Intent vs. Surface: Recovering Acoustic Realization from Modern ASR for Pronunciation Training

Seongjin Park
Speak
360 Spear St. 4F. San Francisco, California. 94105.
seongjin@usespeak.com

Abstract

Pronunciation feedback in language learning depends on accurate detection of learner errors, but it is unclear whether modern ASR systems are suitable for this purpose. Their language models recover intended words rather than what was actually pronounced, systematically masking mispronunciations. This is a tendency we call *intent bias*. By evaluating eight ASR systems spanning three architectures on two L2 English corpora, we find that overcorrection rate correlates inversely with word error rate. In other words, ASR systems with lower WER tend to mask more pronunciation errors. We propose surface-faithful reranking, an inference-time method that uses phoneme-level acoustic similarity to select N-best hypotheses closer to what the learner actually said. Without retraining or access to model internals, the method reduces the false acceptance rate of mispronunciations by 6.0 percentage points on L2-ARCTIC and 5.6 on speechocean762. The improvement is consistent across age groups and first-language backgrounds, though substantial overcorrection remains, pointing to the need for pronunciation-aware ASR objectives.

1 Introduction

Pronunciation is among the most difficult aspects of second-language acquisition, and automated feedback has long been sought as a scalable complement to human tutors (Eskenazi, 2009; Neri et al., 2002). At the core of most computer-assisted pronunciation training (CAPT) systems lies an automatic speech recognition component whose transcriptions are compared against expected text to identify mispronounced words (Witt and Young, 2000). The accuracy of this detection step is therefore critical for providing useful feedback to learners.

Modern ASR systems, however, are optimized to report what the speaker most likely *meant* rather than what was actually produced, since their language model components assign high probability to grammatically and lexically plausible sequences. This results in effectively correcting non-standard pronunciations before they reach the downstream tasks. For example, when a learner says "she coming" (omitting "is"), the ASR often produces "she is coming". Also, when a learner produces something closer to "sheep" while reading "ship," the system may output "ship" based on the context. We refer to this tendency as *intent bias* throughout the paper. In our analysis of the speechocean762 (Zhang et al., 2021) corpus using Parakeet, 82.0% of utterances contain at least one overcorrected mispronunciation in the top-ranked hypothesis, and the system tends to be most confident when it overcorrects.

Thus, we propose *surface-faithful reranking*, a method that works entirely at inference time to recover what the learner actually pronounced. Given the N-best hypotheses from any ASR system, we obtain a reference phoneme sequence from an independent phoneme recognizer, then select the hypothesis whose phonetic transcription is closest to that reference. This reranked hypothesis drives mispronunciation detection while the original top-1 output is retained for display, allowing a single system to serve both readability and diagnostic accuracy. On L2-ARCTIC (Zhao et al., 2018), the method reduces the false acceptance rate on severe mispronunciations from 80.8% to 74.8%, and on speechocean762 test, from 70.6% to 65.0%. The improvements are consistent across age groups (speechocean762) and across all six L1 backgrounds tested (L2-ARCTIC).

The contributions of this work are threefold. First, we show that overcorrection is not incidental but systematic: across eight ASR systems spanning three architectures (autoregressive, hybrid CTC-RNNT, CTC-only), lower word error rate correlates with higher overcorrection, suggesting that

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 336–346

WER alone is insufficient for evaluating ASR in educational contexts. Second, we propose surface-faithful reranking, an inference-time method that partially mitigates this bias without retraining or white-box access. Third, we evaluate across two corpora, multiple age groups, and diverse L1 backgrounds, showing consistent gains.

2 Related Work

2.1 Mispronunciation Detection in CAPT

Goodness of Pronunciation (GOP), introduced by Witt and Young (2000), remains foundational in CAPT. GOP computes the posterior probability ratio between forced-aligned and freely decoded phone sequences, providing a frame-level pronunciation score. It requires white-box access to the acoustic model and frame-level forced alignment, assumptions that hold for HMM-GMM or hybrid systems but not for modern end-to-end ASR or commercial APIs.

End-to-end approaches train neural networks directly on L2 speech for mispronunciation detection and diagnosis (MDD) (Leung et al., 2019; Li et al., 2017). More recently, self-supervised representations from wav2vec2 (Baevski et al., 2020) have been fine-tuned for this task as well (Xu et al., 2021; Peng et al., 2021), achieving competitive results with less labeled data. Korzekwa et al. (2021) incorporate uncertainty modeling to improve detection robustness. However, all these approaches require annotated non-native speech corpora for training, data that remains scarce for most language pairs, and produce task-specific models that must be retrained for new domains.

Transformer-based assessment systems such as GOPT (Gong et al., 2022) predict pronunciation scores at multiple granularities, but their training on specific corpora limits zero-shot generalization. The method we propose avoids these constraints, as it requires no training on L2 data, no access to ASR internals, and no task-specific model.

