A Privacy-Conscious Federated Reinforcement Learning Framework for Affect-Aware English Listening

N. Sreedevi[1], Dr. V. Saranya[2], Kama Ramudu[3], Dr. M. Madhusudhan Rao[4], Sakshi Malik[5], Elangovan Muniyandy[6], Ahmed I. Taloba[7]

Research Scholar, Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India [1]
Assistant Professor, Department of English, Panimalar Engineering College, Chennai, India [2]
Associate Professor, Department of ECE, Aditya University, Surampalem, Kakinada, Andhra Pradesh [3]
Associate professor, Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India [4]
Assistant Professor, Jindal Global Business School, O.P. Jindal Global University, Sonipat, Haryana, India [5]
Department of Biosciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai - 602 105, India[6]
Department of Computer Science, College of Computer and Information Sciences, Jouf University, Saudi Arabia [7]
Faculty of Computers and Information-Information System Department, Assiut University, Assiut, Egypt [7]

***Abstract*—The rapid growth of digital English listening platforms has intensified the need for intelligent personalization mechanisms that adapt to learner progression while preserving data privacy. Existing adaptive systems primarily rely on static difficulty scaling or centralized learning architectures, often neglecting learner engagement dynamics and raising concerns about sensitive data exposure. To address these limitations, this study proposes PrivAURAL, a privacy-preserving and affect-aware adaptive English listening framework that models listening instruction as a sequential decision-making problem. The objective is to dynamically personalize listening tasks by jointly considering comprehension performance and engagement trends, without transmitting raw learner data. PrivAURAL integrates HuBERT-based semantic–acoustic representations with affective proxy signals derived from learner behavior and employs a Federated Deep Q-Network to adapt task difficulty, playback speed, and assessment frequency. The model is implemented using PyTorch, HuggingFace speech models, and a simulated federated learning environment with secure aggregation. Experiments conducted on the TED-LIUM dataset demonstrate a 32.7% reduction in Word Error Rate over ten sessions, a 21.9% decrease in task completion time, and an improvement in listening accuracy from 86.1% to 87.3% compared with non–affect-aware baselines. Federated training further ensures stable convergence, while maintaining strict privacy constraints. The results confirm that reinforcement-driven, affect-aware personalization can significantly enhance listening efficiency and engagement, positioning PrivAURAL as a scalable, ethical, and privacy-conscious solution for next-generation digital language learning systems.**

Keywords*—Adaptive listening learning; federated reinforcement learning; affective proxy modeling; privacy-preserving AI; HuBERT speech representation; English language learning

I. Introduction

English proficiency, particularly listening ability, is now an essential skill in the global academic and business environments [1]. Students in the majority of non-native English-speaking nations have difficulty understanding spoken English due to discrepancies in pronunciation, accent, rate, and emphasis [2]. Such challenges are yet intensified in web-based learning environments, where students must interpret audio directions and lectures without immediate support [3]. Traditional audio-centric e-learning systems, while in existence, don't typically take variability in learners' pace, their capacity for understanding, and their progression pattern into account [4]. Consequently, weaker listeners may lag behind, causing learner disengagement and sub-average learning performance [5]. Adaptive learning systems are an efficient solution for bridging the limitations by adapting the learning experience to actual learner performance in real-time [5]. The recent progress in deep learning, specifically in Automatic Speech Recognition (ASR), has provided phenomenal gains in speech processing ability [6]. HuBERT and other pre-trained models have proved a strong capacity for phonetic and semantic content extraction from raw audio, to boost speech recognition and comprehension in a large variety of acoustical conditions [7].

At the same time, reinforcement learning (RL) has been increasingly inserting itself in the education technology domain to build agents that learn to dynamically select and propose personalized content [8]. Most RL educational systems have nevertheless focused on the visual or text modality rather than listening tasks. In addition, hand-crafted feature-based methods do not generalize across different speech conditions and learners [9]. Most educational systems using RL have nonetheless concentrated on the text or visual modality instead of listening tasks. Besides, hand-crafted feature-based approaches tend not to generalize across different learners and speech contexts [10]. These facts highlight the requirement of a method that can represent natural speech robustly as well as modulate task difficulty and speed based on multiple learner-specific feedback signals. The study presents PrivAURAL, an adaptive federated reinforcement learning based and privacy-preserving affect-sensitive English listening framework. In contrast to the current systems, that personalization is performance-only based or the

centralized training method, PrivAURAL combines HuBERT-based speech representations with affective proxy signals to support task sequencing through a federated Deep Q-Network. This single design allows engagement-based personalization and provides a high level of data privacy and scalability of deployment.

A. Problem Statement

Online and audio-based English learning systems are rapidly expanding, yet most lack effective personalization, leaving learners disengaged and mentally strained. The available platforms usually offer fixed or slightly adaptive assignments with a lack of focus on personal understanding, behavioral reaction, or variability of engagement and decreased motivation and learning efficiency [11]. A wide variety of adaptive systems are based on centralized data processing, meaning that sensitive learner data, such as audio and performance logs, are transmitted to off-site servers, which challenges the privacy issue [12]. The conventional approaches with the use of manual acoustic characteristics or performance indicators do not reflect semantic complexity and listening dynamics in reality [13]. As a filler to these loopholes, the current research study suggests a privacy-convincing, affect-sensitive adaptive listening model based on contextual speech representations and federated reinforcement learning to teach securely and personalize.

B. Research Motivation

The main driving force of this research is to provide a privacy-conscious and affect-conscious digital English listening tool that will promote personalized learning without breaching learner data. Engagement and emotional regulation of the learner are important factors in learning comprehension, attention, and retention; however, most of the current systems only use the measures of accuracy and response time [14]. The proposed method combines concepts of affective computing with the ideas of reinforcement learning to model cognitive performance patterns and the dynamics of engagement as the means of adjusting adaptive task sequencing [15]. PrivAURAL framework builds upon HuBERT-based contextual speech representations and behavioral affective proxy signals in a Federated Deep Q-Learning framework, which allows local personalization and retains privacy. This design promotes ethical, scalable, and intelligent listening training in line with the current data protection demands.

C. Significance of the Study

The study introduces a smart language learning system, which is a hybrid of affect-sensitive learner modeling, federated reinforcement learning, and secure aggregation within one adaptive system. Representation of contextual speech and affective proxy signal is used as a guide in ordering tasks in response to learner engagement and understanding dynamics. Privacy is not violated in federated training because raw audio/performance/affective information is not shared. The reinforcement learning agent is capable of dynamically changing the difficulty of the task, and the speed of the playback and the frequency of evaluation, which is less cognitively demanding, motivating, and increases the efficiency of listening in scalable, ethical, and privacy-conscious e-learning.

