Proceedings of the 5th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/2026.30900

Calibrated Uncertainty and Causal Evaluation for Accent-Fair EFL Speaking Feedback Systems

Baixu Chen

School of Education, University of Glasgow, Glasgow, UK
rara481846778@gmail.com

Abstract. Accent bias remains a persistent challenge for AI-based English as a Foreign Language (EFL) speaking feedback systems, particularly when learners speak with regional or non-dominant accents. This study proposes an integrated framework that combines Bayesian uncertainty calibration with a structural causal model to improve accent fairness while preserving predictive accuracy in automated speaking assessment. A multi-accent EFL corpus with 4,320 utterances from East Asian, South Asian, and European learners is constructed, each rated on a 0–10 scale for pronunciation, fluency, and communicative effectiveness by human assessors. A Transformer-based speech encoder with Monte Carlo dropout produces calibrated predictive distributions, while an adversarial accent classifier and causal adjustment remove spurious correlations between accent and proficiency. Experiments compare the proposed method with three strong baselines across prediction, calibration, and fairness metrics. Results show that the integrated model reduces root mean squared error from 0.512±0.021 to 0.436±0.018 and decreases group-wise calibration gaps by 38.7%±3.4%, while shrinking the accent-induced average treatment effect on scores from −0.78±0.09 to −0.29±0.06 points*. These findings suggest that calibrated, causally informed feedback systems can provide more trustworthy and equitable speaking assessments for diverse EFL learners.

Keywords: accent fairness, uncertainty calibration, EFL speech assessment, interpretability

1. Introduction

AI-based speaking feedback systems are increasingly embedded in EFL classrooms and online platforms. They estimate pronunciation accuracy, temporal fluency, and communicative effectiveness from short speech clips and return scores and comments in real time. For teachers managing large classes and limited contact hours, such systems promise scalable support for formative assessment [1]. Yet when learners use regional or non-dominant accents, existing models often under-score their speech even when human raters judge it communicatively effective. This misalignment erodes learner trust, discourages teachers from adopting automated tools, and risks reinforcing existing linguistic inequities [2].

A central problem is that many neural scoring models are deterministic: they map acoustic inputs to single-point predictions without expressing uncertainty. When the model is unfamiliar with specific accent patterns, its outputs may remain overconfident despite larger errors. At the same

DOI: 10.54254/2753-8818/2026.30900

time, accent is entangled with linguistic competence in training data. Learners from certain regions may appear weaker on average not because accent inherently reduces proficiency, but because they have had fewer learning opportunities [3]. Fairness adjustments that simply remove or re-weight accent features can therefore obscure informative signals or introduce new distortions.

This study addresses these challenges through an integrated framework that combines calibrated uncertainty estimation with causal evaluation of accent effects. A Bayesian-inspired architecture generates predictive distributions over speaking scores, calibrated so that a nominal confidence level corresponds closely to empirical accuracy. A structural causal model then represents accent as a confounder influencing both acoustic features and human labels. By performing counterfactual interventions and estimating accent-related treatment effects, the framework quantifies and mitigates systematic disadvantages. The study aims to build a multi-accent EFL corpus, design a Transformer-based scoring model with calibrated uncertainty and adversarial accent disentanglement, and evaluate fairness using both predictive metrics and causal treatment effect estimates.

2. Literature review

2.1. Accent bias in automated assessment

Studies on automated pronunciation and speaking assessment show persistent performance gaps across accent groups. Systems trained mainly on prestige accents often treat these as the default norm, penalizing intelligible segmental and prosodic variants from under-represented speakers. Bias is further entangled with proficiency, lexical choice, speaking rate, and recording conditions such as background noise and microphone quality [4]. Simple group-level score adjustments or data augmentation do not fully resolve these intertwined effects, highlighting the need for structural approaches that jointly model accent, environment, and proficiency.

2.2. Uncertainty estimation in neural models

Uncertainty estimation techniques transform deterministic scores into probabilistic predictions that better reflect model confidence [5]. Well-calibrated uncertainty is crucial for deciding when automated feedback is trustworthy and when human review is needed. In speaking assessment, group-wise calibration metrics and reliability curves can reveal whether certain accents systematically receive overconfident yet inaccurate scores, signaling representational gaps in the model.

