Assessing the Reliability of Automatic Speech-Processing Methods for L2 Speech: Evidence from an Automatic Pipeline, Posteriorgrams, and Whisper Probability Scores

Mehak Sohail, Nicolas Ballier
ALTAE, Université Paris Cité , Place Paul Ricoeur F-75013 Paris, France
mehak.mehak@etu.u-paris.fr , nicolas.ballier@u-pariscite.fr
0009-0005-4209-6463, 0000-0003-2179-1043

Abstract

This paper examines the reliability of automatic speech processing methods for L2 speech analysis. The study investigates two phonetic phenomena: vowel realization and affix detection, using speech data from a single Italian speaker of English (SESS0133) from the ISLE corpus. An Automatic Speech Processing Pipeline (ASPP) was used to extract vowel formant measurements and identify final affix realizations, which were evaluated through manual auditory inspection. Posteriorgram-based phonetic predictions served as a benchmark for affix detection. In addition, it explores whether posteriorgram-based predictions can provide complementary phonetic evidence to Whisper `tiny.en` probability scores, particularly for mispronunciations that may not be reflected by those scores. The ASPP performed relatively well for formant extraction but substantially worse for final affix realization, predicting affix presence in all 14 cases, while posteriorgrams agreed with the manual label in 13 cases. The Whisper-posteriorgram analysis demonstrated that the single incorrectly classified token received a higher score than the mean score of correctly classified tokens, although this result was likely influenced by the small and imbalanced dataset. Overall, the findings suggest that the effectiveness of automatic methods for L2 speech analysis varies across phonological tasks and that their outputs should be interpreted with appropriate caution.

1 Introduction

Second language (L2) speech assessment has traditionally relied on human judgment. However, manual analysis is typically time-consuming, expensive, and difficult to scale to a large group of L2 learners. These challenges have motivated the development of automatic approaches for pronunciation training and speech evaluation (Watson et al., 1989). Earlier approaches were based on the Goodness of Pronunciation (GoP) metric (Witt and Young, 1998, 2000), which uses forced alignment to compare learner production against the reference phonemes and calculate phone-level pronunciation scores.

Recent work has sought to move beyond GoP measures by utilizing posteriorgram-based representations (Churchwell et al., 2024), which provide fine-grained phonetic information for learner speech analysis. Posteriorgrams are probability-based phonetic representations of speech that encode the likelihood of phonetic units over time. These approaches have proven effective for examining pronunciation variability, highlighting non-native patterns, and supporting Mispronunciation Detection and Diagnosis (MDD) systems (Mao et al., 2018; Li et al., 2018). More recently, the emergence of large-scale Automatic Speech Recognition (ASR) models such as Whisper has opened up new possibilities for L2 speech analysis.

Whisper is a multilingual ASR model trained on approximately 680,000 hours of audio data and primarily used for transcription and translation tasks (Radford et al., 2023). It has several multilingual models and also .en models that are trained only on English data. Recent research has demonstrated that Whisper can be repurposed for L2 speech assessment (Ballier et al., 2023; Lin et al., 2026; Chao et al., 2026). In particular, lower subtoken (Liang et al., 2025) probability scores have been associated with mispronunciation and lower speech proficiency (Ballier et al., 2024a). Recent studies have shown that Whisper's internal representations can be extracted and incorporated into systems developed to support automatic pronunciation assessment and speech scoring (Chao et al., 2026; Lin et al., 2026).

Despite these developments, the reliability of automatic speech processing approaches for the analysis of learner pronunciation remains uncertain, especially when compared with human auditory judgment. Similarly, although the correlation between Whisper low-probability scores and mispro-

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 99–110, November 19–20, 2026, University of Tartu Library

nunciation has been established (Ballier and Méli, 2024), learners may sometimes say an irrelevant word in context that Whisper still accurately recognizes, so some high-probability scores in Whisper may be ambiguous. This raises the question of whether posteriorgram-based predictions can provide complementary phonetic evidence for interpreting Whisper probability scores, particularly when higher scores do not signal a learner mispronunciation but an accurate pronunciation of an irrelevant word? The current study aims to investigate these issues. More specifically, the study addresses the following research questions:

1. How reliably does the Automatic Speech Processing Pipeline (ASPP)[1] capture vowel realizations in learner speech?

2. How accurately does the ASPP predict final affix realization in L2 English speech when compared to posteriorgram-based predictions?

3. Are Whisper's subtoken probability scores supported by posteriorgram-based phonetic evidence in L2 speech? More specifically, can posteriorgrams be used to detect mispronunciations that were not diagnosed by the Whisper probability scores?

To answer these questions, the present study uses speech data from a single Italian speaker of English (SESS0133) from the ISLE corpus (Menzel et al., 2000; Atwell et al., 2003). The target Italian speaker was chosen because he was associated with the lowest proficiency level in the ISLE metadata (1 on a 5-point scale) and was therefore a good candidate for mispronunciation detection. The study focuses on two phonological phenomena relevant to learner speech: vowel realization and final affix realization. Automatic outputs for both phenomena were generated by an Automatic Speech Processing Pipeline (ASPP) for speaker SESS0133.

The extracted vowel formant measurements were evaluated through a combination of manual auditory inspection and acoustic visualization, using a multi-tier TextGrid that displayed the spectrograms, manual phonetic annotations, and automatic scores together (see Section 3.5).

