Intelligent Technologies for Sustainable Electric and Communications Systems (iTech SECOM)

Spoken Language Evaluation Through AI-Enhanced Pronunciation Feedback Systems

Mukhayyo Abdurakhmonova,Tashkent State University of Oriental Studies, Uzbekistan
muhayabdurahman@yahoo.com,https://orcid.org/my-orcid?orcid=0009-0008-0077-2904

Ma'rufkhoja Nadjimkhodjaev,Lecturer at the Higher School of Japanese Studies,
Tashkent State University of Oriental Studies, Uzbekistan, ORCID: 0009-0009-5246-0117,
Email: marufnadjim@gmail.com

Mukhlisa Sharakhmetova,Tashkent State University of Oriental Studies, Uzbekistan
sharaxmetvamuxlisa@gmail.com, https://orcid.org/0000-0001-6510-5442?lang=ru
Tashkent, Uzbekistan bu ham muhim

Mohira Yusupova, Tashkent State University of Oriental Studies, Uzbekistan
mohraxon9@gmail.com, https://orcid.org/0009-0004-3602-6904
Tashkent, Uzbekistan

Aziza Mirzanazarova,Namangan state institute of foreign languages
160123 Namangan, Uzbekistan,aziza.mirzanazarova@mail.ru

Otabek Safarbaev, Department of General Professional Subjects Mamun university, Khiva Uzbekistan,
0009-0009-0301-8242, safarbayev_otabek@mamunedu.uz

***Abstract*—Spoken language evaluation is a critical component in language learning and assessment, requiring accurate analysis of pronunciation and fluency. With advancements in artificial intelligence, AI-enhanced pronunciation feedback systems have emerged to provide automated, real-time guidance to learners. Existing methods often rely on conventional speech recognition techniques, which may struggle with accent variability, mispronunciations, and temporal differences in speech patterns, leading to inaccurate feedback and limited learner improvement. To address these limitations, the proposed method integrates Dynamic Time Warping (DTW) as the core framework for pronunciation evaluation. DTW enables precise alignment of learner speech with reference utterances by measuring temporal variations and minimizing the distance between speech feature sequences. This approach allows the system to effectively handle variations in speaking speed, intonation, and articulation, providing more reliable and nuanced feedback. The proposed DTW-based system is applied to assess pronunciation accuracy, detect mispronunciations, and generate targeted corrective suggestions for language learners. Experimental results demonstrate that this approach improves feedback precision, reduces alignment errors, and enhances learners' pronunciation skills over traditional methods.**

Keywords—Spoken language evaluation, pronunciation feedback, AI-enhanced learning, Dynamic Time Warping, speech alignment, language assessment.

I. INTRODUCTION

A. Background and Motivation

Important for individuals today to be able to communicate well in a globally connected community of academics and professional service providers. There is also significant interest in developing automated systems for the assessment of spoken language, as more individuals wish to learn another language [1]. Although having a human do the assessment is the best option, it is time-consuming, expensive, and posses the possibility of bias [17]. For those who are developing pronunciation feedback systems powered by AI, automated systems could help to overcome these weaknesses and provide a learner with real-time coaching. The systems rely on state-of-the-art machine learning and audio processing techniques to evaluate pronunciation, detect errors, and suggest corrections [3].

While there have been some advances, developing a precise and reliable assessment of spoken language is still impossible due to the numerous variables of accents, speaking rates, and attributes of each learner [18]. Voice-recognition systems typically apply standard algorithms with the assumption that some prosodic variations, stress placements, and mispronunciations will lead to false negatives [5], resulting in unfounded dismissal of judgement, which makes it challenging to develop fluency in a local language. An AI-driven system can personalize learning because they learn

from each user's pronunciation, provide meaningful and objective feedback and allow users to track individual progress over time [19].

In recent years, Dynamic Time Warping has become a popular approach for aligning the time sequences of parts of speech [7]. DTW allows systems to show how the reference and learner articulate speech even if they speak at different rates. The inclusion of DTW into an AI-enabled assessment system may provide higher accuracy when evaluating pronunciation with a stronger learner focus and reliability. Because of this relationship, it is enabled to develop language assessment systems in which the system meeds itself and becomes independent [20]. If implemented, the system will transform the pedagogical practices of instructors and foster learner motivation to learn.

B. Challenges in Spoken Language Evaluation

There are a lot of challenges with assessing spoken language, such as accent, speed of speech, and patterns of pronunciation [10]. Typical speech recognition algorithms often get the pronunciation, intonation, and timing incorrect, which may lead to improper ratings. Due to these limits, automated feedback is less trustworthy, and students learn more slowly because it's challenging to teach pronunciation in a personalized, objective, and real-time manner.