2.2 ASR for L2 Speech and Language Model Bias

ASR systems trained on native speech perform worse on accented and non-native speech (Zechner et al., 2009). This degradation is well documented, but a subtler problem has received less attention. When ASR *succeeds* on L2 speech, it often does so by correcting the learner's errors rather than faithfully transcribing them. This overcorrection is a natural consequence of optimizing the model for low word error rate, which incentivizes recovering canonical word sequences regardless of acoustic evidence to the contrary.

Cámara-Arenas et al. (2023) formalize this as a conflict between two necessary conditions for pronunciation assessment. ASR must recognize diverse speech (requiring robustness) but must also refrain from normalizing non-standard pronunciations (requiring sensitivity). They show that Google ASR cannot satisfy both conditions simultaneously. Won (2025) provides further empirical evidence that WER correlates poorly with human pronunciation judgments for high-performing ASR systems. Chen et al. (2024) demonstrate that acoustic information is often overridden by language model priors, and propose training-time modifications. Our work shares the diagnosis but differs in approach. Rather than modifying training, we intervene at inference time and provide quantitative evidence of the WER-overcorrection trade-off across architectures.

2.3 N-best Reranking

N-best list reranking has been studied extensively for improving ASR accuracy, typically using language models or neural rescorers to select the most linguistically plausible hypothesis (Salazar et al., 2020). In the present work, we try to invert this paradigm. Instead of reinforcing linguistic plausibility, we use phonetic signals to select hypotheses that best match actual pronunciation. We hypothesize that surface-faithful candidates often already exist in the N-best list but are suppressed by the language model's preference for canonical forms.

3 Method

3.1 Problem Formulation

Standard ASR systems tend to choose the hypothesis that maximizes a combination of acoustic likelihood and language model probability. For CAPT, however, this objective is often misaligned. A hypothesis that restores a dropped word or corrects a vowel substitution scores well by conventional metrics but obscures the pronunciation error a learner needs to know about. In this study, we seek an alternative selection criterion that favors hypotheses reflecting what was acoustically produced (i.e., the *surface realization*) rather than what was linguistically intended.

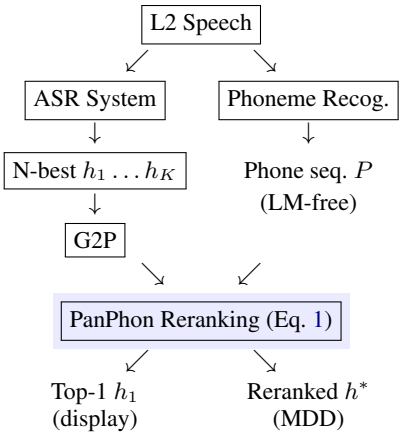

Figure 1: Surface-faithful reranking pipeline. PanPhon distance selects the hypothesis closest to the phoneme recognizer output for MDD; the top-1 hypothesis is retained for display.

3.2 Surface-Faithful Reranking

Given an utterance, we extract two independent signals, which are (1) an N-best list of text hypotheses $\mathcal{H} = \{h_1, \dots, h_K\}$ from the ASR system, and (2) a phoneme sequence P from a separate phoneme recognizer that operates without a language model. The phoneme recognizer provides a language-model-free estimate of the surface realization.

Each hypothesis h is converted to an IPA phoneme sequence using grapheme-to-phoneme (G2P) system via g2p-en (Park and Kim, 2019). We then select the hypothesis whose phonetic representation most closely matches the recognized phoneme sequence:

$$h^* = \arg\min_{h \in \mathcal{H}} \; d_{\text{PanPhon}}\big(\text{G2P}(h),\; P\big) \qquad (1)$$

where d_{PanPhon} is the weighted articulatory feature edit distance described below. The selected hypothesis h^* serves as the surface-faithful transcription for mispronunciation detection, while the original top-1 hypothesis h_1 is retained for display to the learner. This dual-output design allows a single ASR system to provide both a readable transcript and accurate pronunciation diagnostics. Figure 1 illustrates the complete pipeline.

3.3 Phonetic Distance via PanPhon

Standard edit distance treats all phoneme substitutions equally, but pronunciation errors often involve phonetically similar sounds. A learner who produces /t/ instead of /θ/ (dental fricative simplification, common among many L1 backgrounds) has made a smaller error than one who substitutes /m/ for /θ/. PanPhon (Mortensen et al., 2016) captures this by representing each phoneme as a 24-dimensional binary vector encoding articulatory features (place, manner, voicing, nasality, and others). The substitution cost between two phonemes is their Hamming distance in this feature space. This representation provides implicit severity weighting: substitutions between acoustically similar phones incur lower costs than those between dissimilar phones.

All feature weights are fixed based on phonological theory and are not tuned on evaluation data. This avoids overfitting and ensures that the method generalizes across datasets without adaptation.

