IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 July 2025, accepted 19 July 2025, date of publication 25 July 2025, date of current version 31 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3592632

APPLIED RESEARCH

Implementation of an AI English-Speaking Interactive Training System Using Multi-Model Neural Networks

CHING-TA LU[1], (Member, IEEE), **YEN-YU LU**[2], **YI-RU LU**[3], **YING-CHEN PAN**[4], AND **YU-CHUN LIU**[5]

[1]Department of Communications Engineering, Feng Chia University, Taichung 407102, Taiwan
[2]Department of Engineering Science, National Cheng Kung University, Tainan 701, Taiwan
[3]Department of Electronic Engineering, National Taipei University of Technology, Taipei 106344, Taiwan
[4]Department of Electrical Engineering, National Taiwan Normal University, Taipei 11677, Taiwan
[5]Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei 11677, Taiwan

Corresponding author: Ching-Ta Lu (Lucas1@ms26.hinet.net)

This work was supported by the National Science and Technology Council, Taiwan, under Grant NSTC 111-2410-H-035-059-MY3.

ABSTRACT Many people can read and listen to English well but may need help to speak it well. This paper aims to implement an AI English-speaking interactive training (AIESIT) system based on AI for special-purpose English-speaking training research, enabling students to express and communicate in professional English naturally. We will provide new tools for solving the software design in English-speaking training. The proposed AIESIT system integrates generative AI, speech recognition, and body recognition. The AIESIT system uses generative AI to generate an AI agent with mouth shapes that match the English speech, which enhances the user's feeling of being in the real world. The speech recognition system recognizes the voice content of the user's response for passing the evaluation. Since the AIESIT system does not have a natural person online, users can speak English boldly to improve their oral skills. During the learning process, OpenPoseNet is used to recognize whether the user has poor posture or leaves the seat during the learning process. Eye CNN is used to recognize whether the user falls asleep during the learning process and as a reference for continuing the lesson. Finally, the learning trajectory, including the recognition rate and response time, is output to score the performance of the English dialogues. The results of multiple user tests have shown that increasing the number of practice sessions can improve the English speaking, encouraging users to keep practicing, and this system helps improve their English speaking.

INDEX TERMS AI agent, convolutional neural network, interactive English-speaking practice, learning status recognition.

I. INTRODUCTION

Speaking English fluently has become essential for academic advancement, career development, and international communication. With the growing demand for English-speaking competence, learners increasingly seek effective methods to improve their oral language skills. However, despite the abundance of English learning resources, one common obstacle remains: many learners experience anxiety or hesitation when speaking English in front of others. This social pressure discourages active participation, leading to limited practice and slow progress in speaking proficiency [1].

To address this issue, interactive learning systems have emerged as a promising approach to facilitate spoken English training. These systems allow learners to practice speaking in a non-threatening, self-paced environment. For instance, the TOEIC chatbot proposed by Hsu et al. [2] demonstrated how removing human interlocutors can alleviate learners' speaking anxiety. Similarly, Ericsson and Johansson [3] showed that conversational AI agents encourage oral interaction by providing emotionally supportive and engaging environments. However, despite their potential, many existing

The associate editor coordinating the review of this manuscript and approving it for publication was Liang-Bi Chen.

interactive systems have a critical limitation—they provide one-way interaction that lacks awareness of the learner's real-time condition. Specifically, these systems often fail to consider whether the learner is attentive, emotionally engaged, or physically present. For instance, if a learner is distracted, fatigued, or away from the screen, the system may continue delivering content without any adaptive adjustment. This lack of responsiveness can reduce learning effectiveness, learner disengagement, and dropout [4].

In human-to-human instruction, a teacher can dynamically observe and respond to learners' behavior—pausing when they appear confused, adjusting pace based on facial expressions, or re-engaging distracted students. Emulating this adaptive instruction in an automated system is a complex yet crucial challenge in developing intelligent tutoring systems. Recognizing this need, researchers have begun incorporating emotion recognition, pose detection, and speech analysis into learning technologies. For example, Buono et al. [5] proposed a model that uses recurrent neural networks to detect learner engagement based on facial expression and head pose. Lu et al. [6] proposed a virtual English learning system capable of adapting to the learner's behavioral and emotional feedback. These modalities provide valuable insights into the learner's cognitive and emotional states, enabling a system to make real-time instructional decisions, such as pausing content, offering encouragement, or reviewing material.

Motivated by these challenges, we propose an AI English-Speaking Interactive Training (AIESIT) system to enhance spoken English learning through emotion-aware, pose-aware, and speech-enabled interactions. The AIESIT system features a generative AI video agent, implemented via D-iD [7], which simulates a virtual speaking partner with synchronized mouth movements for natural, immersive interactions. Unlike systems that rely solely on speech recognition [6], [8], [9], the AIESIT system integrates real-time eye emotion detection, head pose tracking, and verbal response evaluation to form a holistic view of the learner's engagement. When the system detects that a learner is inattentive, frustrated, or absent, it can pause or adjust the lesson to preserve learning effectiveness.

This approach provides a safe, personalized, and adaptive learning environment for learners who may feel anxious speaking with real people. Moreover, it ensures that learning remains efficient and emotionally supportive—even without human instructors.

The main contributions of this work are as follows:

- A multi-modal computing paradigm is proposed, integrating emotion, pose, and speech recognition to enable real-time assessment of learner states in educational settings.
- An interactive English-speaking system is designed and implemented, capable of adapting to individual learner conditions to enhance engagement and learning outcomes.
- The system simulates realistic human interaction through an AI-generated video agent to reduce speaking anxiety, removing the dependency on human tutors or peers.
- An adaptive oral English training framework is introduced, tailored to different proficiency levels and learning behaviors to promote consistent and effective practice.

The rest of this paper is organized as follows. Section II describes the related works. Section III introduces the proposed AIESIT system. Section IV shows the experimental results, and section V concludes.

II. RELATED WORKS

Recent literature on artificial intelligence (AI) in English language education demonstrates a growing consensus on its pedagogical potential, particularly in enhancing speaking proficiency, fostering learner engagement, and addressing affective factors. This section synthesizes prior studies into three major thematic categories: personalized and adaptive learning, speaking proficiency and pronunciation training, and affective and emotional support.

A. PERSONALIZED AND ADAPTIVE LEARNING

Recently, numerous studies have been dedicated to enhancing English speaking proficiency [1], [2], [3], [10], [11], [12], [13], [14], [15], [16], [17], [18]. Ji et al. [10] systematically reviewed numerous types of research on AI in language education. Semana et al. [11] were concerned about the impact of using smartphones with self-regulated learning in speaking classes. Luo [12] and Wang [13] emphasized how AI-driven tools—such as tutoring systems and intelligent feedback algorithms—tailor content and pacing, thus increasing motivation and engagement. Similarly, Alisoy [14] and Fountoulakis [17] reported significant gains in vocabulary, grammar, and pronunciation, demonstrating how personalization and adaptive pathways can lead to measurable linguistic improvements. Nguyen [16] and Sarnovska et al. [15] further highlighted how adaptive learning paths support differentiated instruction and self-regulated learning, especially in distance education contexts.

Ji et al. [10] systematically reviewed the growing body of research on AI in language education by explicitly addressing the underexplored area of human-computer collaboration. By synthesizing findings from 24 empirical studies published between 2015 and 2021, the review identifies the respective roles of conversational AIs and teachers across different phases of language learning while highlighting existing challenges and offering actionable recommendations. The study's strength lies in its structured, comprehensive analysis of diverse empirical sources, providing a broad perspective on AI-integrated learning environments' current state and potential. The review's emphasis on intelligence amplification and classroom orchestration presents a forward-thinking vision for reducing teacher workload and enhancing learning efficiency. However, a key limitation is the limited availability of empirical evidence on active collaboration between conversational AIs and human teachers. This limitation constrains the depth of practical application.

Semana et al. [11] analyzed the impact of using smartphones with self-regulated learning in the speaking class. The results reveal that the method positively impacted students' English-speaking ability. The study's strength is its structured methodological approach and sizable sample, which support the generalizability of its findings. However, limitations include the lack of a control group, which affects the ability to make definitive causal claims, and the study's limited time frame, which may not reflect long-term effects. Future research should incorporate control comparisons, longitudinal data, and broader language outcomes to strengthen and expand upon these findings. Luo [12] presented a forward-looking approach to enhancing college English listening and speaking instruction through AI-assisted teaching models. The strength of the approach lies in its emphasis on adaptability and personalization, allowing educators to tailor content and feedback to individual learner needs, thereby increasing engagement and learning efficiency. This study also combines theoretical analysis with practical implementation insights, demonstrating how AI technologies can address limitations of traditional teaching, such as rigid pacing and limited interaction, by introducing dynamic, responsive learning environments. However, limitations include a lack of detailed empirical evidence or large-scale experimental validation to substantiate the claimed improvements in student performance. The paper does not deeply explore potential barriers such as infrastructure constraints, teacher readiness, or AI's ethical implications in language education.