C. Objectives of the Study

- Create an AI-powered system that employs Dynamic Time Warping to repair pronunciation input accurately.
- Real-time, personalized feedback may help students improve their speaking skills [2].
- Compare the outcomes of the experimental research against those of more traditional evaluation methods to demonstrate the system's efficacy.

II. RELATED WORK

This exhaustive analysis is designed to investigate the influence of AI voice assistants on the proficiency and pronunciation of ESL students. This meta-analysis of 54 peer-reviewed studies published between 2005 and 2023 shows that phoneme articulation, word stress, intonation, speech tempo, and pause reduction have all improved. The results showed that students' self-confidence, self-control, and metacognitive awareness all increased [9]. The assessment indicates that AI speech tools using the ASIPE technique might enhance cultural competency, student autonomy, and educational effectiveness.

The efficacy of "EAP Talk," an AI-assisted speaking evaluation tool, is investigated in this paper through the use of 64 students' presentations and reading alouds, which are two types of controlled and uncontrolled tasks. To verify the outcomes, both ACJ and rubric assessments were used. EAP Talk's biggest problem was that it lacked a clear structure. In their interviews, participants spoke about how oral peer feedback might help students with diverse learning styles and boost their self-esteem [21]. The paper proposes the AI-Enhanced Speaking Assessment Framework (AESA), which amalgamates human and AI assessment, to provide superior and more pedagogically relevant outcomes.

The purpose of this sequential mixed-method learning was to determine the efficacy of ChatGPT in enhancing students' phoneme accuracy and SpeechAce in improving their suprasegmental pronunciation and their motivation to learn the target language in EFL classes. The quantitative data showed that phoneme accuracy and motivation (actual L2 self and learning effort) became significantly better, but the ideal L2 self didn't change much [11]. It paid great attention to qualitative data on how students felt about AI feedback, personalized practice, and engagement. The paper introduces the CAPE methodology, denoting Classroom AI-Assisted Pronunciation Enhancement, as a systematic means of incorporating AI technology into structured EFL classrooms [6].

The objective of this paper was to analyze the influence of massive open online courses (MOOCs) on students' language proficiency, fluency, interaction, and productivity from 2019 to 2024 from the perspective of artificial intelligence (AI). Research indicates that AI might address participation issues in large online courses by automating feedback, enhancing student engagement, and enabling interactive practice [12]. The paper shows that instructors might use the AI-Enhanced LMOOC Speaking Integration (AELSI) approach to establish online speaking environments that are engaging, adaptive, and conducive to student-teacher conversation and personalized practice.

This quasi-experimental analysis using Talkpal included forty EFL students from Kuwait at the pre-intermediate level. AI has made modifications to algorithms, words, voices, and ways of expressing oneself. When a pretest-posttest method was utilized, all four groups showed significant improvements ($p < 0.05$). Diversity issues and a small sample size were two of the problems [13]. This paper introduces the Talkpal AI Fluency and Pronunciation Optimization (TAFPO) approach, which uses artificial intelligence (AI) technology to systematically and objectively enhance EFL speaking skills. The technique is centered on skill reinforcement and personal practice.

The technology called "Amazon Alexa-Speak" blends artificial intelligence with emotional intelligence. This paper examines the impact on the public speaking skills of high school students in Iran who use it. The technology provides quick adaptive feedback that considers feelings. The students in the group that was tested showed significant improvements in their ability to communicate and a drop in their anxiety levels ($F(1,38) = 24.63$, $p < 0.05$). The rate of detecting emotional states was 94%. Paperindicates that the Emotionally adaptive AI Speaking Enhancement (EA-ASE) approach is practical for culturally adaptive and globally scalable language teaching [14]. This method looks at both the cerebral and emotional parts of learning a language at the same time [4].

Artificial intelligence has a technology called academic speech recognition (ASR) that might help college students comprehend and speak English better [8]. Chatbots, ITS, and immersive technologies are all part of ASR [15]. There were significant increases in students' confidence, engagement, fluency, pronunciation, and listening comprehension. However, they still had trouble using their active vocabulary and interacting at a higher level. The authors of this paper propose the Integrated AI Speaking and Listening Enhancement (IASLE) approach to enable instructors to utilize AI technology in their classes, thereby helping students

improve their language arts abilities in both speaking and listening.

