1. Madrahimov Sh. **Sun’iy intellekt tizimlari: o‘quv qo‘llanma.** - Toshkent: Metodist nashriyoti, 2023. - 54 b. ISBN 978-9910-9843-7-2. BT URSPI
2. Xo‘jaqulov T.A., Malikova N.T. **Sun’iy intellekt: o‘quv qo‘llanma.** - Toshkent: Aloqachi, 2019. - 192 b.
3. Ясницкий Л.Н. **Введение в искусственный интеллект: учебное пособие для студентов вузов.** - М.: Академия, 2005.- 175 с. ISBN 5-7695-1958-4. Search RSL+1
4. Круглов В.В., Борисов В.В. **Искусственные нейронные сети: теория и практика.** - 2-е изд. - М.: Горячая линия–Телеком, 2002. - 381 с. ISBN 5-93517-031-0. Библиотека МГТУ.
5. Галушкин А.И. **Теория нейронных сетей.** - М.: ИПРЖР, 2000.
6. Борисов В.В., Круглов В.В., Федулов А.С. **Нечеткие модели и сети.** - М.: Горячая линия–Телеком.
7. Хасанов Д.Р. **Sun’iy intellekt usullari yordamida qishloq xo‘jaligi takroriy ekinlari hosildorligini bashoratlashning model va algoritmlarini ishlab chiqish.** - Toshkent.
8. Абдусатторова С.Ф. **Raqamli transformatsiya muhitida ijtimoiy jarayonlarni modellashtirish: sun’iy intellektning falsafiy-huquqiy muammolari.** - Toshkent.
9. Bishop C.M. **Pattern Recognition and Machine Learning.** - New York: Springer, 2006.
10. Goodfellow I., Bengio Y., Courville A. **Deep Learning.** - Cambridge: MIT Press, 2016.
11. Russell S., Norvig P. **Artificial Intelligence: A Modern Approach.** - 4th ed. - Pearson, 2021.
12. Hastie T., Tibshirani R., Friedman J. **The Elements of Statistical Learning.** - 2nd ed. - Springer, 2009.

L1-AWARE PHONEME-LEVEL PRONUNCIATION ASSESSMENT OF UZBEK L1–ENGLISH L2 SPEECH USING WAV2VEC 2.0 ACOUSTIC-PHONETIC REPRESENTATIONS

Rohmonqulov Muhammadyusuf Egamberdi o‘g‘li

Jizzakh Branch of the National University of Uzbekistan named after Mirzo Ulugbek, Jizzakh, Uzbekistan

muhammadyusufrohmonqulov501@gmail.com

Abstract. Automatic pronunciation assessment is especially important for Uzbek learners of English because differences between the Uzbek and English phonological systems can produce systematic L1-induced deviations in segmental and prosodic realization. This paper proposes an interpretable pronunciation-assessment model in which wav2vec 2.0 is used as a self-supervised acoustic-phonetic representation extractor, while the assessment stage is formulated as a phoneme-level scoring problem rather than a conventional speech-recognition

problem. For each target phoneme, the model combines four complementary indicators: posterior evidence for the canonical phoneme, distance between learner and reference embeddings, duration/prosodic conformity, and an Uzbek-L1-specific confusion penalty. These components are fused by a bounded logistic scoring function and then aggregated hierarchically from phoneme to word and utterance levels. The mathematical formulation is designed to preserve interpretability: every final score can be decomposed into acoustic, temporal, prosodic, and L1-interference contributions. The proposed formulation is suitable for low-resource Uzbek L1–English L2 corpora because most representation learning is transferred from wav2vec 2.0, while only the calibration and scoring layers require task-specific labeled data. The model provides a technical basis for future computer-assisted pronunciation training systems that can produce both a numerical score and error-oriented feedback.

Keywords: automatic pronunciation assessment; wav2vec 2.0; Uzbek L1; English L2; phoneme scoring; acoustic-phonetic analysis; Goodness of Pronunciation; self-supervised learning.

1. Introduction

Automatic pronunciation assessment (APA) aims to estimate pronunciation quality from speech and to provide scores that are consistent with expert judgments. Classical systems mainly relied on likelihood-based Goodness of Pronunciation (GOP), in which the acoustic likelihood of a target phone is compared with competing phone hypotheses [2]. Modern self-supervised speech models substantially extend this idea because they learn contextual representations directly from raw audio and preserve information relevant to phonetic content, timing, speaker variation and prosody. wav2vec 2.0 is one of the most influential models in this class; it combines a convolutional feature encoder with a Transformer context network and is pre-trained by masked contrastive learning on unlabeled speech [1]. Recent studies have shown that wav2vec 2.0 representations can be used for L2 proficiency assessment, mispronunciation detection and multi-level pronunciation scoring [3–7].