The automatic final affix realizations were evaluated against manual annotations, with posteriorgram-based predictions serving as a benchmark. Furthermore, the C++ implementation of Whisper `tiny.en` model was used to generate subtoken probability scores, as predicted by the last layer of the encoder (Ballier et al., 2024b). These scores were then examined alongside posteriorgram predictions for final affix realization to explore whether the two methods could provide complementary information about phonetic realizations in learner speech.

The study contributes to the ongoing research on the use of automatic speech-processing tools for learner speech analysis. It assesses the performance of the ASPP across two phonological tasks and explores whether posteriorgrams can be used to detect mispronunciations that are not reflected by the Whisper probability scores. It also uses posteriorgrams as a benchmark for comparing the pipeline's final affix-realization output. Insights from such comparisons are important for pronunciation research and the development of future automatic speech assessment technologies. To sum up, the current study provides preliminary insight into the extent to which the automatic methods can support the analysis of L2 speech and, at the same time, highlights when automatic outputs should be interpreted with caution.

The rest of the paper is organized as follows: Section 2 reviews previous MDD research on posteriorgram-based approaches and Whisper for L2 speech analysis. Section 3 describes the data and methodology. Section 4 presents the results, Section 5 discusses them, and Section 6 suggests directions for future research. Finally, Section 7 concludes the study.

2 Previous Research

2.1 Posteriorgram-Based MDD Approaches

In recent years, Mispronunciation Detection and Diagnosis (MDD) studies have moved beyond Goodness of Pronunciation (GoP) methods (Witt and Young, 1998, 2000) and have focused on utilizing posteriorgram-based representations (Churchwell et al., 2024) to support automatic learner analysis.

Influential work in MDD studies has sought to identify L2 phonetic structure from posteriorgram representations to improve L2 speech recognition in ASR-based approaches. For example, one study employed k-means clustering on phonemic "posteriorgrams" (PPGs) representations generated by a DNN-based acoustic-phonemic model (APM) to discover an Extended Phoneme Set in L2 speech (L2-EPS). This set included non-categorical pho-

[1] https://gitlab.huma-num.fr/gwisniewski/speech_representations_analysis

netic units for Cantonese learners of English, which means that it contained more complete information about the pronunciation patterns of L2 speech (Mao et al., 2018).

Building on this idea, Li et al. (2018) made use of unsupervised clustering to find non-native phonetic patterns within each of what they call "native phonemes" (canonical phonetic realisations of the TIMIT corpus they use as reference) for L2 English speakers of Cantonese. They applied clustering to state-based PPGs produced by a DNN trained on L1 data from the TIMIT corpus. To identify the English canonical and non-native phonetic patterns, they used forced alignment to segment both L1 and L2 speech data. They then applied Cluster Sequence Analysis (CSA) to the L1 TIMIT segments to identify (native) Canonical Phonetic Patterns (C-PPs), which were used as reference targets. Next, CSA was applied to L2 segments to identify non-native phonetic patterns (NN-PPs) that deviate from these baseline C-PPs in L2 English produced by speakers in the Chinese University – Chinese Learners of English (CU-CHLOE) corpus. A perception test confirmed that NN-PPs were perceived as distinct from the native phonemes. Although their experiment focused on L2 English speakers of Cantonese, the authors suggest that this approach can be generalized to L2 English speakers of other native languages as well. The expanded acoustic-phonemic coverage provided by this experiment has the potential to improve MDD approaches and support L2 learner pronunciation (Li et al., 2018).

Together, these studies demonstrate that posteriorgram-based approaches can capture fine-grained phonetic variation in learner speech, support the automatic identification of non-native pronunciation patterns, and provide valuable phonetic information for enhancing automatic speech evaluation.

2.2 Whisper Internal Representations for L2 Speech Assessment

More recently, the emerging research has demonstrated the potential of Whisper's internal representations and encoded features for automatic pronunciation assessment.

One recent work (Chao et al., 2026) trained a light classifier on the acoustic and linguistic features extracted from Whisper's hidden representations to evaluate its potential for learner speech assessment. The classifier delivered remarkable performance on the GEPT picture-description dataset[2], outperforming cutting-edge baselines. The study also found that incorporating auxiliary features, such as the Semantic Textual Similarity score and the Image-Text Contrastive, enhanced the system's ability to analyze content relevance. The findings highlight that Whisper's latent capabilities can be used to develop robust models for pronunciation assessment and standardized language scoring.

Similarly, Lin et al. (2026) proposed a unified Multimodal Large Language Model (MLLM) that can provide both part-specific and overall scores by evaluating acoustic and semantic evidence across an entire response session in a single pass. To capture non-lexical cues crucial for delivery (e.g., prosody and fluency), they derived an Acoustic Proficiency Prior (APP) from a frozen Whisper-large-v3 encoder and injected it into the model as a learnable prior-prefix token to ensure robust proficiency assessment. The unified model outperformed previous state-of-the-art systems, including the cascaded pipeline Perezos and Phi-4-STG, on the Speak & Improve Corpus. The results demonstrate how single-pass MLLMs and Whisper can support L2 learner pronunciation analysis and help align automated speech evaluation more closely with human assessment (Lin et al., 2026).

Going beyond posteriorgram-based methods, these studies showcase that Whisper's internal representations contain rich acoustic and linguistic information that can support language models designed for pronunciation assessment and proficiency scoring.

2.3 Whisper Probability Scores for Learner Speech Analysis

Latest research on Whisper probability scores and transcription behavior has sought to determine whether they reflect learner pronunciation quality and proficiency. These studies compare different model sizes, their confidence scores (Ballier et al., 2024b), and subtoken-level probabilities to examine how effectively they can be used for speech assessment and supporting L2 learners' pronunciation.