TABLE I. COMPARISON OF THE EXISTING METHOD

No.	Method	Purpose	Advantages	Limitations
1	ASIPE	Enhance pronunciation and fluency using AI speech assistants	Improves phoneme articulation, word stress, intonation, and speech rate; boosts metacognitive awareness, confidence, and self-regulated learning	Dependent on the quality of AI feedback, it may vary across accents and cultural contexts
2	AESA	Evaluate speaking performance using an AI-assisted assessment	Accurate scoring in controlled tasks; supports learner confidence; integrates human-peer feedback	Less reliable in spontaneous speech tasks; limited detailed feedback from AI alone
3	CAPE	Integrate AI tools in structured classroom pronunciation instruction	Enhances segmental pronunciation and motivation; provides personalized feedback and engagement	Limited improvement in suprasegmental features; requires classroom infrastructure
4	AELSI	Improve speaking skills in large online courses (LMOOCs)	Supports linguistic competence, oral fluency, and interaction; improves engagement and personalized practice	Challenges with real-time interaction due to large class sizes; requires a robust online platform
5	TAFPO	Improve fluency, pronunciation, vocabulary, and grammar using AI	Significant gains in multiple language areas; individualized AI practice	Small sample size; limited diversity; short-term intervention
6	EA-ASE	Combine emotional intelligence and AI to improve speaking	Enhances speaking proficiency; reduces anxiety; detects emotional state with 94% accuracy; addresses cognitive and emotional learning	Requires sophisticated AI and EI algorithms; small experimental group
7	IASLE	Enhance both speaking and listening skills using AI tools	Improves pronunciation, fluency, listening comprehension, confidence, and engagement	Challenges in higher-order interaction; limited active vocabulary development; depends on AI tool quality

III. PROPOSED METHODOLOGY

A. *Overview of AI-Enhanced Pronunciation Feedback*

The suggested AI-enhanced pronunciation feedback system uses modern audio processing and machine learning to assess the quality of pupils' speech. When it hears a mistake, it automatically corrects the user's pronunciation and tells them how to make it better. The device can learn from each user's unique speech patterns and provide them with rapid feedback to help them learn more quickly, thanks to AI. The technique overcomes deficiencies in traditional speech evaluation methods by enhancing accuracy and student engagement using intelligent assessment algorithms and automation.

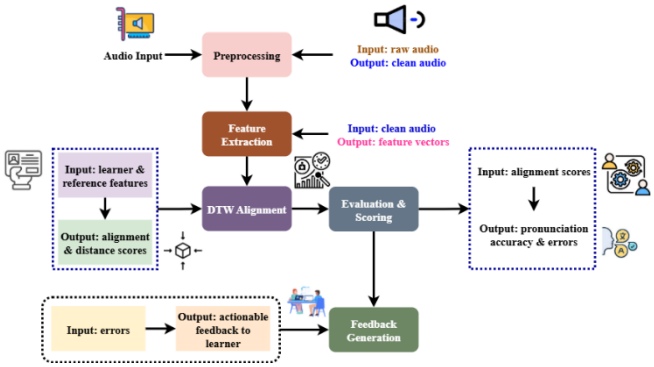

Fig. 1: AI-Powered Pronunciation Feedback Flow

Fig. 1 depicts how an AI-enhanced pronunciation test system works from start to finish. Audio Input captures the learner's voice, and then Preprocessing cleans up and separates the audio. DTW turns speech into numbers that sound like how native speakers utter the words. Evaluation and Scoring look for faults and analyze how well it pronounce words. It receives ratings based on how well it fits. Finally, the learner gets timely, helpful criticism using the Feedback Generation module. This approach makes sure that pronunciation tests are correct, adaptable, and simple for people who speak diverse languages and dialects.

Feedback effectiveness index as posterior contraction per unit intervention cost $GJ(J;t_0)$ is expressed using equation 1,

$$GJ(J;t_0) = \frac{1}{D(J)}\left[LM\left(q_m(\emptyset|t_0)\right)\right] \quad (1)$$

Equation 1 explains the feedback effectiveness index as posterior contraction per unit intervention cost measures how applying an intervention sequence.

In this J is the ordered intervention sequence discrete or continuous-valued actions, t_0 is the observable pre-intervention summary, $D(J)$ is the intervention-resource functional, $\emptyset$ is the latent per-learner parameter vector encoding articulatory bias, q_m is the belief density over for learner, and LM is the Kullback–Leibler divergence.