Wang [13] explored the potential of AI-driven intelligent tutoring systems, specifically Duolingo, in enhancing English language learning through personalized, autonomous instruction. The strength lies in its focus on user-centric adaptability, demonstrating how Duolingo's AI-powered system delivers real-time feedback, adjusts learning pace, and reinforces grammar and vocabulary through spaced repetition. The study also benefits from a sizable participant pool and structured data analysis via SPSS (statistical package for the social sciences), lending credibility to its findings. Results indicate that students perceive the system as significantly improving their learning experience, primarily through immediate corrective feedback and tailored content. However, limitations include the reliance on self-reported data through questionnaires, which may not capture actual language performance gains or long-term retention. Alison [14] presented a study on the effectiveness of AI-powered personalized learning in enhancing ESL (English as a second language) student outcomes, with statistically significant gains in vocabulary retention, grammar accuracy, and pronunciation. The mixed-methods design enriches the analysis by combining quantifiable language improvements with qualitative evidence of increased engagement and self-directed learning, particularly through gamified interfaces. The study's strength is its contextualization within broader educational goals, specifically UN SDG 4 (United Nations Sustainable Development Goals), highlighting AI's potential to promote equitable and inclusive language learning. However, the study also candidly addresses persistent challenges, including algorithmic bias, data privacy concerns, and unequal access to AI technologies, which may exacerbate educational disparities if left unaddressed.

Sarnovska et al. [15] studied how AI technologies reshape university-level foreign language learning in distance education contexts. Its strength lies in its emphasis on personalized and adaptive learning, which significantly departs from traditional, standardized instruction. The study underscores how to optimize individual learning paths for greater engagement and efficiency by leveraging AI algorithms to analyze learner behavior and tailor content. The integration of diverse AI functionalities—intelligent tutoring, conversational agents, automated assessment, gamification, and virtual reality—demonstrates the multifaceted potential of AI to enhance both linguistic competence and learner motivation. Nguyen [16] presented a study on the role of AI technologies, such as chatbots, speech recognition systems, and mobile applications, in enhancing English language speaking skills. Its key strengths lie in highlighting the immediate pedagogical benefits of AI, particularly for novice and intermediate learners, such as improved fluency, more accurate pronunciation, increased engagement, and reduced language anxiety. These technologies support learner independence by enabling real-time feedback and autonomous practice outside the classroom. However, the study also notes significant limitations hindering broader implementation, including inadequate technological infrastructure and teacher training.

B. SPEAKING PROFICIENCY AND PRONUNCIATION TRAINING

Numerous studies specifically addressed the role of AI in improving oral language skills [1], [2], [3], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. Chen et al. [18] introduced a progressive question prompt-based peer-tutoring in virtual reality (PQP-PTVR) approach to improve English-speaking skills. Ericsson and Johansson [3] presented a study on the impact of using a spoken dialogue system for English learning. Fu et al. [1] proposed using mapping recognition results to evaluate English speaking. Hsu et al. [2] proposed an interactive TOEIC practice (TP) chatbot system for English as a foreign language to practice speaking English. Jing [19], Hsu et al. [2], and Dandu et al. [20] demonstrated how AI systems such as speech recognition and mobile apps significantly improve pronunciation, fluency, and grammar. Zhai and Wibowo [21] and Nguyen Huu [22] expanded this focus by emphasizing interactional competence and self-confidence gained through real-time AI feedback. Other experimental studies—including DiCaprio and Diem [23] and Yang et al. [24], offered robust empirical evidence showing that AI-enabled systems yield statistically significant improvements in spoken language across multiple subskills, especially when embedded in authentic and iterative learning environments.

Chen et al. [18] contribute to immersive language learning by introducing a PQP-PTVR approach to improve English-speaking skills. It addresses a critical limitation in conventional VR-based learning—students' passive engagement with pre-designed tasks—by embedding progressive, context-aware prompts that encourage deeper thinking and interaction. The experimental method, comparing PQP-PTVR with a conventional peer-tutoring VR approach, is a key strength, as it enables direct evaluation of the pedagogical enhancement. Results indicating significantly higher gains in speaking ability and self-efficacy for the experimental group affirm the effectiveness of the proposed method. However, limitations include potential variability in tutor competence among peers and the possible influence of novelty effects associated with VR technology. The generalizability of the findings may also be constrained by sample size, duration, or the specific English proficiency level of participants. Nonetheless, the PQP-PTVR model represents a promising advancement in leveraging VR and peer learning strategies to create more meaningful and engaging speaking practice environments. Ericsson and Johansson [3] presented a study on the impact of using a spoken dialogue system for English learning. The students interact orally with conversational AI agents. Experimental results reveal that the students sustained practicing and were emotionally engaged with a positive trend compared to those without using a conversational AI agent in the educational experience. The study's strength is its use of multiple data sources—oral interactions, questionnaires, and logbook reflections—which enrich the validity of the findings. However, the study is limited by its reliance on non-parametric statistical methods. The absence of a comparison group limits conclusions about the system's effectiveness relative to traditional methods.

Fu et al. [1] proposed using mapping recognition results obtained by the deep-belief networks speech model and correlation coefficients to evaluate English speaking. In addition, this method uses a support vector machine to improve phoneme recognition and reliability. Experimental results reveal that the natural language processing-based oral English teaching mode can increase students' overall oral English skills. The study's strength is its multi-layered methodological design, incorporating deep belief networks, support vector machine optimization, and pairwise variability index for rhythm evaluation, all contributing to more precise and individualized feedback. The experimental results significantly improve student engagement, fluency, and vocabulary learning. However, the study's complexity and technical depth may pose limitations in practical implementation, especially for institutions with limited access to advanced computational resources or technical expertise. Hsu et al. [2] proposed an interactive TOEIC practice (TP) chatbot system for English as a foreign language to eliminate their fear of speaking English and enable them to chat with online TP chatbots to practice spoken English. This TP chatbot can recognize the learners' responses through the Google speech recognizer. Because no real persons appear, the TP chatbot helps eliminate learners' anxiety about speaking a foreign language with foreigners. The strength of this study is its real-world application with a clearly defined target group: Taiwanese students with oral TOEIC scores below 100. The four-month experimental period adds validity to the findings, with students reporting increased satisfaction and perceived improvement in their speaking abilities. TPBOT's flexibility for educators to create and adapt content also enhances its pedagogical utility. However, the study's reliance on self-reported satisfaction as the primary outcome limits the objectivity of the findings; a more robust, performance-based assessment of speaking improvement would strengthen the evidence. Furthermore, the generalizability of the results may be limited to similar learner populations and educational contexts.

Jing [19] addressed a critical gap in English language education in China—students' low oral proficiency despite high exam performance—by proposing a speech recognition and synthesis-based system for pronunciation training. The system's strength is its dual use of speech recognition for accurate pronunciation evaluation and speech synthesis for generating model pronunciation, enabling real-time, targeted feedback. This closed-loop mechanism allows learners to identify specific pronunciation errors and receive immediate auditory correction, promoting more effective and individualized learning. Furthermore, the system's automation reduces reliance on teacher intervention, making it scalable for widespread use. However, limitations include potential challenges in accurately assessing pronunciation across diverse accents, speech rates, or dialectal variations, which may affect feedback reliability. Additionally, the study lacks longitudinal data to confirm sustained improvement over time and does not fully address user experience or system usability. Despite these limitations, the research contributes a practical and technically sound approach to improving spoken English, offering a valuable tool for bridging the gap between written proficiency and oral fluency in Chinese college students. Al-husband [2] demonstrated the effectiveness of the English language speech assistant (ELSA) speech analyzer, an AI-powered application, in enhancing English as a foreign language (EFL) university students' speaking skills. The strength of the research is its controlled experimental design, which clearly shows that students using the ELSA app significantly outperformed those receiving traditional rubric-based instruction across all assessed speaking domains—pronunciation, intonation, fluency, vocabulary, and grammar. The ELSA app's ability to provide immediate, personalized feedback and real-life speaking practice contributed to these gains, supporting its value as a dynamic and interactive learning tool. However, the study's limitations include a relatively small sample size and a short intervention period, which may affect the generalizability and long-term applicability of the findings. Additionally, while using rubrics as a control is pedagogically sound, it lacks the interactive affordances of AI, which may have influenced student engagement levels.

Dandu et al. [20] provided robust evidence for the effectiveness of AI-based instruction—specifically through the Rosetta Stone mobile application—in enhancing English-speaking abilities and promoting self-directed learning among ESL engineering students. The strength of the research is its controlled experimental design, which enables clear comparisons between AI-supported and traditional instructional methods. The AI tool's integration of pronunciation feedback, speech recognition, and interactive exercises contributed to the experimental group's significant gains in speaking accuracy, vocabulary, fluency, and pronunciation. However, limitations include a lack of long-term follow-up to assess retention and generalization of speaking skills beyond the controlled environment. Furthermore, while Rosetta Stone is a well-established platform, the study may not capture the variability in effectiveness across different AI applications or learner preferences. Zhai and Wibowo [21] comprehensively assessed the role of AI dialogue systems in enhancing interactional competence among EFL university students. This area has received comparatively less attention than other language skills. The study's strength is its rigorous methodology, following the process of the preferred reporting items for systematic reviews and meta-analyses and analyzing a broad range of sources to identify key dimensions influencing AI integration in EFL contexts. Identifying six overarching dimensions and 25 sub-dimensions offers a structured framework for understanding how AI dialogue systems impact language learning. Notably, the study highlights critical gaps in current AI system design, such as the neglect of debate, problem-solving, and culturally enriched communicative features like humor and empathy—components essential for authentic, interactional competence. These omissions underscore the limitations of existing AI tools. Nguyen Huu [22] presented a study on the effectiveness of AI-powered speaking practice tools in improving English speaking proficiency and learner engagement among Vietnamese university students. The strength lies in its robust 16-week quasi-experimental design involving a large, diverse sample across multiple institutional contexts, enhancing the findings' validity and generalizability. The enormous effect size for speaking proficiency underscores the substantial impact of AI interventions, particularly in enhancing performance-oriented aspects such as fluency and pronunciation. The tools were especially valuable in creating low-stakes, culturally sensitive practice environments that mitigated students' reluctance to speak—an important contribution given the influence of Confucian-heritage educational norms.