3.4 Word-Level Mispronunciation Detection

After selecting the surface-faithful hypothesis h^*, we align it against the canonical prompt (the text the learner was asked to read) using word-level edit distance. Any word that is substituted, deleted, or inserted relative to the prompt is flagged as a detected mispronunciation. This alignment-based approach requires no additional thresholds or classifiers.

4 Experimental Setup

4.1 Datasets

We evaluate on two L2 English corpora that differ in speaker demographics and annotation style.

L2-ARCTIC (Zhao et al., 2018) contains 26,867 read-speech utterances from 24 adult L2 English speakers across six L1 backgrounds (Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese). Each utterance has manual phoneme-level annotations marking substitutions, deletions, and insertions. We derive word-level mispronunciation labels by marking any word containing at least one phone-level error. Following the annotation guidelines, we categorize errors by severity: *severe* errors affect intelligibility (e.g., vowel substitution /ɪ/→/i/ in "ship" heard as "sheep", or consonant deletion making a word unrecognizable), *moderate* errors have uncertain impact on comprehension, and *acceptable* variations are accent features that do not impair intelligibility (e.g., dental fricative simplification /θ/→/t/).

Speechocean762 (SO762) (Zhang et al., 2021) contains 5,000 utterances from Mandarin L1 speakers including children (under 13), teenagers (13–17), and adults (18 and older), with word-level and phone-level accuracy scores on a 0–2 scale. We derive mispronunciation labels from phone-level scores: a word is mispronounced if any of its phones has accuracy below 2.0. For severity analysis, we distinguish *severe* errors (minimum phone accuracy < 1.0, indicating a phone produced so incorrectly that it is not recognizable as the target) from *moderate* errors ($1.0 \leq$ minimum accuracy < 2.0, indicating a noticeable but identifiable deviation). We report main results on the SO762 test split (samples 2500–4999, n=2,497 utterances) for direct comparability with GOPT (Gong et al., 2022), which was trained on the SO762 training split. The full 5,000-utterance set is used only in the WER–FAR correlation analysis of §5.2 where it is explicitly disclaimed.

4.2 ASR Systems and Baselines

Our primary ASR system is NVIDIA Parakeet, a FastConformer-based hybrid CTC-RNNT model (Majumdar and Koluguri, 2024), from which we extract K=10 N-best hypotheses. As a second system, we use Whisper large-v2 (Radford et al., 2023) via faster-whisper (SYSTRAN, 2023), also with K=10 hypotheses from beam search. Testing on two architecturally different systems (Parakeet uses hybrid CTC-RNNT decoding while Whisper uses autoregressive decoding with a stronger internal language model) lets us examine how architectural choices affect both intent bias and the potential for reranking.

The phoneme recognizer is wav2vec2-lv-60-espeak-cv-ft, a multilingual CTC-based model fine-tuned on Common Voice (Baevski et al., 2020). Its language-model-free decoding produces phoneme sequences that reflect acoustic content without linguistic correction.

Our method treats the ASR system as a black box. In other words, only the N-best text hypotheses and their ranks are used, and no internal scores, alignments, or posteriors are used for reranking.

To examine how overcorrection varies with ASR accuracy, we additionally evaluate Whisper tiny, base, small, and medium (39M to 769M parameters), wav2vec2-base-960h and wav2vec2-large-960h (Baevski et al., 2020), both CTC-only models without any language model component. This gives us eight systems spanning three architectures: autoregressive (Whisper family), hybrid CTC-RNNT (Parakeet), and CTC-only (wav2vec2).

We compare against: (1) ASR top-1 baselines across all systems; (2) Kaldi-based GOP (Witt and Young, 2000), which requires white-box access to acoustic model posteriors; (3) GOPT (Gong et al., 2022), a transformer trained on SO762 for pronunciation scoring; (4) word-level ASR confidence thresholding, which flags low-confidence words as mispronounced (oracle threshold); and (5) direct phoneme comparison, which bypasses ASR entirely and compares G2P of canonical text against phoneme recognizer output using PanPhon distance (oracle threshold).

4.3 Evaluation Metrics

We report two complementary metrics following Li et al. (2017). *False Acceptance Rate* (FAR) measures overcorrection, which represents the proportion of mispronounced words that the system incorrectly accepts as correct. High FAR means learner errors go undetected. *False Rejection Rate* (FRR) measures false alarms, and it represents the proportion of correctly pronounced words incorrectly flagged as errors. High FRR annoys learners by providing misleading feedback. All metrics use word-level aggregation across the full evaluation set. Statistical significance is assessed using McNemar's test, and confidence intervals are obtained via bootstrap resampling (10,000 iterations).

5 Results

5.1 Main Results

Table 1 presents the main results across both datasets. The top-1 hypothesis exhibits high FAR on both: 80.8% on L2-ARCTIC severe and 70.6% on SO762 test. Note also the inverse trend within the SO762 column: stronger language models (Whisper large-v2, Parakeet) yield lower WER but higher FAR, foreshadowing the architecture-level correlation in §5.2.

