Master Thesis

Evaluating the Impact of Continuous Pre-Training on ASR Models for Word-Level English Pronunciation Intelligibility

Wayne Kuan

Supervisor Dr. Luis Morgado da Costa
2^{nd} reader Dr. Hennie van der Vliet

a thesis submitted in fulfillment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Linguistics and Text-Mining Lab
Department of Language and Communication
Faculty of Humanities

Date	August 15, 2025
Student number	2801667
Word count	10,231

Abstract

Automatic Speech Recognition (ASR) systems are increasingly being explored for tasks beyond transcription, including intelligibility classification for language learning and assessment. This thesis investigates whether continual pretraining on single-word utterances can enhance the performance of ASR models, specifically OpenAI's Whisper, in recognizing isolated words and classifying their intelligibility. Building on prior work that applied unmodified ASR outputs to intelligibility assessment, this study continually-pretrains Whisper models of varying sizes (Base, Medium, Large, Turbo) using single-word data from the English Massive Open Online Course at the Centre for Global English. Evaluation is conducted using both exact-match and phonetic similarity scoring, alongside confidence thresholding, to capture performance in realistic assessment scenarios. Results show that continual pretraining improves isolated word recognition across model sizes, with performance gains carrying over to intelligibility classification. Error analysis highlights the impact of data labeling inconsistencies and limitations in the current confidence scoring method. The findings suggest that targeted continual pretraining can adapt general purpose ASR models to educational and assessment contexts, and point toward future improvements in confidence estimation and the inclusion of mispronounced speech in training data.

Declaration of Authorship

I, author, declare that this thesis, titled *Evaluating the Impact of Continuous Pre-Training on ASR Models for Word-Level English Pronunciation Intelligibility* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: <August 15, 2025>

Signed: <Wayne Kuan>

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Luis Morgado da Costa, for his time, patience, and invaluable expertise throughout the process of writing this thesis. His guidance and support have been instrumental in shaping this work and in helping me navigate the challenges along the way.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Problem Definition

As English continues to serve as a global lingua franca, clear and intelligible pronunciation remains a cornerstone of effective communication. The Centre for Global English (CGE) supports learners worldwide through its online course *MOOC English Pronunciation in a Global World*, which has attracted over 130,000 participants. Despite its success, the course faces a significant challenge: how to provide timely and meaningful pronunciation feedback at scale.

Currently, CGE relies on a peer review system where students assess each other's speech. While better than no feedback, this approach has notable limitations. Learners may not yet have the skills to evaluate pronunciation accurately, leading to inconsistencies and potentially unhelpful assessments. Meanwhile, instructor-provided feedback is infeasible for most learners due to faculty constraints and delayed response times. These issues limit the course's ability to support improvement in one of the most difficult aspects of language learning.

To address this, CGE is exploring automated approaches for providing pronunciation feedback. A long-term goal is to develop a robust system capable of fine-grained phonetic analysis, but this requires extensive time and expertise. As a more attainable step, this thesis investigates whether continual pretraining of an existing open-source automatic speech recognition (ASR) model, Whisper (Radford et al.), an enhance its performance in assessing isolated English word pronunciations and their intelligibility.

While Whisper performs well on long-form speech recognition, it has been observed to over-generate transcriptions for short inputs, returning multiple words even when only one was spoken (Zou, 2024). This raises concerns about its suitability for single-word input tasks, such as those common in language learning applications. This thesis explores whether continual pretraining on single-word utterances can help adapt Whisper for this specific use case.

1.2 Research Questions

This study is guided by the following research questions:

1. Can continual pretraining on single-word utterances improve the performance of ASR models in recognizing isolated words?

- How does the recognition performance of continually pretrained ASR models compare to their original versions?
- How does the impact of continual pretraining vary across Whisper model sizes (Base, Medium, Large, Turbo)?

2. To what extent does using continually pretrained ASR models improve performance in intelligibility classification?

   - How does intelligibility classification performance of continually pretrained models compare to that of their original counterparts?
   - How does the effect of continual pretraining on intelligibility classification vary across Whisper model sizes?

1.3 Thesis Outline

The remainder of this thesis is structured as follows. Chapter 2 reviews relevant literature on intelligibility classification, including Whisper fine-tuning, pronunciation assessment frameworks, phonetic similarity measures, and sociolinguistic perspectives. Chapter 3 describes the data preparation process, continual pretraining approach, evaluation pipeline, and scoring methods used in this study. Chapter 4 presents the performance outcomes of both original and continually pretrained Whisper models across different sizes and confidence thresholds. Chapter 5 provides a detailed qualitative examination of common error patterns, highlighting systematic challenges in classification. Chapter 6 interprets these findings in relation to the research questions, discusses their implications, and outlines key limitations. Finally, Chapter 7 summarizes the study's contributions, reflects on its broader significance, and identifies directions for future research.

Chapter 2

Related Work

This chapter provides an overview of recent related work in educational Natural Language Processing, with a particular emphasis on the role of Automatic Speech Recognition in language learning and intelligibility assessment. It surveys relevant literature and technologies, highlighting key approaches such as ASR-based intelligibility classification, phonetic similarity scoring, and sociolinguistic perspectives on intelligibility.

2.1 The Whisper Model

Whisper (Radford et al.) is a family of encoder–decoder Transformer-based ASR models trained on 680,000 hours of multilingual and multitask supervised speech data. Unlike domain-specific ASR systems, Whisper is trained on large-scale, weakly supervised datasets collected from the web, enabling strong zero-shot performance across languages, accents, and domains.

Its architecture processes log-Mel spectrograms through a Transformer encoder, with an autoregressive decoder generating transcriptions. While Whisper achieves state-of-the-art performance in transcription accuracy, it is not explicitly optimized for intelligibility classification. As such, adapting Whisper to this task, through continual pretraining on domain-specific data, may enhance its sensitivity to pronunciation deviations.

2.2 ASR-Based Intelligibility Classification

ASR-based intelligibility classification involves adapting speech recognition models to determine whether an utterance is intelligible, focusing on detecting deviations that hinder comprehension rather than maximizing transcription accuracy. Zou (2024) applied an existing ASR model to single-word utterances from an online english course for binary intelligibility classification. Their results showed that even without architectural changes or additional training, ASR outputs can be repurposed effectively for intelligibility assessment. This thesis extends this line of work by investigating whether continual pretraining on single-word data can further improve performance.

2.3 Phonetic Similarity and Scoring Methods

Standard text-based accuracy metrics can misclassify near-correct pronunciations as errors. Phonetic similarity metrics offer a more perceptually aligned evaluation by accounting for small deviations that do not significantly impact intelligibility.

Approaches include weighted phoneme substitution matrices (Hixon et al., 2011) and feature-based edit distances using articulatory representations (Mortensen et al., 2016). The latter, implemented in PanPhon[1], allows for a fine-tuned comparison between the target and the produced forms, recognizing that intelligibility often depends on the overall phonetic proximity rather than exact segmental matches.

2.4 Sociolinguistic Perspectives on Intelligibility

Intelligibility is not determined solely by phonetic accuracy but is influenced by sociolinguistic factors such as listener familiarity, expectations, and communicative context. Hughes et al. (2012) highlight the breadth of English dialectal variation, while Jenkins (2002) promotes English as a Lingua Franca, which focuses on making speech understandable to a wide range of listeners rather than aiming for a perfect native-like accent.

Foundational research by Munro and Derwing (1995) and Derwing and Munro (1997) distinguishes between accent, intelligibility, and comprehensibility, showing that heavily accented speech can still be highly intelligible. These perspectives are particularly relevant for ASR-based systems, which may misinterpret non-native yet intelligible speech as erroneous.

[1] https://pypi.org/project/panphon/0.5/

Chapter 3

Methodology

This chapter describes the approach used to investigate whether continual pretraining on isolated single-word utterances can improve automatic speech recognition (ASR) performance and enhance intelligibility classification for non-native speech. To address the first research question, whether continual pretraining improves recognition of isolated words, the chapter details how the Whisper models were adapted using domain-specific single-word data and how transcription performance was evaluated using Word Error Rate (WER).

To answer the second research question, to what extent these adapted models improve intelligibility classification, the chapter explains how predicted labels were generated, how phonetic information was incorporated to handle homophones, and how standard classification metrics were used to compare performance. This includes an overview of how precision, recall, F1-score, and F0.5-score were calculated, along with the use of confusion matrices for additional insight.

3.1 Data

3.1.1 ASR Data

We began by sourcing isolated single-word utterances from Mozilla Common Voice[1], a large, crowd-sourced speech corpus. However, the dataset contained limited coverage of clearly spoken English words in isolation, especially for the lexical items we aimed to include. To supplement this, we turned to Wiktionary recordings made available via Kaikki.org, which offered higher-quality, dictionary-style pronunciations for a wider range of words. Combining these two sources, we compiled a dataset totaling approximately 7 hours of English audio shown in Table 3.1. The goal was to build a focused collection of short, clearly articulated words to support the model's ability to recognize individual lexical items more effectively. For training and evaluation of the Automatic Speech Recognition system, we performed an 80-20 split on the dataset, allocating 80% of the audio samples for training and the remaining 20% for evaluation to assess the model's performance.

[1] https://commonvoice.mozilla.org/en

Mozilla Common Voice

We incorporated recordings from the *Mozilla Common Voice* project, an open-source initiative designed to crowd source speech data from a global pool of contributors. The platform allows volunteers to read and validate short text prompts, thereby creating a large, multilingual corpus of labeled voice recordings.

Common Voice operates through a community-driven pipeline: sentences are collected and read aloud by users, and these recordings are then validated by others to ensure quality. As of version 21.0, the English portion of the dataset contains over 2,700 validated hours of speech data, contributed by thousands of speakers across various accents and regions. Despite this scale, only a small fraction of the data consists of single-word utterances.

To maintain a high level of transcription quality, we filtered for clips that were validated by multiple users and whose transcriptions consisted of a single English word. After applying these constraints, we extracted 465 recordings, totaling around 19 minutes of usable audio. This strict selection ensured that the resulting subset aligned well with our goal of isolated word recognition, but also made it clear that additional data sources were necessary to meet our coverage needs.

Dataset	**Files**	**Duration**
Wiktionary	21303	6hr 49min 45s
Common Voice	465	19min 10s

Table 3.1: Number of single-word recordings and accumulative duration

Wiktionary

A substantial portion of the data, approximately 6 hours and 40 minutes of audio, came from Wiktionary, an open dictionary project that often includes user-submitted audio recordings of word pronunciations.

We accessed these recordings through Kaikki.org[2], a digital archiving and data mining initiative that gathers and republishes structured linguistic data from Wiktionary and related sources. Kaikki.org is powered by Wiktextract[3], a tool that converts Wiktionary entries into machine-readable formats (Ylonen, 2022). The project aims to make large-scale lexical data more accessible and beneficial to researchers, developers, and language-focused communities. For this work, we used the bulk English pronunciation dataset provided by Kaikki.org, downloaded on April 29, 2025.