Dikaprio and Diem [23] presented a study on the effectiveness of Talkpal.ai, an AI-powered language learning tool, in enhancing English speaking skills. The strength of the study is its robust experimental design, featuring a randomized control group and the use of a validated speaking skills test, which strengthens the reliability of its findings. The significant improvement observed in the experimental group—reflected in both the mean scores and the t-test results—demonstrates the impact of personalized, interactive AI-driven practice on learners' speaking proficiency. Talkpal.ai's features, such as real-time feedback and adaptive dialogue, likely contributed to greater engagement and individualized learning opportunities, which are often limited in traditional classroom settings. However, potential limitations include the study's relatively short duration and its focus on a single AI application within a specific student population, which may affect the generalizability of the findings. Nonetheless, the research offers clear implications for integrating AI tools into higher education curricula. It encourages further investigation into long-term outcomes and cross-platform comparisons to maximize the pedagogical benefits of AI in language learning. Yang et al. [24] highlighted the effectiveness of AI-supported training—specifically through the TalkAI spoken dialog system—in enhancing English-speaking awareness and performance among intermediate-level engineering students. The strength lies in the dual use of quantitative AI-generated feedback and qualitative self-reports, which provide a comprehensive view of both measurable improvements and cognitive shifts in learners. The results indicate that students improved in technical aspects like pronunciation, grammar, and usage and developed greater metacognitive awareness of their speaking abilities. So, the AI system's role is not just as a corrective tool but as a facilitator of self-reflection and strategic learning. However, limitations include the study's focus on a single academic context and discipline, which may limit generalizability and potential over-reliance on AI scoring metrics without triangulation from human evaluation.

C. AFFECTIVE AND EMOTIONAL SUPPORT

AI tools are increasingly recognized for their affective computing benefits [5], [6], [25], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42]. Wu et al. [25], Zhang et al. [32], and Zhang [33] focused on emotional dimensions such as foreign language anxiety, enjoyment, and willingness to communicate, finding that AI-facilitated practice environments often reduce anxiety and increase learner confidence. Yu [34], Sayed et al. [35], and Muthmainnah [36] highlighted AI's potential as an emotionally supportive companion—especially in high-stakes, exam-oriented, or high-anxiety learning contexts—through safe, non-judgmental platforms that encourage active participation and sustained motivation. Pons and Masip [37] proposed a representation of sharing standard features in a loss function for learning. This research shows that learning a detection model with facial muscle movements improves emotion recognition.

Wu et al. [25] presented a study on the effectiveness of integrating AI chatbots into think-pair-share (TPS) activities to address both cognitive and affective challenges in EFL speaking instruction. The strength of the research lies in its holistic approach—examining both speaking performance and key emotional factors like foreign language speaking anxiety and enjoyment. A quasi-experimental pretest and posttest design adds methodological rigor while incorporating quantitative and qualitative data, strengthening the findings' validity. The results show that AI chatbot-assisted TPS activities significantly improved speaking performance, reduced anxiety, and enhanced enjoyment. So, AI can foster more supportive, engaging, and emotionally balanced learning environments. However, limitations include the relatively short six-week duration and focusing only on first-year students at a single university, which may affect generalizability. Zhang et al. [32] presented a study of AI-speaking assistants' emotional and communicative benefits in English language learning grounded in positive psychology. By employing a quasi-experimental design with pretest and posttest intervention measures, the research demonstrates that the AI assistant, Lora, significantly enhanced foreign language enjoyment and willingness to communicate while also reducing foreign language anxiety among Chinese EFL university students. These findings underscore the capacity of AI tools to support linguistic development and positively influence affective variables essential to successful language acquisition. The contrast between the experimental and control groups strengthens the causal inference and highlights the unique value of AI integration. The study's strength is its focus on psychological well-being in language learning—an area often overlooked in traditional pedagogy. However, the six-week duration and reliance on self-report surveys may limit long-term generalizability.

Zhang [33] discussed how generative AI (GenAI) oral coaching tools can impact foreign language learning, particularly by addressing learners' affective experiences. The strength of the research lies in its application of control-value theory to examine anxiety-related dimensions, revealing that GenAI oral coaches significantly reduce various forms of foreign language learning anxiety, including communicative, situational, oral, listening, and cognitive anxiety. The semi-structured interviews with Chinese college students also uncovered unexpected affective benefits, such as emotional support and a sense of companionship, highlighting the potential of GenAI tools not only as language tutors but also as affective learning companions. These outcomes were linked to increased learner self-efficacy, motivation, and emotional comfort—critical factors for sustained engagement and long-term language development. However, the technology is still nascent, and challenges remain, such as ethical risks, emotional misalignment, and value bias in GenAI education. Yu [34] presented a study on the affective and linguistic benefits of integrating AI-assisted tools into EFL instruction. The strength lies in its mixed-methods design, which quantifies changes in anxiety, enjoyment, and proficiency and captures learners' subjective experiences through open-ended questionnaires. The findings indicate that the AI-powered tool, including feedback features, significantly reduced learners' anxiety and increased enjoyment. However, the study's scope is somewhat limited by its short intervention duration and focus on a single learner population. Future research could examine long-term retention effects and explore how different AI tool designs affect learner profiles.

Sayed et al. [35] presented a study on how AI-powered tools—particularly ChatGPT—can influence linguistic development and the psychological dimensions of language learning. Through a concurrent mixed-methods approach, the study demonstrates that integrating AI into EFL speaking assessments leads to measurable improvements in speaking proficiency, learner autonomy, psychological well-being, and academic buoyancy among upper-intermediate students in an Ethiopian university context. Repeated pretest and posttest measures and narrative frames provide statistical rigor and rich contextual insights. The study's strength lies in its multidimensional framework, examining performance outcomes and how AI tools emotionally and cognitively support learners. The results highlight ChatGPT's effectiveness in delivering individualized feedback, fostering learner confidence, and promoting a growth mindset. However, limitations such as the small sample size and context-specific findings may affect generalizability. Muthmainnah [36] presented a study on EFL education by addressing the critical issue of speaking anxiety by integrating AI-CiciBot—a conversational AI tool—into Indonesian EFL classrooms. The strength of the research is its mixed-methods approach, combining pretest and posttest quantitative data with rich qualitative insights from student interviews. The findings show an improvement in speaking performance and a significant reduction in speaking anxiety, underscoring the effectiveness of AI-CiciBot as a low-pressure, judgment-free platform for oral language practice. However, the study's short intervention period and focus on a specific regional context suggest the need for future longitudinal research.

The above studies underscore the transformative potential of AI in English language education, especially for speaking skill development and affective engagement. The benefits are substantial, ranging from adaptive feedback to anxiety reduction. Motivated by these challenges, we propose an AI English-Speaking Interactive Training (AIESIT) system to enhance spoken English learning through emotion-aware, pose-aware, and speech-enabled interactions. The AIESIT system features a generative AI video agent, implemented via D-iD, which simulates a virtual speaking partner with synchronized mouth movements for natural, immersive interactions. Unlike systems that rely solely on speech recognition, the AIESIT system integrates real-time eye emotion detection, head pose tracking, and verbal response evaluation to form a holistic view of the learner's engagement. When the system detects that a learner is inattentive, frustrated, or absent, it can pause or adjust the lesson to preserve learning effectiveness.

The proposed AIESIT system differs from existing AI-based English-speaking training tools in several fundamental ways. While prior studies have demonstrated the value of speech recognition, pronunciation feedback, and personalized learning pathways, most systems focus primarily on linguistic correction or learner engagement through limited modalities, often omitting real-time behavioral and emotional responsiveness. In contrast, the AIESIT system adopts a multi-modal, emotion-aware, and behavior-aware framework by integrating generative AI for realistic avatar interaction, speech recognition for spoken response evaluation, and visual sensing (via OpenPoseNet and Eye CNN) to assess learner posture, attention, and fatigue. Unlike existing applications such as Duolingo, ELSA, or Talkpal.ai, which emphasize gamified feedback or scripted dialogues, the AIESIT system simulates an immersive, adaptive conversational environment with dynamic pacing adjustments based on the learner's attentiveness and affective state. Furthermore, while earlier works highlight AI's potential to reduce speaking anxiety, the AIESIT system uniquely reinforces this benefit by eliminating human interlocutors and employing synchronized visual feedback to enhance realism without pressure. This holistic integration ensures more natural and emotionally supportive interactions. It enables granular performance tracking and continuous adaptation—extending beyond prior models that focus narrowly on speech metrics or static feedback mechanisms.

III. PROPOSED AI ENGLISH-SPEAKING INTERACTIVE

Training System: The AIESIT system uses the OpenPoseNet and an Eye CNN to recognize the user's learning status and interact with the user through an AI agent, ensuring the user is in a good learning status. The system also uses speech recognition to evaluate the correctness of the user's verbal response. Fig. 1 shows the flowchart of the proposed AIESIT system. English tutorial videos are played while sequentially conducting pose and eye emotion recognition. As the user watches the English-learning video, the AIESIT system monitors whether the user is distracted by checking their pose movements or is tired and starting to doze off. The Eye CNN identifies whether his/her eyes are closed. If the system detects these factors, it will automatically display the recognized results and pause the playing class video. The AI agent waits for the user to return to his/her seat and open his/her eyes before continuing the class video playback.