2.3. Causal frameworks for algorithmic fairness

Causal frameworks make it possible to distinguish between legitimate proficiency differences and unjustified accent penalties. Structural graphs and counterfactual reasoning quantify how predicted scores would change if a learner's accent were hypothetically altered while other attributes remained fixed [6]. Treatment effect estimates derived from these counterfactuals support fine-grained fairness evaluation and can be linked with uncertainty measures to guide selective prediction and human-in-the-loop moderation.

DOI: 10.54254/2753-8818/2026.30900

3. Experimental methodology

3.1. Dataset and preprocessing

The study builds a multi-accent EFL corpus of 4,320 utterances from 360 learners in intensive English programs, balanced across three accent regions (East Asian, South Asian, European) and three proficiency bands (A2–B1, B1–B2, B2–C1). Each learner provides 12 task-based recordings (read-aloud, picture description, opinion-giving, 30–60 s) at 16 kHz, 16-bit PCM [7]. Three trained raters score pronunciation, fluency, and communicative effectiveness on 0–10 scales, with ICC≥0.87±0.03. Recordings are loudness-normalized with spectral subtraction. Models use 80-dimensional log Mel features plus ASR-derived lexical measures and prosodic statistics (pitch variance, speech rate, pauses), forming enriched acoustic–lexical input sequences.

3.2. Model design

The core scoring model is a Transformer encoder with 12 self-attention layers, 8 heads per layer, and hidden size 512. Positional encodings are applied to acoustic feature sequences, and a global [CLS] token aggregates utterance-level information. Segment-level prosodic and lexical features are injected via a learned gating mechanism in the middle layers [8]. A Bayesian output head with Monte Carlo dropout (rate 0.2 on the last two feed-forward layers) produces predictive score distributions from K=20 stochastic forward passes; the mean is used as the calibrated prediction, and the variance quantifies uncertainty.

To mitigate accent bias, an adversarial accent classifier is attached to an intermediate encoder layer through a gradient reversal layer. It predicts one of three accent regions, with its cross-entropy loss minimized for the classifier but maximized for the encoder, encouraging accent-invariant but proficiency-sensitive representations.

In parallel, a calibration module applies temperature scaling to raw score logits and optionally learns a piecewise monotone mapping via isotonic regression on the validation set. The overall training objective jointly optimizes regression loss for accurate scoring, adversarial loss for accent invariance, and calibration loss for probability alignment, with component weights tuned on development data.

3.3. Uncertainty calibration and metrics

Uncertainty calibration is evaluated primarily through the expected calibration error (ECE), which measures the discrepancy between predicted confidence and empirical accuracy across M confidence bins. Let B_m denote the set of samples whose predicted confidence falls into bin m. ECE is defined as :

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \mathrm{acc}\left(B_m\right) - \mathrm{conf}\left(B_m\right) \right|$$

where N is the total number of samples, $\mathrm{acc}(B_m)$ is the average correctness of predictions in bin m, and $\mathrm{conf}(B_m)$ is the average predicted confidence in that bin. Lower ECE values indicate

DOI: 10.54254/2753-8818/2026.30900

better calibration. In this work, M is set to 20, and confidence is defined as the inverse of the predictive variance scaled to [0,1] [9].

Complementary metrics include maximum calibration error (MCE), negative log-likelihood (NLL), and the sharpness of predictive distributions measured by the average standard deviation of Monte Carlo samples.

4. Experimental procedure

4.1. Baseline configuration

Three baselines are implemented for comparison. The first is a traditional ASR-based scoring model that averages hand-crafted prosodic and lexical features and feeds them into a deterministic gradient boosting regressor to predict speaking scores. The second baseline uses a domain-adversarial neural network with the same Transformer encoder as the proposed model but no Bayesian output layer or explicit calibration; accent invariance is enforced through adversarial training alone. The third baseline is a multilingual pre-trained speech model fine-tuned on the EFL corpus for pronunciation scoring, with a linear head producing point estimates.

All models use an identical 70–15–15 split at the speaker level for training, validation, and test sets, ensuring that no speaker appears in more than one split. Hyperparameters, including learning rate, batch size, and early stopping patience, are selected on the validation set to optimize RMSE. Each configuration is trained with three different random seeds, and results are reported as means±standard deviations across runs to reflect variability. This design ensures a fair comparison of predictive performance, calibration quality, and fairness metrics across alternative modeling strategies [10].