A previous study (Ballier et al., 2023) used Whisper to examine how model size affects L2 speech transcription. It compared outputs from different model sizes on the metrics of WER and Leven-

[2]It is a spoken English L2 assessment dataset that contains authentic spoken responses (roughly $\approx$ 85 s each) to image-based picture-description prompts.

shtein distance to inspect the transcriptions for potential error cues in L2 speech. They reported that the `tiny` model captured phonetic mistranscriptions most accurately, which could be useful for spotting learner errors. In contrast, the medium model most accurately represented the reference text.

Another work examined Whisper's transcription probabilities, language detection abilities, and subtoken confidence scores to explore how it may be used to classify L2 learners of English and support learner feedback. Using the ANGLISH and ISLE corpora with subtoken–reference alignment, the study found that the `tiny` model was most sensitive to L2 pronunciation errors, and lower subtoken confidence scores were linked to mispronunciations and lower proficiency. In contrast, the language detection for Whisper's larger multilingual models was reported to be mainly useful for identifying less advanced learners (Ballier et al., 2024a).

A more recent work (Ballier and Méli, 2024) examined the Whisper's probability scores against acoustic analysis of the learner speech to investigate whether the scores reflect pronunciation quality. It analyzed the Italian component of the ISLE corpus to compare Whisper's mean subtoken probabilities to Levenshtein distance and to acoustic correlates obtained via forced alignment. The research demonstrated that there was a negative correlation between Levenshtein distance and Whisper's mean probability scores, which suggests that Whisper's confidence scores could be a reliable cue to the learner's phonetic realization. However, the holistic vowel measures failed to predict proficiency accurately.

Hence, previous studies on Whisper probability scores have established that, depending on the task, model, and metrics, its subtoken probabilities can correlate with mispronunciations, Levenshtein distance, and proficiency levels. They highlight that Whisper probability scores could provide phonetically useful cues about pronunciation quality and learner proficiency. Therefore, Whisper has the potential to be repurposed for developing reliable automatic speech assessment approaches.

2.4 Research Gap

Despite recent advances in the development of automatic speech processing approaches, posteriorgrams, and Whisper-based speech assessment, several important aspects remain unexplored. First, the reliability of automatic methods for capturing phonetic features in learner speech, especially against manual auditory analysis, remains underexplored. Second, although Whisper probability scores have been associated with measures of pronunciation quality and learner proficiency, it remains unclear whether posteriorgram-based predictions can provide phonetic information about realizations that may not be adequately reflected by those scores. The present study addresses these issues by evaluating the reliability of the Automatic Speech Processing Pipeline (ASPP) for vowel extraction and final affix detection in learner speech against manual auditory analysis and by examining the relationship between Whisper probability scores and posteriorgram prediction accuracy.

3 Materials and Methods

This study combines automatic speech processing methods with manual auditory analysis to evaluate their reliability for learner speech. It focused in particular on ASPP and its performance for vowel formant extraction and for detecting final affix realization. The outputs of the pipeline were compared against manual auditory annotations to determine how reliably they captured the relevant phonetic features.

In addition, acoustic visualization was combined with auditory analysis to assess how well vowel extraction captured different realizations, and posteriorgram-based predictions were used as an acoustic benchmark for the analysis of final affix realization. Finally, Whisper probability scores were analyzed to investigate whether higher confidence scores were associated with more accurate posteriorgram predictions.

3.1 The ISLE corpus

The graded data from the ISLE (Interactive Spoken Language Education) corpus (Menzel et al., 2000; Atwell et al., 2003) were used in this study. The ISLE corpus consists of English speech data recorded by L2 English speakers in a noise-free environment. It contains 11,484 utterances and 17 hours, 54 minutes, and 44 seconds of speech data recorded by 23 German and 23 Italian native speakers. It is divided into several blocks containing different speech-production tasks. In particular, three blocks (A, B, C) correspond to a reading task, while the remaining blocks contain utterances that were designed to capture common non-native pronunciation issues like weak forms, problem phonemes,

and word-class alternating pairs. This corpus was created to help develop speech recognition models for improving computer-based language learning technologies.

The audio recordings of the Italian speaker SESS0133 from the ISLE corpus were used for this study. The target speaker was selected from the Italian component of the corpus because Italian speakers were generally assigned lower proficiency ratings than German speakers, so they had a better likelihood of producing non-target-like phonetic realizations relevant to the present analysis. The data were reorganized for this study by aggregating audio files from all the blocks for the target Italian speaker SESS0133. This reorganized dataset was used to evaluate the performance of the ASPP and to examine the relationship between Whisper's probability scores and the accuracy of phonetic predictions.

3.2 Automatic Speech Processing Pipeline (ASPP)

The automatic formant extraction measurements and final affix-realization datasets were generated from an ASPP[3]. The pipeline uses speech recordings and reference transcriptions as input. The Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) was used to align the speech signal with the linguistic units. The aligned intervals were used for extracting acoustic measurements and identifying the last phone. Formant extraction was carried out with the help of Parselmouth (Jadoul et al., 2018), a Python interface to Praat (Boersma, 2001). The formant measurements were taken at the vowel midpoints, so a single F1/F2 point represents each vowel. This is useful for broadly visualizing vowel-space patterns, but it has limitations for diphthongs.

The vowel dataset contains vowel labels, temporal information, and midpoint F1/F2 measurements. The final affix-realization dataset contains temporal information of the target word and final phone, and the final phone assigned by the Montreal Forced Aligner (MFA) for each occurrence of the target word. These datasets were checked auditorily to evaluate how reliably the automatic pipeline captures the relevant target vowels and final affix realizations.