After each video has been played, the user begins to answer questions. The AIESIT system captures the speech using a microphone. It proceeds with speech recognition to evaluate the response content and calculate its keyword spotting rate, which is the basis for passing the assessment. As long as the user's response keyword spotting rate reaches the preset passing threshold P, the AIESIT system will play the following video. The assessment score can be computed by (1):

$$S = \sum_{i=1}^{N} g_i \cdot S_i^c \tag{1}$$

where N and g_i denote the number of video clips and the weighting of the i^{th} video clip, respectively. S_i^c represents the keyword spotting rate. g_i can be expressed by

$$S_i^c = N_i^c / N_i \tag{2}$$

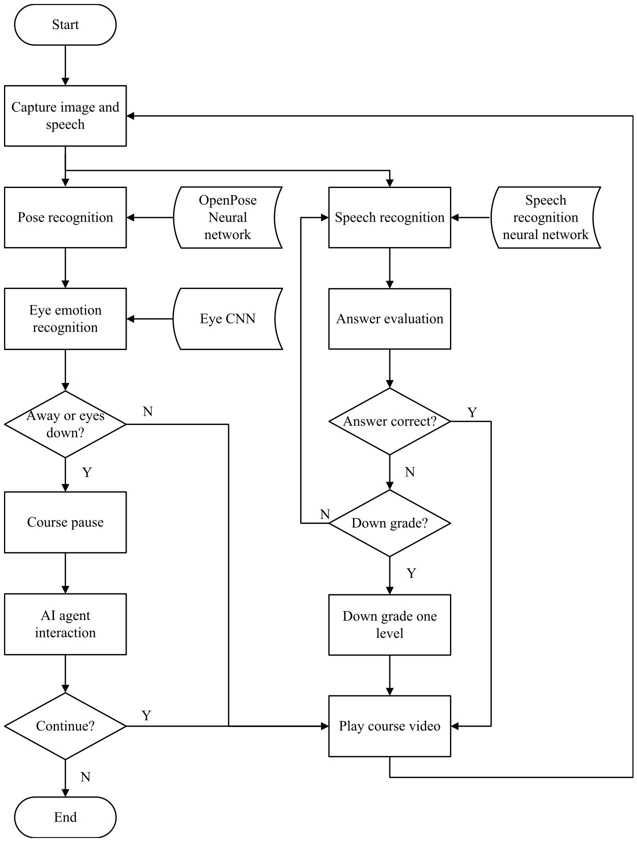


FIGURE 1. Flowchart of the proposed AIESIT system.

where N_i denotes the number of the keywords at the i^{th} video clip.

In (1), S_i^c can be computed by

$$S_i^c = N_i^c / N_i \tag{3}$$

where denotes the number of keywords to be correctly recognized at the i^{th} video clip.

Suppose the keyword spotting rate of the response S obtained by (1) exceeds the preset passing threshold P, i.e., $S \geq P$, the user's response is acceptable to pass the evaluation, and the system plays the following video. The value of P varies with the users' English proficiency. Its range is 0.4 to 0.8 in the experiments. Conversely, if the user's response does not meet the requirement, i.e., $S<P$, for three consecutive attempts, the AIESIT system will ask the users whether they want to downgrade their learning level. If the users verbally respond with 'Yes,' 'Sure,' or 'OK,' the course content will be downgraded, and the video will start playing from the beginning at the lower level.

On the other hand, if the users feel that a downgrade is unnecessary, they can continue answering questions repeatedly until they pass the evaluation. Once all student dialogue has been successfully assessed, the AIESIT system will create a score file, recording the keyword spotting rate of responses and response times during the learning process. This record will serve as feedback on their responses. Users can improve their English-speaking skills by practicing the questions with low scores in the report.

The English-speaking abilities of Chinese learners vary greatly. To avoid the inconvenience of requiring a pretest before using the proposed AIESIT system, students are encouraged to practice sequentially from Level 1 to Level 3, in which case the downgrade feature will not be triggered. Suppose a student with weaker proficiency selects Level 3 directly, with minimal prompts, and repeatedly fails to respond correctly to the AI agent's questions. In that case, the system will then recommend downgrading to Level 2. However, students may reject the suggestion and continue trying to answer correctly. The downgrade feature is typically unnecessary if a placement test is completed before using the system. Therefore, the downgrade mechanism is only activated when students are unsure of their English level and choose to start directly with Level 3.

A. LEARNING POSE RECOGNITION

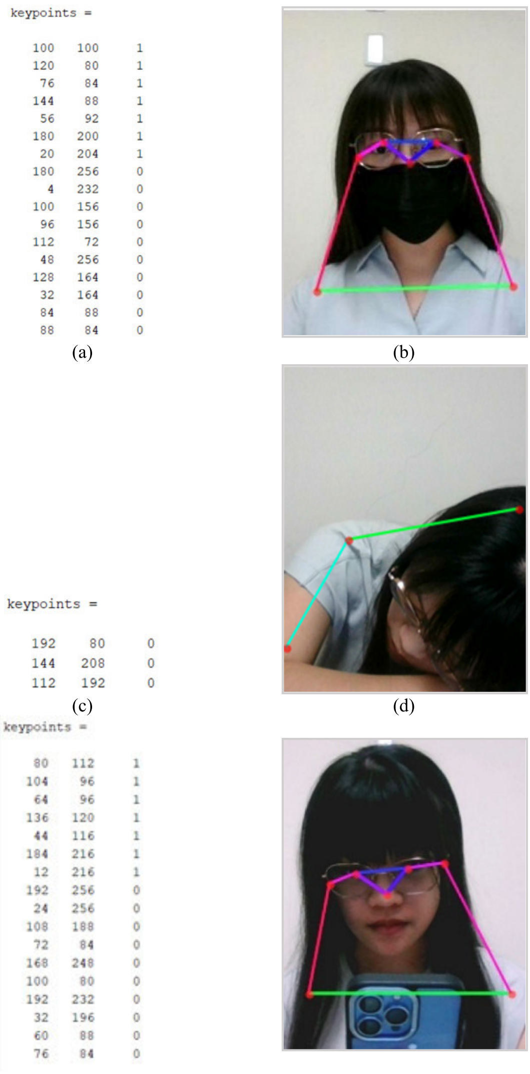
In order to assess whether users maintain good posture during learning sessions, the AIESIT system utilizes a webcam to capture video samples of users while they are watching educational videos. The system applies the OpenPoseNet [43] to recognize the users' pose movements. The network detects seventeen key points, five for the head and twelve for the body. These key points are connected. A key point comprises the coordinates (x, y) and a binary value indicating whether the key point was detected, where "1" represents detected, and "0" indicates not detected.

As shown in Figs. 2(a) and 2(b), the user's key points can be well recognized if the user is in a normal posture for watching the course video. Conversely, as shown in Fig. 2(d), the nose and eyes' key points cannot be detected if she is lying down. So, the values in the third column are 0, as shown in Fig. 2(c). It indicates that the user's facial image has not been detected, so the result denotes that she does not look at the screen. The accuracy rate of this determination reaches 99.8%.

On the other hand, if the recognition results for both the eyes and nose are unity, as shown in the third column of Figs. 2(a) and 2(e), it is determined that the user is sitting in the seat facing the screen. In this scenario, the system will perform emotion recognition on the user's eyes of the captured image. If the user utilizes a cell phone and her face is slightly facing downward, it is still possible to detect the key points of the nose and eyes, as shown in Figs. 2(e) and (f). The Eye CNN can detect this status.

B. EYES STATUS RECOGNITION

The eyes' open/closed status can determine whether the learner is tired or not looking at the screen during real-time learning. The AIESIT system uses the Viola and Jones algorithm (VJA) [44] and an Eye CNN to recognize whether the user's eyes close. These recognized results determine whether the user is tired. The structure of the Eye CNN is shown in Table 1.



```
keypoints =

   100   100     1
   120    80     1
    76    84     1
   144    88     1
    56    92     1
   180   200     1
    20   204     1
   180   256     0
     4   232     0
   100   156     0
    96   156     0
   112    72     0
    48   256     0
   128   164     0
    32   164     0
    84    88     0
    88    84     0
```

(a)

(b)

```
keypoints =

   192    80     0
   144   208     0
   112   192     0
```

(c)

(d)

```
keypoints =

    80   112     1
   104    96     1
    64    96     1
   136   120     1
    44   116     1
   184   216     1
    12   216     1
   192   256     0
    24   256     0
   108   188     0
    72    84     0
   168   248     0
   100    80     0
   192   232     0
    32   196     0
    60    88     0
    76    84     0
```

(e)

(f)

FIGURE 2. **An example of key points for various poses; (a) matrix of the body's key points values for normal posture; (b) the connection of the critical points for (a); (c) matrix of the body's key points values for lying down; (d) the connection of the key points for (c); (e) matrix of the body's key points values for facing down; (f) the connection of the key points for (e).**

C. ENGLISH ORAL TRAINING

Speech recognition is performed to determine whether a user can verbally answer a question correctly, hence comparing the recognized results with target keywords for judgment. The AIESIT system employs a commercial Kamera AI cloud speech recognition. By uploading speech to the cloud, short sentences (up to three words) can be recognized within one second, and long sentences within two seconds. The distance between the microphone and the speaker is within one meter, and the speech recognition accuracy rate reaches 97%. It can recognize about 400 Chinese characters per minute.

