ISSN: 3030-3346

AN EXPLAINABLE NEURO-SYMBOLIC SPEECH MATRIX MODEL FOR PERSONALIZED FOREIGN LANGUAGE PRONUNCIATION FEEDBACK

Kuvandikov Jura Tursunbayevich[1], Rohmonqulov Muhammadyusuf Egamberdi o'g'li[2]

[1]PhD, Jizzakh branch of the National University of Uzbekistan named after Mirzo Ulugbek,
[1]Assistent, Jizzakh branch of the National University of Uzbekistan named after Mirzo Ulugbek,
E-mail: muhammadyusufrohmonqulov501@gmail.com, jorakuvandikov1@gmail.com

K E Y W O R D S

computer-assisted pronunciation training; automatic pronunciation assessment; speech matrix; neuro-symbolic AI; knowledge graph; explainable feedback; foreign language learning; neural networks

A B S T R A C T

Automatic pronunciation support increasingly combines speech recognition, pronunciation assessment, and computer-assisted language learning, yet many systems still return only a transcription or global score. This article proposes the Neuro-Symbolic Speech Matrix Feedback model (NS-SMF), an explainable architecture that converts learner speech into diagnostic and pedagogically actionable feedback. Learner speech is represented through acoustic, phonetic-prosodic, and pedagogical error matrices. A neural encoder extracts acoustic-phonetic evidence, task-specific heads estimate segmental and suprasegmental deviations, reference alignment localizes errors, and a knowledge base or knowledge graph maps detected deviations to linguistically grounded explanations and corrective exercises. The model formalizes posterior-based scoring, dynamic alignment, integrated deviation, explanation confidence, feedback validity, pedagogical priority, learner-profile update, and multi-task training. A pilot protocol for Uzbek-speaking learners of English is outlined using technical, expert-agreement, explainability, and learning-improvement measures. The main contribution is a formal framework that moves pronunciation assessment from black-box scoring toward traceable, rule-grounded, and exercise-oriented feedback.

1. Introduction

Pronunciation strongly affects intelligibility and confidence in foreign-language communication. Persistent segmental errors, misplaced stress, unstable timing, or inappropriate intonation can remain even when learners have adequate vocabulary and grammar. Human teachers can diagnose such problems, but detailed individual correction requires repeated listening, phonetic expertise, and time, which limits scalability in large classes, distance learning, and independent practice.

Computer-assisted pronunciation training (CAPT) addresses this need through speech recognition, automatic pronunciation assessment (APA), and increasingly neural speech models. Reviews show that CAPT can support repeated practice and immediate response, but also emphasize a continuing need for feedback that is individualized, interpretable, and pedagogically meaningful (Amrate & Tsai, 2024; Mohsen et al., 2025). Classical Goodness of Pronunciation (GOP) scoring and modern neural mispronunciation detection can localize weak phones, but a low score is not itself an explanation. A learner who realizes /θ/ as an /s/-like sound needs to know the articulatory difference, the likely contrast involved, the importance of the deviation, and what to practise next.

NS-SMF therefore separates perception from explanation. Neural models estimate what happened in the speech signal, while a symbolic

(cc) BY

knowledge layer explains why it matters and how the learner should respond. The intended processing chain is:

Speech → Matrix → Neural diagnosis
→ Knowledge reasoning
→ Explainable feedback
→ Exercise.

The research problem is to transform uncertain acoustic evidence into feedback that is accurate, transparent, prioritized, and useful for repeated learning. The model contributes a multilayer speech-matrix representation, neural-symbolic diagnosis, formal deviation and confidence functions, pedagogical prioritization, and a pilot evaluation design for Uzbek-speaking learners of English. The article is methodological and does not claim completed large-scale experimental results.

2. Related work and research gap

CAPT has progressed from repetition drills, waveform displays, and auditory comparison to ASR-supported exercises, automatic scoring, and adaptive feedback. Technology-oriented work emphasizes signal processing and error detection, whereas pedagogy-oriented work emphasizes intelligibility, phonological contrasts, learner level, and feedback design. Effective pronunciation support must connect these two strands rather than optimize recognition alone (Neri et al., 2002; Rogerson-Revell, 2021).