[3] https://gitlab.huma-num.fr/gwisniewski/speech_representations_analysis

3.3 Posteriorgrams

For this study, posteriorgram-based phonetic predictions were generated for the final affix-realization dataset using the model described by Churchwell et al. (2024), which was trained on audio data from the Common Voice dataset, CMU Arctic, and TIMIT. The posteriorgrams were compared against manual annotations and used as a benchmark for final affix realization. The posteriorgram output for each target token was evaluated against auditory annotations and classified as either *Affix Pronounced* (AP) when the final affix was realized or *Affix Dropped* (AD) when the final affix was absent. This framework made it possible to compare the performance of posteriorgrams and ASPP for final affix detection. In addition, the relationship between Whisper probability scores and the posteriorgram-based affix detection accuracy was examined.

3.4 Whisper

This study follows the methodology adopted in Ballier et al. (2024a) for the analysis of Whisper probability scores. The C++ implementation (whisper.cpp)[4] was used to run Whisper `tiny.en`[5] for this study. The linguistic data in Whisper is encoded into subtokens; this is known as byte-pair encoding (BPE). The byte-pair encoded units, or subtokens, are associated with probability scores. The C++ implementation allows the export of subtokens and the probability scores associated with them. This data were used to examine how Whisper probability scores align with the accuracy of posteriorgram-based predictions.

3.5 Manual Annotation

Manual annotation was carried out using the recording of speaker SESS0133 and the corresponding reference transcription as input. The annotation was performed by a single annotator, an L2 speaker of English who had been exposed to English during early education and throughout her higher education. The annotator carried out phonetic annotation using TextGrid files. A multi-tier aligned TextGrid was used to compare the manual phonetic annotations with the corresponding spectrograms, automatically generated scores, and outputs. The resulting annotations were used as reference labels

[4] https://github.com/jbyunes/whisper.cpp

[5] Whisper models range from 39M to 1550M parameters; `tiny.en` corresponds to the smallest configuration (~39M) (Radford et al., 2023).

for evaluating the outputs of the ASPP and posteriorgram.

The first dataset contains vowel formant measurements of the target words: *climb*, *climbers*, and *climbing*. These words were selected because the speaker frequently pronounced the diphthong /aɪ/ in a non-target-like manner and often pronounced the word-final /b/. The pipeline extracted a total of 32 instances across the three target words. The formant values and automatic vowel labels were visualized to examine the vowel patterns identified by the pipeline. These patterns were manually checked against the TextGrid files and corresponding audio recordings. The auditory observation for each target occurrence was compared with the vowel labels and formant measurements produced by the pipeline to assess how accurately the pipeline captured vowel realization in L2 speech.

The second dataset records the realization of affixes in past-tense target words: *climbed*, *covered*, *gained*, *opened up*, *reached*, *returned*, and *turned back*. These words were selected because final past tense morphemes are often weakened or omitted by L2 speakers of English. A total of 14 occurrences were extracted for this dataset. The ASPP and posteriorgram outputs for the target occurrences were manually evaluated against manual auditory annotations. The token was considered correct when it matched the manually assigned AP or AD label and incorrect when the two labels differed. These annotations were used as reference labels for evaluating the accuracy of both methods. The agreement between posteriorgram predictions and manual annotations was also used to generate binary correctness labels for the subsequent Whisper probability analysis.

3.6 Descriptive Analysis

The performance of the ASPP and posteriorgram for final affix realization was evaluated by comparing their predictions with manual auditory annotations. Confusion matrices were used to present the counts of correctly and incorrectly predicted affix realizations for each method. Overall prediction accuracy was also calculated for both methods.

To examine the relationship between Whisper probability scores and posteriorgram prediction accuracy, each occurrence in the final affix realization dataset was assigned a binary label reflecting whether the posteriorgram output agreed with the manual annotation. A correct label (T) indicated the posteriorgram output matched the manual annotation, whereas an incorrect label (F) indicated a mismatch between the posteriorgram output and the manual annotation. The Whisper probability scores were examined descriptively for these two groups, because of the small sample size and imbalanced dataset.

4 Results

4.1 Formant Extraction of Keywords

This subsection aims to assess how well the ASPP pipeline captured vowel realization in L2 speech through acoustic visualization and auditory inspection. The following sections discuss the distribution of the automatically extracted vowel tokens shown in Figure 1. The plot outlines vowel patterns identified by the pipeline and serves as the reference for the auditory observations discussed in this section.

4.1.1 *Climb*

In the automatic vowel extraction dataset, all eight occurrences of *climb* are labeled by the pipeline as the expected diphthong /aɪ/. However, the plot in Figure 1 shows noticeable dispersion among the *climb* tokens. Occurrences 87_V1, 150_V1, 209_V1, and 543_V1 appear in the upper part of the vowel space, and several of them are positioned toward the left, while 632_V1, 645_V1, and 1413_V1 are located lower in the plot, with higher F1 values and more central distribution. This suggests acoustic variation in the speaker's production or possible differences in where the midpoint of the diphthong was measured.

Manual auditory checking reveals that this dispersion is not random, as the plotted vowels correspond to different realizations of the target word. Occurrences in the upper part of the cluster are perceived as closer to forms such as *claim*, *clean*, or a more monophthongal realization, rather than as a fully target-like /aɪ/ realization of *climb*. In contrast, 632_V1, 645_V1, 1043_V1, and 1413_V1 are closer to the target-like vowel production. Hence, although the automatic phone label for all occurrences of *climb* is the target-realization /aɪ/, the automatic formant measurements can distinguish between different auditory realizations.