The dataset included thousands of short audio clips in `.mp3` format. Filenames followed a consistent convention, typically containing the language variety and the corresponding spoken word. For example, `en-au-book.mp3` represents a recording of the word *book* in Australian English. Because the transcription consistently appeared at the end of each filename, we were able to extract and align the textual content with the audio files without requiring any manual annotation. The inclusion of regional varieties in this dataset (e.g., Australian, British, American English), shown in Table 3.2, is particularly relevant for downstream intelligibility modeling, as it introduces phonetic variation that may influence how intelligibility is perceived and evaluated.

[2]https://kaikki.org/index.html
[3]https://github.com/tatuylonen/wiktextract

To ensure consistency and suitability for Whisper's training pipeline, we excluded any files whose transcriptions contained non-alphabetic characters. Specifically, filenames including numerals, punctuation, or other special symbols were filtered out, as such inputs may introduce noise and complicate the model's tokenization process. By focusing solely on well-formed alphabetic transcriptions, we aimed to construct a clean, uniform dataset conducive to effective model training.

Country	**Count**
United States	13,686
Australia	5,888
United Kingdom	1,476
New Zealand	7
South Africa	1
N/A	710

Table 3.2: Number of recordings for respective english accents

3.1.2 Intelligibility Classification Data

For model evaluation on Intelligibility Classification, we used two curated dataset of word-level recordings sourced from the *English in a Global World* Massive Open Online Course (MOOC), one created in 2024 and the other in 2025. The goal was to test the model's ability to assess intelligibility in real-world second language learner speech. Each student participant submitted audio recordings as part of a course assignment, in which they were asked to read aloud a list of 53 English words, shown in Table 3.3, adapted from Hughes et al. (2012). In some cases, a short passage was included as well. All submissions were provided in `.mp3` format.

1. bard	15. cot	29. meat	43. pore
2. bay	16. daft	30. meet	44. pot
3. bear	17. dance	31. nose	45. pour
4. bee	18. doll	32. pat	46. pull
5. beer	19. fair	33. Paul	47. put
6. bird	20. farther	34. pause	48. putt
7. board	21. father	35. paw	49. seedy
8. boat	22. fern	36. paws	50. tied
9. boot	23. fir	37. pet	51. tide
10. bout	24. fur	38. pit	52. wait
11. boy	25. half	39. plate	53. weight
12. buy	26. hat	40. pole	
13. caught	27. knows	41. pool	
14. city	28. mate	42. poor	

Table 3.3: Alphabetically ordered list of 53 target words used in the MOOC intelligibility evaluation.

From the course, we have a dataset of 2,631 word-level recordings. This was the initial dataset used to develop and evaluate the **base.en** model by Zou (2024) and

will be referred as **MOOC-2024**. For the purpose of comparing the results between Zou's work and our continually pretrained models, we will use the same test set as a benchmark.

In addition, since the creation of the **MOOC-2024** dataset, the CGE has recorded and labeled 10,200 word-level audio files. We plan to use this **MOOC-2025** dataset to further develop and evaluate our models. To support fair evaluation, we divided the dataset into two equal parts: a development set and a test set, shown in Table 3.4. The split was performed at the level of individual recordings—meaning that for each of the 53 target words roughly half of the available clips were randomly assigned to the development set and the other half to the test set. This ensured that both sets contained all target words, but no individual recording appeared in both. As a result, the model was evaluated on new speaker instances of familiar word types, allowing us to measure generalization while keeping the lexical content consistent across sets.

Dataset	**Intelligible**	**Unintelligible**	**Total Instances**
Development (MOOC-2024)	353 (68.4%)	163 (31.6%)	516
Test (MOOC-2024)	1410 (66.7%)	705 (33.3%)	2115
Development (MOOC-2025)	2877 (56.2%)	2239 (43.8%)	5116
Test (MOOC-2025)	2858 (56.2%)	2226 (43.8%)	5084

Table 3.4: Distribution of Intelligible and Unintelligible Classes

Segmentation and Labeling Procedure

To convert these submissions into usable test data, two student interns at the Centre for Global English (CGE) manually segmented the recordings into isolated word-level clips using the audio editing software *Audacity*. Each student's file was first organized into a folder labeled by course series and student ID (e.g., `series02-s00012`), and then saved as an editable project file for consistency.

Segmentation was performed by zooming into the waveform and isolating each target word. Up to 1.5 seconds of surrounding silence was retained where possible to preserve natural acoustic boundaries. Each word segment was labeled directly within Audacity using its built-in label track functionality and exported as an individual audio file. This process is shown in Figure 3.1.

The exported filenames followed a consistent structure: `seriesXX-sXXXXX-word-label.mp3`, where the final digit denoted intelligibility—`1` for intelligible, `0` for unintelligible. In instances where the same word was pronounced multiple times by the same student, alphabetical suffixes (e.g., `s00001a`, `s00001b`) were added to avoid naming conflicts.

Additional suffixes such as `0audio` were used to flag problematic recordings—clips with excessive background noise, poor microphone quality, or extremely soft volume. Segments featuring abnormal breathing sounds or unrelated speech were also marked as unintelligible to reduce labeling noise and prevent misleading model feedback.

Labeling decisions were based on detailed phonetic criteria developed by Dr. Laura Rupp, head of CGE and a specialist in English phonology. These criteria drew on an intelligibility-focused framework informed by Jenkins' work on English as an International Language (Jenkins, 2002). Ambiguous cases were documented and reviewed in consultation with Dr. Rupp to ensure consistent application of the labeling guidelines.

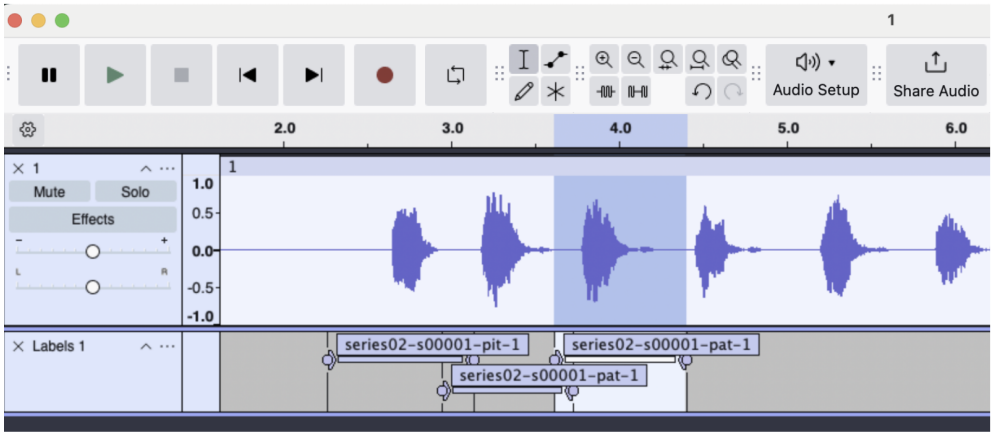

Figure 3.1: Example of word-level segmentation and labeling in Audacity. Each labeled region corresponds to a single word utterance with intelligibility classification.

3.2 Experimental Set up

3.2.1 Model Selection

For automatic transcription, we selected OpenAI's Whisper model due to its proven robustness and ability to generalize across a wide range of speech conditions. Whisper was trained on 680,000 hours of multilingual and multitask audio data, making it well-equipped for zero-shot tasks where no domain-specific fine-tuning is applied Radford et al.. This is particularly relevant for our use case: assessing intelligibility in non-native speaker recordings with considerable variation in accent, pronunciation, and recording quality.

Whisper's zero-shot transfer performance has been shown to rival that of fully supervised models, even on standard speech benchmarks. This capacity allows it to perform reliably on unfamiliar datasets, such as our MOOC recordings, without requiring adaptation or retraining. In addition, Whisper is notably resilient to background noise. Where many ASR systems experience significant performance degradation in noisy environments, Whisper continues to produce high-quality transcriptions, making it suitable for real-world educational contexts where recording conditions are not controlled.

Size	Parameters	English-only	Multilingual	Relative Speed
tiny	39 M	tiny.en	tiny	≈10x
base	74 M	base.en	**base**	≈7x
small	244 M	small.en	**small**	≈4x
medium	769 M	medium.en	**medium**	≈2x
turbo	809 M	N/A	**turbo**	≈8x
large	1550 M	N/A	**large**	1x

Table 3.5: Overview of Whisper models. Bolded and underlined multilingual models were used in the experiments. Relative speed is expressed relative to the large model (1x).

The model is available in six multilingual and four monolingual variants of increas-

ing size, each offering a trade-off between inference speed and transcription accuracy. These range from the smallest **tiny** model to the largest **large** and **turbo** variants. A summary of the available multilingual models is provided in Table 3.5.

In this project, we chose to evaluate multilingual models from the **base** size and up. This decision was influenced by prior work conducted at CGE by Zou (2024), who tested only the English-only **base.en** model. Given our goal of identifying a practical ASR solution that can provide timely, accurate feedback to students, we opted to experiment with a range of multilingual models to find an optimal balance between speed and accuracy. We further prioritized multilingual models on the assumption that learners in our dataset are non-native speakers. Models trained on multilingual data are likely to be more attuned to non-standard or accented pronunciations, increasing their utility in a language learning context.

3.2.2 Continual Pretraining

To further adapt the Whisper models to our specific task, we employed a continual pretraining approach, adapting the multilingual models on English single-word pronunciations. Continual pretraining involves taking a pretrained model and updating its parameters with additional training on new datasets without starting from scratch. This strategy leverages the model's broad knowledge while refining it for more specialized applications.

This process roughly matches the top part of Figure 3.2 from the original Whisper paper (Radford et al.), which shows the encoder-decoder setup used for ASR. In our case, continual pretraining updates the entire encoder-decoder stack, but only for the main *transcription* task; we did not include any additional tasks like language identification. This means that the model keeps learning to map raw audio to its text output, tuning its weights to better handle our specific pronunciation data.

Our continual pretraining methodology followed the procedures outlined by Sanchit Gandhi in their tutorial on Whisper fine-tuning (Gandhi, 2022)[4]. The tutorial highlights practical techniques for adapting Whisper models using the `Trainer` API, focusing on efficient training with limited resources and leveraging carefully curated speech data. Key aspects include setting appropriate learning rates, using mixed precision training to accelerate convergence, and applying data augmentation to improve robustness.

Prior to training, the raw combined Wiktionary & Mozilla Common Voice data was processed into a structured format suitable for use with Hugging Face's `datasets` library. For each segmented audio clip, we extracted the file path along with its corresponding transcription label. These were compiled into a JSON file, which served as the entry point for building our training corpus. This JSON-formatted dataset was loaded with the `load_dataset()` function and automatically split into 80% training and 20% evaluation sets using the `train_test_split()` method, with a fixed seed of 42 to ensure reproducibility. It should be noted that we recognize the methodological limitation of using only a single seed, as it does not capture the potential variability in outcomes that could result from different random initializations or data splits.