We use MATLAB's *detectspeech* function to detect the boundaries of speech in an audio signal. This function effectively identifies segments of the audio signal that contain speech by analyzing characteristics such as energy and

TABLE 1. Detailed layer parameters of the eye CNN.

Layer number	Layer name	Used parameters
1	Image input	Input image size: 75x75x1
2	Convolution	Window size: 3x3, stride: 1, filter number: 8, padding: 1.
3	Batch normalization	
4	ReLU	
5	Maximum pooling	Window size: 2x2, stride: 2
6	Convolution	Window size: 3x3, stride: 1, filter number: 16, padding: 1.
7	Batch normalization	
8	ReLU	
9	Maximum pooling	Window size: 2x2, stride: 2
10	Convolution	Window size: 3x3, stride: 1, filter number: 16, padding: 1.
11	Batch normalization	
12	ReLU	
13	Maximum pooling	Window size: 2x2, stride: 2.
14	Convolution	Window size: 3x3, stride: 1, filter number: 16, padding: 1.
15	Batch normalization	
16	ReLU	
17	Maximum pooling	Window size: 2x2, stride: 2
18	Convolution	Window size: 3x3, stride: 1, filter number: 32, padding: 1.
19	Batch normalization	
20	ReLU	
21	Fully connected	Class number: 2
22	Softmax	
23	Classification	

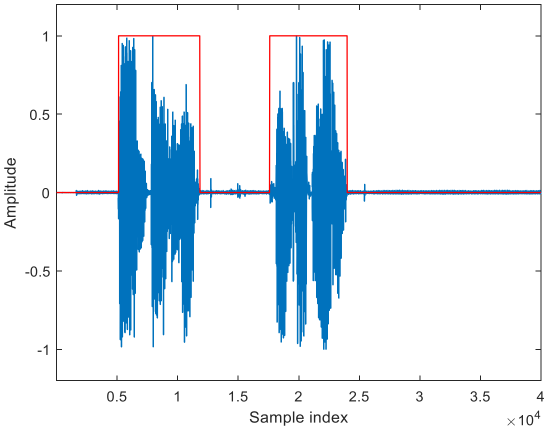


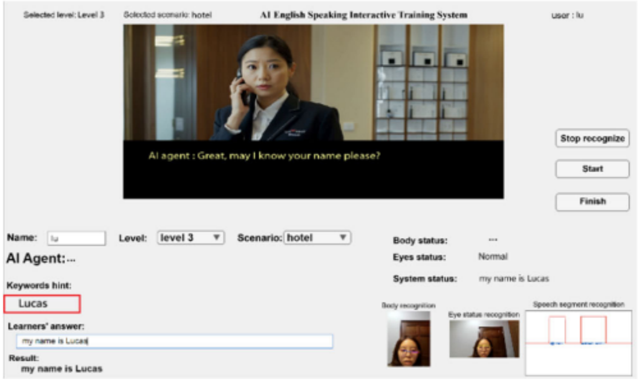
FIGURE 3. An example of speech-activity region detection.

frequency content. An example of speech-activity region detection is shown in Fig. 3.

The AIESIT system evaluates the correctness of the user's verbal response by speech recognition, where the *detect-speech* function detects the speech-activity regions. If the user fails to pass the evaluation, it is suggested that the learning level be lowered so that the user can confidently continue learning. By comparing the user's verbal answer with the target one and making a score, the AIESIT system allows the user to understand whether there is an improvement in the keyword spotting rate and response time of each exercise, which will serve as a motivation to keep practicing. When users cannot keep up with the progress, they can adjust the learning difficulty through dialogues to avoid giving up learning. Because the user does not meet a real person, this situation encourages the user to be confident to speak up in English.

The contents of English oral training are divided into three levels. A user selects the learning level according to the user's English-speaking ability. If the difficulty level is too high for the user to pass the AIESIT system's evaluation, the system will prompt the user to repeat his/her answers. If the user cannot pass the evaluation three times, the AIESIT system will inquire whether he/she feels it is too complicated. The user can also respond in spoken Chinese Mandarin to decide whether to lower the difficulty, thus achieving personalized learning. Level 1 is the most accessible, providing dialogue answers for users to watch and repeat. Its primary purpose is to train users' essential speaking abilities. Level 2 only offers some keywords to assist users in formulating their answers. The primary goal is to train users' fundamental sentence construction and speaking skills. Level 3 is the most challenging, as it only provides a few keywords, requiring users to answer independently. Its main objective is to allow users to demonstrate their proficiency, engage in direct conversation with the AI agent in English, and demand a high level of sentence construction and speaking ability to pass the evaluation. The AIESIT system sets the evaluation criteria for each level based on the user's expected standards. Each dialogue segment must meet the keyword spotting rate P_i % threshold, where the subscript i represents the level number (i = 1-3).

Users can assess their English learning progress and practice repeatedly on weak portions to boost speaking skills. Since there is no real person in the AIESIT system, a user does not need to hesitate to make mistakes or be unable to speak, eliminating the hesitation that often arises during actual conversations. Therefore, the AIESIT system can increase motivation to learn to speak English.



Selected level: Level 3
Selected scenario: hotel
AI English Speaking Interactive Training System
user : lu

AI agent : Great, may I know your name please?

Name: lu
Level: level 3
Scenario: hotel
AI Agent:...
Keywords hint:
Lucas
Learners' answer:
my name is Lucas
Result:
my name is Lucas

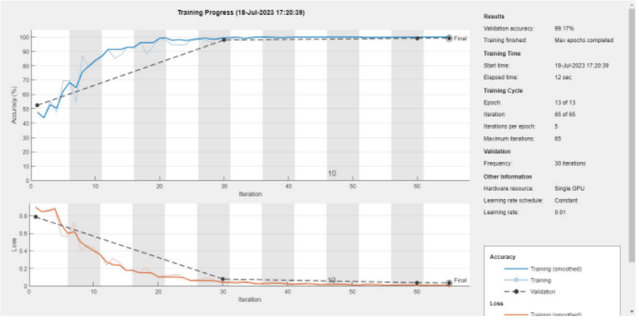
Body status: ...
Eyes status: Normal
System status: my name is Lucas

FIGURE 4. The snapshot of a user using the AIESIT system in a normal status.

Fig. 4 shows the graphic user interface (GUI) of the AIESIT system. First, a user must input his/her name and the desired difficulty level for learning, such as Level 1-3. Hence, the user clicks the "Start" button to interact with the AI agent. In the upper area of the system, the AI agent will engage in an English conversation with the user. It is produced by an AI generator D-iD [7]. At the same time, the webcam captures the user's images, performing pose and eye status recognition. The bottom right area shows the recognized pose status, eye emotions, and speech-activity regions. The bottom left corner displays the AI agent's interaction with the user. The keyword prompt area displays hints to assist the user during Levels 2 and 3. The user's response through speech recognition is displayed in the response area.

IV. EXPERIMENTAL RESULTS

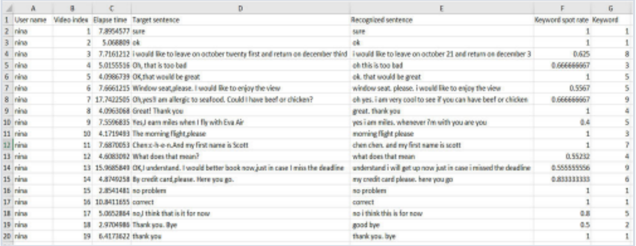
In the Eye CNN training phase, the VJA captured 480 images around the eye with eight people from our laboratory, of which 240 images were taken with the eye open and the others were with the eye closed. The resolution of each image was resized to a resolution of 120 × 60 pixels. Human labeling is applied to categorize emotions into eye closed and eye open. The architecture of the Eye CNN is shown in Table 1. The training accuracy on the validation set reached only 76.67%. Increasing the number of images improves the accuracy of recognition training in the network. We apply slight left-right shifts and horizontal mirroring to the original 480 images, and the number of the training dataset is augmented to 1920. This approach significantly increases the recognition accuracy, reaching 99.17%, as shown in Fig. 5. The Eye CNN can accurately recognize the user's learning status. Accordingly, the AIESIT system can accurately determine whether he/she is engaged in regular learning. The recognition results are demonstrated on the AIESIT system, instantly enabling users to visualize their recognized emotional status. The recognized learning emotion is the decision-making basis for automatically pausing and optionally downgrading the course videos.



Results
Validation accuracy: 99.17%
Training finished: Max epochs completed
Training Time
Start time: 18-Jul-2023 17:20:39
Elapsed time: 12 sec
Training Cycle
Epoch: 13 of 13
Iteration: 65 of 65
Iterations per epoch: 5
Maximum iterations: 65
Validation
Frequency: 30 iterations
Other Information
Hardware resource: Single GPU
Learning rate schedule: Constant
Learning rate: 0.01

FIGURE 5. Training curve of the Eye CNN.

