, 8–9 November 2025*; Darwish, K., Ali, A., Abu Farha, I., Touileb, S., Zitouni, I., Abdelali, A., Al-Ghamdi, S., Alkhereyf, S., Zaghouani, W., Khalifa, S., et al., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 475–479. <https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.66>.
36. Qandos, N.; Sibae, S.; Ahmad, S.; Nacar, O.; Ammar, A.; Boulila, W.; Alhabashi, Y. ANPLers at IqraEval Shared Task: Adapting Whisper-Large-v3 as Speech-to-Phoneme for Qur'anic Recitation Mispronunciation Detection. In *Proceedings of the Arabic Natural Language Processing Conference: Shared Tasks, Suzhou, China, 8–9 November 2025*; Darwish, K., Ali, A., Abu Farha, I., Touileb, S., Zitouni, I., Abdelali, A., Al-Ghamdi, S., Alkhereyf, S., Zaghouani, W., Khalifa, S., et al., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 464–468. <https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.64>.
37. Al-Zaro, S.; Al-Ayyoub, M.; Al-Khaleel, O. Speaker-Independent Phoneme-Based Automatic Quranic Speech Recognition Using Deep Learning. *IEEE Access* **2025**, *13*, 125881–125896. <https://doi.org/10.1109/ACCESS.2025.3589252>.
38. Hyassat, H.; Abu Zitar, R. Arabic Speech Recognition Using SPHINX Engine. *Int. J. Speech Technol.* **2006**, *9*, 133–150. <https://doi.org/10.1007/s10772-008-9009-1>.
39. Lugosch, L.; Likhomanenko, T.; Synnaeve, G.; Collobert, R. Pseudo-Labeling for Massively Multilingual Speech Recognition. In *Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022*; IEEE: New York, NY, USA, 2022; pp. 7687–7691.
40. Unicode Consortium. *The Unicode Standard, Version 17.0.0*; Unicode Consortium: South San Francisco, CA, USA, 2025.
41. Sohn, H.-M. *The Korean Language*; Cambridge University Press: Cambridge, UK, 2001.
42. Choi, S.-K.; Kwon, H.-C.; Kim, M. Combining Autoregressive Models and Phonological Knowledge Bases for Improved Accuracy in Korean Grapheme-to-Phoneme Conversion. *IEEE Access* **2025**, *13*, 107678–107693. <https://doi.org/10.1109/ACCESS.2025.3581981>.
43. Pearce, D.; Hirsch, H.-G. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 16–20 October 2000*; ISCA, 2000; Volume 4, pp. 29–32. <https://doi.org/10.21437/ICSLP.2000-743>.
44. Park, J.-S.; Na, H.-J. Front-End of Vehicle-Embedded Speech Recognition for Voice-Driven Multi-UAVs Control. *Appl. Sci.* **2020**, *10*, 6876. <https://doi.org/10.3390/app10196876>.
45. Levenshtein, V. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Sov. Phys. Dokl.* **1965**, *10*, 707–710.
46. Gillick, L.; Cox, S.J. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK, 23–26 May 1989*; IEEE: New York, NY, USA, Volume 1, pp. 532–535.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.