D. Key Contributions

- Proposes PrivAURAL, a privacy-preserving adaptive English listening framework that models listening instruction as a sequential decision-making process using federated reinforcement learning.
- Introduces affective proxy–aware learner state modeling that integrates comprehension performance and engagement trends without explicit emotion recognition.
- Employs HuBERT-based semantic–acoustic representations to capture contextual listening complexity for informed task adaptation.
- Develops a Federated Deep Q-Network that enables collaborative policy learning across distributed learners without sharing raw audio or behavioral data.
- Demonstrates empirical improvements in listening accuracy, task efficiency, and engagement stability through extensive evaluation and ablation studies under strict privacy constraints.

The rest of the section is structured as follows: Section II provides a literature review of the work, Section III outlines the proposed methodology, Section IV provides experimental results and validation, and discusses results, and Section V provides conclusions and future directions.

II. Literature Review

Ahmed and Hasegawa [16] address the lack of specialized platforms for information and instructional technology students to create online education talking books without the complexity of Web programming. One of the core issues envisioned is the absence of simply accessible tools that integrate pedagogical design with technical simplicity for visual and hearing-impaired students. In response, the research propounds an easy-to-use, web-based platform specially designed for creating educational talking books. The approach involved expert evaluation by fourteen instructional technology professionals in a mixed-method design via an online survey. Results showed that the platform was effective, simple to use, and met the educational needs of the target audience. Future possible enhancements could be enhancing customization levels and adaptive support for different learner profiles.

Valledor et al. [17] analyze the approaches to English as a second language teaching through computer applications and attempt to analyze how these compare to individualized, learner-centered strategies. A key challenge that has been realized is that Audio-Lingual methods are dominant within current applications, which is constraining educational adaptability. Via mixed-method review involving systematic literature search and elicit.org searches, the study reviews various online ESL teaching instruments. Results show that Blended Learning is the most suitable strategy, with an amalgamation of traditional and digital teaching benefits. The tools available are neither adaptive nor conducive to independent learning environments. AI-driven applications based on ASR, TTS, NLU, and DM are proposed to develop digital replicas of teacher-like interaction and enhance ESL instruction tailored to learners.

Hu et al. [18] addresses the challenge of learning focus identification using a combination of cognitive, affective, and behavioral traits, particularly in Virtual Reality (VR) environments. A key challenge discovered is that existing recognition methods lack multimodal fusion of data, decreasing accuracy and contextual information. To reverse the argument, the authors bring in a multimodal feature integration approach combining interaction data (e.g., clickstream, text, and test response) with vision-based inputs (e.g., facial expressions, pupil size, and eye gaze). The performance of the model is tested to perform better in the detection of concentration levels than single-modality methods. Experimental results also show that higher concentration is linked with superior learning outcomes and is heavily influenced by learners' sense of immersion. Subsequent studies can refine modality fusion and immersion measures to further support focus detection in immersive learning contexts.

Hong et al. [19] address the issue that students face when they miss unstructured and lengthy lecture audio while studying online or offline. Traditional skip controls are limited because they operate on the time level and lack semantic awareness, preventing the identification of the current context and position in audio streams. To address this, the authors design HearIt, a system that offers semantic-level skip control through paragraph-based segmentation and auditory feedback. The method uses a combination of positional and topical cues to improve context comprehension without visualizations. A pilot study using an operational prototype was conducted to assess usability and effectiveness. Results indicated that HearIt improves the efficiency and simplicity of browsing auditory information. While promising, additional research is suggested to further refine the design and ascertain its effectiveness in diverse learning environments.

Chaturvedi, Noel, and Satapathy [20] explore sentiment extraction from audio in social media videos on platforms like YouTube and TikTok, especially where there is no language translation involved, like Spanish. The major challenge addressed here is correctly labeling sentiment in noisy environments and in various accents. To correct for this, the authors introduce a novel algorithm that employs a vector space of affective concepts, noting that prefixes to words like "con" or "ab" usually point toward negative sentiments. Unlike typical models based on generic pretrained features, this approach allows for better learning of speech and emotion patterns by neurons. What is new is the use of a novel eigenvalue-based metric to select optimal data augmentations for making the model stronger. The method shows 10–20% improvement over baselines in emotion recognition from YouTube videos.

Chen [21] suggested an English-spoken online dialogue system to overcome the restrictions of static and rule-governed conversation systems that are not responsive to learner interactions. The study is conducted to enhance spoken English skills through adaptive and real-time feedback in online contexts. A reinforcement learning algorithm is used to optimize dialogue strategies, where policy learning controls the system to choose contextually suited responses. The research utilizes the DailyDialog dataset, which offers diverse conversational contexts suitable for oral language learning. Key challenges include handling ambiguous learner inputs, maintaining the natural flow of dialogue, and ensuring personalization of responses. To overcome these issues, the proposed system integrates reinforcement learning with speech recognition, enabling adaptive and context-aware interaction. Evaluation results demonstrated a success rate of 89.6%, with significant improvements in dialogue coherence, as validated through BLEU scores and mean reward metrics. These findings highlight the effectiveness of reinforcement learning in advancing spoken English training systems beyond static, rule-based approaches. Table I shows the summary on literature review.

TABLE. I. SUMMARY OF EXISTING STUDIES

Author	Method	Advantages	Limitations
Ahmed & Hasegawa [16]	Web-based educational talking book platform	Simple and accessible audio-learning content creation	No adaptive personalization or affect-aware learning
Valledor et al. [17]	Blended learning synthesis for ESL applications	Improves learner engagement through mixed instruction	Lacks AI-driven adaptation and individualized learning
Hu et al. [18]	Multimodal affect recognition using VR-based features	Accurate detection of learner concentration and engagement	Requires intrusive sensing and centralized data processing
Hong et al. [19]	Semantic audio segmentation and auditory navigation	Enhances efficiency of lecture audio browsing	No learning adaptation or personalization
Chaturvedi et al. [20]	Audio-based sentiment analysis with affective concepts	Robust emotion detection in noisy speech environments	Not designed for educational adaptation or personalization
Chen [21]	Reinforcement learning–based spoken dialogue system	Enables adaptive and context-aware language interaction	Centralized training; no privacy-aware deployment

Recent advances in educational technology have investigated reinforcement learning as an adaptive content sequence method, affect-conscious modeling as an engagement control method, and federated learning as a privacy protection method; these directions are mostly explored separately. The current reinforcement learning-based learning systems are mainly text-based tutoring or dialogue management and are also based on centralized training pipelines, which is why they are not applicable to the privacy-sensitive listening context. Such approaches that care about affect are also prone to rely on overt emotion detection or multimodal sense, creating ethical issues and limits to deployment. Federated learning research in education focuses mainly on the issues of model scalability and data security, but does not include ways of adaptive choice to affective tasks and sequential decision-making. Conversely, the suggested PrivAURAL system combines federated reinforcement learning and adaptive English listening by using affective proxy-aware learner modeling and contextual speech representations, to ensure the provision of privacy-preserving, engagement-aware adaptive English listening. This has been

combined to portray missing points in the previous literature as critical because it facilitates the personalized progression of listening, emotional control, and decentralized learning without exposing raw learner information.

