SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

An Explainable Neuro-Symbolic Speech Matrix Model for Personalized Foreign Language Pronunciation Feedback

Khamdamov Utkir Rakhmatullayevich[1], Turakulov Olim Xolbutayevich[2], Abdumalikov Akmaljon Abduxoliq o‘g‘li[2], Kayumov Oybek Achilovich[2], Akmuradov Baxtiyor Uralovich[3], Rohmonqulov Muhammadyusuf Egamberdi o‘g‘li[4]

[1] Tashkent University of Information Technologies named after Muhammad al-Khorezmi, 108, Amir Temur Avenue, 100200, Tashkent, Uzbekistan. Email: utkir.hamdamov@mail.ru
[2] Jizzakh Branch of the National University of Uzbekistan named after Mirzo Ulugbek, Jizzakh, Uzbekistan.
Emails: olimxolbotaevich@gmail.com, Akmalabdumalikov6@gmail.com, oybekuzonlined3@gmail.com
[3] Vatan University, 2nd Lane, Shota Rustaveli Street, House 2, Dumbuloq MFY, Yakkasaroy District, Tashkent, Uzbekistan.
Email: b.u.akmuradov@gmail.com
[4] Department of Computer Science and Programming, Jizzakh Branch of the National University of Uzbekistan, 130100, Uzbekistan.
Email: muhammadyusufrohmonqulov501@gmail.com

Abstract
Automatic pronunciation support has become an important direction in natural language processing (NLP), speech technology, and computer-assisted language learning. However, many systems still reduce learner speech to either a transcription result or a global pronunciation score, while learners and teachers need a more diagnostic answer: which sound or prosodic pattern is problematic, why it is problematic, how it is related to articulation and first-language transfer, and what corrective exercise should be attempted next. This article proposes the **Neuro-Symbolic Speech Matrix Feedback model (NS-SMF)**, an explainable architecture for personalized foreign language pronunciation feedback. In the proposed model, learner speech is represented as a multilayer mathematical object consisting of an acoustic matrix, a phonetic-prosodic matrix, and a pedagogical error matrix. A neural speech encoder extracts latent acoustic-phonetic evidence; phone-level and prosody-level diagnosis heads estimate segmental and suprasegmental deviations; a reference-alignment module compares the learner utterance with a target pronunciation; and a knowledge base or knowledge graph transforms neural predictions into linguistically grounded, pedagogically actionable feedback. The article formalizes the model using matrix representations, posterior-based pronunciation scoring, dynamic alignment, deviation functions, knowledge-base rule satisfaction, explanation confidence, feedback utility, and learner-profile update equations. A pilot evaluation protocol is also proposed for Uzbek-speaking learners of English, with technical metrics, expert-agreement measures, and pedagogical improvement indicators. The main contribution is a formal framework that transforms pronunciation assessment from black-box scoring into interpretable, rule-grounded, and exercise-oriented feedback suitable for research in foreign language learning, speech assessment, and explainable educational AI.

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

1. Introduction

Pronunciation is one of the most visible components of oral communication in a foreign language. A learner may know vocabulary and grammar, but persistent segmental errors, misplaced stress, unnatural rhythm, or unstable intonation can reduce intelligibility and confidence. Human teachers can often identify such problems, but detailed individual correction requires time, phonetic expertise, repeated listening, and consistent follow-up. This limitation becomes more serious in large classes, distance learning, independent learning, and resource-constrained education environments where every learner cannot receive detailed corrective pronunciation feedback after every attempt.

Computer-assisted pronunciation training (CAPT) has developed as a response to this need. Recent CAPT research shows a shift from static practice materials and visual displays toward speech-recognition-supported systems that can provide more immediate learner responses. At the same time, systematic and scientometric reviews emphasize that the field still needs feedback that is precise, individualized, pedagogically meaningful, and not limited to surface-level score reporting (Amrate & Tsai, 2024; Mohsen et al., 2025). Automatic speech recognition (ASR) can support pronunciation learning, but ASR alone is not equivalent to pronunciation pedagogy. A transcription error can indicate a pronunciation problem, yet it does not automatically tell the learner which articulatory gesture failed, which phonemic contrast is affected, or what type of corrective practice is appropriate. Automatic pronunciation assessment (APA) has also advanced significantly. Classical methods such as likelihood-based phone scoring and Goodness of Pronunciation (GOP) introduced a way to estimate phone-level pronunciation quality (Witt & Young, 2000). More recent approaches use deep neural networks, self-supervised

speech representations, and end-to-end models for phonemic and prosodic assessment (El Kheir et al., 2023). These models can improve detection accuracy, especially when trained on appropriate data. Nevertheless, a major gap remains between **detecting** a mispronunciation and **explaining** it in a way that is useful for learning. A neural model may correctly identify that a learner has produced an /s/-like realization instead of /θ/, but the learner still needs an explanation involving dental fricative articulation, first-language influence, intelligibility impact, and a targeted exercise.

This article addresses that gap by proposing the **Neuro-Symbolic Speech Matrix Feedback model (NS-SMF)**. The proposed framework is based on the following idea: learner speech should first be transformed into structured acoustic and phonetic-prosodic matrices; neural networks should be used to estimate latent pronunciation evidence and deviation scores; and a knowledge base should translate these numeric and categorical results into explicit, linguistically grounded, and pedagogically actionable feedback. In other words, the system should not stop at:

$$\text{Speech} \rightarrow \text{ASR} \rightarrow \text{Score}. \qquad (1)$$

Instead, it should implement the richer chain:

$$\text{Speech} \rightarrow \text{Matrix} \rightarrow \text{Neural diagnosis} \rightarrow \text{Knowledge reasoning} \rightarrow \text{Explainable feedback} \rightarrow \text{Exercise}. \qquad (2)$$

The central research problem is therefore not only how to classify pronunciation errors, but how to convert neural speech evidence into feedback that is transparent, valid, prioritized, and useful for repeated learning. The article is conceptual and methodological: it presents a formal model, mathematical definitions, a neural-symbolic architecture, and an evaluation protocol rather than claiming completed large-scale experimental results. This is important for research integrity. The model is designed to become the first stage of a broader empirical study in NLP-based foreign language learning.