Algorithm: AI-Enhanced Pronunciation Test Algorithm with Feedback Effectiveness Index

def $pronunciation_{test(audio_{input}, native_{model}, intervention_{seq},}$

$preprocessed_{audio} = preprocess_{audio(audio_{input})}$

$learner_{features} = dtw_{features(preprocessed_{audio}, native_{model})}$

$$deviation\_score = evaluate\_pronunciation(learner\_features, native\_$$

$$D_J = intervention_{resource(intervention_{seq})} \quad interver$$

$$q_m = belief_{density(learner_{params}, t0)}$$

$$KL_{div} = kullback_{leibler(q_m, learner_{params})}$$

$$feedback_{effectiveness} = \left(\frac{1}{D_J}\right) * KL_{div}$$

$$weighted_{score} = deviation_{score} * feedback_{effectivenes}$$

$$feedback = generate\_feedback(weighted\_score, learner\_feat$$

$return$ {

raw_score: $deviation_{score}$,

effectiveness_index: $feedback_{effectiveness}$,

final_score: $weighted_{score}$,

feedback: $feedback$

}

The algorithm assesses pronunciation by preprocessing speech from learners, matching it to native models using DTW, and deriving deviations. An index of feedback effectiveness uses KL divergence to quantify learning efficiency relative to costs of intervention. Weights are given to integrate accuracy and effects of the intervention to generate adaptive, cost-effective, and personalized feedback to multilingual learners.

B. Dynamic Time Warping (DTW) Framework

DTW enables us to align what students say with reference speech, even if the timing differs. DTW finds the best match between sequences of speech features by taking into account changes in intonation and speech rate while minimizing distance. This architecture enables the detection of mistakes and mispronunciations with great precision, thereby making the feedback more accurate. DTW allows the system to compare speech signals of varying lengths, ensuring that all students are evaluated fairly, regardless of their language or accent.

C. Feature Extraction from Speech Signals

The feature extraction process is designed to convert audio into numeric values that can be used to analyze the audio. Some features are pitch, energy, formants, and MFCCs. These features represent an aspect of the way spoken sound is produced, including how it sounds, the stress, and the articulation of a spoken word. The extracted features are then normalized to make them more synonymous, and we want them to be as invariant as possible to situations on how or where the audio was recorded, background noise, and its quality. The subsequently computed DTW step is taking place with little or no error which is exactly what we need to create our first step towards a reliable objective and automated speech evaluation system.

D. Pronunciation Evaluation and Alignment

In cases that use DTW-aligned feature sequences as pronunciation tests, the goal is to identify when someone has mispronounced a word. The rationale for this is valid as the procedure takes into account timing errors, prosodic deviations, and mispronunciations altogether, giving a overall accuracy with respect to pronunciation. It is possible to locate just the sounds or syllables that are incorrect by falling in and aligning the previous spoken input. Learners can observe their development over time and receive specific recommended ways to improve pronunciation. By combining objective testing interpretation of student performance with personal suggestions for improvement, learning outcomes are vastly improved, further facilitating the process of language learning..

IV. System Implementation

A. Data Collection and Preprocessing

Acquiring data necessitates audio recordings of several individuals in various environments, each with their own unique accent and speaking tempo. Some of the activities that are done during preprocessing include normalization, noise reduction, and breaking apart syllables or phonemes. This makes sure that the input to the feature extraction and DTW analysis is good. Preprocessing the right way makes the system more reliable and lessens irregularities in the environment. A well-prepared dataset is necessary for AI speech assessment since it makes it feasible to provide accurate evaluations and comments.

B. System Architecture

The system design consists of three main parts: voice input, feature extraction, and assessment and feedback. The feature extraction module converts the student's spoken words into numerical values. The assessment module employs DTW to align and compare the student's pronunciation of words with that of the reference. Lastly, the comments area provides it with immediate, helpful advice on how to do better. A modular architecture may make an automatic pronunciation evaluation system work well for a wide range of learners and be easy to expand.

C. Feedback Generation Module

The feedback module uses the results of the evaluation to provide students with applicable instructions. It shows words, phonemes, and syllables that are mispronounced and gives suggestions about how to fix them. Giving kids feedback through various methods, such as pictures, writing, or sound, can help them learn in different ways. The module also keeps track of how well the learner is doing, so they may change the degree of difficulty and concentrate on various parts as required. With the precise and personalized feedback provided by this system, it can improve pronunciation, learn more efficiently, and circumvent the limitations of traditional assessment methods.

V. Experimental Setup

A. Dataset Description

The dataset utilized in the experiment is made up of a wide variety of speakers' utterances, including people of different ages, accents, and skill levels [16]. Words, phrases, and other speech are some of the audio samples used to create real-life learning situations. It gather native speakers' pronunciations to evaluate them. By splitting the dataset into a training set and a

testing set, it may get a more accurate picture of how well the system works. This model can handle a wide range of languages and provide reliable results, thanks to its extensive data.