III. Proposed Methodology: Feature Extraction and Adaptive Learning with DQN

The proposed PrivAURAL framework is a privacy-saving, adaptive English listening framework that adapts the difficulty of the tasks and pacing dynamically using reinforcement learning, but under decentralized learning conditions. The system works with the authentic listening materials that it takes as part of the TED-LIUM corpus and frames the interaction between the learners as a problem of successive decisions. Preprocessing of incoming audio is first taken through standardization and then altered to high-level semantic acoustic representations with a pretrained HuBERT encoder that allows robust modelling of the listening complexity in different accents and speech states. Simultaneously, the learner's responsiveness is estimated using affective proxy indicators based on the patterns of behavioral and temporal response, but not the explicit recognition of emotions. The representations are combined into an organized learner state that would reflect the comprehension performance and engagement tendencies. A Deep Q-Network is locally trained on all of the learner machines to decide adaptive actions, such as difficulty adjustments and playback ones, based on observed transitions in the state and observed reward feedback. In order to facilitate joint enhancement without invading privacy, model parameters are aggregated periodically by use of a federated learning process. It is an end-to-end design that guarantees constant personalization, learner scalability, and close guardedness of sensitive learner data. Block Diagram of the proposed PrivAURAL is shown in Fig. 1.

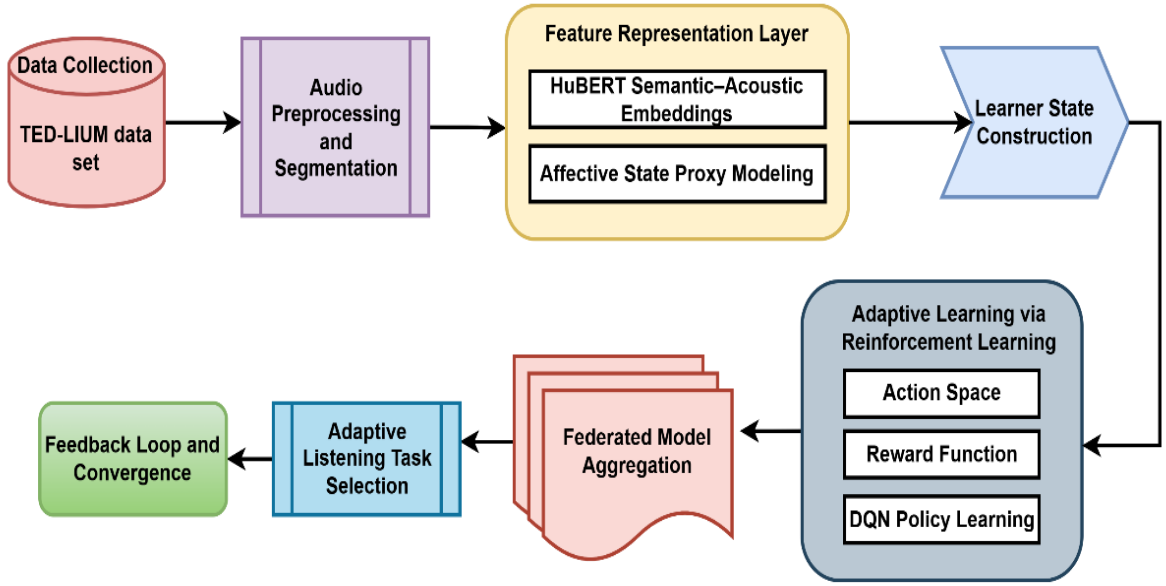

Fig. 1. Block diagram of the proposed PrivAURAL.

A. Data Collection

The TED-LIUM dataset [22], which contains more than 450 hours of English speech given by TED talks with diverse speakers, accented to diverse contexts of the world, was maintained in transcriptions and sampled at 16kHz. TDE-LIUM consists of approximately 118 hours of speech. To enable a generalization to diverse profiles of learning, we considered TED-LIUM particularly appropriate, as it captures variations in accent and speaker factors such as speech rate, intonation and recording conditions, compared to the single-speaker recordings of the LJ Speech corpus. In addition, each audio clip has an aligned transcription, which also supports, through certainty, measurements of benchmarked Word Error Rate (WER) and comprehension-based evaluations of and with the audio clip. The variations in context and speaker performance also expose learners to realistic contexts of variety, noise and accent, which more accurately inform the training of reinforcement learning agents.

B. Audio Preprocessing

- All TDE-LIUM audio samples underwent a comprehensive preprocessing pipeline to ensure they are all in uniform, robust form and demonstrated, or suitable for hybrid feature extraction. Given the variation in record and speaker accent and conditions within TED-LIUM, the natural hearing or listening variability needed to be preprocessed to improve the homogeneity of conditions across the training data. Fig. 2 illustrates the audio-preprocessing pipeline.

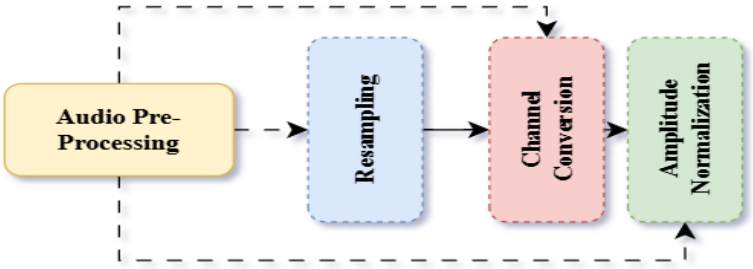

Fig. 2. Audio-preprocessing pipeline.

1) Resampling: All audio is resampled to 16 kHz mono WAV format, which is standard in speech recognition tasks. This preserves the critical 0–8 kHz frequency band while ensuring computational efficiency. The original audio signal is a discrete sequence represented in Eq. (1):

$$x = [x_1, x_2, \ldots . x_T],\ x_i \in \mathbb{R} \quad (1)$$

- where, X is the original sampled waveform, T s the number of samples, x_i is the amplitude at time step i. The original sampling rate is f_s^{orig} and the target rate is f_s =16000 Hz, resampling is mathematically defined by interpolation in Eq. (2):

$$x^{resampled} = Resample(x, f_s^{orig}, f_s) \quad (2)$$

- where, $Resample(\cdot)$ is usually performed with sophisticated sinc-based resampling filters and can be done with either linear interpolators, spline interpolators or sinc filters which provide accurate transformation.