The article makes five contributions. First, it introduces a multilayer **speech matrix representation** for pronunciation feedback, separating acoustic evidence, phonetic-prosodic structure, and pedagogical error objects. Second, it proposes a neural-symbolic architecture that combines speech encoders with a knowledge base or knowledge graph. Third, it formalizes error diagnosis using posterior probabilities, dynamic alignment, deviation functions, confidence calibration, and prosodic distance measures. Fourth, it introduces a feedback-priority and explanation-confidence mechanism so that the system produces only useful and justified feedback. Fifth, it presents a pilot evaluation design for Uzbek-speaking learners of English, with technical and pedagogical metrics.

2. Related work and research gap

2.1 Computer-assisted pronunciation training

CAPT research has long attempted to help learners practise pronunciation outside direct teacher supervision. Early systems often relied on repetition drills, auditory comparison, waveform visualization, spectrograms, or simple perception tasks. More recent systems include mobile tools, ASR-supported exercises, automatic scoring, and feedback generation. Reviews of CAPT show that technology can support learner autonomy, repeated practice, and immediate response, but they also highlight uneven pedagogical quality across systems. Some tools provide general encouragement or a score, while others provide detailed error localization and corrective suggestions.

The field has two major strands. The first is **technology-oriented CAPT**, where the focus is signal processing, speech recognition, and automatic error detection. The second is **pedagogy-oriented CAPT**, where the focus is intelligibility, learner level, feedback type, phonological contrast, and teacher-guided learning. A high-quality pronunciation feedback system must connect both strands. It must be technically accurate enough to detect errors, but also pedagogically grounded enough to explain the errors in a way that learners can act upon.

2.2 Automatic pronunciation assessment and GOP

Automatic pronunciation assessment traditionally treats learner speech as a realization of a target utterance. A target phone sequence is obtained from a prompt or canonical transcription. The learner's acoustic signal is aligned with the expected phone sequence, and a score is produced for each phone, word, or utterance. The classical GOP idea estimates whether the acoustic evidence for an expected phone is strong compared with competing phones.

A simplified phone-level score can be written as:

$$\mathrm{GOP}(p_i) = \frac{1}{d_i} \log \frac{P(p_i \mid X_i)}{\max_{q \in \mathcal{P}, q \neq p_i} P(q \mid X_i)}, \qquad (3)$$

where (p_i) is the expected phone, (X_i) is the aligned acoustic segment, (d_i) is its duration, and () is the phone inventory. A higher score means that the acoustic evidence supports the expected phone more strongly than competing phones. Modern neural models often replace HMM likelihoods with posterior probabilities or contextual embeddings, but the need for alignment and interpretable segment-level scoring remains.

The weakness of a pure GOP-style approach is that a score is not yet a feedback explanation. If the score for /θ/ is low, the system still needs to determine whether the error resembles /s/, /t/, /f/, deletion, or another realization; whether the error is systematic for this learner; whether it affects intelligibility; and which exercise should be

recommended. The proposed NS-SMF model uses GOP-like scoring as only one component inside a wider neural-symbolic feedback process.

2.3 Deep speech representations

Deep learning has changed speech processing by replacing many hand-crafted pipelines with learned representations. Self-supervised speech models such as wav2vec 2.0 learn useful representations from raw audio and can be fine-tuned for speech recognition or related tasks (Baevski et al., 2020). HuBERT uses masked prediction of hidden units to learn speech representations when explicit phone labels are not available during pretraining (Hsu et al., 2021). Whisper demonstrates the value of large-scale weak supervision for robust multilingual speech recognition (Radford et al., 2023). These models are powerful foundations for pronunciation-related systems.

However, general-purpose speech encoders are not automatically pronunciation teachers. Their latent vectors may encode phonetic, speaker, lexical, and acoustic information, but the vectors are not directly understandable to learners. A model can be accurate and still fail pedagogically if it cannot explain the cause of an error. For this reason, the present study treats neural encoders as evidence extractors and diagnosis modules, while reserving final feedback explanation for a symbolic knowledge layer.

2.4 Knowledge bases, knowledge graphs, and neuro-symbolic educational AI

A knowledge base stores explicit facts and rules. A knowledge graph represents entities and relations in a structured graph form, for example: phone /θ/ requires dental fricative articulation; Uzbek/Russian L1 profile may increase the probability of substitution with /s/; substitution /θ/ -> /s/ can trigger minimal-pair practice. Knowledge graphs are widely studied for information integration, recommendation, reasoning, and explainability (Hogan et al., 2021; Ji et al., 2022). In education, knowledge graphs can support personalized learning paths, concept mapping, adaptive recommendations, and transparent explanations.

Neuro-symbolic AI combines neural perception with symbolic reasoning. The neural component is useful for noisy perceptual data such as audio, while the symbolic component is useful for explicit rules, constraints, and explanations. This combination is especially relevant for pronunciation feedback because speech is acoustic and probabilistic, while feedback is conceptual and pedagogical. The proposed model therefore uses neural networks to estimate what happened in the speech signal and a knowledge base to explain why it matters and what to do next.

2.5 Research gap

Table 1 summarizes the gap that motivates the proposed model.

Table 1. Gap analysis of existing pronunciation-feedback approaches

Approach	Main strength	Typical limitation	NS-SMF response
ASR transcription feedback	Fast and scalable recognition of spoken prompts	Transcription mismatch does not explain articulation or phonemic cause	Uses ASR evidence only as one feature and converts it into phone-level diagnosis
GOP and likelihood-based phone scoring	Interpretable phone-level quality score	Score alone does not provide pedagogical correction	Adds error type, confidence, intelligibility impact, and exercise mapping
Deep neural mispronunciation detection	Strong data-driven detection capability	Often black-box; feedback may not be linguistically grounded	Connects neural outputs to knowledge-base rules and explanation templates
Visual feedback systems	Learner can see waveform, pitch, or spectrogram	Visualizations may be difficult for non-specialists to interpret	Converts acoustic/prosodic deviation into simple learner-level feedback
LLM-only explanation	Fluent natural-language explanation	Risk of unsupported or hallucinated feedback if not grounded in speech evidence	Uses rule-supported feedback validity gate and knowledge graph grounding
Teacher-only correction	Rich and context-sensitive human judgement	Time-consuming and hard to scale in large or remote classes	Provides repeatable, personalized practice support while preserving teacher oversight

The gap can be expressed formally. Let (D) be a diagnosis produced by a neural model and (F) be the feedback delivered to a learner. Existing systems often maximize diagnostic accuracy:

$$\max_{\theta} \; Acc(D_{\theta}, D^{*}). \qquad (4)$$

For learning, however, the target should also include explanation validity and feedback utility:

$$\max_{\theta,\phi} \mathcal{J} = \lambda_1 Acc(D_\theta, D^*) + \lambda_2 ExpVal\left(F_\phi, KB\right) + \lambda_3 PedUtil\left(F_\phi, L\right), \qquad (5)$$

where (ExpVal) measures whether the explanation is grounded in the knowledge base, (PedUtil) measures pedagogical usefulness for learner (L), and (_1,_2,_3) are non-negative weights. This shift from diagnosis accuracy alone to diagnosis plus explanation plus usefulness is the core of the proposed work.