4.1.2 *Climbers*

As for *climbers*, the automatic pipeline successfully separates the two vowel positions for each occurrence: V1, which corresponds to the diphthong /aɪ/ in *climb-*, and V2, which corresponds to the reduced vowel /ə/ in *-ers*. The pipeline labeled all

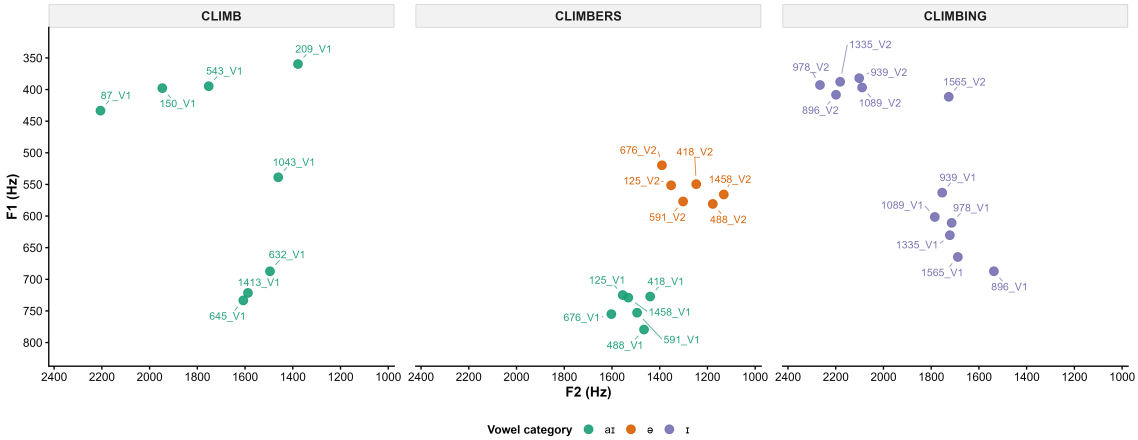


Figure 1: Token-level F1/F2 formant values for speaker SESS0133 across three target words (*CLIMB, CLIMBERS, CLIMBING*). Each point represents a single vowel token, labeled with its absolute token ID. Colors indicate vowel category as predicted by the ASPP (MFA alignment).

six V1 occurrences as /aɪ/ and all V2 occurrences as /ə/.

The plot also supports the separation between two vowel occurrences. The V1 diphthong cluster is located slightly lower and more toward the center-left of the vowel plot, while the V2 monophthong cluster is positioned higher and more toward the right. In addition, both vowel clusters for *climbers* are relatively compact, especially compared to the dispersion observed for *climb*. This suggests that the pipeline extracted the formants consistently for climbers. Manual checking confirms that V1 and V2 are broadly consistent with the vowel transcription of the pipeline. Overall, the automatic vowel extraction appears considerably reliable in separating and labeling the two vowel positions in *climbers*.

4.1.3 *Climbing*

For *climbing*, each occurrence should have two vowel positions, labeled V1 and V2. Phonologically, V1 should correspond to the diphthong /aɪ/ in *climb-* and V2 to the monophthong /ɪ/ in *-ing*. However, the extracted dataset labels both V1 and V2 as /ɪ/. Nevertheless, the plot shows a clear visual separation between V1 and V2. The V1 cluster is located lower and more toward the center of the vowel space, while the V2 cluster is positioned higher and toward the far left side. Manual checking confirms that V1 is auditorily closer to the expected diphthong /aɪ/ and V2 is closer to /ɪ/. Thus, the extracted formant values distinguish the diphthong /aɪ/ from the monophthong /ɪ/ for *climbing*, even though the automatic phone label assigned by the pipeline does not capture this distinction.

Overall, across these three target words, the formant extraction identifies acoustically meaningful differences that are not always captured by the automatic phone labels.

4.2 Automatic Detection of Final-Affix Realization: Comparison of ASPP and Posteriorgram Outputs

The final-affix realization dataset extracted from the ASPP had several issues. The most prominent issue was that the temporal information for the target word and final phone was incorrect for all 14 occurrences. Moreover, the pipeline predicted some realization of the final affix for all target instances. This suggests that the pipeline is heavily dependent on the reference transcription for the final affix prediction.

In addition, the outputs of the ASPP and posteriorgram were compared with manual annotations to evaluate their accuracy in capturing affix realization. Each token was grouped into two categories: either Affix Pronounced (AP) or Affix Dropped (AD). The resulting confusion matrices are presented in Tables 1 and 2.

Table 1: Posteriorgram Confusion Matrix

	Posteriorgram Prediction	
Manual Annotation	Affix Dropped	Affix Pronounced
Affix Dropped	10	1
Affix Pronounced	0	3

Table 2: ASPP Confusion Matrix

	ASPP Prediction	
Manual Annotation	Affix Dropped	Affix Pronounced
Affix Dropped	0	11
Affix Pronounced	0	3

The posteriorgram correctly identified 13 of the 14 affix occurrences, whereas the ASPP correctly identified only three. The posteriorgram identified all 3 instances of AP and 10 of the 11 instances of AD. In contrast, the ASPP predicted AP for all 14 occurrences, so it failed to identify any instances of AD[6]. To sum up, in the present dataset, posteriorgram predictions for affix detection aligned better with manual annotation, especially compared to the ASPP.

4.3 Whisper Subtoken Probability Scores and Posteriorgram Prediction Accuracy

The general tendency observed from the dataset showed that the single incorrectly classified posteriorgram token received a higher score than the average score of the correctly classified tokens. The mean Whisper probability score for the 13 correctly classified tokens was 0.699, while the single incorrectly classified token received a score of 0.903.