Following this, we applied preprocessing using a custom `prepare_dataset` function, which handled tasks such as feature extraction and input formatting. For the small and medium Whisper models, a single feature extractor was used, and the resulting dataset

[4] `https://huggingface.co/blog/fine-tune-whisper`

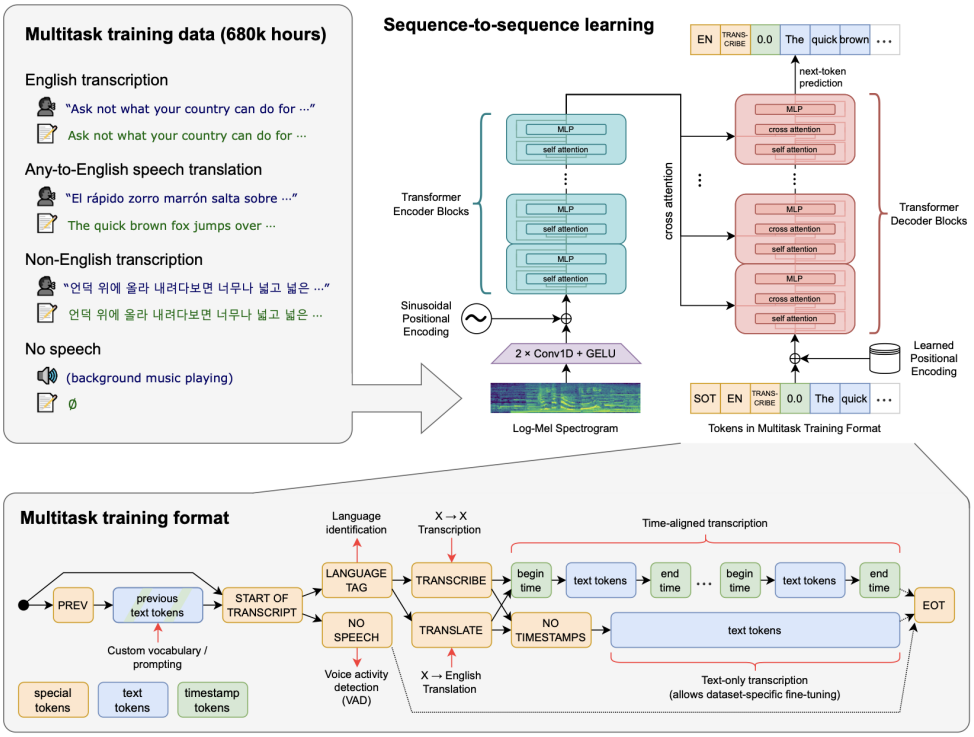

Figure 3.2: Overview of OpenAI Whisper's Initial Training

was saved to disk for reuse. For the large and turbo models, which require a different feature configuration (e.g., specific spectrogram channel settings), we explicitly loaded the `WhisperFeatureExtractor` to ensure compatibility before processing.

All audio inputs were also preprocessed to conform to Whisper's expected sampling rate of 16 kHz. Matching the sampling rate is essential, as audio signals with different sampling rates have very different distributions. If not properly aligned, passing mismatched audio to the model can lead to degraded performance or unexpected behavior. For example, a 16 kHz audio played back at 8 kHz will sound like it is in slow motion—similarly, Whisper's feature extractor will misinterpret input that does not match its expected format. To ensure model consistency and avoid inadvertently training on distorted audio, all audio files were resampled to 16 kHz prior to training.

Specifically, the continual pretraining was conducted using the **Seq2SeqTrainingArguments** interface, with hyperparameters suggested by Gandhi (2022) to balance training efficiency and model performance. The training employed a batch size of 16 per device and a learning rate of 1×10^{-5}, optimized over 15 epochs. Gradient checkpointing and mixed precision (fp16) were enabled to reduce memory consumption and speed up training. The evaluation was performed every 1,000 steps using Word Error Rate (WER) as the primary metric, with the best-performing model saved automatically. Additional strategies such as gradient accumulation, warmup steps, and checkpointing were left as default. Logging was integrated with TensorBoard for detailed monitoring, and the trained model was pushed to the Hugging Face Hub[5] for reproducibility and future access. A visualization of the training arguments is provided in the appendix as Figure 1.

[5] `https://huggingface.co/WayneTK`

3.2.3 Transcribing MOOC Data

The binary intelligibility classification system follows the design and implementation proposed by Zou (2024). This system uses three key features derived from ASR outputs, transcriptions, confidence scores, and temperature, to make decisions about whether a spoken word is intelligible or not. To support this classification task, it was first necessary to obtain transcriptions and accompanying confidence scores derived from Whisper's raw outputs. This process consisted of three main stages: retrieving the raw transcription data, post-processing the outputs to produce cleaned segment-level transcriptions and interpretable confidence scores, and aligning these outputs with ground truth labels to be used for intelligibility classification. We describe later how the confidence scores are derived and used to quantify Whisper's certainty in its predictions, as well as how they are incorporated into the intelligibility classification system.

Obtaining Raw Transcription Data

For each word-level audio file, we generated transcriptions using ten different OpenAI Whisper models — five **original** models and five **adapted** models. The original models produced structured JSON outputs containing multiple fields, including language, duration, text, and segments. The segments field stored a list of dictionaries, each detailing a portion of the audio with information such as start and end times, the transcribed text, the temperature used, and the average log probability.

Since the adapted models continually pretrained with Hugging Face's **Trainer API** and used through **Transformers**, it is not currently supported to calculate average log probabilities directly. They were computed manually by running the same input features back through the decoder to get output logits for each position. For each predicted token, the logits at position i–1 were used to get the probability of the token at position i, applying a softmax with a fixed temperature of 0.5 and then taking the log of the resulting probability. Special tokens and empty strings were ignored, since they don't reflect the spoken word. Finally, the remaining log probabilities were averaged to get one score per transcription, which we used to compare confidence levels across models and thresholds.


```
"series03-s000057-bear-1": {
    "studentID": "s000057",
    "word": "bear",
    "trueLabel": "intelligible",
    "segments": [
        {
            "text": " Bear?",
            "temperature": 0.5,
            "avg_logprob": -1.3270183563232423,
            "compression_ratio": 0.38461538461538464,
            "no_speech_prob": 0.12474017590284348
        }
    ]
}
```


Figure 3.3: Example of Raw Transcription Data

Each transcribed entry additionally included metadata, shown in Figure 3.3, derived directly from the file naming convention of the input audio. As a result, the raw transcription record for each file consisted of four key fields: studentID, word, trueLabel, and segments. Here, studentID identifies the speaker, word denotes the target word, and trueLabel indicates whether the pronunciation was labeled as intelligible or unintelligible. The segments field contained Whisper's segment-level outputs, which were later used to generate features for classification.

Post-Processing Raw Text

In its raw form, Whisper's text output may include leading spaces, punctuation marks, and inconsistent casing. To ensure a consistent basis for comparison with the true transcription, all text outputs were cleaned through a standard post-processing pipeline. This involved stripping leading spaces, removing any punctuation, and converting all characters to lowercase.

Additionally, although each word-level audio file was typically only one to two seconds in duration, Whisper occasionally produced multiple segments for a single file. In these cases, the text fields from all segments were cleaned individually and then concatenated into a single, unified string representing the model's final transcription for that file.

Post-Processing Raw Average Log Probabilities for Confidence Scores

Beyond the raw text, Whisper provides an average log probability for each segment, which reflects the model's internal estimate of confidence. More negative values indicate lower confidence, while values closer to zero suggest greater certainty. To make these scores interpretable for downstream use, the logarithmic probabilities were converted back into linear probabilities in the range [0, 1]. This was done by applying the exponential function, math.exp(), in Python, as done in Zou (2024). Converted linear probabilities were then rounded to three decimal places for reporting.

If Whisper produced multiple segments for a single word-level file, we set its confidence score to zero. Experiments done by Zou showed that averaged probabilities across segments often gave misleadingly high scores for incorrect multi-word transcriptions. For example, in the development set, over half of the segments in multi-segment cases showed high local probabilities despite not matching the target word. Setting these to zero ensured that only single-segment transcriptions contributed valid confidence values, keeping the scoring aligned with true word-level outputs.

It should be noted that the confidence scores in this study are derived solely from the model's top-1 transcription hypothesis. This means that only the single most likely output, as determined by Whisper, is considered, without accounting for the probability distribution across alternative candidates. In some cases, other top-n hypotheses may have probabilities very close to the top-1 result, indicating a level of uncertainty that is not reflected in our current scoring method. As a result, the top-1 confidence score may overstate the model's certainty in borderline cases. The implications of this limitation, and possible approaches to incorporating broader probability information, will be discussed further in a later chapter.

3.2.4 Applying Confidence Thresholds for Automatic Intelligibility Classification

The derived confidence scores were then used as part of an automatic filtering step. By defining a threshold value, the system could distinguish between reliable and unreliable transcriptions. If a transcription's confidence score exceeded this threshold, its predicted word was compared to the true label to classify it as intelligible or unintelligible. If the score fell below the threshold, or if multiple segments were produced, the spoken word was automatically classified as intelligible by default. This precaution helped avoid penalizing genuine pronunciations that the model struggled to transcribe confidently, thereby reducing false negatives. To find an optimal balance between sensitivity and specificity, a range of thresholds (0.3, 0.4, 0.5, 0.7, and 0.8) was evaluated, and their effects on downstream classification performance were assessed.

3.2.5 Classification Algorithms

After post-processing, predicted intelligibility labels (which Zou refers to as **Whisper's Judge**) were determined by comparing the model's transcription to the target word. If the predicted transcription matched the true word exactly, the instance was labeled *intelligible*; otherwise, it was labeled *unintelligible*.

However, relying solely on exact spelling introduces a limitation. For example, if the model transcribes *meat* as *meet*, the spelling does not match, so it is marked *unintelligible*, even though both words are homophones. Since the word list contains isolated words with no context, such variation is acceptable.

To address this, Zou also defined a second label type, **Whisper's Judge Phonetic**, which relies on phonetic comparison. For this, phonetic representations were sourced from the CMU Pronouncing Dictionary[6] (CMUDict), a widely used resource that provides ARPAbet[7] transcriptions for over 134,000 American English words. CMUDict uses ASCII-based symbols, which are practical for speech tasks that require computer-readable output.

When the model produced outputs not listed in CMUDict, such as non-words or misspellings, a phonetic value of *N/A* was assigned and the prediction marked as *unintelligible*. The same rule applied to outputs containing multiple words: since CMUDict provides phonetic forms for single words only, multi-word transcriptions could not be compared directly and were treated as *unintelligible*. For instance, if the target word *bard* was transcribed as *by heart*, the combined phrase does not match the single-word phonetic form for *bard* and was therefore labeled *unintelligible*.