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

3.1 Overall architecture

The proposed NS-SMF architecture is shown in Figure 1. The input is learner speech recorded from a prompted word, sentence, or short dialogue. The output is a structured feedback object containing error type, evidence, explanation, correction instruction, recommended exercise, and retry prompt.

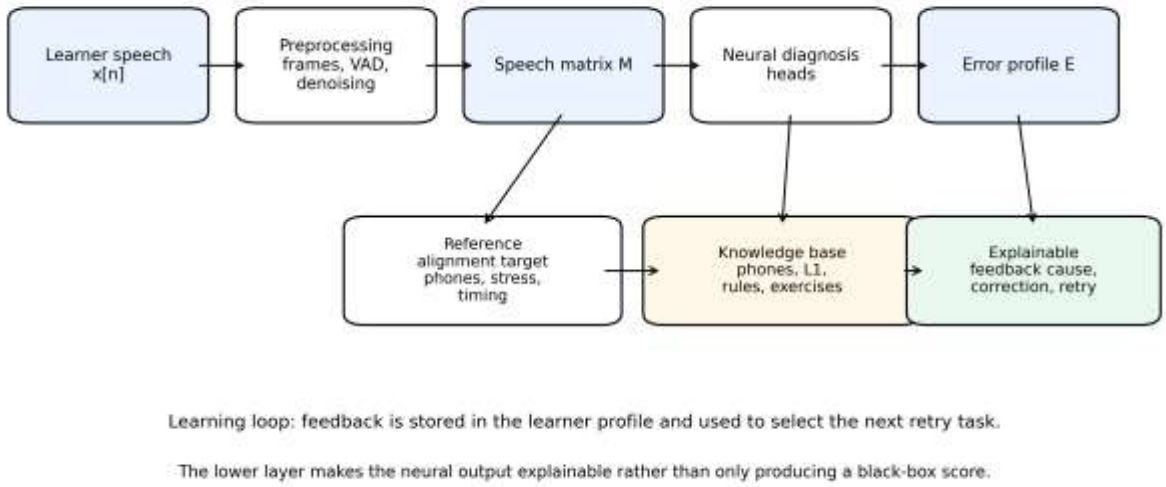


Figure 1. General architecture of the proposed NS-SMF model.

The system is designed as a closed diagnostic-learning loop rather than a one-time speech-recognition engine. Its main modules are: preprocessing, speech matrix construction, neural speech analysis, reference alignment, knowledge-base reasoning, feedback generation, and learner-profile update. The architecture assumes that reliable feedback must satisfy three conditions: it must be supported by acoustic evidence, it must be interpretable through linguistic knowledge, and it must be pedagogically useful for the learner's current level.

3.2 Speech signal and preprocessing

Let the learner's speech waveform be represented as a discrete-time signal:

$$x[n], \quad n = 0,1, \dots, N-1. \qquad (6)$$

The signal is divided into overlapping frames using a window function (w[n]). For frame index (t), frame length (L), and hop size (h), the framed signal is:

$$x_t[n] = x[n + th]w[n], \quad 0 \le n < L. \qquad (7)$$

The short-time Fourier transform is then computed as:

$$X(t,k) = \sum_{n=0}^{L-1} x\,[n + th]w[n]e^{-j2\pi kn/L}. \qquad (8)$$

The power spectrum is converted into a Mel-filterbank representation:

$$M_{t,f}^{a} = log\left(\varepsilon + \sum_{k=0}^{K-1} |X(t,k)|^2\, B_{f,k}\right), \qquad (9)$$

where (M^a) is the acoustic matrix, (B_{f,k}) is the Mel-filter coefficient for Mel band (f) and frequency bin (k), and () is a small constant preventing undefined logarithms. Additional acoustic features can be concatenated, including pitch (F_0(t)), short-time energy (E(t)), duration features, and spectral tilt.

A compact acoustic feature vector for frame (t) can be defined as:

$$a_t = \left[M^a_{t,1}, \dots, M^a_{t,F}, F_0(t), E(t), \Delta M^a_t, \Delta^2 M^a_t\right]. \quad (10)$$

The full acoustic matrix is:

$$A = [a_1; a_2; \dots; a_T] \in \mathbb{R}^{T \times d_a}. \quad (11)$$

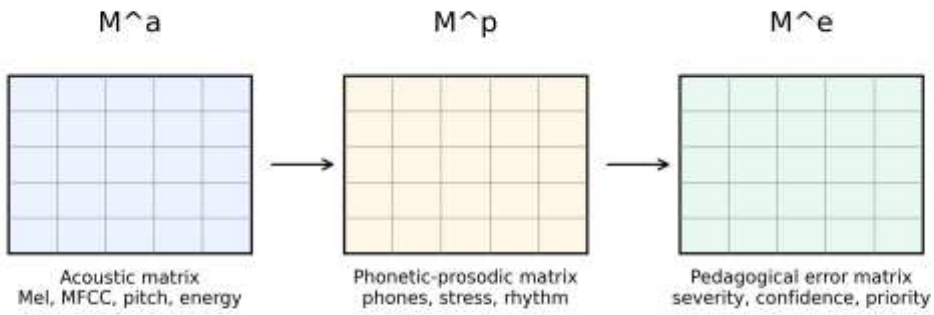
This formulation is important because it turns a variable acoustic waveform into a structured mathematical object suitable for neural analysis.

3.3 Multilayer speech matrix representation

The proposed model does not use a single feature table only. It defines a multilayer speech representation:

$$M = \{M^a, M^p, M^e\}. \quad (12)$$

Here (M^a) is the acoustic matrix, (M^p) is the phonetic-prosodic matrix, and (M^e) is the pedagogical error matrix. Figure 2 illustrates this transformation.