These observations show that Whisper probability scores and posteriorgram-based predictions did not follow the expected pattern, but because of the imbalance and small dataset, it is impossible to regard these observations as conclusive.

5 Discussion

5.1 Validity of Automatic Speech Processing Pipeline (ASPP)

The results indicate that the Automatic Speech Processing Pipeline (ASPP) performs relatively well for vowel formant extraction. However, its performance for automatic affix detection is considerably poor, especially compared to the posteriorgram. This suggests that the performance of the pipeline depends on the linguistic or phonetic feature being investigated.

[6]The posteriorgram achieved an accuracy of 92.9%, precision of 75.0%, recall of 100%, and F1-score of 85.7%. The corresponding values for the ASPP were 21.4%, 21.4%, 100%, and 35.3%, respectively. Though given the small dataset, these metrics should be interpreted with caution.

5.1.1 Evaluation of Automatic Vowel Formant Extraction

The pipeline performs quite well for *climbers*, as it separates and labels the two vowel positions correctly. So for *climbers*, the automatic phone label and formant values both effectively distinguish between monophthongs and diphthongs. As for *climb*, the formant values mark the distinction between different auditory realizations, but the automatic phone label assigns /aɪ/ to all occurrences. In the case of *climbing*, the formant measurements separate the two vowel positions as V1 and V2, but the phone label for both vowel positions is /ɪ/.

This highlights both the effectiveness and the limitations of the automatic pipeline. The extracted formants from the pipeline can be acoustically informative, but the phone labels for vowels are not fully reliable with the formant extraction granularity we chose here. Hence, manual auditory checking is necessary, especially when interpreting automatic vowel labels for words with more than one vowel position or words that contain diphthongs.

However, it is important to highlight that the formants were extracted at the vowel midpoint. This approach works well for monophthongs, but it appears less suitable for diphthongs because it captures only one point within a fluid and dynamically changing vowel. This may account for some of the dispersions observed among the *climb* tokens. Consequently, it is impossible to properly explain the observed variation without taking into account the effects of the timing of formant extraction on the measurements.

5.1.2 Evaluation of Automatic Final-Affix Detection

The ASPP performed poorly in predicting the realization of final affixes. It reported low agreement with manual annotation and incorrectly predicted all manually annotated cases of Affix Dropped (AD) as Affix Pronounced (AP). In addition, the temporal information in the affix dataset was completely inaccurate across all occurrences. The comparison with the posteriorgram benchmark provides further insight into the limitations of the pipeline for affix deletion. The posteriorgram demonstrated substantially higher agreement with manual annotation and correctly predicted nearly all cases of affix deletion.

One possible explanation for the weak performance of the pipeline could be that it relies heavily on reference transcription when predicting affix re-

alization. If the reference transcription includes the affix, the pipeline may be biased towards predicting its presence regardless of the actual pronunciation. In contrast, the posteriorgram predictions showed closer agreement with manual annotations and are suitable for detecting surface-level phonetic variations such as final affix deletion.

Overall, the pipeline performs better for extracting vowel formant measurements than for identifying final affix realization. The formant values produced by the pipeline provide a more accurate representation of the observed vowel variation than the automatic phone label. However, the affix deletion analysis raises concerns about the reliability of the pipeline for tasks that depend on accurate segment identification and temporal alignment. Hence, the validity and performance of the pipeline depend upon the task and the phonological feature that is being analyzed: the precision of the time frame might be sufficient to capture diphthongs but not the plosive closures.

5.2 Whisper Probability Scores and Posteriorgram-Based Phonetic Evidence

The relationship between Whisper probability scores and posteriorgram prediction accuracy was explored to examine whether higher probability scores corresponded to correctly classified phonetic realizations. Whisper probability scores reflect the model's confidence in its predicted subtokens (Ballier et al., 2024b), so correctly classified tokens would be expected to receive higher scores than incorrectly classified tokens. The analysis focused on examining whether these scores showed a consistent pattern with posteriorgram predictions evaluated against manual annotations.

The single incorrectly classified token received a higher Whisper probability score than the mean score of the correctly classified tokens. This observation does not provide evidence of a relationship between Whisper probability scores and posteriorgram prediction accuracy, nor does it provide sufficient evidence to determine whether posteriorgrams can detect mispronunciations that Whisper probability scores fail to identify. However, these observations should be interpreted carefully. The dataset for this analysis consisted of only 14 tokens and a limited set of target words. All 14 tokens were drawn from a single speaker and focused on a narrow set of phonetic features (i.e., final affix realization). The imbalance between the two groups is another important factor: 13 out of 14 tokens were correctly classified by the posteriorgram, which means only one token was in the incorrect group. Taken together, these factors make it difficult to assess whether Whisper probability scores are associated with posteriorgram prediction accuracy.

It is therefore more plausible to attribute the current findings to the limitations of the dataset instead of concluding that no association exists between Whisper probability scores and posteriorgram prediction accuracy. The observations drawn from the present dataset should be regarded as preliminary. A future investigation with a larger dataset, multiple speakers, and a wide range of suitable target words could help evaluate the relationship between Whisper probability scores and prediction accuracy.

6 Future Research

The findings of this study provide several directions for future research, including improving the formant extraction procedure and expanding the dataset to assess the performance of the ASPP pipeline. The following subsections discuss the most relevant areas for future work.