Surface-faithful reranking reduces Parakeet's FAR by 6.0pp on L2-ARCTIC severe and 5.6pp on SO762 test ($p < 0.001$ for both), at a modest FRR cost (1.7pp and 3.9pp). Residual FAR remains high (74.8% on L2-ARCTIC severe), indicating reranking is one component of a complete CAPT pipeline rather than a standalone solution.

To contextualize this gain, we examined paired word-level decisions on SO762 test. For every word that reranking newly accepts as correct, it

Table 1: Mispronunciation detection results (%). L2-ARCTIC: severe errors (n_{misp}=4,183); SO762 test: n_{utt}=2,497, n_{misp}=6,066. Top-1 WER varies inversely with FAR across architectures, consistent with intent bias (§5.2). *Kaldi GOP requires white-box AM access; no transcription, so WER not applicable. †GOPT was trained on the SO762 train split. ‡Oracle thresholds. + *Ours* = surface-faithful reranking via PanPhon distance. Dashes indicate not evaluated on that dataset.

	L2-ARCTIC severe			SO762 test		
Method	WER	FAR	FRR	WER	FAR	FRR
Kaldi GOP*	–	20.4	74.8	–	6.0	3.0
GOPT†	–	–	–	–	34.0	54.8
Confidence‡	–	56.9	28.5	–	50.4	35.0
Phoneme-only‡	–	22.7	59.6	–	15.8	67.4
wav2vec2-base-960h	22.7	63.1	14.0	49.0	41.0	32.1
wav2vec2-large-960h	18.9	67.2	10.9	45.0	44.8	28.4
Whisper tiny	21.7	66.6	12.3	35.9	55.8	20.0
Whisper large-v2	9.2	82.7	5.1	16.9	74.2	8.4
Whisper + Ours	9.8	80.9	5.3	20.0	72.4	8.2
Parakeet Top-1	8.0	80.8	5.9	19.2	70.6	8.4
Parakeet + Ours	11.8	**74.8**	7.6	24.9	**65.0**	12.3

correctly rejects 3.7 previously accepted mispronunciations (Fixed:Hurt = 3.66:1, McNemar exact OR = 3.66, 95% CI $[3.01, 4.49]$, $p < 0.001$). Detection recall rises from 29.4% to 35.0%, a 19% relative increase that closes about 41% of the K=10 oracle gap (5.6 / (70.6 − 56.9)).

The cluster-bootstrap 95% CI for ΔFAR, resampling by utterance over 10,000 iterations, is $[-6.4, -4.9]$ on SO762 test, compared with $[-6.8, -5.0]$ on L2-ARCTIC severe. The trade-off is asymmetric: each percentage point of FAR reduction comes at roughly 0.7pp of FRR increase, concentrated in short function words (§6.6). Whisper shows a smaller 1.8pp improvement on both datasets, likely due to limited N-best diversity (§6.3).

On L2-ARCTIC, GOP achieves low FAR (20.4%) but extremely high FRR (74.8%), flagging three out of four correctly pronounced words as errors. For CAPT, persistent false alarms would erode learner trust and motivation (Neri et al., 2002). GOP also requires white-box access to acoustic model internals and frame-level forced alignment, which makes it incompatible with modern API-based ASR services. Our method offers the opposite profile within the black-box paradigm, moderate FAR with low FRR.

5.2 Overcorrection Scales with ASR Accuracy

To examine whether overcorrection is a general property of ASR optimization, we measured WER and FAR for eight systems spanning three architectures on the full SO762 set (Figure 2; numbers in this subsection use the full set for correlation power, in contrast to the test-only Table 1). We found a very strong trend between WER and FAR (r=−0.99 (p<0.001), Spearman ρ=−1.00 (p<0.001)). Wav2vec2-base-960h, a CTC-only model with no language model, has the highest WER (50.0%) but the lowest FAR (40.1%), making many transcription errors but rarely masking mispronunciations. Whisper large-v2, with its strong autoregressive decoder, achieves the lowest WER (17.7%) but the highest FAR (72.1%). Parakeet (hybrid CTC-RNNT, WER 20.3%, FAR 68.6%) and four smaller Whisper variants (tiny, base, small, medium) fall in between.

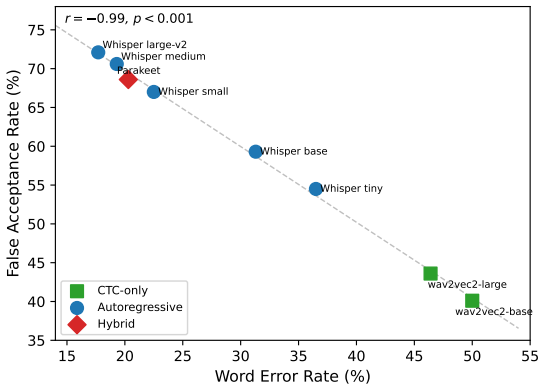

Figure 2: WER vs. overcorrection rate (FAR) on the full SO762 set (n_{misp} ≈12,235) across eight ASR systems spanning three architectures. Computed on the full set for correlation power; the test-split subset yields the same direction with wider CI. Lower WER correlates strongly with higher overcorrection (r=−0.99, p<0.001).