The model separates raw acoustic evidence, interpretable phone-prosody evidence, and learner-facing error objects.

Figure 2. Multilayer speech matrix representation in NS-SMF.

The phonetic-prosodic matrix is defined over linguistic units such as phones, syllables, words, and intonation groups:

$$M^p_i = [p_i, s_i, d_i, z_i, \pi_i, \rho_i], \quad (13)$$

where (p_i) is the expected phone or prosodic unit, (s_i) is the predicted phone confidence, (d_i) is duration, (z_i) is stress evidence, (_i) is pitch-contour evidence, and (_i) is rhythm or timing evidence.

The pedagogical error matrix stores learner-oriented diagnostic objects:

$$M^e_i = [type_i, sev_i, conf_i, imp_i, recur_i, Pr_i]. \quad (14)$$

Here (type_i) is the error type, (sev_i) is severity, (conf_i) is diagnostic confidence, (imp_i) is intelligibility impact, (recur_i) is recurrence in the learner profile, and (Pr_i) is feedback priority. This third matrix is the bridge between machine diagnosis and educational action.

Table 2 describes the mathematical objects used in the model.

Table 2. Core mathematical objects of the proposed NS-SMF model

Symbol	Object	Dimension or type	Role in the model
(x[n])	Speech waveform	(N) samples	Raw learner speech input
(M^a)	Acoustic matrix	(T F) or (T d_a)	Time-frequency and acoustic evidence
(H)	Neural latent representation	(T' d_h)	Contextual speech embedding
(M^p)	Phonetic-prosodic matrix	(I d_p)	Phones, stress, duration, rhythm, pitch
(M^e)	Pedagogical error matrix	(I d_e)	Severity, confidence, intelligibility, priority
(KB)	Knowledge base/graph	entities, relations, rules	Explains and grounds feedback
(F_i)	Feedback object	structured text/action	Learner-facing correction and exercise

3.4 Neural speech analysis layer

The neural layer maps the acoustic matrix into a latent representation:

$$H = Enc_\theta(M^a), \quad H \in \mathbb{R}^{T' \times d_h}. \qquad (15)$$

The encoder can be a CNN-BiLSTM, a Transformer, or a pretrained self-supervised model such as wav2vec 2.0 or HuBERT. If a Transformer encoder is used, self-attention can be written as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \qquad (16)$$

For pronunciation diagnosis, the encoder output is passed to several task-specific heads. A phone-posterior head estimates:

$$P_\theta(p \mid h_t) = softmax\left(W_p h_t + b_p\right). \qquad (17)$$

A stress head estimates word or syllable stress:

$$P_\theta(z_i = 1 \mid H_i) = \sigma\left(w_z^T \bar{h}_i + b_z\right). \qquad (18)$$

A duration head estimates expected normalized duration:

$$\hat{d}_i = ReLU\left(w_d^T \bar{h}_i + b_d\right). \qquad (19)$$

A prosodic-contour head can estimate pitch or intonation representation:

$$\hat{\pi}_i = W_\pi \bar{h}_i + b_\pi. \qquad (20)$$

where ({h}_i) is a pooled representation for the aligned phone, syllable, or word segment. The model is multi-task because pronunciation is not only a phone-recognition problem; it also involves stress, timing, and prosody.

3.5 Reference alignment and phone-level deviation

For prompted pronunciation practice, the target utterance is known. Let the expected phone sequence be:

$$P^{ref} = \left(p_1^{ref}, p_2^{ref}, \ldots, p_I^{ref}\right). \qquad (21)$$

The learner's neural evidence produces a sequence of predicted phone distributions:

$$\hat{P} = \left(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_J\right). \qquad (22)$$

Alignment between the reference and learner sequence can be formulated using dynamic programming:

$$D(i,j) = min\begin{cases} D(i-1,j) + c_{del}\left(p_i^{ref}\right), \\ D(i,j-1) + c_{ins}\left(\hat{p}_j\right), \\ D(i-1,j-1) + c_{sub}\left(p_i^{ref}, \hat{p}_j\right). \end{cases} \qquad (23)$$

The resulting alignment map () assigns learner acoustic segments to target units:

$$\mathcal{A}: i \mapsto \left[t_i^{start}, t_i^{end}\right]. \qquad (24)$$

For each aligned unit, a posterior-based deviation score is computed:

$$E_i^{seg} = 1 - P_\theta\left(p_i^{ref} \mid X_i\right). \qquad (25)$$

A contrastive phone-confusion score can be added:

$$C_i^{conf} = \max_{q \neq p_i^{ref}} P_\theta(q \mid X_i) - P_\theta\left(p_i^{ref} \mid X_i\right). \qquad (26)$$

If (C_i^{conf}>0), the most likely competing phone is more probable than the target phone. This makes it possible to identify not only that a phone is weak, but also which replacement it resembles.

3.6 Prosodic deviation model

Pronunciation feedback should not focus only on individual consonants and vowels. Stress, duration, rhythm, and intonation also affect intelligibility and naturalness. Word stress can be evaluated using a target stress label (z_i^{ref}) and predicted stress probability (_i):

$$E_i^{stress} = \left|z_i^{ref} - \hat{z}_i\right|. \qquad (27)$$

Duration deviation can be normalized against a reference duration (d_i^{ref}):

$$E_i^{dur} = \left|\frac{d_i - d_i^{ref}}{d_i^{ref} + \varepsilon}\right|. \qquad (28)$$

Pitch-contour deviation can be computed using a distance between normalized contours:

$$E_i^{pitch} = \frac{1}{K}\sum_{k=1}^{K}\left|\tilde{F}_{0,i}(k) - \tilde{F}_{0,i}^{ref}(k)\right|. \qquad (29)$$

The total pronunciation deviation for unit (i) is then:

$$E_i = \beta_1 E_i^{seg} + \beta_2 E_i^{stress} + \beta_3 E_i^{dur} + \beta_4 E_i^{pitch} + \beta_5 max\left(0, C_i^{conf}\right), \qquad (30)$$

where (_1,,_5) are non-negative weights and (_k_k=1). This equation is central to the proposed model because it integrates segmental and suprasegmental evidence in a single deviation measure.