2) Channel conversion: Channels in an audio clip can be more than one (e.g., stereo — left and right). Mono channel audio is good enough for speech recognition and similar embedding models. The stereo audio be represented in Eq. (3):

$$x(t) = [x_L(t), x_R(t)] \quad (3)$$

where, $x_L(t), x_R(t)$ are the left and right channel amplitudes at time t. The mono signal is obtained in Eq. (4):

$$x_{mono}(t) = \frac{1}{2}(x_L(t), +x_R(t)) \quad (4)$$

where, x_{mono} is mono signal. It reduces the dimensionality of the audio without losing meaningful speech information.

3) Amplitude normalization: Amplitude normalization is performed to normalize the loudness across different recordings and reduce variability as a result of the recording context. Given a waveform x(t), normalization scales the signal to a predefined dynamic range (typically [-1,1]), as shown in Eq. (5):

$$x_{norm}(t) = \frac{x(t)}{\max(|x(t)|)+\mathcal{E}} \quad (5)$$

where, $\mathcal{E}$ is a small constant, $\max(|x(t)|)$ is the normal maximum amplitude of the signal. This prevents signal clipping or less of the amplitude range training.

4) Data augmentation: To improve the level of robustness and to resemble the situation in real-life, pitch shifting and tempo variation was done. These changes enhanced the difference in the dataset that would enable the model to do better in accent and speech rate. Such a pipeline will make sure every input will be well structured and balanced in the process and prepared to the next stage of hybrid feature extraction through MFCCs and HuBERT embeddings.

C. Feature Representation Layer

The Feature Representation Layer maps the trivial but informative state of each listening audio fragment to a compact but informative representation that is both linguistically challenging and trended on the learner affects and therefore is important in adaptive task sequencing in the proposed PrivAURAL framework. In each listening task, the raw audio signal x_t is processed by a pretrained HuBERT encoder which processes the waveform directly with no feature engineering. HuBERT projects the temporal acoustic input into a high-level contextual embedding $h_t = \text{HuBERT}(x_t)$, which h_t represents phonetic transitions, speaking rate, pronouncing variability and semantic coherence that exist within the speech fragment. Within the frames of the current study, greater variance and entropy in h_t correlate with a greater listening difficulty, which makes such a representation appropriate in the context of modeling comprehension load across different accents and speech conditions in TED-LIUM.

The system uses the affective state proxy modeling to supplement linguistic modeling, without making claims of explicit emotion recognition. Affective response a tis, which is a normalized composite signal based on prosodic instability in response to learners, latency of response to checked comprehension tasks, and trend of error in successive tasks. The aggregation and scaling of these signals into a restricted affective proxy score $a_t \in [-1,1]$, such that a_t negative signals increasing frustration, a_t positive signals stable engagement are calculated. This proxy is not the emotion of psychological nature but a control signal that allows the reinforcement learning agent to control the difficulty and pacing of the task. The end result of this layer is an integrated representation $\{h_t, a_t\}$, which directly drives the adaptive decision process so that tasks orderings are guided by both speech complexity and learner responsiveness in a privacy-sensitive learning process.

D. Learner State Construction

After the stage of feature representation, the PrivAURAL frameworks model adaptive listening as a Markov Decision Process where each interaction between learners is modeled as a structured state-vector. The learner state at time step t is defined, as in Eq. (6):

$$s_t = \{\text{WER}_t, \text{QuizScore}_t, \text{ResponseTime}_t, a_t\} \quad (6)$$

The Word Error Rate WER_t is computed by aligning the learner's transcribed response with the reference TED-LIUM transcript, reflecting phonetic decoding accuracy under diverse accents and speech conditions. The score of comprehension, denoted as the score of the QuizScore_t, is taken to represent the task-specific listening questions, and it is scaled to reflect the meaning of the understanding without the influence of the task length. ResponseTime_t is the interval between the completion of the audio and the learner responding, which is a measure of cognitive load, with higher values indicating more processing difficulty. The affective proxy a_t, calculated as a response latency trend, prosodic variation, and error progression, is the measure of engagement or frustration within the limits of $[-1,1]$ range (to ensure learning stability). The resulting condition s_t brings together cognitive functionality and affective sensitivity in a single representation so that the reinforcement learning agent can enable context-sensitive adaptation decisions to the task so as to maximize the listening progression with emotional balance and privacy.

E. Adaptive Learning via Reinforcement Learning

Adaptive personalization in the PrivAURAL framework is PrivAURAL framework uses reinforcement learning to obtain adaptive personalization, in which the adjustment of listening tasks is modeled and defined as a sequence decision-making problem. In the context of every interaction step t, the agent chooses an act a t that manages the presentation of the next listening task to the learner. Action space can be stated as a joint set of the difficulty of the task and the speed of playback:

$$a_t \in \{\text{Easy,Medium,Hard}\} \times \{\text{Speed}\uparrow, \text{Speed}\downarrow\} \quad (7)$$

In Eq. (7), the difficulty levels are associated with the differences in the complexity of lexical, sentence length, and speech rate based on the TED-LIUM segments, whereas the demands of temporal processing are controlled with the playback speed changes. This type of action formulation enables the agent to adjust the intensity of challenges at a discrete level without presenting the learner with rapid cognitive change that might break the program.

The performance of every action is measured by a composite reward function balancing an improvement in comprehension and affective stability:

$$r_t = \lambda_1 \Delta\text{WER}_t + \lambda_2 \Delta\text{QuizScore}_t + \lambda_3 \Delta a_t \quad (8)$$

In Eq. (8), ΔWER_t captures the change in listening accuracy between successive tasks, $\Delta\text{QuizScore}_t$ reflects gains in semantic understanding, and Δa_t represents the variation in the affective proxy, ensuring that emotional engagement is preserved. The weighting coefficients $\lambda_1, \lambda_2, \lambda_3$ are constrained such that their sum equals one, allowing controlled emphasis on accuracy, comprehension, and affect regulation depending on pedagogical priorities. This rewarding method eliminates excessive oversight to the accuracy of the learner at the cost of frustration.

Policy learning is implemented using a Deep Q-Network, which approximates the action-value function $Q(s_t, a_t)$ that estimates the expected long-term reward of executing action a_t in state s_t. The network parameters are updated using the Bellman optimality principle:

$$Q(s_t, a_t) \leftarrow r_t + \gamma \max_a Q(s_{t+1}, a) \quad (9)$$

In Eq. (9), γ denotes the discount factor controlling the influence of future rewards. The DQN is able to reach an adaptive policy through repeated cycles of interaction to dynamically select listening tasks that follow learning proficiency progression and emotional balance in order to achieve long-term and individualized learning skills.