When the model produced transcriptions that could not be found in CMUDict, such as non-words, misspellings, or outputs containing multiple words, a phonetic value of *N/A* was assigned, and the result was marked as a *no match* at the phonetic level. Since CMUDict provides phonetic forms only for single English words, multi-word outputs could not be directly compared to the target pronunciation and were also treated as *no match*. However, the final intelligibility label was not determined by phonetic match alone. In Zou's system, confidence score takes precedence: if the model's transcription had a confidence score below a predefined threshold, the utterance was automatically labeled *intelligible*, regardless of whether the phonetic match was valid or not. This

[6] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`
[7] `https://en.wikipedia.org/wiki/ARPABET`

precedence simplifies the classification process but introduces a limitation: all transcriptions with confidence scores below the threshold, whether correct or incorrect, are automatically labeled as intelligible. Consequently, this can result in false negatives within the unintelligible class and reduce recall the higher the confidence threshold, as genuinely unintelligible utterances with low confidence may be consistently missed.

Both text-based and phonetic labels were retained for evaluation. The final processed outputs, including transcriptions, true labels, both label types, phonetic forms, confidence scores, and temperature settings, were saved in structured JSON files to support detailed analysis across conditions.

3.2.6 Evaluation Metrics

Calculating Word Error Rate

To address the research question of whether continual pretraining on single-word utterances can improve the performance of ASR models in recognizing isolated words, we first evaluate transcription quality using Word Error Rate (WER). WER is a standard metric in automatic speech recognition for quantifying the difference between a predicted transcription and its corresponding reference text. It measures the minimum number of insertions, deletions, and substitutions required to transform the predicted output into the correct reference, normalized by the total number of words in the reference.

Formally, WER is defined as:

$$\text{WER} = \frac{S + D + I}{N}$$

where S is the number of substitutions, D the number of deletions, I the number of insertions, and N the total number of words in the reference transcription.

In this project, each input is an isolated single-word utterance. As a result, WER in this setting effectively reduces to a simple measure of transcription accuracy: the output is either exactly correct or incorrect. A perfect match between the predicted word and the reference results in a WER of 0, while any mismatch (due to a substitution, deletion, or insertion) results in a WER of 1 for that instance. By averaging WER across all test samples, we obtain a clear and interpretable metric that directly reflects the model's word-level recognition accuracy.

Comparing WER scores between the original and continually pretrained Whisper models therefore provides direct evidence of whether additional domain-specific training on single-word pronunciations improves the base model's performance on this specific task.

Standard Classification Metrics

To compare the performance of each model configuration on the binary intelligibility classification task, standard metrics including precision, recall, F1-score, and F0.5-score were used. These metrics were calculated separately for each class — *intelligible* and *unintelligible* — treating each class independently. This aligns with macro averaging in that it does not weight metrics by the class distribution, ensuring that both classes contribute equally to the final evaluation.

Confusion matrices were also generated to visualize prediction errors and to provide additional insight into model performance across different temperature settings.

Precision measures how reliable the model's predictions are for a given class. Formally, precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP denotes true positives and FP denotes false positives.

Recall measures how well the model identifies all actual instances of a class. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

where FN denotes false negatives.

The F1-score is the harmonic mean of precision and recall and balances both metrics equally:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

To emphasize precision more strongly in this project's context, the F0.5-score was also used. It adjusts the harmonic mean to weight precision higher than recall:

$$F_{0.5} = (1 + 0.5^2) \times \frac{\text{Precision} \times \text{Recall}}{(0.5^2 \times \text{Precision}) + \text{Recall}}.$$

High precision is prioritized to ensure that most words marked as *unintelligible* are genuinely unintelligible, avoiding misleading feedback for learners. Both text-based and phonetic-based labels were used for evaluation, allowing the metrics to capture how well the models handle homophones and pronunciation variation.

Chapter 4

Results

4.1 Transcription Accuracy

To examine the effects of continual pretraining and model size, ten models were evaluated on the **MOOC-2025** development set. When calculating Word Error Rate (WER), instances labelled as *unintelligible* were excluded, as mispronounced words cannot reasonably be expected to produce fully accurate transcriptions.

Model Size	**Original**	**Pretrained**	**Difference**
Base	40.05%	36.01%	4.04%
Small	37.02%	33.15%	3.87%
Medium	37.65%	31.16%	6.49%
Turbo	33.09%	33.45%	-0.35%
Large	35.49%	31.39%	4.10%

Table 4.1: MOOC-2025 Development – Average WER/Accuracy Scores

Table 4.1 shows the average WER for each model size, comparing the original off-the-shelf Whisper models with their continually pretrained versions. For the Base, Small, Medium, and Large models, continual pretraining on single-word utterances consistently reduced WER, with improvements ranging from about 4% to over 6%. The Medium model showed the largest drop overall. In contrast, the Turbo model had a slight increase in WER after continual pretraining, which may be due to how Turbo differs from the Large version: the Turbo[1] variant is a pruned version of the large-v3 model with the number of decoding layers reduced from 32 to 4. While this pruning makes the model much faster, it comes with a small quality trade-off that likely limits the benefits of extra training on isolated words. Overall, these results suggest that targeted continual pretraining can help improve single-word recognition for most model sizes tested.

[1] https://huggingface.co/openai/whisper-large-v3-turbo

4.2 Intelligibility Classification

4.2.1 Results: Intelligible Class

Average Metrics Without Confidence Threshold

The results in Table 4.2 highlight clear performance trends across model sizes for the intelligible class, providing a baseline before any confidence thresholding is applied. In general, larger model sizes achieve stronger results across precision, recall, F1, and F0.5, regardless of whether the model is original or adapted. This pattern is most apparent when moving from the Base and Small variants to the Medium, Turbo, and Large models, where improvements are seen in both precision and balanced metrics. Adapted models consistently outperform their original counterparts at every size, but the relative gains become more substantial as the model size increases. For example, the Large Adapted model achieves the highest scores overall, with an F1-score of 0.599 and an F0.5-score of 0.652, reflecting both improved transcription quality and a stronger balance between precision and recall. Similarly, the Medium Adapted model surpasses all smaller variants, achieving an F1-score of 0.591, showing that scaling up the model provides meaningful benefits for intelligibility classification.

While the base and small variants benefit from adaptation, their limited capacity appears to constrain performance gains, particularly in recall. In contrast, the larger models leverage their greater representational capacity to produce more accurate and contextually aligned transcriptions, leading to higher intelligibility detection rates without relying heavily on threshold-based filtering. This suggests that both model scaling and adaptation contribute to better performance, but scaling has a pronounced impact on the model's inherent classification ability.

Results with Varying Confidence Thresholds

To complement the average metrics presented in Table 4.2, Figures 4.1 and 4.2 visualize how performance on the intelligible class evolves across different confidence thresholds (0.3–0.8). These graphs are split by scoring method: Whisper's Judge and Whisper's Judge Phonetic. Each graph tracks four metrics across all model variants: Precision, Recall, F1, and F0.5. Across both scoring methods, precision tends to decrease as the confidence threshold increases, while recall rises, especially for original models. This phenomenon reflects the design of the classification system: predictions with lower confidence are more likely to be filtered out and labeled as intelligible by default, inflating recall while lowering precision. This is most visible in the original models, where there are steep curves in both recall and precision across all sizes.

With Phonetic Whisper's Judge (Figure 4.2), adapted models consistently outperform their original counterparts in precision and consequently F0.5 scores, particularly at lower thresholds. The medium-adapted and large-adapted models especially achieve higher F0.5 scores across the 0.3–0.8 range, suggesting that adaptation improves overall balance between capturing intelligible cases and avoiding false positives when prioritizing precision.

Overall, these results reinforce the notion that both model size and adaptation improves the model's confidence in its predictions and consequently intelligibility classification, especially when combined with a calibrated confidence threshold. While continual pretraining does not uniformly improve performance across all settings, it does offer benefits in mid-range thresholds and for specific model sizes—most notably

Model	Variant	Judge Type	Precision	Recall	F1-Score	F0.5-Score
Base	Original	Whisper Judge	0.625	0.264	0.327	0.411
		Whisper Judge Phonetic	0.706	0.373	0.430	0.508
	Adapted	Whisper Judge	0.711	0.337	0.423	0.527
		Whisper Judge Phonetic	0.741	0.464	0.536	0.616
Small	Original	Whisper Judge	0.637	0.319	0.383	0.466
		Whisper Judge Phonetic	0.725	0.449	0.498	0.576
	Adapted	Whisper Judge	0.731	0.386	0.454	0.538
		Whisper Judge Phonetic	0.743	0.506	0.561	0.627
Medium	Original	Whisper Judge	0.671	0.309	0.374	0.462
		Whisper Judge Phonetic	0.767	0.441	0.498	0.588
	Adapted	Whisper Judge	0.699	0.414	0.474	0.547
		Whisper Judge Phonetic	0.727	0.542	0.591	0.650
Turbo	Original	Whisper Judge	0.643	0.385	0.437	0.507
		Whisper Judge Phonetic	0.722	0.531	0.567	0.625
	Adapted	Whisper Judge	0.712	0.389	0.455	0.546
		Whisper Judge Phonetic	0.728	0.499	0.554	0.625
Large	Original	Whisper Judge	0.685	0.354	0.414	0.500
		Whisper Judge Phonetic	0.744	0.486	0.530	0.605
	Adapted	Whisper Judge	0.680	0.424	0.480	0.552
		Whisper Judge Phonetic	0.729	0.559	0.599	0.652

Table 4.2: MOOC-2025 Dev Average Results for Intelligible Class

small-adapted and medium-adapted, which show improved F0.5 performance at thresholds between 0.5 and 0.7.

4.2.2 Results: Unintelligible Class

Average Metrics Without Confidence Threshold

When evaluating the unintelligible class, Table 4.3 shows a consistent pattern of high recall and comparatively lower precision across all models. The most prominent trend is the improvement observed when moving from original to adapted models, while differences between model sizes remain minimal. For instance, in the Medium model, the F0.5 score increases from 0.566 (original) to 0.577 (adapted), and in the Large model from 0.552 to 0.578. These gains, though measurable, are modest compared to the more substantial improvements seen in the intelligible class. The persistent gap in precision highlights a disparity in the ability to correctly identify unintelligible pronunciations even in the absence of confidence-based filtering. Possible explanations for this trend are explored further in the Error Analysis chapter, where specific error patterns and contributing factors are examined in greater detail.