3.7 Knowledge-base and knowledge-graph layer

The knowledge base is defined as:

$$KB = (\mathcal{V}, \mathcal{R}, \mathcal{T}), \qquad (31)$$

where () is a set of entities, () is a set of typed relations, and () is a set of feedback templates and pedagogical actions. A knowledge-graph triple has the form:

$$(v_s, r, v_o) \in \mathcal{R}, \qquad (32)$$

for example:

$$(/\theta/, requires, dental_fricative_articulation). \qquad (33)$$

A rule for a specific error can be represented as:

$$Rule_k: \ condition_k(E_i, L_i, context_i) \Rightarrow action_k(F_i). \qquad (34)$$

where (L_i) denotes learner profile information, including first-language background, proficiency level, and previous errors. Figure 3 shows how the graph transforms a neural diagnosis into a grounded explanation.

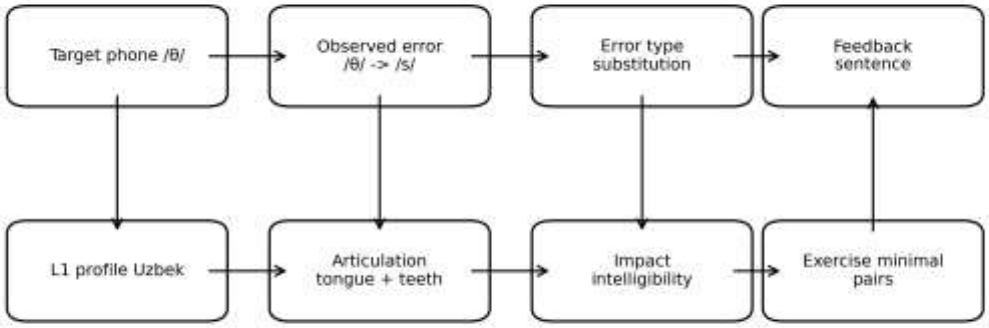


Figure 3. Knowledge-graph reasoning for explainable pronunciation feedback.

The knowledge base should contain at least six entity groups: phonemes, articulatory features, acoustic correlates, common L1-transfer patterns, error types, and corrective exercises. Table 3 gives a compact structure.

Table 3. Knowledge-base entities and relations for pronunciation feedback

Entity group	Example entities	Example relations	Feedback role
Phoneme	/θ/, /ð/, /æ/, /ɪ/, /iː/	contrasts_with, often_confused_with	Defines target contrasts
Articulation	tongue tip, teeth, lips, jaw opening, voicing	requires, differs_by	Explains physical speech production
Acoustic cue	/F2, duration, frication, pitch contour	indicates, deviates_from	Links signal evidence to phonetic interpretation
Error type	substitution, deletion, insertion, stress shift	diagnosed_as, affects	Categorizes learner error
L1 profile	Uzbek, Russian, Arabic, Korean	may_bias, lacks_contrast	Explains likely transfer effects
Corrective exercise	minimal pair, shadowing, slow repetition	recommended_for	Converts diagnosis into action
Feedback template	cause, correction, warning, retry task	verbalizes	Produces learner-facing explanation

3.8 Explanation confidence and feedback validity

A key problem in AI-generated educational feedback is the risk of confident but unsupported explanation. NS-SMF therefore uses a feedback-validity gate. Let (c_i^{NN}) be neural confidence, (c_i^{KB}) be knowledge-base support, and (c_i^{align}) be alignment reliability. Overall explanation confidence is:

$$Conf_i = \gamma_1 c_i^{NN} + \gamma_2 c_i^{KB} + \gamma_3 c_i^{align}, \qquad \sum_{m=1}^{3} \gamma_m = 1. \qquad (35)$$

Knowledge-base support can be computed as:

$$c_i^{KB} = \frac{|\{r \in KB: r\text{supports}type_i\}|}{\left|\{r \in KB: r\text{is relevant to}p_i^{ref}\}\right| + \varepsilon}. \qquad (36)$$

A feedback item is allowed only if:

$$Valid(F_i) = \mathbb{I}[Conf_i \geq \tau_c \wedge E_i \geq \tau_e \wedge c_i^{KB} \geq \tau_{KB}], \qquad (37)$$

where (_c), (e), and ({KB}) are confidence, deviation, and knowledge-support thresholds. If this condition fails, the system should avoid strong correction and use a cautious message such as "please repeat the word more clearly" or route the item to teacher review.

3.9 Error priority and pedagogical utility

Learners should not receive feedback on every small imperfection. Too much feedback can overload the learner and reduce motivation. NS-SMF ranks errors using a priority function:

$$Pr_i = \alpha_1 E_i + \alpha_2 Imp_i + \alpha_3 Rec_i + \alpha_4 Level_i + \alpha_5 Goal_i, \qquad \sum_{m=1}^{5} \alpha_m = 1. \qquad (38)$$

Here (Imp_i) is intelligibility impact, (Rec_i) is recurrence in previous attempts, (Level_i) measures relevance to the learner's level, and (Goal_i) measures relevance to the current lesson objective. The feedback generator selects only the top (K) valid errors:

$$\mathcal{F}_{out} = TopK\{F_i: Valid(F_i) = 1 \text{ sorted by } Pr_i\}. \qquad (39)$$

Pedagogical utility is estimated as:

$$PedUtil(F_i) = \eta_1 Clarity_i + \eta_2 Actionability_i + \eta_3 Specificity_i + \eta_4 LevelFit_i + \eta_5 ExerciseLink_i. \qquad (40)$$

This separates technically correct diagnosis from useful feedback. A feedback sentence may be scientifically correct but not pedagogically useful if it is too long, too technical, not connected to a practice task, or above the learner's level.

3.10 Learner-profile update

A feedback system must remember recurring patterns. Let (r_{i,t}) be the recurrence value of error (i) after attempt (t). An exponential moving average update can be used:

$$r_{i,t} = \lambda r_{i,t-1} + (1 - \lambda)\mathbb{I}[E_{i,t} \geq \tau_e]. \qquad (41)$$

where (0) controls memory. Improvement after feedback can be measured as:

$$\Delta_i = E_{i,t}^{before} - E_{i,t}^{after}. \qquad (42)$$

If (_i>0), pronunciation error decreased after feedback. A learner-level progress score over (N) target units can be defined as:

$$Progress_t = \frac{1}{N}\sum_{i=1}^{N} max\left(0, \Delta_{i,t}\right). \qquad (43)$$

The feedback loop is shown in Figure 4.