GOP and related likelihood-based methods provide interpretable phone-level evidence. A simplified posterior-based phone score is:

$$GOP(p_i) = \frac{1}{d_i} log \frac{P(p_i \mid X_i)}{\max\limits_{q \in \mathcal{P}, q \neq p_i} P(q \mid X_i)}, \qquad (1)$$

where p_i is the expected phone, X_i is its aligned acoustic segment, d_i is duration, and P is the phone inventory. Neural encoders such as wav2vec 2.0 and HuBERT improve speech representation learning, while large-scale systems such as Whisper demonstrate robust multilingual recognition (Baevski et al., 2020; Hsu et al., 2021; Radford et al., 2023). However, latent embeddings and phone labels do not automatically become learner-facing explanations.

Knowledge bases and knowledge graphs represent explicit entities, relations, and rules and are well suited to grounding explanations (Hogan et al., 2021; Ji et al., 2022). Neuro-symbolic AI is especially appropriate here because speech evidence is acoustic and probabilistic, whereas corrective feedback is linguistic, articulatory, and pedagogical. Existing ASR feedback is scalable but often lacks articulatory explanation; GOP gives phone-level scores without exercise mapping; deep neural detectors can be accurate but opaque; visualization may be difficult for non-specialists; and ungrounded language-model explanations risk unsupported feedback. NS-SMF addresses these limitations by combining neural evidence, explicit rules, confidence gating, and corrective actions.

Accordingly, the learning objective extends beyond diagnostic accuracy to explanation validity and pedagogical utility:

$$\max_{\theta,\phi} \mathcal{J} = \lambda_1 Acc(D_\theta, D^*) + \lambda_2 ExpVal\big(F_\phi, KB\big) + \lambda_3 PedUtil\big(F_\phi, L\big), \qquad (2)$$

Here ExpVal measures whether feedback is grounded in the knowledge base, PedUtil measures usefulness for learner L, and the λ weights are non-negative.

3. Materials and methods: the proposed NS-SMF model

3.1 Speech representation and neural analysis

The input is a learner recording produced for a known word, sentence, or short passage. After framing and spectral analysis, the waveform x[n] is transformed into a log Mel representation. The main preprocessing relations are:

$$x[n], \quad n = 0,1, \dots, N-1. \qquad (3)$$

$$x_t[n] = x[n + th]w[n], \quad 0 \leq n < L. \qquad (4)$$

(cc) BY

$$X(t,k) = \sum_{n=0}^{L-1} x\,[n + th]w[n]e^{-j2\pi kn/L}. \qquad (5)$$

$$M_{t,f}^{a} = log\left(\varepsilon + \sum_{k=0}^{K-1} |X(t,k)|^2\, B_{f,k}\right), \qquad (6)$$

Additional pitch, energy, and dynamic features may be concatenated. The complete model uses a multilayer speech representation:

$$M = \{M^a, M^p, M^e\}. \qquad (7)$$

M^a contains acoustic evidence; M^p contains phone, confidence, duration, stress, pitch-contour, and rhythm evidence; and M^e stores pedagogical error objects such as type, severity, confidence, intelligibility impact, recurrence, and priority. The two diagnostic matrices are formalized as:

$$M_i^p = [p_i, s_i, d_i, z_i, \pi_i, \rho_i], \qquad (8)$$

$$M_i^e = [type_i, sev_i, conf_i, imp_i, recur_i, Pr_i]. \qquad (9)$$

A neural encoder, implemented as a CNN-BiLSTM, Transformer, or pretrained self-supervised speech model, maps the acoustic matrix to contextual representations:

$$H = Enc_\theta(M^a), \quad H \in \mathbb{R}^{T\prime \times d_h}. \qquad (10)$$

For a Transformer, self-attention is expressed as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \qquad (11)$$

A phone-posterior head then estimates target-phone probabilities:

$$P_\theta(p \mid h_t) = softmax\big(W_p h_t + b_p\big). \qquad (12)$$

Additional heads predict stress, normalized duration, and prosodic contours. Multi-task analysis is necessary because pronunciation quality includes both segmental and suprasegmental behavior.