The trend holds across all three architecture families. Even within the CTC-only pair, the larger wav2vec2 model achieves lower WER (46.4% vs. 50.0%) but higher FAR (43.6% vs. 40.1%), consistent with the overall pattern. This suggests that overcorrection is not driven solely by explicit language models. Model capacity itself may contribute through implicit linguistic biases learned from large-scale training data. For CAPT, the practical implication is that better ASR as measured by WER may produce worse pronunciation feedback regardless of architecture.

5.3 Direct Phoneme Comparison

We also examined whether the phoneme recognizer alone, without ASR, can perform mispronunciation detection. We compared G2P-converted canonical text directly against the phoneme recognizer output using PanPhon distance, sweeping the detection threshold to find the oracle operating point. At the best threshold, this baseline achieves FAR of 15.8% but FRR of 67.4% (Table 1). This means that the system catches most mispronunciations but flags two-thirds of correctly pronounced words as errors. Because the phoneme recognizer has a $\sim$15% error rate on L2 speech, it is too noisy for word-level detection on its own. Our method combines both signals: ASR provides word-level candidates and the phoneme recognizer selects among them.

5.4 Severity-Stratified Analysis

Tables 2–4 present detailed analyses for Parakeet, which shows the larger reranking gain; Whisper's limited N-best diversity (§6.3) produces smaller improvements that do not yield informative subgroup patterns. Table 2 shows the results broken down by error severity on both datasets. On L2-ARCTIC, the improvement scales with severity. The differences were 3.1pp for acceptable-level variations, 4.3pp for moderate errors, and 6pp for severe errors. The high residual FAR on acceptable variations (92.9%) is expected and arguably desirable, since accent features like dental fricative simplification do not impair intelligibility and need not be flagged in most educational contexts.

On SO762, a different pattern emerges. Severe errors (phone accuracy below 1.0) have a much lower baseline FAR (21.5%) than moderate errors (81.9%). Pronunciations that are very different from any recognizable phone are so acoustically distinct that even a strong language model cannot map them back to the intended word. Moderate errors, by contrast, are subtle enough for the language model to override the acoustic evidence and produce the canonical word. This suggests that intent bias may be most problematic precisely for the errors that are hardest to detect by other means.

Table 2: FAR (%) by error severity (Parakeet). L2-ARCTIC categories follow annotation guidelines; SO762 categories are based on minimum phone accuracy scores.

	L2-ARCTIC		
	Accept.	Moderate	Severe
Baseline	96.0	88.4	80.8
Ours	92.9	84.1	74.8
Δ	-3.1	-4.3	-6.0
	SO762		
		Moderate	Severe
Baseline		81.9	21.5
Ours		76.7	19.0
Δ		-5.2	-2.4

5.5 Learner Demographics: Age Groups

Table 3 presents results stratified by speaker age on the SO762 test split (n=2,497 utterances). Children show the largest FAR reduction (6.8pp), followed by teens (5.1pp) and adults (5.0pp). This pattern may reflect greater acoustic variability in younger speakers' speech, where the language model's intent-oriented bias is most pronounced. The consistency of improvement across all age groups is relevant for educational deployment, where CAPT systems must serve diverse learner populations without per-demographic tuning. FRR rises by 3.8–4.8pp across strata; the modest cost is within bootstrap CI overlap and consistent with the overall trade-off in Table 1.

Table 3: FAR / FRR (%) by speaker age on SO762 test (Parakeet, n=2,497). *Ours* cells show the reranked value with Δ vs. Base.

		FAR (%)		FRR (%)	
Age	n	Base	Ours (Δ)	Base	Ours (Δ)
Child	1,039	67.3	60.5 (-6.8)	11.1	14.9 ($+3.9$)
Teen	240	84.4	79.3 (-5.1)	7.2	12.0 ($+4.8$)
Adult	1,218	70.6	65.6 (-5.0)	6.9	10.7 ($+3.8$)
All	2,497	70.6	65.0 (-5.6)	8.4	12.3 ($+3.9$)

5.6 L1 Background Generalization

Table 4 shows that the method improves detection across all six L1 backgrounds in L2-ARCTIC,

with FAR reductions ranging from 3.8pp to 4.9pp. Korean and Vietnamese speakers show the largest and smallest baseline FAR respectively (96.0% vs. 81.5%), possibly reflecting differences in how well the ASR language model can normalize each L1's characteristic error patterns. FRR rises by 1.1–3.1pp across L1 strata; Vietnamese has the smallest FRR cost (1.1pp), Chinese the largest (3.1pp). The absence of any L1 group showing degradation suggests that the method's reliance on universal articulatory features, rather than L1-specific error models, provides robust cross-linguistic generalization.