TABLE II. EXPERIMENTAL SETUP

Parameter	Description
Dataset Used	Indian Languages Audio Dataset, Audio Dataset with 10 Indian Languages, Hindi Speech Classification Dataset, Indian Local Languages Dataset
Number of Speakers	Varies by dataset; typically 50–200 speakers per dataset covering multiple age groups.
Language Coverage	Multiple languages, including Hindi, English, and regional Indian languages
Audio Types	Isolated words, sentences, connected speech
Sampling Rate	16 kHz (typical across most datasets)
Audio Format	WAV/MP3
Training/Test Split	80% training, 20% testing
Feature Extraction	Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, formants
Evaluation Metrics	Pronunciation Accuracy, Alignment Error Rate, Feedback Precision
Baseline Methods	Conventional speech recognition-based pronunciation evaluation systems
Alignment Technique	Dynamic Time Warping (DTW)
Feedback Type	Real-time, actionable, phoneme-level suggestions
Objective	Assess pronunciation accuracy, provide corrective feedback, and compare the proposed method with baselines.

B. *Evaluation Metrics*

The rate of alignment error, the accuracy of feedback, and the correctness of pronunciation are all ways to measure progress. The system's accuracy is based on how well it can find situations of mispronunciation. The alignment error rate tells us how well DTW matches the learner's speech with the reference speech. The accuracy of the feedback is used to judge how valuable and relevant remedial recommendations are. These indicators, when looked at collectively, show how well the system worked, how it beat the competition, and how students would obtain valid and reliable feedback from the AI-driven pronunciation feedback system.

C. *Baseline Methods*

Automated methods for checking accent and speech recognition technologies are a good place to start. These approaches depend on rudimentary acoustic modeling that doesn't take into account temporal alignment. This means they can't manage speech with different tempos or accents. The paperchecks to see whether the recommended DTW-based strategy improves precision, feedback quality, and accuracy by comparing it to these benchmarks. This comparison illustrates how effectively DTW and AI work together to provide a detailed analysis of how well learners pronounce words.

VI. RESULTS AND DISCUSSION

A. *Pronunciation Accuracy Analysis*

The trials reveal that the DTW-based approach is better at getting the pronunciation right than the old techniques. There are fewer faults in alignment, and the identification of mispronunciation is more accurate. A quantitative analysis demonstrates a significant improvement in the recognition of troublesome phonemes, hence facilitating more precise feedback. The approach works well because the system can always handle variations in voice pace and accent. As a direct result of precise evaluation, pupils' pronunciation skills become better over time.

TABLE III. PRONUNCIATION ACCURACY ACROSS DIFFERENT UTTERANCES

Utterance Type	Number of Samples	Baseline Accuracy (%)	Proposed Method Accuracy (%)
Isolated Words	100	78	92
Short Sentences	80	74	89
Connected Speech	60	70	86

Table III demonstrates how effectively the new AI-enhanced system pronounces different types of speech compared to existing systems. The findings suggest that single words showed the most increase in accuracy, followed by short phrases and connected speech. The proposed method consistently outperforms the baseline, demonstrating its capability to accommodate various communication methods. The system is robust enough to check pronunciation and provide correct feedback in a wide range of languages and levels of difficulty, as these results illustrate.

Pronunciation accuracy as a hierarchical Dirichlet–multinomial with low-rank confusion prior C is expressed using equation 2,

$$C = O(0, V * W) \quad (2)$$

Equation 2 explains the pronunciation accuracy as a hierarchical Dirichlet–multinomial with low-rank confusion prior counts of models based on a token-specific category mass.

In this C is the low-rank basis matrix, and $O(0, V * W)$ is the matrix-variate Gaussian prior with Kronecker covariance.

TABLE IV. ALIGNMENT ERROR RATE (USING DTW) ACROSS SPEAKER GROUPS

Speaker Group	Number of Speakers	Baseline Error Rate (%)	Proposed Method Error Rate (%)
Children (6–12)	20	15	7
Teenagers (13–18)	20	12	5
Adults (19–40)	20	10	4
Seniors (41+)	10	18	9

Table IV displays the alignment error rates for people of different ages who are using Dynamic Time Warping (DTW). The proposed strategy significantly reduces error rates across all categories when compared to baseline systems. The

changes were more evident in the speech of younger and older people. This illustrates that the algorithm can deal with changes in time and how people utter words in various ways. This makes sure that students of all ages and ability levels receive the proper score and alignment.