3.2 Reference alignment and integrated deviation

Because prompted practice has a known target, the expected phone sequence is aligned with learner evidence using dynamic programming. The alignment cost is:

$$D(i,j) = min\begin{cases} D(i-1,j) + c_{del}\big(p_i^{ref}\big), \\ D(i,j-1) + c_{ins}\big(\hat{p}_j\big), \\ D(i-1,j-1) + c_{sub}\big(p_i^{ref}, \hat{p}_j\big). \end{cases} \qquad (13)$$

The alignment localizes learner segments for target units. For each aligned unit, segmental deviation and competing-phone evidence are estimated as:

$$E_i^{seg} = 1 - P_\theta\big(p_i^{ref} \mid X_i\big). \qquad (14)$$

$$C_i^{conf} = \max_{q \neq p_i^{ref}} P_\theta(q \mid X_i) - P_\theta\big(p_i^{ref} \mid X_i\big). \qquad (15)$$

Stress, duration, and pitch deviations are also measured. The integrated pronunciation deviation is:

$$E_i = \beta_1 E_i^{seg} + \beta_2 E_i^{stress} + \beta_3 E_i^{dur} + \beta_4 E_i^{pitch} + \beta_5 max\big(0, C_i^{conf}\big), \qquad (16)$$

where β1…β5 are non-negative weights that sum to one. This function is central to NS-SMF because it combines target-phone mismatch, stress, timing, pitch, and confusion evidence in a single unit-level quantity while retaining interpretable components.

3.3 Knowledge-base reasoning and feedback validity

The symbolic layer is defined as a knowledge base or graph containing entities, typed relations, and feedback templates:

$$KB = (\mathcal{V}, \mathcal{R}, \mathcal{T}), \qquad (17)$$

A graph relation follows the triple form:

$$(v_s, r, v_o) \in \mathcal{R}, \qquad (18)$$

Typical entities include phonemes, articulatory features, acoustic correlates, error

types, L1-transfer patterns, and corrective exercises. A rule connects neural diagnosis, learner profile, and context to a pedagogical action:

$$Rule_k:\ condition_k(E_i, L_i, context_i) \Rightarrow action_k(F_i). \quad (19)$$

For example, evidence supporting /θ/→/s/ can activate a rule explaining dental-fricative articulation and recommending a minimal-pair exercise. L1 information is used only as contextual support; a rule is not activated without corresponding acoustic evidence.

To prevent confident but unsupported correction, NS-SMF combines neural confidence, knowledge-base support, and alignment reliability:

$$Conf_i = \gamma_1 c_i^{NN} + \gamma_2 c_i^{KB} + \gamma_3 c_i^{align}, \quad \sum_{m=1}^{3} \gamma_m = 1. \quad (20)$$

A feedback item is accepted only when confidence, deviation magnitude, and knowledge support satisfy the validity gate:

$$Valid(F_i) = \mathbb{I}[Conf_i \geq \tau_c \wedge E_i \geq \tau_e \wedge c_i^{KB} \geq \tau_{KB}], \quad (21)$$

If the condition fails, the system requests a clearer repetition or routes the case to teacher review instead of generating a strong correction. This conservative design is important because misleading pronunciation feedback can be more harmful than withholding uncertain feedback.

3.4 Feedback priority, learner adaptation, and training

Learners should not receive every detectable imperfection. NS-SMF ranks valid errors using deviation magnitude, intelligibility impact, recurrence, learner-level relevance, and lesson-goal relevance:

$$Pr_i = \alpha_1 E_i + \alpha_2 Imp_i + \alpha_3 Rec_i + \alpha_4 Level_i + \alpha_5 Goal_i, \quad \sum_{m=1}^{5} \alpha_m = 1. \quad (22)$$

Only the top K valid items are returned:

$$\mathcal{F}_{out} = TopK\{F_i : Valid(F_i) = 1 \ sorted\ by\ Pr_i\}. \quad (23)$$