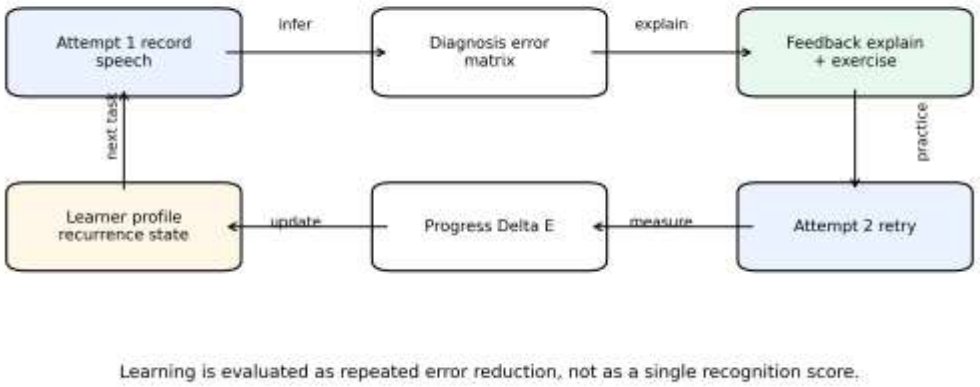


Figure 4. Closed-loop feedback and learner-profile update.

3.11 Model training objective

For supervised or semi-supervised training, the model can use a multi-task loss:

$$\mathcal{L} = \lambda_p \mathcal{L}_{phone} + \lambda_s \mathcal{L}_{stress} + \lambda_d \mathcal{L}_{dur} + \lambda_e \mathcal{L}_{error} + \lambda_c \mathcal{L}_{conf} + \lambda_f \mathcal{L}_{fair}. \quad (44)$$

The phone loss is cross-entropy:

$$\mathcal{L}_{phone} = -\sum_i \sum_{p \in \mathcal{P}} y_{i,p} \, log P_\theta(p \mid X_i). \quad (45)$$

The error-type loss is also cross-entropy:

$$\mathcal{L}_{error} = -\sum_i \sum_{k \in \mathcal{K}} y_{i,k}^{err} \, log P_\theta(k \mid X_i). \quad (46)$$

The duration loss can be mean absolute error:

$$\mathcal{L}_{dur} = \frac{1}{I} \sum_i \left| d_i - \hat{d}_i \right|. \quad (47)$$

A confidence-calibration loss can be defined through expected calibration error:

$$ECE = \sum_{b=1}^{B} \frac{|S_b|}{n} |acc(S_b) - conf(S_b)|. \quad (48)$$

Calibration is important because invalid high-confidence feedback can harm learning. The model should know when it is uncertain.

3.12 Algorithmic description

Algorithm 1 describes the proposed inference and feedback process.

Table 4. Algorithm 1: NS-SMF inference and feedback generation

Step	Operation	Formal output
1	Record learner speech for a known prompt	(x[n])
2	Preprocess, frame, and extract acoustic features	(M^a)
3	Encode speech with neural encoder	(H=Enc_(M^a))
4	Align learner signal with target phone/prosody sequence	()
5	Estimate phone, stress, duration, and pitch deviations	(E_i)
6	Detect error type and competing realization	(type_i, q_i)
7	Query knowledge base for explanation rules	(Rule_k)
8	Compute explanation confidence and feedback validity	(_i, (F_i))
9	Rank valid errors by pedagogical priority	(Pr_i)
10	Generate feedback and corrective exercise	(F_{out})
11	Store post-feedback result and update learner profile	(r_{i,t}, _t)

The algorithm is intentionally conservative. It does not attempt to provide fluent feedback for every uncertain event. Instead, it prioritizes feedback that is simultaneously supported by neural evidence, knowledge-base rules, and pedagogical relevance.

4. Pilot study design and evaluation protocol

4.1 Research questions

The first empirical study can be organized around the following research questions:

- Can the proposed NS-SMF model detect selected segmental and prosodic pronunciation errors with acceptable agreement against expert annotations?
- Does knowledge-base-grounded feedback improve learner understanding compared with score-only or ASR-only feedback?
- Does prioritized feedback reduce repeated pronunciation errors across attempts?
- Does the explanation-confidence gate reduce unsupported or misleading feedback?
- How does the model perform for Uzbek-speaking learners of English on sounds and prosodic patterns that are known to be challenging?

4.2 Participants and speech material

A feasible first study should use a narrow population and a controlled task design. The recommended initial group is 40-80 Uzbek-speaking learners of English at A2-B2 levels. Each participant records a set of prompted items: isolated words, minimal pairs, short sentences, and short read-aloud passages. A public corpus such as L2-ARCTIC can be used for pretraining, external comparison, or methodological benchmarking because it contains non-native English speech from 24 speakers with several first-language backgrounds and includes manually annotated mispronunciation labels for part of the corpus (Zhao et al., 2018). However, a local Uzbek-speaking learner corpus is needed for the target study because L1-specific transfer patterns are central to the proposed knowledge-base layer.

The initial target set should be limited but theoretically meaningful. For English, it may include dental fricatives /θ/ and /ð/, tense-lax vowel contrasts /i:/-/ɪ/, low front vowel /æ/, final consonant clusters, word stress, and sentence prominence. The limited scope makes it possible to collect high-quality expert annotations and to validate whether the knowledge-base explanations are correct.

4.3 Experimental groups

The pilot can compare three groups:

Table 5. Suggested pilot-group design

Group	Feedback condition	Main purpose
Control group	Teacher's ordinary classroom correction or delayed feedback	Baseline learning comparison
ASR/GOP group	Automatic score or phone-level error indication without explanation	Tests the value of automatic detection alone
NS-SMF group	Neural-symbolic explanation plus prioritized exercise	Tests the added value of explainable feedback

All groups should perform pre-test, practice sessions, immediate post-test, and delayed post-test. This design allows the researcher to evaluate not only immediate improvement but also retention.

4.4 Technical metrics

Mispronunciation detection can be evaluated using precision, recall, and -score:

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}. \qquad (49)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \qquad (50)$$

Phone-level agreement with experts can be measured using Cohen's kappa:

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \qquad (51)$$

where (p_o) is observed agreement and (p_e) is chance agreement. Pronunciation score correlation with expert ratings can be measured by Spearman's rank correlation:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}. \qquad (52)$$

These metrics are necessary but not sufficient, because a technically accurate system may still produce poor feedback.