F. Local On-Device Training

Adaptive Learning in the PrivAURAL model is conducted by the local training process on the device to make sure there is a personalization and privacy of the data. To every learner, the Deep Q-Network is realized and trained on the learner-side with local observed state-action-reward transitions that are produced during listening tasks. The representation of state, rewards and the changes in policies are calculated without any raw audio recording, the inferenced affective signal, or performance feedback served to any third-party server. The learned model parameters are only stored locally to be used in decision-making during task adaptation. This design enables the DQN to learn the speech rate or difficulty escalation sensitivity of individual listening patterns and to make future task sequencing based on those. The framework also removes the risks linked with centralized data storage by ensuring that all interactions between learners are limited by the device, and effective personalization can be maintained in an environment of strict privacy preservation requirements in continuous and learner-specific devices.

G. Federated Model Aggregation

The PrivAURAL framework will assume a federated model aggregation approach to facilitate collaborative adaptation of adaptive listening policies and be strict in guaranteeing the privacy of users. Following a specified training step locally, each learner device forwards its new DQN model parameters, which are represented as θ_i, to a central aggregation server. No raw audio files, comprehension answers, affective signals, or learner data are sent, and as such, sensitive data is only stored on the local machine. The server then calculates an overall model by doing a weighted average summation of the parameters it gets:

$$\theta_g = \sum_{i=1}^{N} \frac{n_i}{N} \theta_i \quad (10)$$

In Eq. (10), N represents the total number of participating learners, and n_i denotes the relative contribution of learner i, proportional to the number of local interactions or training samples observed during the aggregation round. Such a weighted formulation helps to keep off the dominance of

sparsely trained models and stabilizes convergence of the global policy. Aggregated parameters, denoted as θ_g, were then reallocated to all the learners, and the following cycle of local adaptation was initiated. Secure aggregation or encryption can be provided as an option to enhance the protection of privacy, but the main defense is the lack of data exchange. This federated process enables PrivAURAL to enjoy collective learning patterns among different learners, retain personalization, and avoid exposing individual listening behavior.

H. Adaptive Listening Task Selection

Once the federated aggregation has been completed, the new reinforcement learning policy is applied to choose real-time listening tasks to be executed by the individual learners. In each interaction step, the agent compares the state of the learner and chooses the next task configuration with a maximum estimated long-term reward. This choice directly determines the level of difficulty, the playback rate and the intensity of assessment of the next listening section. The policy increases task complexity over time as the learner becomes more proficient with respect to vocabulary, speech rate or length of utterances, without making sudden levels of difficulty that might lead to cognitive overload. On the other hand, the policy increases or decreases the task requirements in response to the indications of diminishing understanding or interest to maintain the learning balance. PrivAURAL, with this active dynamic adjustment process, automatically adjusts the difficulty of the tasks in accordance with the progression of the learners, so that the activities involved in listening remain both challenging but attainable. Such a balance is maintained and encourages continuous growth in skills, as well as the support of stable learning paths that are stable and emotional.

I. Feedback Loop and Convergence

PrivAURAL framework is a closed-loop system of learning that allows for maintaining adaptability and continuous policy improvement throughout time. Every iteration starts with the introduction of an adaptive listening task chosen by the existing policy, and the answer of the learner is the spoken or comprehension-based one. This response elicits quantifiable feedback indicators such as the accuracy of listening, the end outcome of comprehension, response time, and affective proxy changes. Such signals are directly added to the learner state representation to calculate the reward of reinforcement that is a reflection of performance change and emotional stability. The Deep Q-Network subsequently proceeds to update its action-value estimations with the help of this feedback, which enables the policy to change its future task selections in accordance to the changing behavior of the learner.

With the progress of learning, convergence indicators are monitored by the system to identify policy stability. When there is a little variation in the Word Error Rate across the consecutive sessions, then convergence is observed, showing the stability of the listening accuracy under similar levels of difficulty. At the same time, the reward cumulative curve starts to flatten, which means that the agent has found an efficient task challenge/learner ability balance. Moreover, a decrease in variance in the affective proxy signal is an indicator of stabilized engagement as well as the lack of frustration during interactions. All these criteria are used together to be sure that adaptation is not swamping and adapting to the immediate responses of learners. PrivAURAL successfully develops efficient and stable learning trajectories through convergence in line with cognitive and affective consistency to facilitate the long-term development of listening skills and ensure emotional balance and privacy protection.

Algorithm 1: PrivAURAL – Privacy-Preserving Adaptive Listening via Federated Reinforcement Learning

```
Input:
   Audio tasks from TED-LIUM
   Initial global DQN parameters θ_g
   Learning rate α, discount factor γ
   Reward weights λ1, λ2, λ3
   Number of learners N
   Local training episodes E
Output:
   Adaptive listening policy π*
Initialize global Q-network parameters θ_g
Broadcast θ_g to all learner devices
Repeat for each federated round:
   In parallel, for each learner i:
      Load local listening tasks
      Initialize local Q-network θ_i ← θ_g
      Initialize experience buffer B_i
      Repeat for each episode:
         Receive current listening task
         Preprocess audio and extract HuBERT embedding h_t
         Estimate affective proxy a_t
         Construct learner state s_t = {WER_t, QuizScore_t,
ResponseTime_t, a_t}
         While task interaction is active:
            Select action a_t using ε-greedy policy from Q(s_t, ·; θ_i)
            Apply adaptive action (difficulty, speed, quiz frequency)
            Observe learner response and compute WER_t+1,
QuizScore_t+1, ResponseTime_t+1
            Update affective proxy a_t+1
            Compute reward r_t = λ1ΔWER + λ2ΔQuizScore + λ3Δa
            Store transition (s_t, a_t, r_t, s_t+1) in B_i
            Update Q-network θ_i using mini-batch gradient descent
            Update state s_t ← s_t+1
      Compute model update Δθ_i = θ_i − θ_g
      Send Δθ_i to federated server
   Aggregate updates to obtain global model:
      θ_g ← Σ (n_i / N) · θ_i
   Broadcast updated θ_g to all learners
Until convergence criteria satisfied
Return final adaptive policy:
   π*(s) = argmax_a Q(s, a; θ_g)
```

Algorithm 1 applies a privacy-preserving adaptive listening model based on federated reinforcement learning. The local training of a Deep Q -Networks by each learner on semantic-acoustic features and affective proxy feedback means that task difficulty and pacing are altered. Instead of using the learner data, model parameters are periodically aggregated across devices, so that collaborative policy refinements are made and personalized adaptation and strong protection of sensitive learner information are guaranteed.