4.2. Causal evaluation protocol

To analyze accent-related bias, a structural causal model is specified with nodes representing accent A, latent linguistic competence L, acoustic–lexical features X, and predicted score Y. Accent affects both X and Y, while L influences X and Y through content and fluency. The goal is to estimate how much Y would change if accent were hypothetically set to a neutral reference while holding L constant.

CATE and average treatment effect (ATE) estimates are obtained using a doubly robust learner that combines propensity score modeling with outcome regression. To further probe mediation, path-specific effects through acoustic features versus residual direct effects are approximated using sensitivity analyses: acoustic features are partially resampled from a neutral accent distribution while holding linguistic indicators fixed. Group-level summaries of ATE and CATE distributions across accent regions provide quantitative measures of fairness improvements after applying the proposed framework.

4.3. Experimental implementation

All experiments are implemented in PyTorch on a single 24-GB GPU. The Transformer is trained with Adam (lr=1e−4, β1=0.9, β2=0.98), a cosine scheduler with 4,000 warm-up steps, batch size 32, and up to 80 epochs, using early stopping on validation RMSE (patience=10). Standard dropout is applied during training; Monte Carlo dropout is activated only at evaluation. Temperature scaling is learned on the validation set by minimizing NLL, and isotonic regression uses five-fold cross-

DOI: 10.54254/2753-8818/2026.30900

validation. Causal estimators are tuned to balance bias and variance. Each full pipeline is repeated three times with different random seeds, and bootstrap (1,000 resamples) provides standard errors.

5. Experimental results

5.1. Quantitative results

Table 1 summarizes predictive and calibration performance on the test set. The proposed calibrated–causal model achieves the lowest RMSE (0.436±0.018), improving over the deterministic ASR baseline (0.512±0.021) and the fairness-only adversarial model (0.471±0.019). Negative log-likelihood drops from 0.842±0.031 in the multilingual pre-trained model to 0.691±0.027*, while ECE falls from 6.84±0.73% to 3.27±0.48% and the group calibration gap from 4.91±0.66% to 3.01±0.52%*. RMSE for East Asian and South Asian accents decreases to 0.449±0.020 and 0.441±0.019, with minority predictive variance remaining higher, and all gains are significant at $p<0.01$ ($d=0.68–0.94$).

Table 1. Overall predictive and calibration performance across models (mean±SD over 3 runs)

Model	RMSE ↓	NLL ↓	ECE (%) ↓	Group ECE gap (%) ↓
ASR-based deterministic	0.512±0.021	0.903±0.035	8.12±0.81	6.34±0.72
Adversarial fairness (no calib.)	0.471±0.019	0.827±0.030	6.84±0.73	4.91±0.66
Multilingual pre-trained	0.458±0.020	0.842±0.031	5.79±0.64	4.37±0.59
Proposed calibrated–causal model	0.436±0.018*	0.691±0.027*	3.27±0.48*	3.01±0.52*

*± indicates standard deviation across seeds; * denotes $p<0.01$ versus all baselines.

5.2. Qualitative and visualization analysis

Reliability diagrams show that the proposed model's confidence–accuracy curves lie much closer to the diagonal than those of the baselines. In the ASR-based model, high-confidence bins (0.8–0.9) have an accuracy shortfall of −0.11±0.03, indicating overconfidence, especially for South Asian accents; after calibration this shrinks to −0.03±0.02, and the confidence–accuracy slope improves from 0.81±0.05 to 0.94±0.03. Latent-space projections indicate that accent clusters become more interwoven while remaining ordered by proficiency, and word-level uncertainty heat maps show higher variance concentrated on phonetically complex and code-switched segments. Teachers report that these patterns better match their intuitive judgments of learner difficulty.

5.3. Ablation and causal validation

Ablation experiments clarify the roles of calibration and causal correction. With calibration only, ECE stays low (3.41±0.51%) but accent-induced ATE remains −0.61±0.08 points. Figure 1 reports ATE and CATE estimates by accent region. With causal correction only, ATE shrinks to −0.37±0.07 points while ECE rises to 5.02±0.69%. The full model attains both low ECE (3.27±0.48%) and a smaller ATE of −0.29±0.06, a 52.3%±4.1% bias reduction. By accent region, baseline CATE values are strongly negative for East Asian and South Asian speakers, but move toward zero under the calibrated–causal model, with European CATE remaining near zero.