6.1 Refined Automatic Formant Extraction and Time Resolution

To address the limitations of midpoint formant extraction discussed in Section 5.1.1, future studies could adopt a multi-point extraction approach. Specifically, formants could be extracted at multiple points within diphthongs, such as at 20% and 80% of their duration, to capture a more complete representation of vowel trajectories. This approach could be useful for analyzing variations in the realization of the diphthong in words like *climb* in learner speech.

Another way to analyze accuracy in automatic predictions for mispronunciation diagnosis is to question the time resolution. Whisper time frames depend on the subtoken predicted; posteriorgrams operate at a 20 ms resolution in the implementation we used and the accuracy of the ASSP pipeline depends on the phenomena we analysed : it was validated for diphthongs but not with plosives, it may be because the time window was too small to predict or to align with plosive closures accurately.

6.2 The Encoder-Decoder Division of Labor

The current analysis can be extended by examining the division of labour between Whisper's encoder and decoder. This would also connect to the

broader research on the acoustic space hypothesis, the idea that different acoustic realizations may map differently to Whisper subtokens IDs (Liang et al., 2026). At the end of the encoder, Whisper does not provide a phonetic transcription but an interpretation of the acoustic frame by the Whisper tokenizer (one for the multilingual model, structured after the token inventory of ChatGPT2, used for the English models like `tiny.en`). Predictions from the encoder's last layer could highlight how the acoustic signal is represented before the decoder processes it. The final transcription produced by Whisper can be attributed to the decoder. Comparing these two predictions could help understand the degree to which the decoder overrides, modifies, or preserves the acoustic cues present at the encoder level and how acoustic realizations map onto Whisper subtokens.

This is especially relevant for learner speech because non-target-like pronunciations may be present in the encoder-level acoustic representation, while the decoder may transform the acoustic non-target-like cues into the intended target word. Hence, the decoder is likely to normalize the acoustic signal according to lexical and contextual probabilities. Therefore, understanding where in the encoder-decoder processing the phonetic information is lost or preserved can help determine if encoder subtokens and probability scores can better represent the learner's pronunciation. This leads to a more complex question: how does Whisper encoder interpret the acoustic space, assigning a Whisper token ID to phonetic realisations?

6.3 Whisper Models and the Tessellation of the Acoustic Space

In the current dataset, the word *ascent* was pronounced [əˈʃent] but predicted as "ancient" by the `tiny` and `tiny.en` Whisper models, but sometimes the two tokenizer model predictions differ. The multilingual (`tiny`) model properly captures "the kingdom of Nepal" (with a 0.71 probability) whereas the English `tiny.en` model has "the kingdom of Naples" (0.67 probability). Hence, acoustic realisations are mapped to Whisper subtoken inventory differently. The diagnostic posited by the Whisper models depends on the Whisper inventory of sub-tokens and on the training data. The "South Col" was pronounced with an Italian-like [kol] vowel, transcribed as the "south coal" (probability 0.48) by the `tiny.en` model and by "the south cold" (0.15) by the `tiny` model. As pointed out in previous research, probability scores are higher with the multilingual model, and lower for the English model, which is more sensitive to phonetic variability.

Further research will be needed to disentangle textual probability[7] and the acoustic features that account for the prediction. For example, "height" /ha t/ was transcribed as "hate" in the reading of "Some of them had died at that great height" whereas *heights* was more likely to be predicted in the *at great heights* context in a text describing the Everest. For a potential implementation in a computer assisted pronunciation training (CAPT) system, another step will be to map the Whisper prediction (*ancient*) to the diagnosis (*ascent* realised as [əˈʃent]). The last stage would be feedback generation (ie, the <sc> sequence in <ascent> reads /s/, not /ʃ/) and the interpretation of the Whisper subtoken transcription: "ancient" was the closest subtoken available in Whisper, even if no nasal or diphthong was pronounced by the learner.

7 Conclusion

The present study evaluated the reliability of automatic speech-processing methods for L2 speech analysis. The results suggest that the ASPP produced acoustically informative formant measurements, but showed limitations in detecting final affix realization. For formant extraction, although the automatic phone labels were not always accurate, the extracted formant values captured the distinction between different vowel realizations in the learner's speech. In contrast, the performance of the pipeline was underwhelming for final affix realization, whereas the posteriorgram performed exceptionally well, demonstrating strong agreement with manual annotation.

In the current dataset, there was no evidence of an association between Whisper probability scores and accuracy of posteriorgram predictions. However, the analyzed dataset was both limited in size and highly unbalanced, with only one token incorrectly classified by the posteriorgram. Therefore, these observations cannot be considered a comprehensive evaluation of the relationship between

[7] On the contrary, it should be noted that the posteriorgrams are produced frame by frame, regardless of conditional probability. We observed a posteriorgram prediction which is, technically speaking, phonologically impossible. For "reached a", the values predicted by the posteriorgrams were [ɹiɪtʃd], and the final posteriorgram prediction is phonologically impossible: from a phonotactic standpoint, a voiceless affricate cannot be followed by a voiced plosive.

Whisper probability scores and posteriorgram predictions accuracy, nor do they provide a decisive assessment of whether posteriorgram-based predictions can provide complementary phonetic evidence for interpreting Whisper probability scores.

Overall, the findings indicate that the effectiveness of automatic speech-processing tools depends upon the phonological feature that is being investigated. The ASPP, posteriorgrams, and Whisper probability scores seem to capture different aspects of learner speech, but the limited dataset does not permit broader conclusions about their comparative reliability. Future research with multiple speakers, a balanced dataset, and a wider range of phonological phenomena could help determine how these methods can be effectively combined for the analysis of L2 speech.