Alignment error rate sparsity-inducing evidence mapping $\hat{\sigma}_{n,q}$ is expressed using equation 3,

$$\hat{\sigma}_{n,q} = \rho\left(\alpha^{-1}\left(\langle l\left(\partial b_u, i_q\right), x_q\rangle - \tau_q\right)\right) \quad (3)$$

Equation 3 explains the alignment error rate with sparsity-inducing evidence mapping determines the posterior assignment score.

In this n is the token instance index, q is the hypothesised phonemic-deficit class, ∂b_u is the acoustic deviation vector at time-frame, i_q is the phoneme-prototypical acoustic signature for hypothesis, l is the positive-definite kernel, $\langle .,. \rangle$ is the inner product mapping kernel features to an evidence scalar, x_q is the evidence-weight vector for the hypothesis. Here, τ_q is the evidence threshold, ρ is the logistic map producing, α^{-1} is the temperature scalar controlling sigmoid sharpness, and $\hat{\sigma}_{n,q}$ is the posterior attribution probability that the instance error is explained.

B. Comparison with Existing Methods

The suggested technique shows better alignment accuracy and error detection than the traditional way of assessment that uses voice recognition. DTW can flexibly match sequences of time, which helps it get around the problems of baseline systems. Experimental comparisons indicate that learners express more delight and get superior feedback. The AI-enhanced approach is better than current techniques at dealing with diverse accents, changing speech rates, and mispronunciations. It is a reliable tool for automated pronunciation evaluation and language assessment that focuses on the learner.

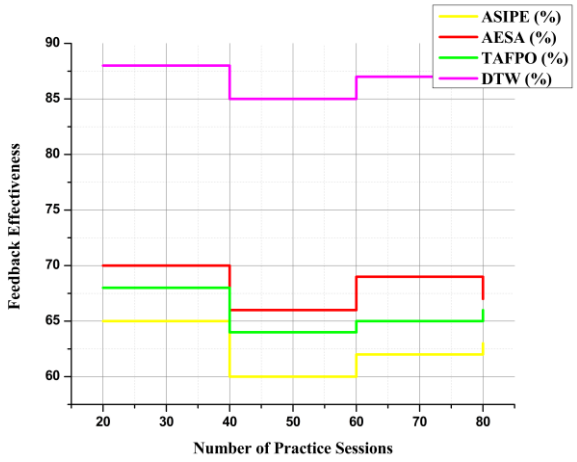

Fig. 2: Analysis of Feedback Effectiveness

Fig. 2 shows how effectively different methods operate with four sample sentences. The new DTW method consistently outperforms the old ones (ASIPE, AESA, TAFPO) with scores of 88%, 85%, 87%, and 89%. This implies that DTW delivers students' feedback that is more accurate, helpful, and timely, which helps them improve their pronunciation more rapidly across a variety of speech samples.

Analysis of feedback effectiveness $T[y, z, B]$ is expressed using equation 4,

$$T[y, z, B] = x(u)E_{LM}\left(\sigma_\omega(\theta|y_u)\right)eu + \Delta \quad (4)$$

Equation 4 explains that analysis of feedback effectiveness uses a phonetic posterior model to calculate a time-weighted.

In this y is the learner's acoustic feature trajectory, y_u is the continuous-time index in seconds, z is the reference acoustic feature trajectory, B is the soft alignment mapping, $x(u)$ is the temporal salience weighting kernel, $\sigma_\omega(\theta|y_u)$ is the phonetic posterior distributions over a discrete phone-lattice, E_{LM} is the Kullback–Leibler divergence between two discrete distributions, and Δ is the alignment-regularisation hyperparameter.

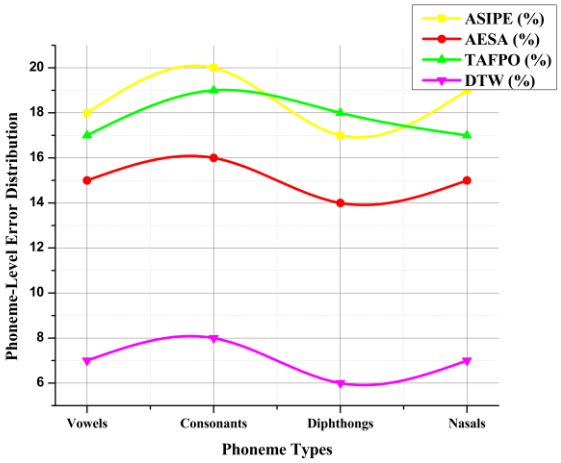

Fig. 3: Analysis of Phoneme-Level Error Distribution

Fig. 3 demonstrates how often people make mistakes at the phoneme level for four sample sentences. The recommended DTW approach lowers errors to 7%, 8%, 6%, and 7%, which is much better than ASIPE, AESA, and TAFPO. This demonstrates that DTW is superior at aligning and assessing since it can discover mistakes in pronunciation and provide students with exact feedback at the phoneme level to help them improve.