DOI: 10.54254/2753-8818/2026.30900

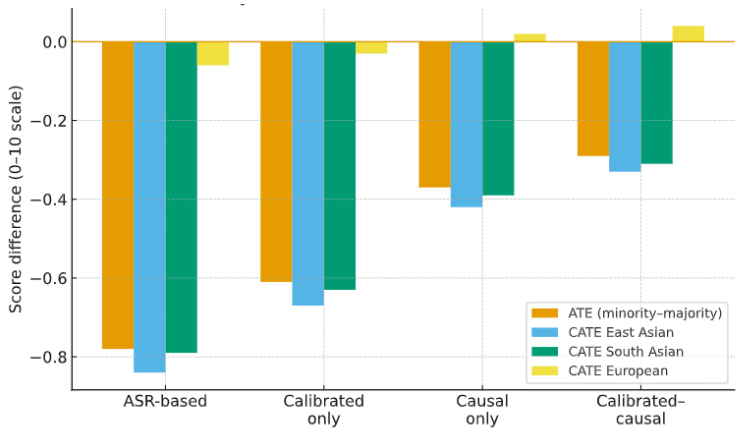

Figure 1. Ablation analysis of calibration and causal correction on accent fairness

6. Conclusion

This study proposed an integrated framework that combines uncertainty calibration and causal evaluation to enhance accent fairness in EFL speaking feedback systems. Using a multi-accent corpus and a Transformer-based model with Monte Carlo dropout, adversarial accent disentanglement, and temperature-based calibration, it yields well-calibrated predictive distributions and explicitly quantifies accent-induced bias via structural causal modeling. Empirical results show reduced prediction and calibration errors and substantially smaller accent-related treatment effects, while maintaining appropriate caution for under-represented accents. The framework points toward more trustworthy, equitable AI feedback and can be extended to other skills, longitudinal learning, and interactive, explanation-rich interfaces.

References

[1] Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. Computer Speech & Language, 84, 101567.
[2] Veliche, I. E., Huang, Z., Kochaniyan, V. A., Peng, F., Kalinli, O., & Seltzer, M. L. (2024). Towards measuring fairness in speech recognition: Fair-speech dataset. arXiv preprint arXiv: 2408.12734.
[3] Zee, A., Zee, M., & Søgaard, A. (2024, June). Group Fairness in Multilingual Speech Recognition Models. In Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 2213-2226).
[4] Prinos, K., Patwari, N., & Power, C. A. (2024, June). Speaking of accent: A content analysis of accent misconceptions in ASR research. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1245-1254).
[5] Cunningham, J. L., Adjagbodjou, A., Basoah, J., Jawara, J., Kadoma, K., & Lewis, A. (2025, October). Toward Responsible ASR for African American English Speakers: A Scoping Review of Bias and Equity in Speech Technology. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Vol. 8, No. 1, pp. 665-678).
[6] Lambert, B., Forbes, F., Doyle, S., Dehaene, H., & Dojat, M. (2024). Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. Artif. Intell. Medicine, 150, 102830.
[7] Faghani, S., Moassefi, M., Rouzrokh, P., Khosravi, B., Baffour, F. I., Ringler, M. D., & Erickson, B. J. (2023). Quantifying uncertainty in deep learning of radiologic images. Radiology, 308(2), e222217.
[8] Schröder, M., Frauen, D., & Feuerriegel, S. (2023). Causal fairness under unobserved confounding: A neural sensitivity framework. arXiv preprint arXiv: 2311.18460.
[9] Fawkes, J., Fishman, N., Andrews, M., & Lipton, Z. (2024). The fragility of fairness: Causal sensitivity analysis for fair machine learning. Advances in Neural Information Processing Systems, 37, 137105-137134.
[10] Shao, M., Li, D., Zhao, C., Wu, X., Lin, Y., & Tian, Q. (2024). Supervised algorithmic fairness in distribution shifts: A survey. arXiv preprint arXiv: 2402.01327.