Pedagogical utility is modeled through clarity, actionability, specificity, level fit, and exercise linkage:

$$PedUtil(F_i) = \eta_1 Clarity_i + \eta_2 Actionability_i + \eta_3 Specificity_i + \eta_4 LevelFit_i + \eta_5 ExerciseLink_i. \quad (24)$$

The learner profile stores recurring error patterns. Recurrence is updated by an exponential moving average:

$$r_{i,t} = \lambda r_{i,t-1} + (1-\lambda)\mathbb{I}[E_{i,t} \geq \tau_e]. \quad (25)$$

Improvement after feedback is quantified as the reduction in deviation:

$$\Delta_i = E_{i,t}^{before} - E_{i,t}^{after}. \quad (26)$$

For supervised or semi-supervised learning, the neural component uses a multi-task objective combining phone, stress, duration, error-type, confidence-calibration, and fairness losses:

$$\mathcal{L} = \lambda_p \mathcal{L}_{phone} + \lambda_s \mathcal{L}_{stress} + \lambda_d \mathcal{L}_{dur} + \lambda_e \mathcal{L}_{error} + \lambda_c \mathcal{L}_{conf} + \lambda_f \mathcal{L}_{fair}. \quad (27)$$

The resulting inference cycle is compact: record speech; extract M^a; encode H; align with the target; estimate segmental and prosodic deviations; infer the error type; query the knowledge base; compute explanation confidence and validity; rank valid errors; generate a correction and exercise; and update the learner profile after the retry.

4. Pilot study design and evaluation protocol

The first empirical study should test whether NS-SMF detects selected segmental and prosodic errors in agreement with experts, whether grounded explanations improve understanding beyond score-only feedback, whether prioritized

(cc) BY

feedback reduces repeated errors, and whether confidence gating reduces unsupported corrections. A feasible design includes 40–80 Uzbek-speaking learners of English at A2–B2 level. Participants would record controlled words, minimal pairs, short sentences, and read-aloud passages. Public corpora such as L2-ARCTIC may support pretraining or benchmarking, but a local Uzbek-speaking learner corpus is required to validate L1-sensitive rules (Zhao et al., 2018).

The initial target inventory should remain narrow and theoretically meaningful: English /θ/ and /ð/, /iː/–/ɪ/, /æ/, final consonant clusters, word stress, and sentence prominence. Three feedback conditions can be compared: ordinary or delayed teacher correction as a baseline, ASR/GOP-based automatic scoring without explanation, and NS-SMF neural-symbolic feedback with a prioritized exercise. Pre-test, practice, immediate post-test, and delayed post-test allow evaluation of both immediate correction and retention.

Technical evaluation uses precision, recall, F1, Cohen's kappa against expert labels, and Spearman correlation with expert pronunciation ratings. Explainability and pedagogical evaluation measure rule-grounded faithfulness, learner/teacher clarity ratings, corrective success after retry, and repeated-error reduction across sessions. A combined score can integrate these dimensions:

$$Eval = \omega_1 F1 + \omega_2 \kappa + \omega_3 Faithfulness + \omega_4 Clarity + \omega_5 CSR + \omega_6 RER, \quad \sum_m \omega_m = 1. \qquad (28)$$

The evaluation must therefore distinguish a model that detects errors accurately from one that also explains them correctly and supports learning. Calibration is particularly important because a high-confidence but incorrect explanation can undermine trust and learning.

5. Analytical behavior of the model

The methodological model implies several testable behaviors. First, if a learner pronounces think with an /s/-like onset, NS-SMF should not return only a numeric score; it should identify the substitution, relate it to the target dental fricative, provide a concise articulatory instruction, and recommend a minimal-pair retry. Second, when alignment or neural confidence is low because of noise, the validity gate should suppress strong correction and request a clearer repetition. Third, when several deviations occur, the priority function should select only the most intelligibility-relevant and recurrent items. Fourth, L1-aware explanations should be generated only when both an L1 rule and actual speech evidence support the pattern. These behaviors make the system auditable: each feedback item can be traced from learner speech to neural evidence, alignment, symbolic rule, and corrective action.