Results with Varying Confidence Thresholds

Figures 4.3 and 4.4 depict the performance of all model variants on the unintelligible class across a range of confidence thresholds (0.3–0.8), under both Whisper's Judge

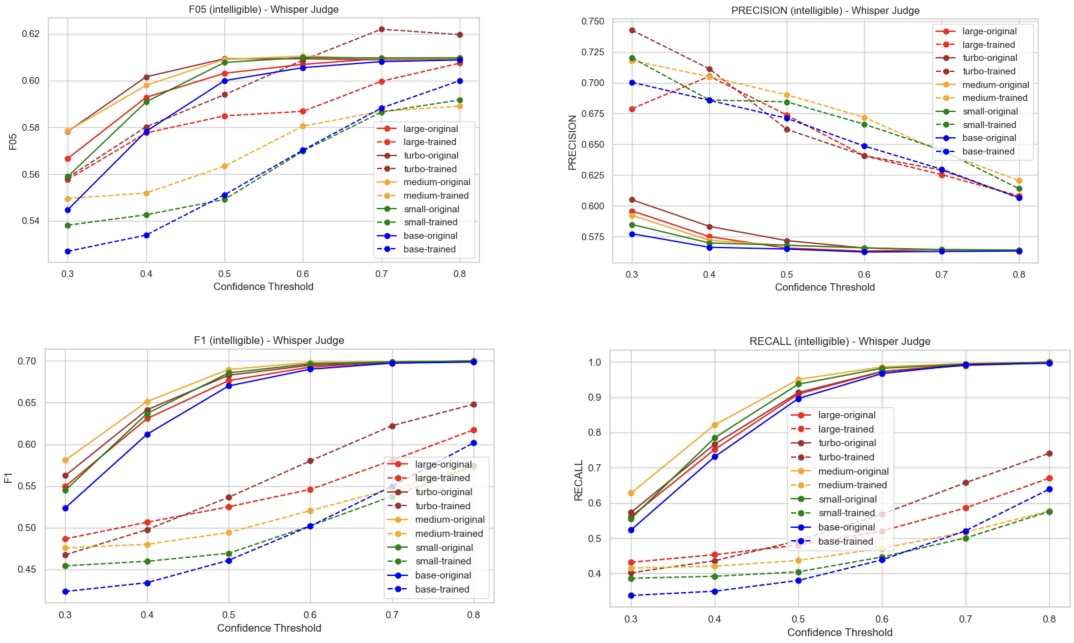

Figure 4.1: Performance metrics for Intelligible class classification – Whisper Judge

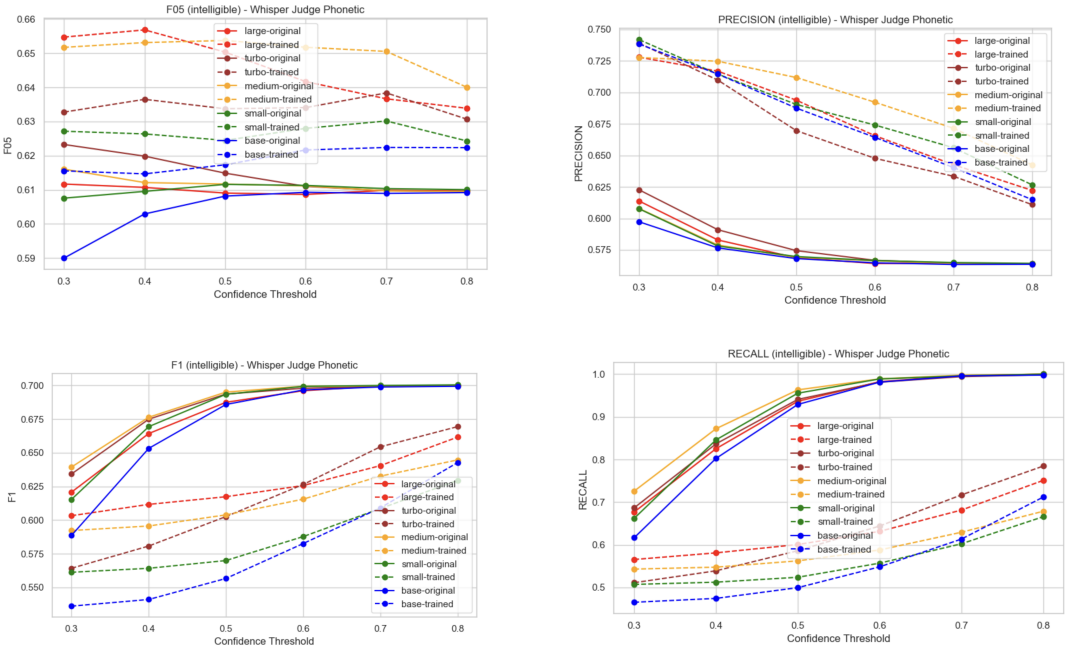

Figure 4.2: Performance metrics for Intelligible class classification – Whisper Judge Phonetic

Model	Variant	Judge Type	Precision	Recall	F1-Score	F0.5-Score
Base	Original	Whisper Judge	0.483	0.869	0.592	0.519
		Whisper Judge Phonetic	0.522	0.804	0.602	0.546
	Adapted	Whisper Judge	0.502	0.852	0.608	0.537
		Whisper Judge Phonetic	0.542	0.788	0.620	0.568
Small	Original	Whisper Judge	0.499	0.842	0.595	0.529
		Whisper Judge Phonetic	0.554	0.769	0.605	0.564
	Adapted	Whisper Judge	0.511	0.821	0.602	0.540
		Whisper Judge Phonetic	0.548	0.751	0.611	0.569
Medium	Original	Whisper Judge	0.502	0.855	0.599	0.533
		Whisper Judge Phonetic	0.545	0.784	0.606	0.560
	Adapted	Whisper Judge	0.526	0.807	0.609	0.553
		Whisper Judge Phonetic	0.560	0.734	0.614	0.577
Turbo	Original	Whisper Judge	0.515	0.795	0.591	0.538
		Whisper Judge Phonetic	0.567	0.710	0.597	0.571
	Adapted	Whisper Judge	0.519	0.817	0.602	0.545
		Whisper Judge Phonetic	0.548	0.760	0.613	0.568
Large	Original	Whisper Judge	0.504	0.824	0.589	0.530
		Whisper Judge Phonetic	0.545	0.742	0.588	0.552
	Adapted	Whisper Judge	0.527	0.802	0.605	0.551
		Whisper Judge Phonetic	0.565	0.726	0.611	0.578

Table 4.3: MOOC-2025 Dev Average Results for Unintelligible Class

and Whisper's Judge Phonetic scoring. Across both scoring methods, the most striking pattern is the stability of adapted models across increasing thresholds, especially when compared to the steep drop-offs exhibited by original models. Unlike original models, which are likely to produce less confident predictions, adapted models continue to assign the unintelligible label with higher precision even at stricter thresholds. This indicates that the adapted models predict with higher confidence, and thus their predictions are recognized as unintelligible speech more reliably because only high-confidence predictions are eligible to be labeled as unintelligible.

This is most evident in the precision and F0.5 graphs under both scoring strategies, where adapted models such as medium-adapted, turbo-adapted, and large-adapted sustain high and consistent values throughout. In contrast, the performance of original models deteriorates sharply beyond threshold 0.5, often approaching zero by 0.8. This suggests that the original models either fail to detect unintelligibility at higher thresholds or are not confident enough to trigger the label, resulting in many incorrect default assignments to the intelligible class.

4.2.3 MOOC-2024 Benchmark Comparison

To further examine the effect of continual pretraining, we compare our adapted base model, here referred to as **adp-base**, with Zou's **base.en** model on the **MOOC-2024** test set. Zou used a temperature of 0.5 and a confidence threshold of 0.3 as the optimal settings for this dataset, so we adopted the same configuration to keep the comparison

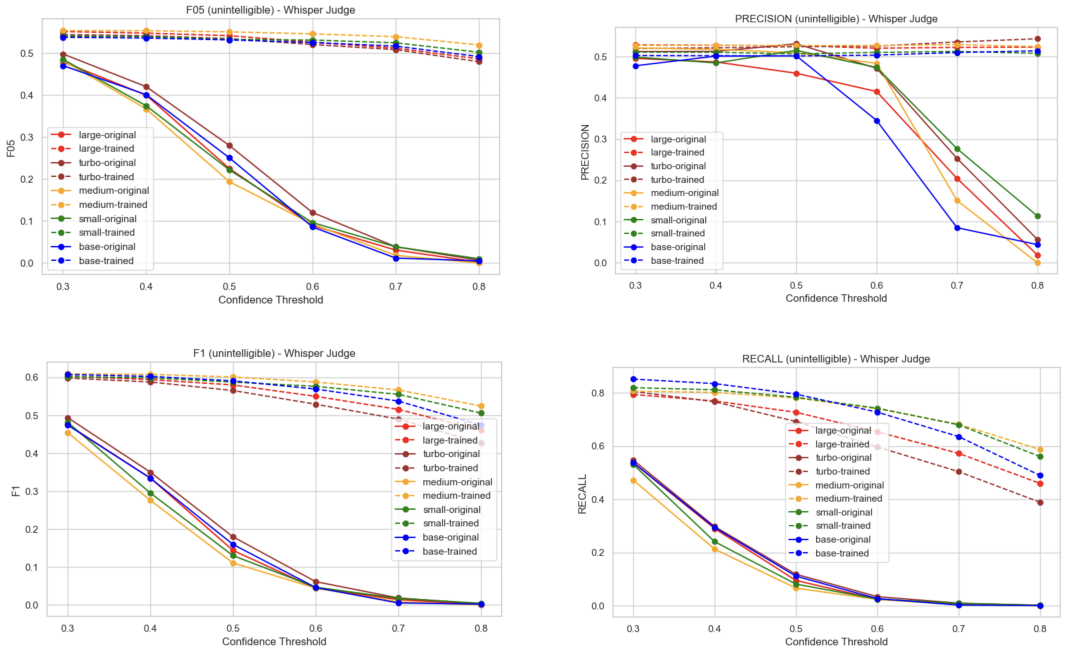

Figure 4.3: Performance metrics for Unintelligible class classification – Whisper Judge

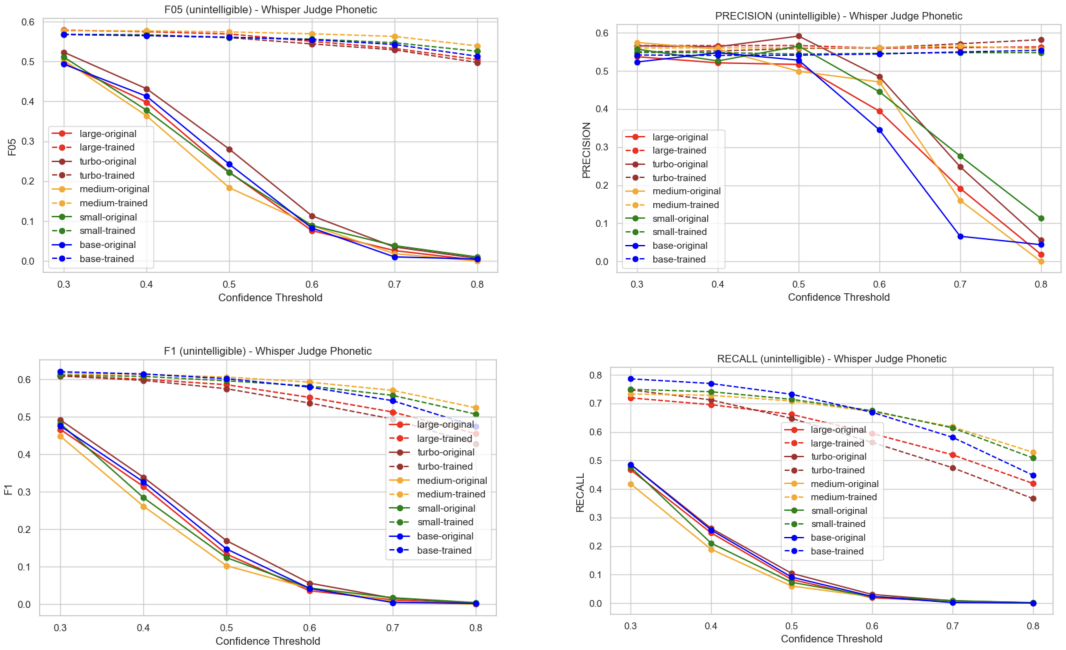

Figure 4.4: Performance metrics for Unintelligible class classification – Whisper Judge Phonetic

consistent.