For Uzbek L1 speakers learning English as L2, an additional problem is systematic cross-linguistic interference. English phonemes and contrasts that are absent or realized differently in Uzbek, together with differences in vowel reduction, consonant clusters and stress organization, can generate predictable error patterns [8]. Typical high-value confusions include interdental fricatives /θ, ð/, the /w/–/v/ contrast, tense–lax vowel distinctions such as /iː/–/ɪ/, and stress-related temporal deviations. Therefore, a pronunciation scorer trained only to imitate a generic English acoustic model may penalize accented but intelligible speech inconsistently, or fail to identify which acoustic deviation is pedagogically important.

The objective of this work is to formulate a technically transparent assessment stage for Uzbek L1–English L2 speech after acoustic-phonetic representations have been obtained by wav2vec 2.0. The main contribution is an L1-aware phoneme-level scoring function that integrates (i) canonical phoneme posterior evidence, (ii) latent embedding similarity to reference speech, (iii) duration and prosodic conformity, and

(iv) an Uzbek-specific confusion penalty. In contrast with a single confidence score, the proposed formulation yields a decomposable 0–100 score and can support diagnostic feedback at phoneme, word and utterance levels.

2. Materials and Methods

2.1. Input representation and wav2vec 2.0 encoding

Let a normalized mono speech waveform sampled at 16 kHz be

$$\mathbf{x}=[x_1,x_2,\ldots,x_N]^T, \qquad x_n\in[-1,1].$$

The convolutional feature encoder of wav2vec 2.0 maps the waveform to a shorter latent sequence

$$\mathbf{z}_t=f_{\text{enc}}(\mathbf{x})_t, \qquad t=1,\ldots,T,$$

and the Transformer context network produces contextual acoustic-phonetic vectors

$$\mathbf{h}_t=f_{\text{ctx}}(\mathbf{z}_{1:T})_t, \qquad \mathbf{h}_t\in\mathrm{R}^D.$$

A phoneme classification head converts each contextual vector into posterior probabilities over an English phoneme inventory P (plus the CTC blank symbol):

$$P(q|\mathbf{h}_t)=\frac{\exp\left(\mathbf{w}_q^T\mathbf{h}_t+b_q\right)}{\sum_{r\in\mathrm{P}\cup\{\varnothing\}}\exp\left(\mathbf{w}_r^T\mathbf{h}_t+b_r\right)}.$$

For a prompted English utterance, the canonical phoneme sequence $\mathbf{p}=(p_1,\ldots,p_K)$ is obtained from the reference transcription and pronunciation lexicon. CTC-based forced alignment associates each target phone p_i with a frame set $\Omega_i=\{t_i^{(s)},\ldots,t_i^{(e)}\}$.

This produces the mapping required for phone-specific scoring without manually labeling every frame.

2.2. Phoneme-level assessment variables

Posterior evidence. A confidence margin is computed between the canonical phone and its strongest competitor:

$$m_i=\frac{1}{|\Omega_i|}\sum_{t\in\Omega_i}\left[\log P(p_i|\mathbf{h}_t)-\max_{q\neq p_i}\log P(q|\mathbf{h}_t)\right].$$

A positive m_i indicates that the target phone dominates competing hypotheses; a negative value indicates that another phone is acoustically more plausible. This posterior-margin formulation is a neural analogue of the classical GOP principle [2].

Embedding similarity. The learner phone embedding is obtained by temporal pooling,

$$\bar{\mathbf{h}}_i=\frac{1}{|\Omega_i|}\sum_{t\in\Omega_i}\mathbf{h}_t.$$

For each English phoneme p, a reference centroid $\boldsymbol{\mu}_p$ and covariance $\boldsymbol{\Sigma}_p$ are estimated from correctly pronounced reference speech. The Mahalanobis distance and its normalized similarity are

$$d_i^{(e)}=\sqrt{\left(\bar{\mathbf{h}}_i-\boldsymbol{\mu}_{p_i}\right)^T\boldsymbol{\Sigma}_{p_i}^{-1}\left(\bar{\mathbf{h}}_i-\boldsymbol{\mu}_{p_i}\right)},$$

$$s_i^{(e)}=\exp\left[-\frac{1}{2}\left(d_i^{(e)}\right)^2\right].$$

Thus, $0<s_i^{(e)}\leq 1$; values closer to one indicate that the learner realization lies nearer to the reference region in the contextual wav2vec 2.0 space.