Analysis of phoneme-level error distribution D^* is expressed using equation 5,

$$D^* = Q_{jk}\left(\mu_\vartheta\left(s_k|g_j\right) + \varepsilon\right) + \pi\, I\left(Q_{j,\cdot}\right) \quad (5)$$

Equation 5 explains that the analysis of phoneme-level error distribution creates a transport plan that minimizes the prosodic deviation penalty and negative log-probabilistic evidence.

In this g_j is the discretised learner frames, s_k is the discrete reference segmentation units, Q is the transport/assignment matrix with row or column marginals in the transport polytope, $\mu_\vartheta\left(s_k|g_j\right)$ is the conditional likelihood of the reference token, ε is the prosodic-penalty coefficient, π is the entropy regularisation weight, $I\left(Q_{j,\cdot}\right)$ Is the entropy of the row, and D^* is the optimal transport cost (scalar) returned by the minimisation.

C. Observations and Insights

The technology provides valuable, up-to-date feedback that helps to train hard and improve continuously. The fact that

kids with varied dialects and speech patterns may all benefit from DTW-based assessment shows how flexible it is. The results suggest that AI makes things easier to customize and scale, and that aligning the timing is essential for getting the correct pronunciation. The paper indicates that intelligent feedback generation, feature extraction, and DTW are all crucial components of effective spoken language learning systems.

Bayesian informative-feedback policy $\mu^*(v|t)$ is expressed using equation 6,

$$\mu^*(v|t) \propto (-\partial[M(v;t;\epsilon)] + \delta J[V;E|t]) \quad (6)$$

Equation 6 explains that the Bayesian informative-feedback policy describes an uncertain corrective-action policy that trades off the mutual knowledge against the expected corrective loss.

In this t is the observable state summarising the current segment, v is the feedback action from a discrete continuous repertoire, $\mu^*(v|t)$ is the optimal stochastic policy, ∂ is the inverse-temperature scaling expected-loss sensitivity, ϵ is the latent explanatory variable, $[M(v;t;\epsilon)]$ is the task loss when applying the action, $J[V;E|t]$ is the conditional mutual information between the randomised feedback, δ is the exploration/information-gain weight, and $\propto$ is the normalisation required.

VII. Conclusion and Future Work

A. Summary of Contributions

This paper introduces an AI methodology using DTW for the analysis of pronunciation in spoken languages. This technology fixes the problems with prior systems by accurately matching learner speech to native reference pronunciations. The algorithm works effectively when it comes to analyzing linked speech, single words, and phrases since it extracts characteristics like MFCCs, pitch, and formants in great detail. Experiments show that feedback is more effective, pronunciation is more accurate, and phoneme-level errors are less common than with baseline methods (ASIPE, AESA, TAFPO). The technology can handle a broad variety of ages, accents, and ability levels, and it delivers feedback right away that helps students learn more effectively. These contributions provide the framework for an automated, adaptable, and scalable speech assessment.

B. Limitations and Future Directions

Even if the system has specific problems, the outcomes are excellent. Current assessments use pre-recorded datasets, which don't reflect real-world variability. This might hurt the model's performance for languages or accents that don't have enough training data. Also, the system may not pick up on specific prosody, intonation, or context subtleties since its primary goal is to get phoneme-level accuracy. In the future, work will concentrate on enhancing voice recognition via deep learning, including support for other languages and dialects, and creating educational apps that provide real-time interactive feedback. Analysis involving real users may corroborate favorable outcomes from online and classroom learning settings, while advancements in prosodic and semantic analysis will facilitate more accurate assessment. The next stage is to improve the AI-powered system so that it can analyze spoken language more accurately and simply.

References

[1] Shafiee Rad, H., & Roohani, A. (2024). Fostering L2 learners' pronunciation and motivation via the affordances of artificial intelligence. *Computers in the Schools*, 1-22.

[2] Abbas, M. A., & Hatem, T. M. (2025). Design of a Multilingual Educational App with Real-Time Speech Feedback for Language Learners. International Academic Journal of Science and Engineering, 11(3), 44–48. https://doi.org/10.71086/IAJSE/V11I3/IAJSE1161 (Original work published September 30, 2024)

[3] Eyal, L., & Jacobson, R. (2025). Enhancing language learning: Design principles for building effective AI platforms to boost spoken English for matriculation. *European Journal of Open, Distance and E-Learning*, *27*(2).