Table 4.4 summarizes the phonetic-based classification results for both the intelligible and unintelligible classes. Compared to **base.en**, the adapted model shows higher precision for the intelligible class (0.878 vs. 0.732) but at the cost of lower recall (0.382 vs. 0.565). This indicates that **adp-base** is more conservative in predicting intelligible words, preferring to avoid false positives, which aligns with the aim of prioritizing precision in an educational context. For the unintelligible class, the continually pretrained model outperforms **base.en** on both precision (0.420 vs. 0.403) and recall (0.894 vs. 0.586), leading to a notable improvement in the model's ability to detect unintelligible pronunciations.

Method	**Metric**	**Intelligible Class**		**Unintelligible Class**	
		base.en	**adp-base**	**base.en**	**adp-base**
Whisper Judge Phonetic	Precision	0.732	0.878	0.403	0.420
	Recall	0.565	0.382	0.586	0.894
	F1-Score	0.638	0.533	0.477	0.571
	F0.5-Score	0.691	0.697	0.429	0.470

Table 4.4: MOOC-2024 Test – Metric Score Comparison

While the F1-score for the intelligible class is lower for **adp-base**, the F0.5-score, which weights precision more heavily, is slightly higher (0.697 vs. 0.691), reflecting that the model better meets the intended emphasis on high precision. For the unintelligible class, both the F1-score and F0.5-score are higher for the pretrained model, suggesting more reliable detection of mispronounced words overall.

It should be noted that a direct comparison between **base.en** and our **adp-base** model does introduce an additional factor: **base.en** is a dedicated English-only variant of Whisper, whereas our **adp-base** model is based on the multilingual Whisper base model with additional domain-specific training. This difference means that the observed improvements cannot be attributed solely to continual pretraining but may also partly reflect architectural or training data differences between the original English-only and multilingual base versions.

Overall, these results suggest that continual pretraining on isolated word data can enhance the detection of unintelligible pronunciations, while maintaining strong precision for intelligible outputs, which is critical for reliable pronunciation feedback.

4.2.4 Results: MOOC-2025 Test Set

Table 4.5 reports the average intelligible and unintelligible class performance of the Large Adapted model on the MOOC-2025 test set using Whisper's Judge Phonetic scoring at a confidence threshold of 0.3. This configuration was identified in the development experiments as the best overall performer and was therefore selected for final evaluation. For the intelligible class, the model reached a precision of 0.741, with recall at 0.562. This imbalance is reflected in the F1-score (0.654) and F0.5-score (0.600), indicating that while the model was effective at correctly labeling many intelligible utterances, it still failed to capture a substantial proportion of them. These results suggest a stronger emphasis on precision, leading to fewer false positives but at the cost of missing some true positives.

The unintelligible class shows the inverse trend: recall (0.726) is notably higher than precision (0.581), indicating that the model is more effective at capturing unintelligible utterances than at avoiding false positives. The F0.5 score for the unintelligible class (0.616) is slightly lower than for the intelligible class, though the gap is narrow, reflecting a relatively balanced but not highly accurate performance across both classes.

Model	Class	Precision	Recall	F1-Score	F0.5-Score
Adapted Large	Intelligible	0.741	0.562	0.654	0.600
	Unintelligible	0.581	0.726	0.616	0.588

Table 4.5: Adapted Large Whisper Judge Phonetic. MOOC-2025 Test Average Results. Confidence Threshold 0.3.

Per-Word Performance Analysis

Table 4.6 presents the per-word performance of the Large Adapted model on the MOOC-2025 test set using Whisper's Judge Phonetic scoring at a confidence threshold of 0.3. Overall, the intelligible class exhibits generally higher precision scores across most words, with particularly strong performance on items such as *bee* (0.904), *boot* (0.917), and *mate* (0.959). However, recall for the intelligible class is more variable, with certain words such as *putt* (0.037) and *poor* (0.059) demonstrating severe underdetection, suggesting that the model struggles to recognize these pronunciations as intelligible even when they are correctly pronounced or labeled as such.

Conversely, the unintelligible class tends to show higher recall than precision. For example, words like *Paul* (0.955 recall) and *pat* (0.912 recall) are frequently identified as unintelligible when they should be, but at the cost of higher false positive rates, reflected in lower precision scores (0.708 and 0.403, respectively).

Certain minimal pairs reveal asymmetric performance patterns between classes. For instance, *caught* shows strong results for the unintelligible class (precision 0.938, recall 0.803) while performing less consistently for the intelligible class (precision 0.559, recall 0.826). Similarly, the *fir–fur* pair both display high unintelligible-class recall (0.952 and 0.932, respectively), but recall drops substantially in the intelligible class (0.441 and 0.450). These discrepancies suggest that subtle vowel quality differences remain challenging for the model to classify reliably in both directions.

The intelligible class words that perform the poorest, such as *poor* (F5 = 0.098) and *putt* (F1 = 0.161), are characterized by low recall or very low precision, indicating inconsistent recognition behavior. On the unintelligible side, certain words such as *bee* (precision 0.333) and *pet* (precision 0.231) suffer from excessive false positives.

In summary, while the aggregate metrics offer a broad indication of performance, the per-word analysis reveals that the model's behavior is far from consistent across the word list. High recall in the unintelligible class often comes at the expense of precision, reflecting a tendency to over-predict unintelligibility. Conversely, several intelligible words are markedly under-recognized, pointing to cases where the model's predictions diverge from the gold standard labels. In many of these low performing cases, the predicted transcriptions also fail to trigger the confidence threshold, set at a relatively low value of 0.3, meaning they are not automatically filtered as intelligible and instead remain as misclassifications. These patterns highlight that, beyond

overall scores, word-level variation remains a key source of error and warrants further investigation in targeted error analyses.

	Precision		Recall		F1		F0.5	
Word	**In**	**Un**	**In**	**Un**	**In**	**Un**	**In**	**Un**
Paul	0.667	0.708	0.188	0.955	0.293	0.813	0.441	0.746
bard	0.514	0.828	0.621	0.757	0.562	0.791	0.533	0.813
bay	0.828	0.405	0.658	0.630	0.733	0.493	0.787	0.436
bear	0.611	0.711	0.717	0.604	0.660	0.653	0.630	0.687
bee	0.904	0.333	0.955	0.182	0.929	0.235	0.914	0.286
beer	0.697	0.727	0.836	0.545	0.760	0.623	0.721	0.682
bird	0.632	0.867	0.750	0.788	0.686	0.825	0.652	0.850
board	0.318	0.862	0.840	0.357	0.462	0.505	0.363	0.672
boat	0.885	0.473	0.371	0.921	0.523	0.625	0.693	0.524
boot	0.917	0.600	0.775	0.828	0.840	0.696	0.884	0.635
bout	0.865	0.710	0.640	0.898	0.736	0.793	0.808	0.741
boy	0.853	0.240	0.771	0.353	0.810	0.286	0.836	0.256
buy	0.867	0.542	0.855	0.565	0.861	0.553	0.864	0.546
caught	0.559	0.938	0.826	0.803	0.667	0.865	0.597	0.908
city	0.870	0.365	0.548	0.760	0.672	0.494	0.778	0.408
cot	0.762	0.457	0.716	0.516	0.738	0.485	0.752	0.468
daft	0.829	0.318	0.392	0.778	0.532	0.452	0.678	0.361
dance	0.768	0.333	0.606	0.519	0.677	0.406	0.729	0.359
doll	0.688	0.412	0.524	0.583	0.595	0.483	0.647	0.438
fair	0.790	0.657	0.803	0.639	0.797	0.648	0.793	0.653
farther	0.880	0.770	0.564	0.950	0.688	0.851	0.791	0.801
father	0.615	0.909	0.982	0.222	0.757	0.357	0.665	0.562
fern	0.786	0.618	0.611	0.791	0.688	0.694	0.743	0.646
fir	0.833	0.759	0.441	0.952	0.577	0.845	0.708	0.792
fur	0.818	0.714	0.450	0.932	0.581	0.809	0.703	0.749
hat	1.000	0.228	0.488	1.000	0.656	0.371	0.827	0.270
knows	0.671	0.565	0.831	0.351	0.742	0.433	0.698	0.504
mate	0.959	0.636	0.897	0.824	0.927	0.718	0.946	0.667
meat	0.789	0.684	0.789	0.684	0.789	0.684	0.789	0.684
meet	0.862	0.467	0.778	0.609	0.818	0.528	0.843	0.490
nose	0.843	0.429	0.897	0.316	0.870	0.364	0.854	0.400
pat	0.870	0.403	0.303	0.912	0.449	0.559	0.633	0.453
pause	0.423	0.783	0.423	0.783	0.423	0.783	0.423	0.783
paw	0.545	0.807	0.261	0.934	0.353	0.866	0.448	0.829
paws	0.529	0.831	0.409	0.889	0.462	0.859	0.500	0.842
pet	0.955	0.231	0.259	0.947	0.408	0.371	0.621	0.272
pit	0.850	0.312	0.236	0.893	0.370	0.463	0.559	0.359
plate	0.917	0.500	0.667	0.846	0.772	0.629	0.853	0.545
pole	0.857	0.533	0.125	0.980	0.218	0.691	0.395	0.587
pool	0.657	0.682	0.523	0.789	0.582	0.732	0.625	0.701
poor	0.118	0.595	0.059	0.758	0.078	0.667	0.098	0.622
pore	0.571	0.541	0.417	0.688	0.482	0.606	0.532	0.565
pot	0.841	0.304	0.487	0.708	0.617	0.425	0.734	0.343
pour	0.656	0.554	0.420	0.766	0.512	0.643	0.590	0.586
pull	0.450	0.539	0.205	0.788	0.281	0.641	0.363	0.576
put	0.886	0.375	0.437	0.857	0.585	0.522	0.735	0.423
putt	1.000	0.735	0.037	1.000	0.071	0.847	0.161	0.776
seedy	0.800	0.670	0.114	0.984	0.200	0.797	0.364	0.716
tide	0.472	0.733	0.515	0.698	0.493	0.715	0.480	0.726
tied	0.514	0.729	0.543	0.705	0.528	0.717	0.519	0.724
wait	0.850	0.429	0.810	0.500	0.829	0.462	0.842	0.441
weight	0.901	0.692	0.889	0.720	0.895	0.706	0.899	0.698

Table 4.6: Adapted Large Whisper Judge Phonetic MOOC-2025 Test results per word. In = Intelligible, Un = Unintelligible. Confidence Threshold 0.3.