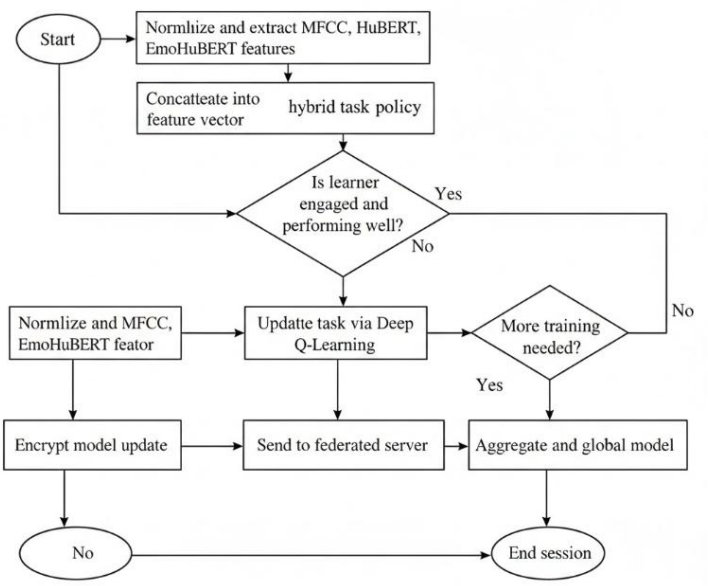

Fig. 3. Overall workflow of the study.

The novelty of this study is in its methodological approach that allows unified fusion of affect-conscious adaptive reinforcement learning with federated training of English listening personalization without being based on centralized data on learners or overt emotion recognition. PrivAURAL is a listening system that uses affective proxy signals to balance the cognitive load and engagement applied to tasks in sequencing, unlike other existing systems that change depending on accuracy or fixed levels of proficiency. Moreover, whereas the previous methods of reinforcement learning in education are centralized or performance-based, the framework is provided to allow decentralized policy learning with safe model aggregation, which does not violate privacy but allows gaining the advantage of collective adaptation patterns. The general workflow of the study is presented in Fig. 3. Such a mixture creates a scalable and ethically sound methodology of intelligent listening systems that allow personalization, emotional stability, and data privacy.

IV. Results and Discussion

This section presents a comprehensive evaluation of the proposed PrivAURAL framework, focusing on its effectiveness in adaptive listening personalization, affect-aware policy learning, and privacy-preserving federated training. The experimental analysis examines learning accuracy, task efficiency, engagement trends, and reinforcement learning convergence across multiple sessions. Performance is assessed using Word Error Rate, task completion time, cumulative reward progression, and personalization behavior, and is compared against centralized and non–affect-aware baseline models. In addition, ablation studies and federated training analyses are conducted to isolate the contribution of reinforcement learning and affective proxy integration, demonstrating the stability and scalability of the proposed approach under decentralized learning constraints. Table II shows the hyperparameters and training configurations.

TABLE. II. Hyperparameters and Training Configuration

Parameter	Value	Description
Learning Rate	3e−4	Step size for DQN updates
Batch Size	64	Number of samples per training batch
Discount Factor (γ)	0.99	Weight of future rewards
Federated Rounds	50	Communication cycles between nodes
Embedding Dimension	768	HuBERT + affective proxy features
Maximum Timesteps	500,000	Total agent–environment interactions
Encryption Scheme	Paillier HE	Privacy protection for model updates

A. WER Reduction Across Sessions

Fig. 4 shows the downward trend of WER throughout the sessions. WER in Session 1 was 27.9, and it decreased to 26.4 after adaptation. By the end of Session 5, the WER decreased to 18.9% (20.6% improvement), and by the end of Session 10, to 14.6% (32.7% improvement). The stable declining trend denotes proper policy streamlining by the DQN, which is motivated by the feedback of the reward in the form of emotion and comprehension rates. The findings confirm the hypothesis that the individual sequence of content taught by the system produces quicker acquisition of listening skills than traditional techniques.

Fig. 5 demonstrates the time spent engaged between sessions. The time spent by learners in a session was initially around 13 minutes, which decreased to 8 minutes at a point of Session 10. The decreasing curve represents a more effective performance in the tasks without any understanding loss- the students accomplished the listening tasks with reduced time and at high feelings. Emotion-aware pacing enabled the system to balance engagement stability and cognitive effort.

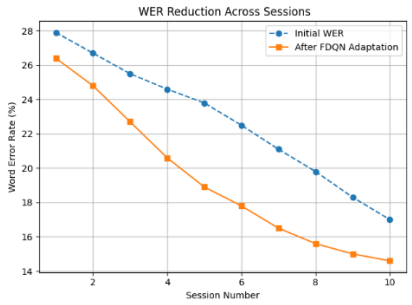

Fig. 4. WER progressions over sessions.

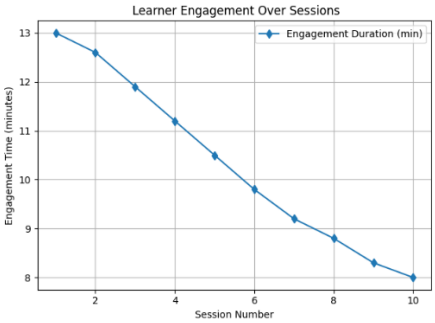

Fig. 5. Learner engagement.

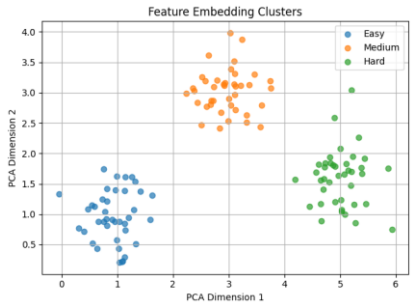

Fig. 6. Feature embedding clusters.

In Fig. 6, Principal Component Analysis (PCA) visualizations indicate that the hybrid embedding space contains very clear clusters of easy, medium, and hard tasks. This isolation exhibits that the MFCC + HuBERT + EmoHuBERT feature combination is a powerful solution to distinguish the changes in linguistic complexity and tone of voice by enabling the DQN agent to allocate the duties to the learners based on their capabilities.

TABLE. III. PERFORMANCE OVER WER

Session	Avg WER (Initial)	Avg WER (After DQN)	Improvement (%)
1	27.9%	26.4%	5.4%
5	23.8%	18.9%	20.6%
10	21.7%	14.6%	32.7%

Table III demonstrates that PrivAURAL increasingly decreased the WER of learners in ten sessions. In the first instance, there was only a slight improvement in the average comprehension, but through federation of adaptation, WER declined steadily- between 27.9 and 26.4 in Session 1 (5.4% change). WER was lowered to 14.6 by Session 10, which is a reduction of 32.7%. These results attest to the fact that emotions-sensitive DQN learns to adapt task sequencing to the progress of individual learners in a dynamic manner, which leads to an increase in listening comprehension. All the findings were statistically significant ($p < 0.05$).