Duration and prosody. Let the aligned duration of phone p_i be d_i and the reference duration statistics be $\left(\mu_{p_i}^{(d)},\sigma_{p_i}^{(d)}\right)$. A duration conformity score is

$$s_i^{(d)}=\exp\left[-\frac{1}{\tau_d}\left|\log\frac{d_i}{\mu_{p_i}^{(d)}}\right|\right],$$

where $\tau_d>0$ controls tolerance. For stressed vowels and syllable nuclei, a compact prosodic vector can be defined as $\mathbf{v}_i=[F0_i,E_i,R_i]^T$, containing normalized pitch, energy and local rhythm/stress indicators. With reference statistics $\left(\boldsymbol{\mu}_{p_i}^{(v)},\mathbf{R}_{p_i}\right)$,

$$s_i^{(v)}=\exp\left[-\frac{1}{2}\left(\mathbf{v}_i-\boldsymbol{\mu}_{p_i}^{(v)}\right)^T\mathbf{R}_{p_i}^{-1}\left(\mathbf{v}_i-\boldsymbol{\mu}_{p_i}^{(v)}\right)\right].$$

These terms prevent the assessment from reducing pronunciation to segment identity alone, which is important because current pronunciation assessment research treats segmental, suprasegmental and temporal features as complementary dimensions [4,6].

2.3. Uzbek-L1 interference penalty and score fusion

Let $\mathrm{C}_{UZ}(p_i)$ denote a set of likely Uzbek-L1-induced competitors for the target English phone. Examples include $\mathrm{C}_{UZ}(\theta)=\{s,t\}$, $\mathrm{C}_{UZ}(\eth)=\{z,d\}$, $\mathrm{C}_{UZ}(w)=\{v\}$ and a bidirectional confusion set for /ɪ/ and /iː/. The confusion penalty is

$$c_i=\sum_{q\in\mathrm{C}_{UZ}(p_i)}\lambda_{p_iq}\left(\frac{1}{|\Omega_i|}\sum_{t\in\Omega_i}P\left(q|\mathbf{h}_t\right)\right),$$

where $\lambda_{p_iq}\geq 0$ is estimated from annotated Uzbek-accented English data. This term makes the scorer sensitive to recurrent L1 transfer patterns while keeping the acoustic evidence data-driven.

The final phoneme score is defined by a bounded logistic fusion:

$$S_i=100\,\sigma\left(\beta_0+\beta_1 m_i+\beta_2 s_i^{(e)}+\beta_3 s_i^{(d)}+\beta_4 s_i^{(v)}-\beta_5 c_i\right),$$

where $\sigma(a)=1/(1+e^{-a})$ and $\beta_j\geq 0$ for $j=1,\ldots,5$. Word-level and utterance-level pronunciation scores are obtained hierarchically:

$$S_{word}=\frac{\sum_{i\in word}\omega_i S_i}{\sum_{i\in word}\omega_i},\qquad S_{utt}=\alpha S_{seg}+(1-\alpha)S_{supra},$$

where ω_i can reflect phoneme functional load or reliability, S_{seg} aggregates phoneme scores, and S_{supra} summarizes stress, rhythm, intonation and temporal fluency. For calibration against expert labels y_j, the scoring parameters are estimated using a robust Huber objective,

$$L=\frac{1}{M}\sum_{j=1}^{M}\text{Huber}\left(S_{utt}^{(j)}-y_j\right)+\lambda\|\boldsymbol{\beta}\|_2^2.$$

2.4. Parameter estimation and validation protocol

For an Uzbek L1–English L2 corpus, each learner utterance should be scored independently by at least two trained raters. Phone-level labels can distinguish *correct*, *substitution*, *deletion/insertion* and *distorted but intelligible* realizations, while utterance-level pronunciation can be represented on an ordinal scale and linearly mapped to 0–100. Inter-rater reliability should be measured before model training; disagreements are resolved by consensus only after the initial independent ratings have been stored. The fusion coefficients are constrained to non-negative values by the re-parameterization $\beta_j=\text{softplus}\left(\gamma_j\right)$, which preserves the intended monotonic direction of every score component during optimization.