4.5 Pedagogical and explainability metrics

The NS-SMF model should also be evaluated through feedback-specific metrics. Explanation faithfulness can be defined as the proportion of feedback items grounded in valid knowledge-base rules:

$$Faithfulness = \frac{|\{F_i : Valid(F_i) = 1 \wedge ExpertAccept(F_i) = 1\}|}{|\{F_i : Valid(F_i) = 1\}| + \varepsilon}. \qquad (53)$$

Feedback clarity can be rated by learners and teachers on a Likert scale:

$$Clarity = \frac{1}{N} \sum_{i=1}^{N} r\,ating_i^{clarity}. \qquad (54)$$

Corrective success rate can be measured after a retry:

$$CSR = \frac{\left|\{i : E_{i,t+1} < E_{i,t}\}\right|}{|\{i : F_i \in \mathcal{F}_{out}\}| + \varepsilon}. \qquad (55)$$

Repeated-error reduction measures whether systematic problems decrease over sessions:

$$RER = 1 - \frac{\sum_i r_{i,T}}{\sum_i r_{i,0} + \varepsilon}. \qquad (56)$$

A combined evaluation score can be defined as:

$$Eval = \omega_1 F1 + \omega_2 \kappa + \omega_3 Faithfulness + \omega_4 Clarity + \omega_5 CSR + \omega_6 RER, \quad \sum_m \omega_m = 1. \qquad (57)$$

Table 6 maps the metrics to research purposes.

Table 6. Evaluation metrics for NS-SMF

Metric	Formula or scale	What it measures	Why it matters
//	Equations (49)-(50)	Error-detection quality	Detects false alarms and missed errors
Cohen's kappa	Equation (51)	Agreement with experts	Validates annotation-level consistency
Spearman correlation	Equation (52)	Score/rank relationship with human ratings	Checks whether model ranking matches expert judgement
	Equation (53)	Knowledge-base grounding	Prevents unsupported explanations
	Equation (54)	Learner/teacher interpretability	Ensures feedback is understandable
Corrective success rate	Equation (55)	Immediate effect after retry	Measures whether feedback leads to action
Repeated-error reduction	Equation (56)	Longitudinal improvement	Shows learning rather than one-time scoring
Combined evaluation	Equation (57)	Overall model value	Integrates technical and pedagogical dimensions

5. Analytical scenarios and expected comparative behavior

Because this article presents a methodological model, the following analytical scenarios define how the model should behave before full empirical validation. They are not reported as completed experimental results; they are formal test cases for implementation.

5.1 Scenario A: score-only system versus explainable feedback

A learner pronounces think with an /s/-like initial sound. A score-only system may return 62/100 or mark the word as weak. NS-SMF should return a structured diagnosis:

$$type_i = substitution, \quad p_i^{ref} = /\theta/, \quad \hat{q}_i = /s/, \quad E_i > \tau_e. \qquad (58)$$

If the knowledge base contains the dental-fricative rule, the system generates a grounded feedback item:

$$F_i = \{substitution, /\theta/ \rightarrow /s/, cause, articulation, minimal\ pair\ exercise\}. \qquad (59)$$

The advantage is not merely more text. It is that the feedback is traceable to an acoustic diagnosis and an explicit knowledge-base rule.

5.2 Scenario B: overcorrection prevention

Suppose the alignment confidence is low due to background noise. A neural model may still produce a phone prediction, but strong feedback would be risky. NS-SMF blocks such feedback using Equation (37):

$$Conf_i < \tau_c \Rightarrow Valid(F_i) = 0. \qquad (60)$$

In this case, the system should ask for a clearer repetition or flag the item for teacher review. This behavior is important because incorrect pronunciation feedback can be more harmful than no feedback.

5.3 Scenario C: feedback prioritization

Suppose a learner has five minor deviations and one recurring phonemic contrast error that affects intelligibility. The model should not list all errors. It should prioritize the important one:

$$F_{out} = TopK(F_i, Pr_i), \quad K \in \{1,2,3\}. \qquad (61)$$

This aligns the system with pedagogical practice. Learners usually benefit more from a small number of actionable corrections than from a long list of technical observations.

5.4 Scenario D: L1-aware explanation

For Uzbek-speaking learners, some English contrasts may not map directly onto familiar L1 phonological categories. The knowledge base can include L1-transfer relations:

$$(L1_{Uzbek}, may_bias, /\theta/ \rightarrow /s/). \qquad (62)$$

This does not mean that every Uzbek-speaking learner will make the same error. Rather, it allows the system to generate a plausible explanation when neural evidence confirms the pattern. The model therefore avoids stereotyping by requiring both L1-rule support and actual speech evidence.

Table 7 summarizes the expected comparative behavior.

Table 7. Expected comparative behavior of pronunciation feedback models

Situation	Score-only ASR/GOP	Neural MDD only	Proposed NS-SMF
Correct word but weak phone	May miss local error	May detect phone error	Detects, explains, and assigns exercise

Situation	Score-only ASR/GOP	Neural MDD only	Proposed NS-SMF
Low confidence due to noise	May still output a score	May output uncertain label	Blocks strong feedback and requests retry
Multiple errors in one sentence	May provide global score	May list many labels	Ranks by intelligibility and learner priority
L1-related error	Usually not explained	Detected but not pedagogically interpreted	Explains using L1-aware knowledge-base rules
Prosodic weakness	Often underdeveloped	Possible if trained	Integrates stress, duration, and pitch deviation
Teacher audit	Hard to interpret	Model-dependent	Traceable from feedback to rule and evidence

6. Discussion

6.1 Scientific novelty

The proposed model is novel not because it uses neural networks in speech processing; that is now common. Its novelty is the formal integration of three levels: mathematical speech representation, neural pronunciation diagnosis, and symbolic pedagogical explanation. The speech matrix (M={M^a,M^p,M^e}) makes the model more structured than a direct waveform-to-score system. The deviation function in Equation (30) combines segmental and suprasegmental evidence. The knowledge-base rules in Equations (31)-(37) make feedback traceable rather than black-box. The priority function in Equation (38) prevents feedback overload. The learner-profile update in Equation (41) transforms the system from one-time assessment into adaptive learning support.

This combination is important for Scopus-level research because it addresses both technological and educational validity. A system may perform well as a classifier but fail as a learning tool. Conversely, a system may produce friendly feedback but lack diagnostic evidence. NS-SMF attempts to satisfy both requirements: evidence-based diagnosis and pedagogically grounded explanation.