The user progresses through all dialogues, and then the system generates a learning record, as illustrated in Fig. 6, which includes the user's name, question number, response time for each dialogue, target answer for the video, user's response, keyword spotting rate for each question, and the number of keywords. The results reveal that short phrases are more accessible for users to answer, so the keyword spotting rate reaches 100%, such as video indices 1, 2, 15, 16, and 19. Longer sentences with particular nouns are more difficult for users to answer and, therefore, have a low keyword spotting rate (≤40%), such as video index 9. Some target sentences are long, but he/she can still get an acceptable keyword spotting rate (>50%) if the user answers well, such as video indices 3, 7, and 13. If the user's answering habits are inconsistent with the target sentence, the user may get a low score, even for some short phrases, such as video index 18. Accordingly, the keyword spotting rate threshold for passing the evaluation should be set at a reasonable level so that users can pass the evaluation flexibly when they answer the question correctly. Users can improve their English-speaking skills by practicing particular questions with low scores.

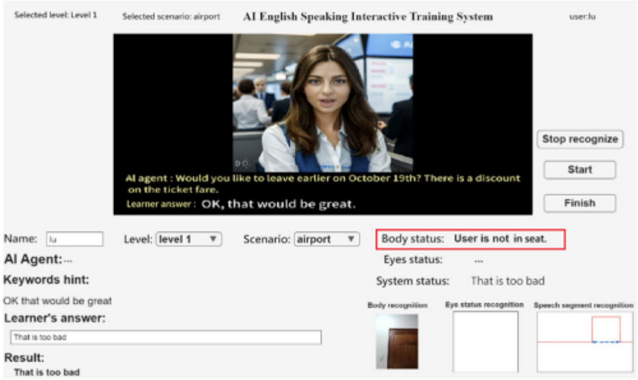


	A	B	C	D	E	F	G
1	User name	Video index	Elapse time	Target sentence	Recognized sentence	Keyword spot rate	Keyword
2	nina	1	7.8954577	sure	sure	1	1
3	nina	2	5.068809	ok	ok	1	1
4	nina	3	7.7161212	i would like to leave on october twenty first and return on december third	i would like to leave on october 21 and return on december 3	0.625	8
5	nina	4	5.0155516	Oh, that is too bad	oh this is too bad	0.666666667	3
6	nina	5	4.0986739	OK,that would be great	ok. that would be great	1	5
7	nina	6	7.6661215	Window seat,please. I would like to enjoy the view	window seat. please. i would like to enjoy the view	0.5567	5
8	nina	7	17.7422505	Oh,yes!I am allergic to seafood. Could I have beef or chicken?	oh yes. i am very cool to see if you can have beef or chicken	0.666666667	9
9	nina	8	4.0963068	Great! Thank you	great. thank you	1	4
10	nina	9	7.5596835	Yes,I earn miles when I fly with Eva Air	yes i am miles. whenever i'm with you are you	0.4	5
11	nina	10	4.1719493	The morning flight,please	morning flight please	1	3
12	nina	11	7.6870053	Chen:c-h-e-n.And my first name is Scott	chen chen. and my first name is scott	1	7
13	nina	12	4.6083092	What does that mean?	what does that mean	0.55232	4
14	nina	13	15.9685849	OK,I understand. I would better book now,just in case I miss the deadline	understand i will get up now just in case i missed the deadline	0.555555556	9
15	nina	14	4.8749258	By credit card,please. Here you go.	my credit card please. here you go	0.833333333	6
16	nina	15	2.8541481	no problem	no problem	1	1
17	nina	16	10.8411655	correct	correct	1	1
18	nina	17	5.0652884	no,I think that is it for now	no i think this is for now	0.8	5
19	nina	18	2.9704986	Thank you. Bye	good bye	0.5	2
20	nina	19	6.4173622	thank you	thank you. bye	1	1

FIGURE 6. A snapshot of learning progress file.

OpenPoseNet [43] can recognize whether the user is in the seat and the key points of the eyes. It cannot recognize whether the user is head down during the learning process. The Eye CNN improves the correctness of the OpenPoseNet on whether the learner is looking at the screen or not, which is the reference of whether the learner is attentive. OpenPoseNet and Eye CNN recognize the result as regular learning when the user's eyes look at the screen. Accordingly, combining OpenPoseNet, Eye CNN, and speech recognition can instantly identify the user's learning status, interact with the user, and adjust the difficulty level of the content promptly to avoid situations in which the user's learning status is not good.

A user must select the dialogue level of difficulty for which he/she wants to challenge. The AIESIT system will play interactive videos sequentially and provide keywords according to the level. Hence, the user has to respond to the content according to the hint keywords. The AIESIT system compares the recognized text with the desired answer. If the user achieves a keyword spotting rate exceeding $Pi\%$ for level i, he/she will smoothly progress to the following video. Conversely, the system will inquire whether the user wants to downgrade the difficulty if the user fails to respond three times. Fig. 7 shows a screenshot of the GUI interface of the AIESIT system in the airport scenario. A user leaves the seat, and the AIESIT system shows that the user is no longer on the screen for body recognition and eye state. Although the user answered correctly, the AI agent will stop the lesson until the user returns to the seat in front of the screen, and then the AI agent will continue interacting with the user. On the contrary, if the user's pose is typical, the eyes are open, and the user answers the AI agent correctly, the lesson goes normally, as shown in Fig. 4.


Selected level: Level 1
Selected scenario: airport
AI English Speaking Interactive Training System
user:lu

AI agent : Would you like to leave earlier on October 19th? There is a discount on the ticket fare.
Learner answer : OK, that would be great.

Name: lu
Level: level 1
Scenario: airport
Body status: User is not in seat.
AI Agent: ...
Eyes status: ...
Keywords hint:
System status: That is too bad
OK that would be great
Learner's answer:
That is too bad
Result:
That is too bad

FIGURE 7. A snapshot of the AIESIT system detecting the user leaving the seat.

The demo video link of the AIESIT system is shown below: https://www.youtube.com/watch?v=F4eXWuJS7Ds

The time spent using the AIESIT system varies from person to person. If a user can answer questions smoothly, the time spent per question is about 5-15 seconds, depending on the questions' length and difficulty level. In this demo video, a level contains 19 questions. A user can complete it in about five minutes if he/she answers it smoothly.

In the interaction with the AIESIT system, there is no limit to the number of times a user can try to answer a question. If the user fails three times, the system will ask the user in text whether he/she wants to downgrade. If the user decides to downgrade, he/she can answer verbally with affirmative sentences such as "OK, " "Sure, " or "Yes, " and the system will downgrade the course contents. In contrast, if the user does not want to downgrade, he/she can keep attempting to answer the questions until he/she passes. Failing to answer a question verbally does not affect the keyword spotting score of answering the question, but the time taken to pass the evaluation is poor.

A. EXPERIMENTAL DESIGN

A randomized controlled experimental design was used to assess the system's training effectiveness in English speaking. A total of 100 university students were recruited at the beginning of the study, and all of them were given a pretest of English language proficiency to select a sample of students with similar levels of language proficiency. Based on the test results, 52 participants were selected to fall within the same range to ensure a consistent learning starting point and minimize individual differences' effect on the results.

The 52 students were randomly segmented into 26 experimental and 26 control groups. The experimental group used the AIESIT system for multiple practice sessions during the study period. In contrast, the control group did not use the system and traditionally conducted their language learning, such as reading textbooks, writing notes, or self-reviewing. To ensure control of the experimental conditions, the two groups of students maintained the same learning time, topic content, and test format, using the AIESIT system as the only significant variable.

B. LEARNING OUTCOME ANALYSIS

Users can inquire about their learning performance after learning sessions, as shown in Fig. 8. Users must input the desired user name and level they want to query. The system will display the user's learning performance. The left-hand side of Fig. 8 shows the learning trajectory of the average keyword spotting rate (y-axis) versus the number of practice sessions (x-axis). The user's average keyword spotting rate is 57.2% in the first and 67.2% in the second. Although the results dropped slightly in the seventh practice sessions, the average keyword spotting rate increased to 86.7% with ten practice session. The improvement of the average keyword spotting rate reaches 51.57% ((86.7-57.2)/57.2∗100%) in the average keyword spotting rate of the user's verbal English responses. The right-hand side of Fig. 8 shows the learning trajectory of the average time (in seconds) required to pass each question versus the number of practice sessions. The average time to pass each question assessment is 9.6 seconds for the first practice session. Although the average time for passing a question increased in the seventh and eighth practice sessions, the average time for passing a question was shortened to 5.5 seconds in the 10^{th} practice session. The improvement in the average time for passing a question was 42.71% ((9.6-5.5)/9.6∗100%), representing improved English answering proficiency. Therefore, the average keyword spotting rate increases, and the time to pass the questions is shortened by practicing many times. The user's English-speaking ability improves.