Chapter 5

Error Analysis

5.1 Confusion Matrix

To better understand the performance of the Large Adapted model on the MOOC-2025 test set, Figure 5.1 presents the confusion matrix at a confidence threshold of 0.3 using the phonetic scoring method. This visualization allows for inspection of both correct classifications and misclassifications, offering insight into the model's strengths and weaknesses in distinguishing between intelligible and unintelligible speech.

From the perspective of the unintelligible class, the definitions of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) follow those outlined in subsection 3.2.6. In this case, the model achieved 1,678 TPs, meaning it correctly identified these utterances as unintelligible.

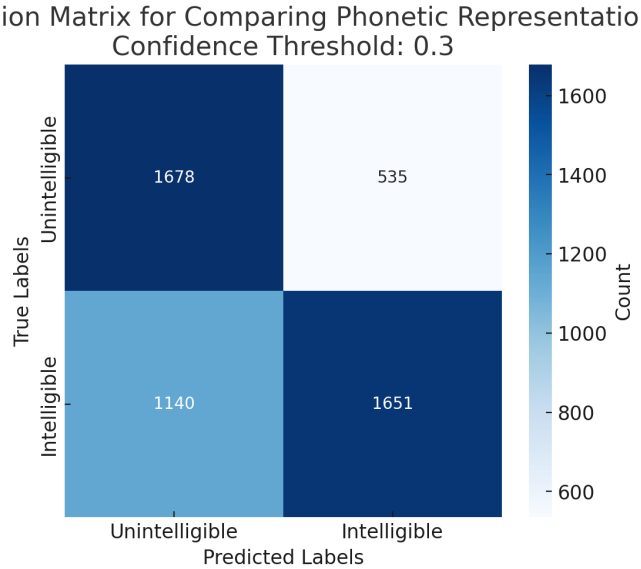

Figure 5.1: Example of word-level segmentation and labeling in Audacity. Each labeled region corresponds to a single word utterance with intelligibility classification.

The number of FPs was 1,140, representing cases where intelligible words were incorrectly labeled as unintelligible. This relatively high figure indicates a tendency toward over-predicting unintelligibility, which directly lowers the precision for this class. Such behavior suggests that while the model is effective at capturing actual errors, it

may also penalize borderline or accented pronunciations that would be intelligible to a human listener.

The 535 FNs correspond to unintelligible words misclassified as intelligible. These errors reduce recall, as they represent missed opportunities to identify actual pronunciation problems. Although the number is lower than that of false positives, it still highlights a limitation in the model's ability to capture all unintelligible cases.

Lastly, the model correctly classified 1,651 intelligible words as intelligible (TNs). While this does not directly affect the unintelligible-class precision and recall, it indicates that the model retains a reasonable ability to affirm correctly pronounced words, which is essential for maintaining learner confidence in feedback scenarios.

Overall, the confusion matrix underscores a consistent pattern: the model is adept at detecting unintelligible speech but tends to over-predict it, leading to excessive false positives. This imbalance suggests that further calibration of the decision threshold, or there may be need of a reassessment of gold standard label definitions.

5.1.1 Error Distribution Across Words

Table 5.1 presents the distribution of TPs, TNs, FPs, and FNs for each word in the MOOC-2025 test set. These counts offer a granular view of where the classification pipeline succeeds and where it fails, revealing patterns that are not immediately visible in the aggregated metrics.

Some words stand out for having far more errors than others. For example, *pet*, *pit*, *pat*, and *pole* each have over 40 false positives, making them among the most frequently misclassified. These words share similar sounds at the beginning or end, particularly voiceless stop consonants, that may be harder for the model to capture accurately, especially when produced by L2 speakers with varying degrees of aspiration or clarity. In contrast, words like *bee*, *meat*, and *caught* show relatively low error counts, suggesting that their sound patterns are clearer and less likely to be confused within this classification task.

High false negative counts are less common but still notable for certain words, such as *father* (35 FNs), *board* (45 FNs), and *bear* (21 FNs). These cases suggest that the model sometimes struggles to spot mispronunciations for words with more complex sound patterns, even when the pronunciation is off enough that a human listener might notice. In these situations, the ASR system often produces the correct word with high confidence, masking the actual errors.

Looking across the full set of results, the distribution of mistakes points to a few recurring trouble spots, particularly differences in vowel sounds, the way some consonants are pronounced at the start of words, and the clarity of final consonants. These patterns are explored further in the next subsections, where we look more closely at specific examples to see how they might be influencing the intelligibility classification.

Table 5.1: Per-Word Prediction Outcomes for MOOC-2025 Test Set

Word	True Positive	True Negative	False Positive	False Negative
bard	53	18	11	17
bay	17	48	25	10
bear	32	33	13	21

Continued on next page

Table 5.1: Per-Word Prediction Outcomes for MOOC-2025 Test Set

Word	True Positive	True Negative	False Positive	False Negative
bee	2	85	4	9
beer	24	46	9	20
bird	52	24	8	14
board	25	21	4	45
boat	35	23	39	3
boot	24	55	16	5
bout	44	32	18	5
boy	6	64	19	11
buy	13	65	11	10
caught	61	19	4	15
city	19	40	33	6
cot	16	48	19	15
daft	21	29	45	6
dance	14	43	28	13
doll	21	33	30	15
fair	23	49	12	13
farther	57	22	17	3
father	10	56	1	35
fern	34	33	21	9
fir	60	15	19	3
fur	55	18	22	4
hat	13	42	44	0
knows	13	49	10	24
mate	14	70	8	3
meat	26	45	12	12
meet	14	56	16	9
nose	6	70	8	13
pat	31	20	46	3
Paul	63	6	26	3
pause	54	11	15	15
paw	71	6	17	5
paws	64	9	13	8
pet	18	21	60	1
pit	25	17	55	3
plate	11	22	11	2
pole	48	6	42	1
pool	45	23	21	12
poor	47	2	32	15
pore	33	20	28	15
pot	17	37	39	7
pour	36	21	29	11
pull	41	9	35	11
put	24	31	40	4

Continued on next page

Table 5.1: Per-Word Prediction Outcomes for MOOC-2025 Test Set

Word	True Positive	True Negative	False Positive	False Negative
putt	72	1	26	0
seedy	63	4	31	1
tide	44	17	16	19
tied	43	19	16	18
wait	6	34	8	6
weight	18	64	8	7
Total	1678	1651	1140	535

5.2 Notable Errors

To identify systematic weaknesses in the model's predictions, a per-word confusion matrix was generated for each item in the word list, comparing the predicted labels to the gold-standard annotations across all utterances. The results for each word, along with their associated true labels, were compiled into a spreadsheet to facilitate visual inspection of potential trends or recurring patterns. From these observations, a subset of words exhibiting unusually high misclassification rates was selected for a more detailed qualitative error analysis. This process allowed for targeted investigation into specific phonetic and acoustic factors that may challenge the model or interact unfavorably with the intelligibility classification pipeline.

5.2.1 Voiceless Stop Aspiration

A recurring pattern was observed for certain monosyllabic words beginning with voiceless stops—specifically *pat*, *pit*, *pot*, and *put*. These items exhibited disproportionately high false positive rates, with the model frequently classifying otherwise clear pronunciations as unintelligible. A plausible explanation for this trend lies in the role of aspiration in English. In stressed syllable, initial position, /p/, /t/, and /k/ are typically produced with a burst of air following the stop release, known as aspiration. Variation in aspiration duration is common among speakers from different L1 backgrounds, and while native English listeners can generally accommodate such variation without a loss of intelligibility, English learners may have difficulty doing so.

Manual inspection of the audio confirmed that many of these so-called errors were, in fact, comprehensible. However, the gold standard labeling process for the test data may have contributed to the discrepancy. Initial annotations did not explicitly consider aspiration when determining intelligibility, meaning that several tokens labeled as intelligible lacked this feature. In subsequent review, Dr. Laura Rupp identified aspiration as an essential criterion for intelligibility, suggesting a potential mismatch between labeling guidelines and model evaluation. This inconsistency may partly explain the elevated false positive counts, as the model's learned decision boundaries could have been influenced by similar patterns present in the training data.

5.2.2 Vowel Substitution

Another notable pattern emerged with the word *daft*, which showed a high rate of FP misclassifications. Listening to the audio revealed that most speakers produced an arguably clear and intelligible version of the word, yet the intelligibility pipeline had difficulty applying the correct label.

Table 5.2 shows the distribution of phonetic representations for these cases. The most frequent variants occur in the vowel position, with the target vowel (represented as **AE1** in CMUDict) in *daft* often replaced by **EH1**, as in *deft*. Rows marked "N/A" indicate instances where CMUDict did not contain a phonetic representation for the model's transcription. In these cases, the model output is shown in parentheses, and in most of them, the only difference from the target is the vowel.

Phonetic Representation	**Frequency**
D AE1 F T	35
D EH1 F T	24
N/A (Duft)	12
N/A (Nonsensical)	8
N/A (Doft)	3
DH AE1 T	3
D EH1 T	2
D EH1 TH	2
D AH1 F	1
D AA1 R T	1
D AA1	1
S T AH1 F T	1
D EH1 V	1
D AH1 B	1
D AO1 TH	1
AW1 T	1
AE1 F T	1
HH AE1 V	1
D AH1	1
D EH1 F	1

Table 5.2: Frequency of target word *daft*'s CMUDict representations on MOOC-2025 test set.

From a phonetic standpoint, this is a straightforward vowel substitution. In human-to-human communication, such a shift rarely affects intelligibility, particularly in isolated words. However, within the classification pipeline, the system's strict reliance on CMUDict's canonical forms treats this vowel change as a completely different word, triggering an error. This highlights a broader challenge in automated intelligibility assessment: the system applies categorical phoneme boundaries, while human listeners often tolerate or adapt to minor vowel variation. In the case of *daft*, this results in otherwise clear pronunciations being labeled unintelligible, inflating false positive counts and disproportionately impacting the model's apparent performance on vowel-sensitive items.

5.2.3 Word-Final Consonant Reduction

Within the test set, *board* was notable for generating a relatively high number of false negatives. In many recordings, the final /d/ was either absent or so weakly articulated that it was difficult to detect. This reflects a well-documented phenomenon in second-language English: final stop consonants are prone to reduction, devoicing, or complete deletion, particularly when the learner's first language does not require full closure or voicing in word-final position. What distinguishes *board* in this dataset is that, despite the absence or weakening of the final consonant, the model frequently transcribed the word correctly and even produced a phonetic representation that included the final /d/. Annotators, however, appeared to place significant weight on the audible presence and clarity of the consonant when assigning gold labels. As a result, tokens where the consonant was not clearly audible were often marked as unintelligible by human raters, even though the model treated them as accurate matches.