6. Discussion

The scientific contribution of NS-SMF is the formal integration of three levels that are often separated: mathematical speech representation, neural pronunciation diagnosis, and symbolic pedagogical explanation. The multilayer matrix M={M^a,M^p,M^e} structures acoustic, linguistic, and educational information; the integrated deviation function combines segmental and prosodic evidence; the knowledge-base rule layer makes explanations traceable; the validity gate prevents unsupported feedback; and the priority mechanism reduces feedback overload. The learner-profile update turns one-time assessment into adaptive practice support.

The mathematics is functional rather than decorative. It specifies the objects that are measured, the transformations applied to them, and the points at which decisions are made. This allows future studies to replace the neural encoder, dataset, or feedback templates while retaining the core NS-SMF structure. Pedagogically, each learner-facing feedback item should state what is wrong, why it matters, how to correct it, and what to retry. The goal is not native-like imitation but intelligibility, communicative clarity, and learner-relevant improvement.

The framework has several limitations. Reliable evaluation requires carefully annotated learner speech, and a local Uzbek-speaking corpus is necessary for L1-sensitive analysis. Knowledge-

(cc) BY

base construction requires phonetic and pedagogical expertise. Microphone quality, background noise, and alignment errors can reduce confidence. The initial implementation should therefore focus on a restricted phone and prosody inventory before expanding to spontaneous speech. Pronunciation norms must also avoid penalizing harmless accent variation. Fairness can be incorporated through a group-error penalty:

$$\mathcal{L}_{fair} = \sum_{g \in G} \left| ErrorRate_g - ErrorRate_{all} \right|. \quad (29)$$

Future implementation should verify that error rates and feedback severity are not systematically harsher for particular learner groups and should preserve teacher oversight for uncertain or high-impact cases.

7. Conclusion

This article presented NS-SMF, an explainable neuro-symbolic speech matrix model for personalized foreign-language pronunciation feedback. The framework transforms learner speech into layered acoustic, phonetic-prosodic, and pedagogical representations; uses neural speech analysis and reference alignment to detect deviations; and grounds explanations and corrective exercises in an explicit knowledge base. Its main methodological contribution is to treat diagnosis accuracy, explanation validity, and pedagogical utility as connected objectives rather than independent components. The model also includes confidence gating, error prioritization, learner-profile adaptation, and a multi-dimensional evaluation protocol. The next research stage is to construct and annotate a pilot corpus of Uzbek-speaking learners of English, implement the neural-symbolic pipeline, and compare NS-SMF with score-only and neural-only feedback systems using both technical and learning-oriented measures.

References

1. Amrate, M., & Tsai, P.-H. (2024). Computer-assisted pronunciation training: A systematic review. ReCALL, 1–21. https://doi.org/10.1017/S0958344024000181
2. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33.
3. El Kheir, Y., Ali, A., & Chowdhury, S. A. (2023). Automatic pronunciation assessment: A review. Findings of ACL: EMNLP 2023, 8304–8324. https://doi.org/10.18653/v1/2023.findings-emnlp.557
4. Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. Computer Assisted Language Learning, 27(1), 70–105.
5. Hogan, A., et al. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), Article 71. https://doi.org/10.1145/3447772
6. N.Raximov, J.Kuvandikov, D.Khasanov, “The importance of loss function in artificial intelligence”, International Conference on Information Science and Communications Technologies (ICISCT 20222), DOI: 10.1109/ICISCT55600.2022.10146883
7. Rahimov Nodir, Khasanov Dilmurod. (2022). The Mathematical Essence Of Logistic Regression For Machine Learning. https://doi.org/10.5281/zenodo.7239169
8. Khasanov Dilmurod, Tojiyev Ma’ruf, Primqulov Oybek., “Gradient Descent In Machine”. International Conference on Information Science and Communications Technologies (ICISCT), https://ieeexplore.ieee.org/document/9670169
9. Rohmonqulov , M. (2025). DEVELOPMENT OF AN INTERACTIVE SOFTWARE TOOL FOR AUTOMATIC ASSESSMENT AND IMPROVEMENT OF PRONUNCIATION IN FOREIGN LANGUAGE LEARNING. *Scientific Practical Conference*, *1*(1), 344-349.

(cc) BY *dtai.uz*