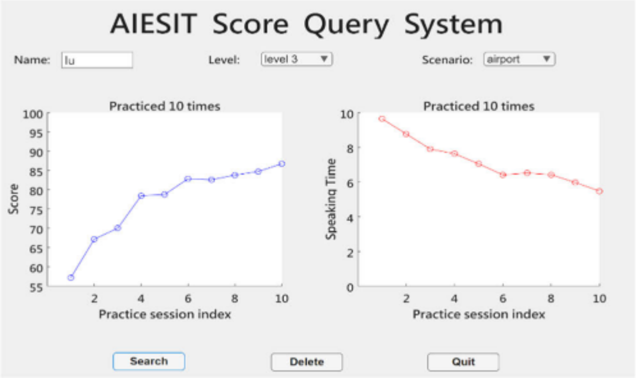
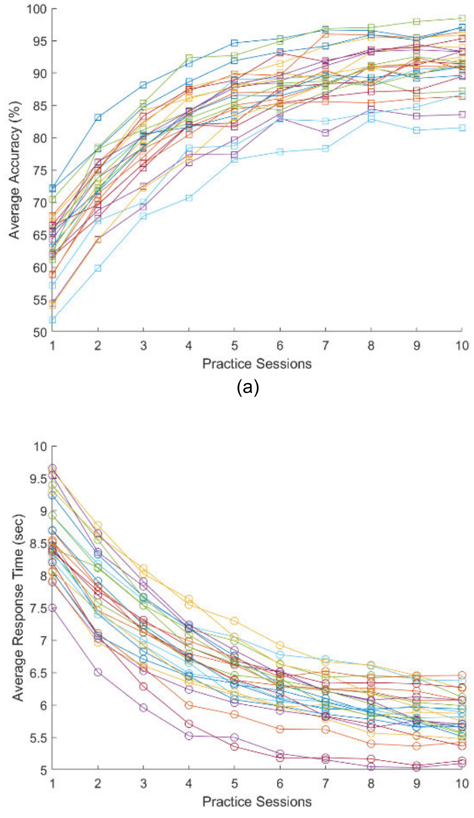

AIESIT Score Query System
Name: lu
Level: level 3
Scenario: airport

FIGURE 8. Screenshot of the learning record and grade query system interface.

Fig. 9(a) shows the changes in the accuracy rates of all the students in the experimental group over the ten practice sessions.

Most students showed an upward trend, and although some showed slight fluctuations in certain stages of the practice sessions, the overall curve still showed a gradual increase in the accuracy rates. This trend confirms that the training



(a)

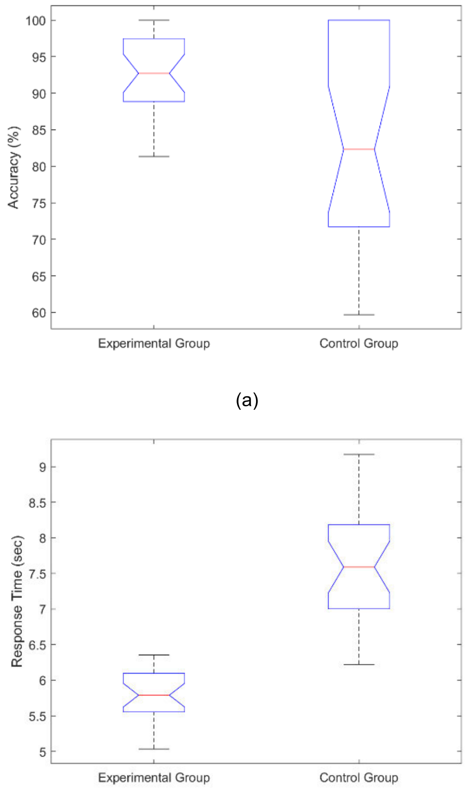
(b)

FIGURE 9. Graphs of the student's learning outcomes in the experimental group; (a) the relationship between the accuracy rate and the number of practice sessions; (b) the relationship between the response time and the number of practice sessions.

system is effective for students of different levels and shows that the system has significant benefits in enhancing language comprehension and application skills. Fig. 9(b) presents the changes in students' response time in all experimental groups in each exercise. From Fig. 9(b), students' response time decreased as practice sessions increased. This result indicates that the practice process effectively promotes fluency in verbalization. Although some students showed stabilization or a slight increase in time in the later stages of the practice, the overall trend was still decreasing. Through systematic and repetitive training, students could progressively reduce the time required for answering questions, thus demonstrating greater fluency.

Fig. 10 shows the boxplot of the accuracy rate and response time of the experimental and control groups in the posttest stage, which can be used to observe the distribution and median difference of the data between the two groups.

As shown in Fig. 10(a), the overall distribution of the accuracy rate of the experimental group is significantly higher than that of the control group, with the data concentrated in the range of 90% to 100%, and the median is close to 95%. In contrast, the median of the control group is about 83%, with a broader distribution range, which indicates that the performance variations are larger. Fig. 10(b) shows that in terms of response time, the overall speed of students in the experimental group is significantly faster than that of the control group, with a median of only about 5.8 seconds, while the control group reaches 7.6 seconds. The distributions of the two groups are almost non-overlapping, reflecting that using the AIESIT system in the experimental group can improve the response effectiveness.



(a)

(b)

FIGURE 10. Plot of ANOVA analysis; (a) response accuracy rate; (b) response time.

Fig. 11(a) presents the boxplots of the experimental and control groups regarding the percentage of accuracy improvement (i.e., posttest accuracy minus pretest accuracy).

It can be observed from the figure that the overall accuracy improvement of the experimental group is significantly higher than that of the control group, with a median of about 28% for the experimental group and about 18% for the control group. The result shows that the students using the proposed AIESIT system improved significantly in answering accuracy after training, significantly better than the control group who did not use the system. Fig. 11(b) compares the mean response times of the experimental group and the control group in the posttest. The distribution of students' response time in the experimental group is between 5 and 6.5 seconds, with a median of about 5.8 seconds, significantly lower than that of the control group, with a median of about 7.6 seconds. The overall distribution of the control group was also more spread out, with an upper limit of nearly 9.2 seconds,

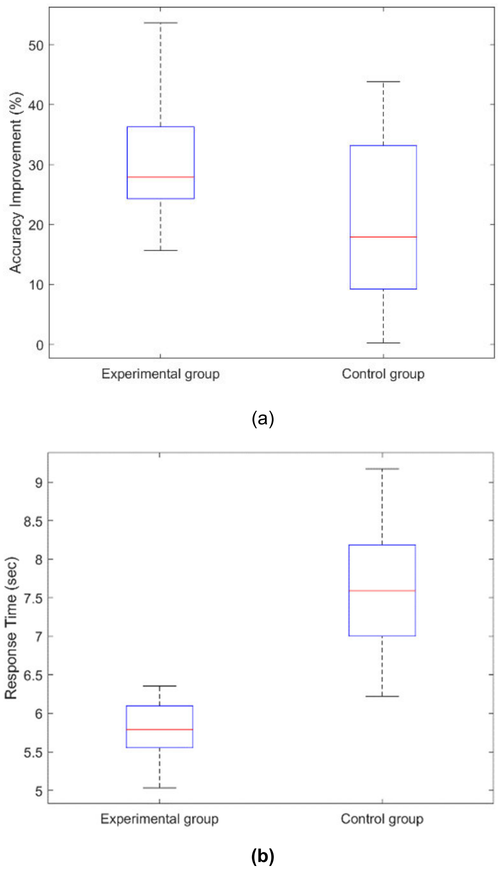


FIGURE 11. Boxplots of accuracy improvement and response time; (a) accuracy improvement; (b) response time.

indicating that the students in the experimental group were faster in answering the questions from the AI agent. This result demonstrates that the system can improve the response speed and fluency of answering English questions.

C. DISCUSSION

In the analysis of variance (ANOVA), the F-statistic can be used to compare the variance between the experimental group using the AIESIT system and the control group with the variance within the same group in order to determine whether the difference in learning outcomes between the two groups is statistically significant or not. F-statistic is defined as shown in (4):

$$F = \frac{\mathrm{MS}_{\mathrm{between}}}{\mathrm{MS}_{\mathrm{within}}} \tag{4}$$

where $\mathrm{MS}_{\mathrm{between}}$ and $\mathrm{MS}_{\mathrm{within}}$ denote the mean squares between and within groups, respectively.

In (4), $\mathrm{MS}_{\mathrm{between}}$ can be computed by

$$\mathrm{MS}_{\mathrm{between}} = \mathrm{SS}_{\mathrm{between}}/\mathrm{df}_{\mathrm{between}} \tag{5}$$

where $\mathrm{SS}_{\mathrm{between}}$ denotes the sum of squares between groups. $\mathrm{df}_{\mathrm{between}}$ is the degrees of freedom between groups; it is usually the number of groups minus 1.

Similarly, S_{within} in (4) can be computed by

$$\mathrm{MS}_{\mathrm{within}} = \mathrm{SS}_{\mathrm{within}}/\mathrm{df}_{\mathrm{within}} \tag{6}$$

where $\mathrm{SS}_{\mathrm{within}}$ represents the sum of squares within groups. $\mathrm{df}_{\mathrm{within}}$ represents the degrees of freedom within groups; it is usually the total number of samples minus the number of groups.

Tables 2 and 3 present the corresponding one-way ANOVA statistical tables computed by (4)-(6). Table 2 shows that the between-group variation in accuracy reached a statistically significant level, with an F-value of 11.76 and a p-value of 0.0012, well below the significance threshold of 0.05. Table 3 presents the analysis of variance for response time, revealing an F-value as high as 108.82 and a p-value less than 3.81 × $10^1$4, indicating a highly significant difference between groups. These statistical results further confirm that the experimental group using the proposed AIESIT system achieved significantly better language learning outcomes than the control group using traditional learning methods.

TABLE 2. Analysis of variance for accuracy.

Source	Groups	Between	Total
SS	1190.99	5061.84	6252.83
df	1	50	51
MS	1190.99	101.24	-
F-value	11.76	-	-
p-value	1.2e-3	-	-

TABLE 3. Analysis of variance for response time.