Table 4: L1 generalization on L2-ARCTIC (Parakeet, all severity levels). *Ours* cells show the reranked value with Δ vs. Base.

		FAR (%)		FRR (%)	
L1	n_{misp}	Base	Ours (Δ)	Base	Ours (Δ)
Arabic	1,406	89.9	85.5 (−4.4)	5.1	7.7 (+2.6)
Chinese	2,250	91.0	87.2 (−3.8)	5.9	8.9 (+3.1)
Hindi	1,984	95.6	91.3 (−4.2)	3.1	5.5 (+2.4)
Korean	1,236	96.0	91.2 (−4.9)	3.1	6.2 (+3.0)
Spanish	2,451	90.8	86.4 (−4.4)	5.3	8.1 (+2.9)
Vietnamese	4,138	81.5	76.9 (−4.6)	4.0	5.1 (+1.1)

5.7 Reranking Strategy Comparison

Table 5 compares alternative reranking strategies on L2-ARCTIC. PanPhon and the *phoneme-recognizer log-posterior* (PR Log-Posterior) use the same phoneme recognizer in different ways: PanPhon measures the articulatory distance between the hypothesis IPA and the recognizer's argmax phone sequence, while PR Log-Posterior scores the hypothesis IPA against the recognizer's frame-wise posterior using a proportional frame alignment. Neither signal is Parakeet's ASR decoder score, and PR Log-Posterior is not a conventional acoustic-model likelihood. On L2-ARCTIC severe errors, the two signals achieve comparable FAR (74.6 vs. 74.8), but PR Log-Posterior incurs substantially higher FRR (10.5 vs. 7.6) and WER (15.6 vs. 11.8). On SO762 test, PR Log-Posterior wins on FAR (McNemar exact OR $= 0.78$, 95% CI $[0.68, 0.90]$, $p < 0.001$), but PanPhon remains better on FRR and WER (12.3 vs. 15.9 FRR; 24.9 vs. 28.3 WER). We adopt PanPhon because its CAPT trade-off is more favorable, it avoids the proportional frame alignment that PR Log-Posterior depends on, and its failure modes are more transparent through the surface-anchor audit (§6.2). PanPhon is therefore a Pareto-efficient operating point on the WER and FRR axes.

Replacing PanPhon with unweighted Levenshtein edit distance on the same IPA sequences yields 2.9pp less FAR reduction (77.7% vs. 74.8%), confirming that articulatory feature weighting meaningfully contributes to the reranking quality. Equal-weight combination of ASR and phoneme scores yields intermediate results. The WER Oracle row shows what happens when the hypothesis closest to the canonical text is selected as an output (i.e., optimized with WER). FAR *increases* to 86.1%, indicating that optimizing for conventional ASR accuracy works against mispronunciation detection. The MDD Oracle shows the best achievable FAR from the N-best list (56.9%), representing the ceiling imposed by beam search pruning.

Table 5: Reranking strategies on L2-ARCTIC severe errors (Parakeet, n=4,183). PR Log-Posterior scores each hypothesis under the phoneme recognizer's posterior, not an ASR decoder likelihood (see §5.7).

Method	WER	FAR	FRR
WER Oracle	5.1	86.1	2.7
ASR Top-1	8.0	80.8	5.9
Levenshtein	9.5	77.7	6.3
Equal Weight	10.1	76.3	6.4
PanPhon (Ours)	11.8	74.8	7.6
Log-Likelihood	15.6	74.6	10.5
MDD Oracle	13.8	56.9	5.4

5.8 N-best List Size

Increasing K from 1 to 5 reduces FAR by 3.6pp (80.8%→77.2%), and extending to K=10 provides an additional 2.4pp improvement (77.2%→74.8%), with FRR rising modestly from 5.9% to 7.6%. We use K=10 throughout as a practical trade-off.

6 Analysis and Discussion

6.1 Quantifying Intent Bias

To understand the relationship between ASR confidence and overcorrection, we examined how overcorrection rates vary across N-best ranks on SO762 (Table 6). Rank 1, the hypothesis the system is most confident about, has the highest overcorrection rate (69.4%), with the rate decreasing at lower ranks. This is consistent with language model's preference for canonical forms being strongest at the top of the N-best list, and suggests that surface-faithful alternatives exist at lower ranks but are systematically suppressed.

This finding also explains why word-level confidence thresholding (i.e., flagging low-confidence

Table 6: Overcorrection by N-best rank on SO762 (Parakeet). % Utts = utterances with at least one overcorrection.

Rank	Overcorrection	% Utts
1 (highest)	69.4%	82.0%
3	62.5%	78.3%
10 (lowest)	59.0%	74.7%

words as mispronounced) is not an effective detection strategy. At its oracle operating point on SO762, confidence thresholding achieves only 50.4% FAR with 35.0% FRR (Table 1), which is worse than all other methods. Confidence reflects the language model's certainty about the *intended* word, not the acoustic fidelity of the transcription.