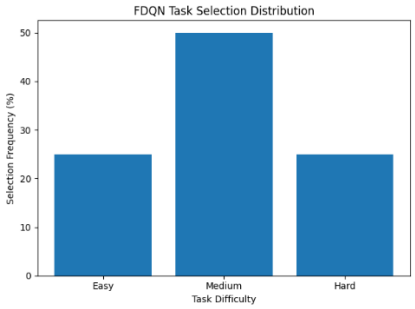

Fig. 7. Distribution of DQN task choices.

Fig. 7 represents the distribution of the task difficulties that the DQN agent chose. Most of the tasks with the middle-level of difficulty were selected, with easy and hard tasks coming next. This dynamic equilibrium helps avoid getting bored and frustrated among learners and instead achieves a progressive learning curve with emotionally maximized engagement.

B. Federated Training Performance

Fig. 8 shows the training of the federated Deep Q-Network agent. The cumulative reward is progressively growing, and it becomes stable at 240 following 400,000-time steps, which demonstrates effective convergence of the federated policy. This tendency proves that cooperative learning among distributed learners can provide a high level of optimization of policies and preserve privacy on the basis of the federated structure.

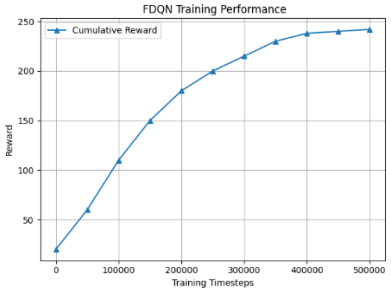

Fig. 8. Federated training performance.

C. Task Completion Time Improvement

Table IV shows the decrease in the time spent on completing tasks in PrivAURAL. By the 10th session, the learners were solving listening tasks 21.9% faster than in the instance of the baseline system. Emotion-sensitive adaptation assists in balancing cognitive load and being motivated so that a learner can advance effectively with specialized challenges.

TABLE. IV. TASK COMPLETION OVER TIME

Session	Baseline System	PrivAURAL System	Time Saved (%)
1	9.6 sec	9.1 sec	5.2%
5	8.8 sec	7.6 sec	13.6%
10	8.2 sec	6.4 sec	21.9%

D. Personalization Efficiency

According to the heatmap in Fig. 9, the first sessions were based on simple tasks, and as the skills of learners improved, the tasks became either medium or hard. This adaptive scaling is also an indication that PrivAURAL has an internal policy of reinforcing over time and balancing challenges and confidence with lower lexical complexity, which confirms their presence as major features to adaptive decision-making.

Fig. 9. Task personalization heatmap.

E. Ablation Study (Without vs. With DQN)

The ablation experiment in Table V indicates the presence of definite performance enhancements when DQN is incorporated. The federated RL agent improved the performance of the baseline by 7.1% in WER and 1.8 seconds in task completion. This confirms the relevance of reinforcement-based personalization of enhancing understanding and effectiveness without compromising privacy.

TABLE. V. ABLATION STUDY

Metric	Without DQN	With DQN	Improvement
Avg WER (Final)	21.7%	14.6%	+7.1%
Task Completion Time	8.2 sec	6.4 sec	−1.8 sec

F. Performance Comparison

Table VI compares three systems in terms of WER, accuracy and task-time reduction. HuBERT base model delivers the best WER with good self-supervised representations but mediocre efficiency improvements. Transformer ASR model is more accurate but does not adapt to emotions. Compared with other, PrivAURAL performs better in terms of overall learning efficiency with more accuracy and largest decreasing task time, proving the benefit of a combination of reinforcement learning, emotional awareness, and privacy-preserving federated updates. HuBERT Base reports lower raw WER due to offline ASR optimization, whereas PrivAURAL prioritizes adaptive learning efficiency and engagement-aware task sequencing rather than standalone recognition accuracy.

TABLE. VI. PERFORMANCE COMPARISON

Method	WER (%) ↓	Accuracy (%) ↑	Task Time Reduction (%)
HuBERT Base (Self-Supervised Speech Model) [23]	10.1	84.7	14.3
Transformer ASR (No Emotion Modeling) [24]	16.7	86.1	17.0
Proposed PrivAURAL (DQN + Emotion + Privacy)	14.6	87.3	21.9

G. Discussion

The experimental outcomes prove that the suggested PrivAURAL framework is efficient in facilitating the adaptive listening to English allowing to combine the reinforcement learning, personalization based on affect, and federated training. The steady declining Word Error rate between sessions suggests that the modeling of instruction can be represented as a progressive decision-making problem which facilitates the system to match the task complexity with that of the individual learner. Contrary to the baselines, which are usually more accurate and somewhat static, PrivAURAL modulates the content pacing and difficulty dynamically and achieves better understanding results and lower completion time of the task. The subsequent stabilization of cumulative rewards and policy entropy further affirms that the federated Deep Q-Network approach will approach sound strategies of adaptation without oscillations.

Analysis of engagement indicates that affective proxy signals can be used to control cognitive load, which enables learners to accomplish tasks more effectively whilst ensuring performance is steady. The visualization of feature embedding reveals that HuBERT-based representations together with behavioral cues are useful in distinguishing the level of listening difficulty, which can guide the policy to choose an informed task. Notably, federated aggregation allows all learners to gain the advantages of collective learning but keeps the privacy of learners intact since the raw audio or behavior data are not transferred. This confirms that the observed improvements in accuracy, efficiency, and engagement are realized under strict federated privacy constraints, supporting the practical deployment of the framework in real-world adaptive learning environments. Ablation study supplements the central importance of reinforcement learning in attaining these gains by demonstrating evident degradation when adaptive policy learning is eliminated. On the whole, the results suggest that PrivAURAL can be used as a scalable and privacy-sensitive alternative to centralized listening systems and is especially applicable in distributed and personalized language learning systems.

V. CONCLUSION AND FUTURE WORKS

This study introduced PrivAURAL, a privacy-conscious and affect-sensitive adaptive listening model, which trained English instructional learning through federated reinforcement learning as a series of sequential decisions. The system dynamically adjusted the difficulty and the speed of tasks to each individual learning curve by combining HuBERT-based semantic acoustic representations with affective proxy indicators based on the

behavior of learners. The experimental findings showed that there were consistent reductions in Word Error Rate, enhanced task completion efficiency, stabilized engagement patterns and consistent policy convergence between distributed learners. Compared to centralized or accuracy-only listening systems, PrivAURAL was able to perform personalization that did not require sending raw audio, performance logs or affect-related signals thus balancing adaptive learning goals against the hard privacy limits. It was demonstrated that the federated Deep Q-Network allows policy refinement by collaboratively relying on local autonomy, which supports the scalability of adaptive listening systems based on data accountability. Altogether, the results indicate that personalization, reinforced by reinforcement, together with affect-conscious regulation and decentralized training, can contribute to the improvement of the listening comprehension outcomes in the privacy-constraining educational context under a considerable degree. Importantly, these gains are achieved under strict federated privacy constraints, demonstrating that scalable and privacy-compliant deployment can be realized without compromising adaptive listening performance.