This mismatch between the human labeling criteria and the model's transcription behavior helps explain the high false negative count for *board*. The model's reliance on broader acoustic and contextual cues, such as vowel quality, length, and residual articulatory noise, allowed it to identify the intended word even when the final consonant was diminished or absent. While this supports robust recognition, it can obscure underlying segmental inaccuracies, highlighting the difficulty of aligning ASR-based predictions with perceptual criteria for intelligibility.

Chapter 6

Discussion

This chapter discusses the implications of the results presented in Chapter 4, addressing the research questions, considering limitations of the study, and outlining directions for future work.

6.1 Discussion of Results

The evaluation of original and continually pretrained Whisper models across a range of sizes revealed clear patterns in their binary intelligibility classification ability of single-word English utterances. Two main factors, model size and continual pretraining, emerged as key influences on performance.

For the *intelligible* class, continual pretraining consistently improved performance, particularly for the *medium-adapted* and *large-adapted* models, which achieved higher precision under Whisper Judge phonetic scoring than their original counterparts. These adapted models also demonstrated greater stability in F1 and F0.5 scores across varying confidence thresholds, suggesting an improved balance between precision and recall when compared to the more volatile performance of smaller or unadapted models. Larger models generally maintained stronger overall performance than smaller ones, although the gap narrowed once continual pretraining was applied, highlighting that domain adaptation can partially offset the limitations of smaller models.

For the *unintelligible* class, the pattern was more pronounced. Adapted models such as *medium-adapted*, *turbo-adapted*, and *large-adapted* maintained consistently lower, but more stable, precision across the full threshold range. This contrasted with the sharp precision drop-offs in original models beyond a confidence threshold of 0.5. Crucially, adapted models not only recognized unintelligible speech more reliably, but did so with greater confidence. Because the classification pipeline assigns the unintelligible label only when prediction confidence exceeds the set threshold, this behavior suggests that continual pretraining improved confidence calibration for these cases. Larger adapted models, in particular, benefited from this effect, producing more robust predictions across a range of decision boundaries.

6.2 Answering the Research Questions

The experiments in Chapter 4 provide clear evidence that continual pretraining on domain-specific, single-word utterances can improve both word recognition and intelligibility classification in Whisper models. Across the range of model sizes tested,

adapted models generally demonstrated stronger and more stable performance than their original counterparts, though the extent of improvement varied by task and model capacity.

For the first research question, whether continual pretraining improves isolated word recognition, the results show a consistent advantage for adapted models, particularly in challenging cases with non-canonical pronunciations. The large-adapted model, for example, maintained high recognition accuracy across a range of confidence thresholds, while the original large model was more sensitive to threshold changes. Even smaller models, such as the base-adapted version, showed measurable gains over their original forms, though the most pronounced improvements appeared in the medium and large-sizes. This suggests that while continual pretraining benefits all model scales, larger architectures are better able to integrate the phonetic and lexical patterns present in the target domain. Interestingly, the turbo-adapted model performed competitively in some cases, but did not consistently surpass the large-adapted model, hinting that adaptation benefits are not solely determined by parameter count.

Turning to the second research question, whether continual pretraining improves intelligibility classification. the strongest gains were observed in the unintelligible class. Adapted models not only identified unintelligible speech more reliably but also maintained this performance across a wider range of thresholds. This contrasts with the original models, which often saw steep drops in performance as the threshold increased. These results indicate that continual pretraining improves confidence calibration for unintelligible cases, a crucial factor given that the classification pipeline applies the unintelligible label only when model confidence exceeds the set threshold.

Model size again played a role in these outcomes. The medium-adapted and large-adapted models offered the most consistent balance between intelligible and unintelligible classification, while the base-adapted model, though improved over its original version, had more difficulty maintaining performance across both classes. The turbo-adapted model performed well in certain metrics but did not consistently match the large-adapted model's robustness. Taken together, these findings confirm that continual pretraining can meaningfully improve performance on both recognition and intelligibility classification tasks, with the largest benefits emerging in models that combine higher capacity with targeted domain adaptation.

6.3 Limitations

While the results presented in this thesis demonstrate measurable improvements through continual pretraining, several limitations should be acknowledged when interpreting these findings.

First, the quality and consistency of the gold-standard labels in the MOOC-2025 dataset may have introduced noise into the evaluation. As discussed in Chapter 5, gold labels were created by human annotators, primarily interns, who sometimes applied differing criteria for determining intelligibility. In some cases, this led to discrepancies between what a trained phonetician might consider intelligible and the assigned label. For example, aspiration in voiceless stops was not consistently considered during initial labeling, which may have inflated false positive counts for words like *pat* and *pit*. Similarly, in cases such as *board*, annotators appeared to place heavy weight on final consonant articulation, even when other cues would allow a human listener to infer the intended word. These inconsistencies mean that the evaluation sometimes reflects

alignment with annotator criteria rather than a purely phonetic or perceptual standard.

Second, the method used to derive the model's confidence scores may have limited the precision of the intelligibility classification step. The current pipeline considers only the probability of the top-1 predicted token sequence when comparing against the set threshold. This approach does not account for situations in which the model assigns moderately high probability to multiple plausible transcriptions. In practice, this can lead to overconfident assignments when the top-1 prediction is only marginally more likely than the next-best alternative, or underestimation of intelligibility in cases where a correct alternative exists just below the top prediction. A top-n or cumulative probability approach might yield more stable and informative confidence measures, potentially reducing borderline misclassifications.

Finally, while multiple Whisper model sizes were evaluated, the study was limited to the versions and training configurations available at the time. It remains possible that alternative fine-tuning strategies, more diverse domain-adaptation data, or different decoding approaches could lead to further gains, particularly for smaller models that benefited less consistently from continual pretraining.

In sum, the results should be interpreted with these constraints in mind. Some performance differences may reflect characteristics of the evaluation framework and annotation practices rather than the model's inherent ability to capture intelligibility in a listener-like fashion.

6.4 Future Work

Several promising directions emerge from the findings and limitations of this study. One immediate avenue for improvement lies in refining how model confidence scores are derived for intelligibility classification. The current pipeline relies solely on the probability of the top-1 predicted transcription, which may not fully capture the model's uncertainty. Adopting a top-n or cumulative probability approach could provide a more nuanced representation of confidence, particularly in cases where multiple transcriptions are plausible. Such methods might reduce borderline misclassifications and yield more stable performance across varying confidence thresholds.

A second promising direction involves expanding the continual pretraining dataset to include both correctly pronounced and mispronounced audio. At present, the adaptation data primarily consists of intelligible speech, which may bias the model toward recognizing intelligible pronunciations more. Introducing carefully annotated mispronounced tokens, covering common segmental errors, vowel shifts, and consonant reductions, could help the model better distinguish between intelligible and unintelligible cases during inference. This would align the model's learned decision boundaries more closely with the types of variation it encounters in the classification task.

Taken together, these directions aim to build on the strengths demonstrated in this study while addressing its current constraints, moving toward an intelligibility classification system that is both more accurate and reliable.

Chapter 7

Conclusion

This thesis investigated the use of continual pretraining to adapt Whisper ASR models for binary intelligibility classification of single-word English utterances. Building on Whisper's robust zero-shot capabilities, the study explored whether targeted domain adaptation could improve both recognition accuracy and the classification of speech as intelligible or unintelligible.

The experiments demonstrated that continual pretraining led to consistent improvements in several areas. For ASR recognition, adapted models—particularly in the larger configurations—achieved more stable performance across confidence thresholds, with notable gains in correctly identifying unintelligible speech. This effect was most pronounced in the *large-adapted* model, which emerged as the best-performing configuration for the final evaluation. The benefits extended to balanced metrics such as F1 and F0.5 scores, suggesting better calibration between precision and recall compared to the original models.

However, the results also revealed that improvements were not uniform across both classes. While unintelligible speech recognition benefited from adaptation, the intelligible class showed mixed results, with original models sometimes retaining higher raw precision. These differences highlight the complexity of balancing both classes within the same classification pipeline.

The error analysis provided further insight into the system's behavior. Patterns such as vowel substitutions, word-final consonant reduction, and over-prediction of unintelligibility revealed that certain phonetic deviations, while acceptable to human listeners, were penalized by the ASR–phonetic matching pipeline. This was compounded by potential inconsistencies in the gold-standard annotations and by the confidence scoring method, which relied exclusively on the top-1 prediction. Such factors likely inflated false positive and false negative rates in specific cases.

In addressing the research questions, the findings confirm that continual pretraining can enhance Whisper's performance in both recognition and intelligibility classification, particularly for unintelligible speech, and that these gains are more substantial in larger model sizes. Nevertheless, the study also underscores that model adaptation alone cannot fully address the limitations inherent in the current confidence scoring and evaluation framework.

Overall, this work contributes to the growing body of research on domain-adapted ASR for educational contexts and points to concrete directions for improvement. Refining confidence scoring, incorporating a broader range of pronunciation variations in training, and aligning annotation practices with perceptual intelligibility judgments

stand out as promising avenues for future development.

Appendix

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./whisper-small-singleword",
    per_device_train_batch_size=16,
    gradient_accumulation_steps=1,
    learning_rate=1e-5,
    warmup_steps=500,
    num_train_epochs=15,
    gradient_checkpointing=True,
    fp16=True,
    eval_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    save_steps=1000,
    eval_steps=1000,
    logging_steps=25,
    report_to=["tensorboard"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False,
    push_to_hub=True,
)
```

Figure 1: Training Arguments for Continually Pretraining Whisper Small

References

T. M. Derwing and M. J. Munro. Accent, intelligibility, and comprehensibility: Evidence from four l2 learners. *Studies in Second Language Acquisition*, 19(1):1–16, 1997.

S. Gandhi. Fine-tune whisper for multilingual ASR with hugging face transformers, 2022. URL `https://huggingface.co/blog/fine-tune-whisper`.

B. Hixon, B. Shitrit, and S. Epstein. Phonemic similarity metrics to compare pronunciation methods. In *Interspeech 2011*, 2011.

A. Hughes, P. Trudgill, and D. Watt. *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles*. Routledge, 5th edition, 2012. doi: 10.4324/9780203784440. URL `https://doi-org.vu-nl.idm.oclc.org/10.4324/9780203784440`.

J. Jenkins. A sociolinguistically based, empirically researched pronunciation syllabus for english as an international language. *Applied Linguistics*, 23(1):83–103, 2002. doi: 10.1093/applin/23.1.83.

D. R. Mortensen, S. Dalmia, and P. Littell. Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proceedings of COLING 2016*, pages 3475–3484, 2016.

M. J. Munro and T. M. Derwing. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1):73–97, 1995.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. URL `http://arxiv.org/abs/2212.04356`.

T. Ylonen. Wiktextract: Wiktionary as machine-readable structured data. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France, June 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.140/`.

F. Zou. Exploring an existing ASR model for a binary classification of intelligibility on MOOC english speech data. 2024.