Speaker-independent train/validation/test partitions must be used so that the same speaker never occurs in more than one subset. Three baselines are recommended: classical GOP, direct wav2vec 2.0 utterance-level regression, and a phoneme-level scorer without the Uzbek-L1 penalty. The complete model should additionally be evaluated by ablation, removing m_i, $s_i^{(e)}$, $s_i^{(d)}$, $s_i^{(v)}$ or c_i one at a time.

3. Results

Because the present contribution is a mathematical assessment formulation rather than a claim of completed corpus-scale experimentation, the results are stated as analytical properties of the scorer and as its expected diagnostic behavior. First, the logistic fusion rule guarantees a bounded and interpretable score $0<S_i<100$. If $\beta_1,\ldots,\beta_5>0$, the score is monotonic in the intended direction. For example,

$$\frac{\partial S_i}{\partial m_i}=100\beta_1\sigma(a_i)[1-\sigma(a_i)]>0, \qquad \frac{\partial S_i}{\partial c_i}=-100\beta_5\sigma(a_i)[1-\sigma(a_i)]<0,$$

where a_i is the logit of the phoneme-level fusion rule. Therefore, stronger evidence for the target phoneme always raises the score, whereas stronger evidence for an Uzbek-specific confusion always lowers it.

Second, the model uses multiple independent error channels. If an Uzbek learner realizes English /θ/ close to /s/, the posterior margin decreases, the embedding distance increases, and the L1 penalty rises through the $\lambda_{\theta s}$ term. In contrast, a temporally unusual but spectrally correct realization can preserve m_i and $s_i^{(e)}$ while receiving a reduced $s_i^{(d)}$. This decomposition allows the system to distinguish substitution errors from duration/stress errors instead of collapsing both into one acoustic confidence value.

Third, the hierarchical aggregation prevents a few local phoneme errors from automatically determining the entire utterance score. Segmental accuracy and suprasegmental intelligibility can be calibrated separately and then combined. This is consistent with recent work showing that multi-granular representations are useful in automatic pronunciation assessment [4,6].

A fourth analytical property is computational separability. After the wav2vec 2.0 encoder has produced $\mathbf{h}_{1:T}$, the assessment layer requires only posterior aggregation, pooled embeddings and a small number of scalar operations per phone. If diagonal reference covariance is used, the additional scoring complexity is approximately $O(K(D+|P|))$ for K aligned phones, embedding dimension D and phoneme inventory size $|P|$. Consequently, the L1-aware assessment stage adds relatively little computation compared with the Transformer encoder and is compatible with near-real-time feedback.

For explanation, the model can also expose a contribution vector $\mathbf{r}_i = \left[\beta_1 m_i, \beta_2 s_i^{(e)}, \beta_3 s_i^{(d)}, \beta_4 s_i^{(v)}, -\beta_5 c_i\right]$. The largest negative component identifies the principal reason for a low phone score. This converts the numerical output into an actionable diagnostic signal without requiring a separate black-box explanation model.

4. Discussion

The proposed model extends three major lines of prior work. Classical GOP contributes the idea of comparing a target phone against competing acoustic hypotheses [2]; wav2vec 2.0 supplies transferable contextual speech representations that reduce dependence on handcrafted features [1,3]; and recent pronunciation-scoring systems motivate multi-level and phonological-feature-aware assessment [4–7]. The new element for the Uzbek L1–English L2 setting is that recurrent cross-linguistic confusions are incorporated explicitly as a calibrated penalty rather than being left implicit in a generic acoustic model.

This design is particularly relevant to a low-resource setting. The high-capacity encoder can be initialized from a pretrained wav2vec 2.0 model, while the Uzbek-specific components require substantially fewer labeled examples: reference centroids/covariances, confusion weights λ_{pq} and fusion coefficients β_j. The resulting system is also explainable: a low score can be traced to posterior competition, embedding mismatch, abnormal duration/prosody or a known L1-interference pattern. Such decomposition is preferable for computer-assisted pronunciation training because the learner needs not only a scalar rating but also information about the source of the error.

The principal limitation is calibration. Human expert ratings are still required to determine whether model features correspond to perceived intelligibility and to avoid over-penalizing legitimate accent variation. Recent automatic-scorer research also reports sensitivity to language-family groups and audio quality, which means that device conditions, age/gender balance and accent diversity must be controlled during validation [6]. Future work should therefore construct an annotated Uzbek L1–

English L2 corpus, estimate the confusion matrix empirically, compare the proposed model with GOP-only and wav2vec-only baselines, and perform ablation tests for each score component.