6.2 Why mathematics is necessary in this model

The mathematical component is not decorative. It clarifies which objects are being measured, how they are transformed, and where decisions are made. For example, without Equation (12), the word "matrix" would remain vague. Without Equation (30), the relation between phone errors and prosodic errors would be unspecified. Without Equation (37), the system could generate feedback without sufficient confidence. Without Equation (57), the evaluation might collapse into only model accuracy and ignore feedback faithfulness and learning improvement.

The formulas also make the model reproducible. Researchers can change neural encoders, datasets, or feedback templates while preserving the core formal structure. This is essential for a publishable methodological article because it allows future work to test, compare, and extend the model rather than merely describe an idea.

6.3 Pedagogical implications

The model supports a practical feedback style: short, specific, evidence-based, and exercise-linked. A useful learner-facing feedback item should contain four elements:

$$F_i = (what_i, why_i, how_i, retry_i). \qquad (63)$$

The component (what_i) states the error, (why_i) explains the phonetic or prosodic reason, (how_i) gives an articulation or practice instruction, and (retry_i) provides the next speaking task. For example, for a dental-fricative substitution, the system may say: "In think, the first sound was close to /s/. The target /θ/ requires the tongue tip near the teeth and continuous airflow. Practise thin-sin and then repeat the sentence."

This feedback is different from generic advice such as "your pronunciation is not good" or "repeat again." It gives the learner a specific correction path. It also allows teachers to review learner profiles and identify common classroom patterns.

6.4 Limitations

Several limitations must be acknowledged. First, the model requires high-quality annotated speech data for reliable evaluation. Public corpora are useful, but an Uzbek-speaking learner corpus is needed if the study focuses on Uzbek learners. Second, knowledge-base construction requires linguistic and pedagogical expertise. Third, acoustic environments, microphone quality, and background noise may affect neural confidence and alignment. Fourth, pronunciation norms should be selected carefully; the goal should be intelligibility and communicative effectiveness rather than unrealistic native-speaker imitation. Fifth, the first implementation should focus on a limited set of sounds and prosodic targets before expanding to open conversation.

6.5 Ethical and fairness considerations

Pronunciation feedback systems can unintentionally reinforce harmful norms if they treat native-like pronunciation as the only acceptable target. The proposed model should therefore define pronunciation goals in terms of intelligibility, communicative clarity, and learner objectives. The knowledge base should distinguish serious intelligibility-related deviations from harmless accent variation. A fairness-aware penalty can be included in the training objective:

$$\mathcal{L}_{fair} = \sum_{g \in G} \left| ErrorRate_g - ErrorRate_{all} \right|. \qquad (64)$$

where (G) is a set of learner groups. The purpose is not to erase accents, but to avoid systematically harsher or less accurate feedback for specific learner populations.

7. Conclusion

This article proposed NS-SMF, an explainable neuro-symbolic speech matrix model for personalized foreign language pronunciation feedback. The model transforms learner speech into a multilayer mathematical representation, uses neural speech analysis to detect segmental and suprasegmental deviations, aligns learner productions with target pronunciation units, and uses a knowledge base to generate grounded feedback. Its central claim is that pronunciation learning systems should not stop at black-box scoring. They should explain errors through phonetic, prosodic, articulatory, and pedagogical knowledge and should recommend targeted practice based on learner-specific priority.

The proposed formalism includes acoustic feature matrices, neural latent representations, alignment equations, pronunciation deviation functions, knowledge-graph rules, feedback-validity gates, priority scoring, learner-profile updates, and combined technical-pedagogical evaluation metrics. These components make the model suitable for future implementation and empirical testing. The next research step is to build a pilot corpus of Uzbek-speaking learners of English, annotate selected segmental and prosodic errors, implement the neural-symbolic pipeline, and compare NS-SMF with score-only and neural-only feedback systems.

Declarations

Funding. No external funding has been declared for this conceptual manuscript draft.

Conflict of interest. The author declares no conflict of interest.

Data availability. No new dataset is reported in this manuscript draft. The proposed pilot study may use public corpora for benchmarking and a locally collected learner corpus for target evaluation.

Ethical note. Future human-subject data collection should obtain informed consent, anonymize learner recordings, protect personal data, and avoid feedback that stigmatizes accent variation unrelated to intelligibility.

References

- Amrate, M., & Tsai, P.-H. (2024). Computer-assisted pronunciation training: A systematic review. ReCALL, 1-21. https://doi.org/10.1017/S0958344024000181
- 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.
- Ben Chaabene, N. E. H., & Hammami, H. (2026). Neuro-symbolic synergy in education: A survey of LLM-knowledge graph integration for explainable reasoning and emotion-aware student support. Smart Learning Environments, 13, Article 6.
- El Kheir, Y., Ali, A., & Chowdhury, S. A. (2023). Automatic pronunciation assessment: A review. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 8304-8324). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.557
- 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. https://doi.org/10.1080/09588221.2012.700315
- Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), Article 71. https://doi.org/10.1145/3447772
- Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451-3460.

- Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems, 33(2), 494-514. https://doi.org/10.1109/TNNLS.2021.3070843
- Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. TESOL Quarterly, 39(3), 369-377.
- Mohsen, M. A., Mahdi, H. S., AlThebi, S. H., Alkhammash, R., Alsagour, R., Alsalem, A., Almudawi, A., & Alshahrani, A. (2025). A scientometric study of computer-assisted pronunciation training in second language acquisition: Technological affordances and research trends. Humanities and Social Sciences Communications, 12, Article 438. https://doi.org/10.1057/s41599-025-04474-y
- Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. Computer Assisted Language Learning, 15(5), 441-467. https://doi.org/10.1076/call.15.5.441.13473
- Pennington, M. C., & Rogerson-Revell, P. (2019). English pronunciation teaching and research: Contemporary perspectives. Palgrave Macmillan.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (pp. 28492-28518). PMLR.
- Rogerson-Revell, P. M. (2021). Computer-assisted pronunciation training (CAPT): Current issues and future directions. RELC Journal, 52(1), 189-205. https://doi.org/10.1177/0033688220977406
- Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication, 30(2-3), 95-108. https://doi.org/10.1016/S0167-6393(99)00044-8
- Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2018). L2-ARCTIC: A non-native English speech corpus. In Proceedings of Interspeech 2018 (pp. 2783-2787). https://doi.org/10.21437/Interspeech.2018-1110