6.2 Reliability of the Surface Anchor

Our reranker selects the hypothesis whose IPA transcription is closest to a fixed reference: the G2P transcription of the prompt, which we call the *surface anchor*. We write $d_{\text{canonical}}$ for the PanPhon distance from a candidate hypothesis to this anchor. If the anchor itself disagrees with what a correctly-pronouncing learner produces, the reranker has nothing to recover from, and the residual FRR in Table 1 reflects anchor noise rather than reranker error. We test this directly.

We isolate anchor noise on L2-ARCTIC utterances annotated as fully correct (n=1,083), where the canonical anchor should agree with the phoneme-recognizer output. Even in this no-error stratum, the median normalized PanPhon distance between anchor and recognizer is 0.52 (IQR=0.49), nearly matching the error-bearing stratum's median of 0.57. The anchor is therefore systematically noisy across utterances, not selectively noisy for mispronounced speech.

Two stratifications show that this noise propagates into false rejection. First, word-level FRR scales monotonically with anchor distance: from 3.2% in the lowest $d_{\text{canonical}}$ quartile to 11.7% in the highest. Second, comparing utterances by whether their top-1 hypothesis contains an inserted word, the reranker's helped-to-hurt ratio — newly correct accepts versus newly incorrect rejects — drops from 4.7:1 without insertions to 1.5:1 with insertions (Fisher exact OR=3.04, 95% CI $[1.70, 5.46]$, $p < 0.001$). We therefore suggest single-pronunciation G2P as a measured limitation, and motivate multi-pronunciation lattices as the principled remedy (§6.5).

6.3 ASR Architecture and N-best Diversity

Whisper and Parakeet achieve similar baseline FAR on L2-ARCTIC (82.7% vs. 80.8%), yet Parakeet benefits substantially more from reranking (Table 1). The explanation lies in N-best diversity. Parakeet's hybrid decoding produces an average of 10 unique hypotheses per utterance, while Whisper's stronger autoregressive language model constrains beam search to only 4.6 unique hypotheses. In 82.7% of Whisper utterances, the top-1 hypothesis is already the closest to the oracle, leaving little room for reranking to improve. For CAPT applications, this suggests that ASR systems with weaker language models may be preferable when paired with surface-faithful reranking, since they preserve more diverse candidates.

Increasing candidate diversity through generic Whisper decoding does not close this gap: doubling the beam to K=20 raises Whisper's unique-hypotheses-per-utterance only from 2.8 to 4.3 and yields a reranked FAR improvement of just 1.4pp on SO762 test, capturing only 30% of the oracle headroom (4.6pp). The added candidates are mostly linguistic variants rather than acoustically faithful alternatives. The bottleneck is candidate generation under an intent-oriented decoder, not search width; pronunciation-aware decoding (§6.5) is the principled extension.

6.4 Deployment and Educational Implications

An alternative to reranking would be to use a CTC-only model without a language model, which exhibits lower overcorrection (Figure 2). However, such models produce transcripts with 46–50% word error rate, making them unsuitable for learner display. Our approach preserves the benefits of a strong ASR system while mitigating its overcorrection: a single ASR pass produces two outputs, where the top-1 hypothesis provides a readable transcript for display while the reranked hypothesis drives mispronunciation detection internally. Reranking adds negligible compute beyond N-best extraction.

The FAR/FRR distinction maps onto two failure modes in pronunciation feedback: high FAR leaves errors undetected, while high FRR frustrates learners with false alarms (Neri et al., 2002). Our method's profile—lower FAR with a modest FRR rise—is best suited to intermediate and advanced learners; the age-group results (Table 3) suggest particular benefit for younger learners.

Table 7: Examples from L2-ARCTIC. The top-1 hypothesis corrects the learner's pronunciation and the reranked hypothesis preserves the surface form.

Type	Top-1 (intent)	Reranked (surface)
Omission	. . . she **is** coming. . .	. . . she coming. . .
Subst.	. . . previous **wives**. . .	. . . previous wise. . .
Acoustic	. . . **was streaming**. . .	. . . were trimming. . .

6.5 Future Work

WER alone is insufficient for evaluating ASR in CAPT (Figure 2); we suggest the *Pronunciation Preservation Rate* (PPR $= 1 -$ FAR) as a complementary metric (range 25.8%–59.0% on SO762 test). Four directions remain open. **(i) Multi-pronunciation lattices** replacing the single canonical G2P, motivated by the surface-anchor audit (§6.2) showing 3.7× higher FR rate in the top $d_{\text{canonical}}$ quartile—a likely route to reducing the dominant FRR cost. **(ii) Latency-aware two-pass diagnosis**, running a wider diagnostic beam (K=100–200) alongside the display ASR; better offline detection at the cost of user-facing latency. **(iii) Pronunciation-aware decoding** objectives that balance LM probability against acoustic faithfulness, motivated by the §6.3 finding that wider beams (K=20) do not recover surface forms. **(iv) Learning-outcome user studies** to verify that the 5.6pp absolute / 19% relative recall gain translates to learning benefit.