5. Conclusion

A phoneme-level mathematical framework for automatic assessment of Uzbek L1–English L2 pronunciation has been proposed. wav2vec 2.0 is used to obtain contextual acoustic-phonetic representations, while assessment is performed by combining target posterior evidence, embedding similarity, temporal/prosodic conformity and an Uzbek-L1 confusion penalty. The resulting 0–100 score is bounded, monotonic and decomposable, which makes it suitable for both numerical assessment and diagnostic feedback. The framework provides a technically grounded basis for subsequent corpus construction, parameter estimation and experimental validation within intelligent speech-processing and computer-assisted pronunciation-training systems.

References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[2] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000. doi: 10.1016/S0167-6393(99)00044-8.

[3] S. Bannò and M. Matassoni, "Proficiency Assessment of L2 Spoken English Using Wav2Vec 2.0," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1088–1095. doi: 10.1109/SLT54892.2023.10023019.

[4] R. C. M. C. Shekar, M. Yang, K. Hirschi, S. Looney, O. Kang, and J. H. L. Hansen, "Assessment of Non-Native Speech Intelligibility using Wav2vec2-based Mispronunciation Detection and Multi-level Goodness of Pronunciation Transformer," in *Proc. Interspeech 2023*, pp. 984–988, 2023. doi: 10.21437/Interspeech.2023-2371.

[5] S. Gupta, A. Unnam, K. Yadav, and V. Aggarwal, "Towards Building a Language-Independent Speech Scoring Assessment," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, pp. 23200–23206, 2024. doi: 10.1609/aaai.v38i21.30366.

[6] D. Cai, B. Naismith, M. Kostromitina, Z. Teng, K. P. Yancey, and G. T. LaFlair, "Developing an Automatic Pronunciation Scorer: Aligning Speech Evaluation Models and Applied Linguistics Constructs," *Language Learning*, vol. 75, Suppl. 1, pp. 170–203, 2025. doi: 10.1111/lang.70000.

[7] M. Shahin, J. Epps, and B. Ahmed, "Phonological level wav2vec2-based Mispronunciation Detection and Diagnosis method," *Speech Communication*, vol. 173, Art. no. 103249, 2025. doi: 10.1016/j.specom.2025.103249.

[8] D. A. Yuldasheva, “Phonetic interference in Uzbek learners’ production of English sounds,” *Advances in Science and Humanities*, vol. 2, no. 4, pp. 63–66, 2026. doi: 10.70728/human.v02.i04.016.

[9] 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

[10] Rahimov Nodir, Khasanov Dilmurod. (2022). The Mathematical Essence Of Logistic Regression For Machine Learning. https://doi.org/10.5281/zenodo.7239169

[11] 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

[12] Kuvandikov, J., o‘g‘li, R. M. E. (2026). An explainable neuro-symbolic speech matrix model for personalized foreign language pronunciation feedback. Raqamli Transformatsiya va Sun’iy Intellekt, 4(4), 153-158. https://dtai.uz/articles/an-explainable-neuro-symbolic-speech-matrix-model-for-personalized-foreign-language-pronunciation-feedback/

[13] o‘g‘li, A. A. A., o‘g‘li, R. M. E. (2026). Development of an interactive ai-based tool for automatic pronunciation assessment in foreign language learning. Raqamli Transformatsiya va Sun’iy Intellekt, 4(1), 38-45. https://dtai.uz/articles/development-of-an-interactive-ai-based-tool-for-automatic-pronunciation-assessment-in-foreign-language-learning /

[14] 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. https://d-pressa.com/index.php/spc/article/view/258

ALGORITHMS FOR EVALUATING SYMPTOMS FROM MAMMOGRAPHIC IMAGES IN THE DETECTION OF BREAST CANCER

O‘ngboyeva Muqaddas Yo‘ldoshevna

Jizzakh branch of the National University of Uzbekistan named after Mirzo Ulugbek, Jizzakh, Uzbekistan

E-mail: mungboyeva@gmail.com

Abstract

Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide, making early and accurate diagnosis critical for improving survival rates. Mammography is the primary imaging modality for early detection; however, visual analysis can be challenging due to dense tissue overlap and subtle radiological signs, such as micro calcifications, mass lesions, and architectural distortions. This paper presents an automated algorithmic framework designed to evaluate and classify pathological symptoms from mammographic images.