Limitations

The findings of this study should be interpreted in light of several limitations. First, the entire analysis is based on data from a single speaker and a small number of tokens. Consequently, the findings may reflect characteristics of the individual speaker instead of broader patterns of L2 speech. Additionally, the dataset focused on a limited set of phonological features, such as final affix realization. This further limits how broadly the findings can be applied to other speakers or linguistic contexts.

There are also a number of methodological limitations that should be discussed. First, the formant values were extracted at the vowel midpoint, which is less suitable for diphthongs than for monophthongs. Second, the temporal information was incorrect for all 14 tokens in the final affix dataset extracted by the ASPP pipeline. Another limitation is that the pipeline consistently predicted the presence of a final affix for all tokens, which suggests it relies heavily on the reference transcription instead of the speech signal for predicting the final affix.

Lastly, the analysis of Whisper probability scores was conducted on an unbalanced dataset, with only one incorrectly predicted token and 13 correctly predicted tokens. This reduced the statistical power of the analysis and made it difficult to draw a firm conclusion about the relationship between Whisper probability scores and posteriorgram prediction accuracy.

These limitations suggest that the present findings should be interpreted carefully. Nevertheless, the current work provides useful initial insight into the performance of the ASPP pipeline and the reliability of Whisper probability scores for the analysis of learner speech. Future research can build upon the findings of the current study to further evaluate the applicability of the introduced methods and tools for L2 speech.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Detail: ChatGPT was used for language editing, including grammar, spell-checking, sentence structure correction, tone adjustment, and improving overall clarity of the manuscript. It was also used for code development and refinement. All AI-assisted code was reviewed and validated by the authors, who retain full responsibility for the implementation, scientific accuracy, and findings presented in this study.

Acknowledgments

We thank the two anonymous reviewers for their insightful comments and Dr Jean-Baptiste Yunès for his implementation of the Whisper scoring[8]. We also thank Dr Guillaume Wisnieswki for his Speech Representations Analysis gitlab[9].

References

Eric Atwell, Peter Howarth, and Clive Souter. 2003. The ISLE Corpus: Italian and German Spoken Learner's English. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, 27:5–18.

Nicolas Ballier, Taylor Arnold, Adrien Méli, Tori Thurston, and Jean-Baptiste Yunès. 2024a. Whisper for L2 speech scoring. *International Journal of Speech Technology*, 27(4):923–934.

Nicolas Ballier, Léa Burin, Behnoosh Namdarzadeh, Sara B Ng, Richard Wright, and Jean-Baptiste Yunès. 2024b. Probing whisper predictions for french, english and persian transcriptions. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 129–138.

[8] https://github.com/jbyunes/whisper.cpp
[9] https://gitlab.huma-num.fr/gwisniewski/speech_representations_analysis

Nicolas Ballier and Adrien Méli. 2024. Investigating Acoustic Correlates of Whisper Scoring for L2 Speech Using Forced Alignment with the Italian Component of the Isle corpus. In *Swedish Language Technology Conference and NLP4CALL*, pages 20–32.

Nicolas Ballier, Adrien Méli, Maelle Amand, and Jean-Baptiste Yunès. 2023. Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech: A Case Study with French Learners of English. In *6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, volume 6, pages 282–292.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Fu-An Chao, Bi-Cheng Yan, and Berlin Chen. 2026. Probing the Hidden Talent of ASR Foundation Models for L2 English Oral Assessment. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4056–4060. IEEE.

Cameron Churchwell, Max Morrison, and Bryan Pardo. 2024. High-Fidelity Neural Phonetic Posteriorgrams. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 823–827. IEEE.

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing Parselmouth: A Python Interface to Praat. *Journal of Phonetics*, 71:1–15.

Xu Li, Shaoguang Mao, Xixin Wu, Kun Li, Xunying Liu, and Helen Meng. 2018. Unsupervised Discovery of Non-native Phonetic Patterns in L2 English Speech for Mispronunciation Detection and Diagnosis. In *Interspeech 2018*, pages 2554–2558.

Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. 2025. Beyond WER: Probing Whisper's Sub-token Decoder Across Diverse Language Resource Levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31225–31235.

Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. 2026. The limits of data scaling: Sub-token utilization and acoustic saturation in multilingual ASR. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, volume 11, pages 5785–5793.

Hong-Yun Lin, Jhen-Ke Lin, Chung-Chun Wang, Hao-Chien Lu, and Berlin Chen. 2026. Session-Level Spoken Language Assessment with A Multimodal Foundation Model Via Multi-Target Learning. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17327–17331. IEEE.

Shaoguang Mao, Xu Li, Kun Li, Zhiyong Wu, Xunying Liu, and Helen Meng. 2018. Unsupervised Discovery of an Extended Phoneme Set in L2 English Speech for Mispronunciation Detection and Diagnosis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6244–6248.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner for English: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, and Clive Souter. 2000. The ISLE corpus of non-native spoken English. In *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2*, pages 957–964. European Language Resources Association.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Charles S Watson, Daniel J Reed, Diane Kewley-Port, and Daniel Maki. 1989. The Indiana Speech Training Aid (ISTRA) i: Comparisons between human and computer-based evaluation of speech quality. *Journal of Speech, Language, and Hearing Research*, 32(2):245–251.

Silke M Witt and Steve J Young. 1998. Performance Measures for Phone-Level Pronunciation Teaching in CALL. In *Proc. of the Workshop on Speech Technology in Language Learning*, pages 99–102.

Silke M Witt and Steve J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3):95–108.