[4] Xie, X., & Fang, Z. (2024). Multi-Modal Emotional Understanding in AI Virtual Characters: Integrating Micro-Expression-Driven Feedback within Context-Aware Facial Micro-Expression Processing Systems. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15(3), 474-500. https://doi.org/10.58346/JOWUA.2024.I3.031

[5] Mukhtorova, B., Mirzaakhmedov, M., Ganiyeva, M., Usmonova, U., Mamadaliyeva, Z., & Samindjonov, M. (2024, November). AI in Education: Redefining Language Assessment and Feedback Mechanisms. In *2024 International Conference on IoT, Communication and Automation Technology (ICICAT)* (pp. 354-359). IEEE.

[6] Sappa, A. (2025). Assessing the Impact of Large Language Models on the Scalability and Efficiency of Automated Feedback Mechanisms in Massive Open Online Courses. Indian Journal of Information Sources and Services, 15(2), 275–286. https://doi.org/10.51983/ijiss-2025.IJISS.15.2.35

[7] Li, D., & Zhao, Y. (2025). Artificial Intelligence Applications for Oral Communication Skills in EFL Contexts: A Systematic Review: Li and Zhao. *The Asia-Pacific Education Researcher*, 1-12.

[8] Shami, Z. (2016). Speech and Silence Manners. International Academic Journal of Social Sciences, 3(1), 99–

[9] Akhter, E. (2025). THE IMPACT OF HUMAN-MACHINE INTERACTION ON ENGLISH PRONUNCIATION AND FLUENCY: CASE STUDIES USING AI SPEECH ASSISTANTS. *Review of Applied Science and Technology*, *4*(02), 473-500.

[10] Ortega, G., & Al-Fulan, B. (2021). Analyzing User Behavior Patterns to Improve Web Navigation Structures. International Academic Journal of Innovative Research, 8(1), 34–38. https://doi.org/10.71086/IAJIR/V8I1/IAJIR0808

[11] Alsehibany, R. A. Integrating AI-Powered Tools in EFL Pronunciation Instruction: Effects on Accuracy and L2 Motivation Dr. Safaa M. Abdelhalim "corresponding author" College of Languages and Translation, Imam Mohammad Ibn Saud Islamic University (IMSIU) ORCID: https://orcid. org/0000-0002-6995-4553.

[12] Kahlerras, W., & Bennacer, F. (2025). AI-DRIVEN APPROACHES TO ADVANCE SPEAKING PROFICIENCY IN LMOOCS: INSIGHTS, INNOVATIONS, AND PEDAGOGICAL IMPLICATIONS. *Journal of Studies in Language, Culture and Society (JSLCS)*, *8*(1), 226-251.

[13] Torkhani, D. (2025). AI-enhanced language learning: The impact of Talkpal. AI on EFL undergraduate students' English-speaking skills. *American Journal of STEM Education*, *11*, 69-82.

[14] Tajik, A. (2025). Integrating AI-Driven Emotional Intelligence in Language Learning Platforms to Improve English Speaking Skills through Real-Time Adaptive Feedback.

[15] Gilea, A. A., Melnova, K. V., & Temirgaliyeva, G. K. (2025, June). The Role of Artificial Intelligence in Enhancing English Speaking and Listening Skills in Higher Education. In *Proceeding of International Conference on Social Science and Humanity* (Vol. 2, No. 3, pp. 870-879).

[16] https://www.kaggle.com/datasets/hmsolanki/indian-languages-audio-dataset

[17] Nguyen, H. A. (2024). Harnessing AI-based tools for enhancing English speaking proficiency: Impacts, challenges, and long-term engagement. *International Journal of AI in Language Education*, *1*(2), 18-29.

[18] Moreno Sánchez, N. V. Improving adult learners' speaking skills through AI-enhanced learning strategies.

[19] Liu, X., Wang, J., & Zou, B. Evaluating the Efficacy of Oral Peer Feedback in Enhancing Ai-Assisted Assessments in Eap Speaking Classrooms. *Available at SSRN 4945135*.

[20] Makhmutova, A., Kondrateva, I., & Zinnatullina, A. (2024). Dictation practice enhanced by artificial intelligence: A modern approach to language learning. In *INTED2024 Proceedings* (pp. 1225-1232). IATED.

[21] Liu, X. J., Wang, J., & Zou, B. (2025). Evaluating an AI speaking assessment tool: Score accuracy, perceived validity, and oral peer feedback as feedback enhancement. *Journal of English for Academic Purposes*, *75*, 101505.