6.6 Qualitative Examples

Table 7 illustrates three common patterns where the top-1 hypothesis masks a pronunciation error that the reranked hypothesis exposes. In the first example, the learner omits the verb, a common pattern for speakers whose L1 lacks obligatory verb marking. Here, the ASR inserts "is" because it is grammatically expected. In the second, the learner produces a word closer to "wise" than "wives," but the language model selects the contextually appropriate word. A more complex case appears in the third example, where multiple phonetic confusions lead to a substantially different surface form.

The method can also introduce errors. The 1.7–3.0pp FRR increase (Table 1) occurs when the phoneme recognizer output happens to be closer to an incorrect hypothesis than to the canonical text, and it often happens for short function words (e.g., "a," "the") where phoneme sequences are too brief for PanPhon distance to discriminate reliably. These false alarms are concentrated in high-frequency, low-information words that are less relevant for pronunciation feedback.

7 Conclusion

The results of the present study suggest that WER alone is insufficient for evaluating ASR systems intended for pronunciation training. Across eight systems spanning three architectures, lower WER consistently predicts higher overcorrection of mispronunciations. We propose Pronunciation Preservation Rate as a complementary metric that captures how well an ASR system preserves learner errors in its output.

Surface-faithful reranking provides a practical mitigation. By selecting N-best hypotheses that are phonetically closer to what was actually pronounced, the method reduces false acceptance of mispronunciations by 6.0 and 5.6pp on two datasets, across diverse age groups and L1 backgrounds, without retraining or access to ASR internals.

The method's ceiling is bounded by beam-search pruning, and direct phoneme comparison alone produces too many false alarms; pronunciation-aware decoding (§6.5) is the principled extension.

Limitations

Several limitations remain. (1) Residual FAR is high (74.8% on L2-ARCTIC severe), and the K=10 oracle ceiling (56.9%) shows the N-best list lacks surface-faithful hypotheses for many utterances—a beam-search limitation. (2) We evaluate on read speech with known prompts; spontaneous L2 speech would require a different alignment formulation. (3) Our evaluation is limited to L2 English, and the G2P component is language-specific. (4) The phoneme recognizer has a 19.2% error rate on L2-ARCTIC (15.0%–24.5% across L1s), which propagates into reranking. (5) The single-canonical-G2P anchor is an approximation: §6.2 shows it disagrees with the recognizer on correctly-pronounced words and predicts false rejection; multi-pronunciation lattices (§6.5) are the principled fix. (6) We have not conducted user studies; the 5.6pp absolute / 19% relative recall gain is a necessary but not sufficient indicator of learning benefit.

Acknowledgments

We thank the BEA 2026 reviewers for their constructive feedback, which led to the additional analyses and clarifications presented in this revision.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Enrique Cámara-Arenas, Cristian Tejedor-García, Carlos J. Tomás-Vázquez, and David Escudero-Mancebo. 2023. Automatic pronunciation assessment vs. automatic speech recognition: A study of conflicting conditions for L2-English. *Language Learning & Technology*, 27(1):1–19.

Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, EngSiong Chng, and Chao-Han Huck Yang. 2024. It's never too late: Fusing acoustic information into large language models for automatic speech recognition. In *The Twelfth International Conference on Learning Representations*.

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844. Spoken Language Technology for Education.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7262–7266.

Daniel Korzekwa, Jaime Lorenzo-Trueba, Szymon Zaporowski, Shira Calamaro, Thomas Drugman, and Bozena Kostek. 2021. Mispronunciation detection in non-native (l2) English with uncertainty modeling. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7738–7742.

Wai-Kim Leung, Xunying Liu, and Helen Meng. 2019. CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136.

Kun Li, Xiaojun Qian, and Helen Meng. 2017. Mispronunciation detection and diagnosis in l2 English speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):193–207.

Somshubra Majumdar and Nithin Rao Koluguri. 2024. Pushing the boundaries of speech recognition with NVIDIA NeMo parakeet ASR models. NVIDIA Technical Blog.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Ambra Neri, Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 15(5):441–467.

Kyubyong Park and Jongseok Kim. 2019. g2pE. `https://github.com/Kyubyong/g2p`.

Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan. 2021. A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis. In *Interspeech 2021*, pages 4448–4452.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

SYSTRAN. 2023. faster-whisper. `https://github.com/SYSTRAN/faster-whisper`.

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.

Yongkook Won. 2025. Assessing the efficacy of word error rate as a proxy for pronunciation quality. *Journal of Second Language Pronunciation*, 11(3):394–422.

Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. 2021. Explore wav2vec 2.0 for Mispronunciation Detection. In *Interspeech 2021*, pages 4428–4432.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment. In *Interspeech 2021*, pages 3710–3714.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-ARCTIC: A Non-native English Speech Corpus. In *Interspeech 2018*, pages 2783–2787.