The present research will be expanded in future to include PrivAURAL to actual classroom and mobile learning situations in order to confirm its applicability with real-life learners and their various levels of proficiency. The proposed approach can also be evaluated by incorporating multi-lingual and code-switching speech datasets to assess the overall applicability of the proposed approach. More sophisticated federated optimization techniques, including adaptive client weighting and asynchronous aggregation can also be beneficial to convergence efficiency where there is heterogeneous participation of learners. In the modeling perspective, a policy-gradient or multi-agent reinforcement learning formulation would be of value to long-term curriculum planning. Lastly, more detailed affective proxy modeling with multimodal behavioral data, without ethical or privacy breaches, is a prospective direction to enhance the understanding of learner engagement without bringing explicit emotion detection and intrusive data acquisition.

REFERENCES

[1] S. Panchyshyn, S. Dobrovolska, and M. Opyr, "ENGLISH PROFICIENCY IS THE KEY TO SUCCESS IN GLOBAL BUSINESS," Актуальні проблеми сучасного бізнесу: обліково-фінансовий та управлінський, 2023.

[2] S. M. Islam, "Segmental errors in English pronunciation of non-native English speakers," Journal of Education and Social Sciences, vol. 16, no. 1, pp. 14–24, 2020.

[3] J. C. Mosquera Feijóo, F. Suárez, I. Chiyón, and M. G. Alberti, "Some web-based experiences from flipped classroom techniques in aec modules during the covid-19 lockdown," Education Sciences, vol. 11, no. 5, p. 211, 2021.

[4] K. Brown, "Integration and evaluation of virtual reality in distance medical education," Master's Thesis, Colorado State University, 2022.

[5] R. Liu, "Simulation of E-learning in English personalized learning recommendation system based on Markov chain algorithm and adaptive learning algorithm," Entertainment Computing, vol. 51, p. 100719, Sept. 2024, doi: 10.1016/j.entcom.2024.100719.

[6] R. Isaeva, N. Karasartova, K. Dznunusnalieva, K. Mirzoeva, and M. Mokliuk, "Enhancing learning effectiveness through adaptive learning platforms and emerging computer technologies in education," Jurnal Ilmiah Ilmu Terapan Universitas Jambi, vol. 9, no. 1, pp. 144–160, 2025.

[7] C. Ren, "Enhancing User Experience through Improving the User Interface of Phonetics Tools and Studies on Phone-level ASR-based Automation through Deep Learning Techniques," PhD Thesis, Auburn University, 2023.

[8] X. Zhang, X. Zhang, W. Chen, C. Li, and C. Yu, "Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments," Scientific Reports, vol. 14, no. 1, p. 9543, 2024.

[9] D. Nimma, V. Arunadevi, M. Manu, K. S. Punithaasree, and F. Ruban Alphonse, "Deep Learning Approaches for Improving English Listening Comprehension in E-Learning Environments for Adult Learners," in 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Oct. 2024, pp. 924–930. doi: 10.1109/I-SMAC61858.2024.10714786.

[10] K. Schäfer, J.-E. Choi, and S. Zmudzinski, "Explore the world of audio deepfakes: A guide to detection techniques for non-experts," in Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation, 2024, pp. 13–22.

[11] A. B. FRANKLIN and M. C. CA, "THE USE OF MOBILE DEVICES IN ENGLISH LANGUAGE LEARNING AMONG ADVANCED LEARNERS: INSIGHTS FROM INTERVIEW DATA," The Online Journal of Distance Education and e-Learning, vol. 13, no. 1, p. 58, 2025.

[12] F. M. Alshammary and W. S. Alhalafawy, "Digital platforms and the improvement of learning outcomes: Evidence extracted from meta-analysis," Sustainability, vol. 15, no. 2, p. 1305, 2023.

[13] M. Lichouri, R. F. A. Embarek, K. Lounnas, and R. Djeradi, "Cross-Linguistic Speaker Profiling: Evaluating Monolingual and Multilingual Recognition through Machine Learning and Mel-Frequency Cepstral Coefficients," 2023.

[14] P. Zhou, "Real time feedback and E-learning intelligent entertainment experience in computer English communication based on deep learning," Entertainment Computing, vol. 51, p. 100752, Sept. 2024, doi: 10.1016/j.entcom.2024.100752.

[15] I. Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, and P. Tudorache, "Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review," Education Sciences, vol. 13, no. 12, Art. no. 12, Dec. 2023, doi: 10.3390/educsci13121216.

[16] M. E. Ahmed and S. Hasegawa, "An Expert Usability Evaluation of a Specialized Platform for Designing and Producing Online Educational Talking Books," Applied System Innovation, vol. 7, no. 5, p. 74, 2024.

[17] A. Valledor, A. Olmedo, C. J. Hellín, A. Tayebi, S. Otón-Tortosa, and J. Gómez, "The eclectic approach in English language teaching applications: A qualitative synthesis of the literature," Sustainability, vol. 15, no. 15, p. 11978, 2023.

[18] R. Hu, Z. Hui, Y. Li, and J. Guan, "Research on learning concentration recognition with multi-modal features in virtual reality environments," Sustainability, vol. 15, no. 15, p. 11606, 2023.

[19] J. Hong, H. Jeon, H. Lee, D. Kim, and M. Ko, "HearIt: Auditory-Cue-Based Audio Playback Control to Facilitate Information Browsing in Lecture Audio," Applied Sciences, vol. 11, no. 9, p. 3803, 2021.

[20] I. Chaturvedi, T. Noel, and R. Satapathy, "Speech emotion recognition using audio matching," Electronics, vol. 11, no. 23, p. 3943, 2022.

[21] T. Chen, "Design and Application of English Oral Online Dialogue System Based on Reinforcement Learning Algorithm," Procedia Computer Science, vol. 261, pp. 716–723, 2025.

[22] TensorFlow, "tedlium | TensorFlow Datasets," TensorFlow. Accessed: Sept. 18, 2025. [Online]. Available: https://www.tensorflow.org/datasets/catalog/tedlium

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM transactions on audio, speech, and language processing, vol. 29, pp. 3451–3460, 2021.

[24] M. Xu, S. Li, and X.-L. Zhang, "Transformer-based end-to-end speech recognition with local dense synthesizer attention," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 5899–5903.