Source	Groups	Between	Total
SS	41.85	19.23	61.09
df	1	50	51
MS	41.86	0.38	-
F-value	108.82	-	-
p-value	3.81e-14	-	-

The AIESIT system innovatively integrates generative AI, speech recognition, body posture tracking (OpenPoseNet), and eye-tracking (Eye CNN) to provide a comprehensive, immersive training environment for spoken English practice. By eliminating live human interaction, the system reduces learners' anxiety and fosters confidence in speaking English, a common language learning barrier. The system focuses on special-purpose or professional English, which may limit its applicability for general conversational fluency or cultural pragmatics. While the system offers pedagogical innovation, integrating emotion recognition may introduce ethical dilemmas—particularly regarding user surveillance, data privacy, and informed consent. Practical application requires obtaining user consent.

V. CONCLUSION

This article presents an AIESIT system to train users' English-speaking proficiency. The system utilizes multi-functional detection to recognize users' learning emotions, including speech recognition, OpenPoseNet, and Eye CNN. Experimental evidence shows that multiple usages of the AIESIT system can improve the keyword spotting rate of a user's oral answers, reaching 94%. In addition, the improvement in the average time for answering a question can reach 57%. The user's English-speaking proficiency improves. Accordingly, the AIESIT system is an effective tool for practicing English speaking. Future versions of AIESIT can incorporate AI-driven automatic scenario generation to enhance flexibility and scalability. Using large language models and user proficiency data, the system could dynamically create context-aware dialogues, role-play exercises, and domain-specific simulations (e.g., medical, business, travel) tailored to the learner's goals and current level.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable comments, which have significantly improved the quality of this article. They also thank Dr. I.-Tien Chu's group for their help in making interactive video by the D-iD generator. The gratitude goes to Bo-Yen Chen for his help in managing complex code.

REFERENCES

[1] Y. Fu, Z. Zhang, and H. Yang, "Design of oral English teaching assistant system based on deep belief networks," *Soft Comput.*, vol. 27, no. 22, pp. 17403–17418, Nov. 2023.

[2] M.-H. Hsu, P.-S. Chen, and C.-S. Yu, "Proposing a task-oriented chatbot system for EFL learners speaking practice," *Interact. Learn. Environments*, vol. 31, no. 7, pp. 4297–4308, Oct. 2023.

[3] E. Ericsson and S. Johansson, "English speaking practice with conversational AI: Lower secondary students' educational experiences over time," *Comput. Educ., Artif. Intell.*, vol. 5, Feb. 2023, Art. no. 100164.

[4] X. Li, "A systematic study of oral English practice in computer multimedia networks," in *Proc. IEEE Int. Conf. Image Process. Comput. Appl. (ICIPCA)*, Aug. 2023.

[5] P. Buono, P. B. D. Carolis, F. D'Errico, N. Macchiarulo, and G. Palestra, "Assessing Student engagement from facial behaviour in online learning," *Multimedia Tools Appl.*, vol. 82, pp. 12859–12877, Mar. 2023.

[6] C.-T. Lu, I.-T. Chu, Y.-C. Pan, and Y.-C. Liu, "An interactive English learning system with image and speech recognition," in *Proc. IET Int. Conf. Eng. Technol. Appl.*, 2023, pp. 1–11.

[7] *D-iD*. Accessed: Jan. 26, 2025. [Online]. Available: https://www.d-id.com/

[8] S. Kobashikawa, A. Odakura, T. Nakamura, T. Mori, K. Endo, T. Moriya, R. Masumura, Y. Aono, and N. Minematsu, "Does speaking training application with speech recognition motivate junior high school students in actual classroom—A case study," in *Proc. 8th ISCA Workshop Speech Lang. Technol. Educ. (SLaTE)*, Sep. 2019.

[9] Y. Qian, R. Ubale, P. Lange, K. Evanini, V. Ramanarayanan, and F. K. Soong, "Spoken language understanding of human-machine conversations for language learning applications," *J. Signal Process. Syst.*, vol. 92, no. 8, pp. 805–817, Aug. 2020.

[10] H. Ji, I. Han, and Y. Ko, "A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers," *J. Res. Technol. Educ.*, vol. 55, no. 1, pp. 48–63, Jan. 2023.

[11] S. Menggo, H. C. Darong, and I. L. Semana, "Self-regulated learning method through smartphone assistance in promoting speaking ability," *J. Lang. Teaching Res.*, vol. 13, no. 4, pp. 772–780, Jul. 2022.

[12] Y. Luo, "Innovative research on AI-assisted teaching models for college English listening and speaking courses," *Appl. Comput. Eng.*, vol. 69, no. 1, pp. 155–160, Jul. 2024.

[13] Q. Wang, "AI-driven autonomous interactive English learning language tutoring system," *J. Comput. Methods Sci. Eng.*, vol. 25, no. 2, pp. 1155–1166, Mar. 2025.

[14] H. Alisoy, "The impact of AI-powered personalized learning on ESL student outcomes," *EuroGlobal J. Linguistics Lang. Educ.*, vol. 2, no. 2, pp. 89–98, Mar. 2025.

[15] N. Sarnovska, J. Rybinska, and Y. Mykhailichenko, "Enhancing university remote language learning through innovative applications of artificial intelligence technologies amidst global challenges," *Teaching Lang. Higher Inst.*, vol. 44, pp. 151–165, May 2024.

[16] H. A. Nguyen, "Harnessing AI-based tools for enhancing English speaking proficiency: Impacts, challenges, and long-term engagement," *Int. J. AI Lang. Educ.*, vol. 1, no. 2, pp. 18–29, Dec. 2024.

[17] M. S. Fountoulakis, "Evaluating the impact of AI tools on language proficiency and intercultural communication in second language education," *Int. J. 2nd Foreign Lang. Educ.*, vol. 3, no. 1, pp. 12–26, Oct. 2024.

[18] C.-Y. Chen, S.-C. Chang, G.-J. Hwang, and D. Zou, "Facilitating EFL learners' active behaviors in speaking: A progressive question prompt-based peer-tutoring approach with VR contexts," *Interact. Learn. Environments*, vol. 31, no. 4, pp. 2268–2287, May 2023.

[19] W. Jing, "Speech recognition sensors and artificial intelligence automatic evaluation application in English oral correction system," *Meas., Sensors*, vol. 32, Apr. 2024, Art. no. 101070.

[20] G. Dandu, G. M. Charyulu, and K. L. Kumari, "AI-driven language learning: The impact of Rosetta stone on ESL students' speaking proficiency and self-control," *Rupkatha J. Interdiscipl. Stud. Humanities*, vol. 16, no. 4, p. 2, Dec. 2024.

[21] C. Zhai and S. Wibowo, "A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university," *Comput. Educ., Artif. Intell.*, vol. 4, Jan. 2023, Art. no. 100134.

[22] H. N. Huu, "AI for English speaking practice: A study of effectiveness and engagement among Vietnamese university learners," *Australas. J. Educ. Technol.*, to be published.

[23] V. Dikaprio and C. Dahlan Diem, "How effective is Talkpal.Ai in enhancing English proficiency? Insights from an experimental study," *Lang., Technol., Social Media*, vol. 2, no. 1, pp. 48–59, Jun. 2024.

[24] F. Yang, "AI in language education: Enhancing learners' speaking awareness through AI-supported training," *Int. J. Inf. Educ. Technol.*, vol. 14, no. 6, pp. 828–833, 2024.

[25] T.-T. Wu, I. P. Hapsari, and Y.-M. Huang, "Effects of incorporating AI chatbots into think–pair–share activities on EFL speaking anxiety, language enjoyment, and speaking performance," *Comput. Assist. Lang. Learn.*, vol. 2025, pp. 1–39, Mar. 2025.

[26] N. Hu, "English listening and speaking ability improvement strategy from artificial intelligence wireless network," *Wireless Netw.*, vol. 31, no. 2, pp. 1071–1080, Feb. 2025.

[27] Q. Zhou, H. Hashim, and N. A. Sulaiman, "Supporting English speaking practice in higher education: The impact of AI chatbot-integrated mobile-assisted blended learning framework," *Educ. Inf. Technol.*, vol. 30, no. 10, pp. 14629–14660, Jul. 2025.

[28] K. Phanwiriyarat, K. J. Anggoro, and T. Chaowanakritsanakul, "Exploring AI-powered gamified flipped classroom in an English-speaking course: A case of duolingo," *Cogent Educ.*, vol. 12, no. 1, Dec. 2025, Art. no. 2488545.

[29] S. Setyawan and R. F. Pramudita, "Enhancing academic speaking skills through AI chatbots: A study on English language learning in higher education," *Int. J. Appl. Educ. Res.*, vol. 3, no. 2, pp. 129–140, 2025.

[30] D. Darmawansah, G.-J. Hwang, C.-J. Lin, and F. Febiyani, "An artificial intelligence-supported GFCA learning model to enhance L2 students' role-play performance, English speaking and interaction mindset," *Educ. Technol. Res. Develop.*, vol. 73, no. 3, pp. 1451–1479, Jun. 2025.

[31] L. Liu, "Impact of AI gamification on EFL learning outcomes and nonlinear dynamic motivation: Comparing adaptive learning paths, conversational agents, and storytelling," *Educ. Inf. Technol.*, vol. 30, no. 8, pp. 11299–11338, Jun. 2025.

[32] C. Zhang, Y. Meng, and X. Ma, "Artificial intelligence in EFL speaking: Impact on enjoyment, anxiety, and willingness to communicate," *System*, vol. 121, Apr. 2024, Art. no. 103259.

[33] M. Zhang, "The impact of artificial intelligence virtual oral tutoring APPs on Chinese youth's anxiety in oral English learning—Interview research based on users of artificial intelligence speaking tutoring APPs," *J. Lang. Culture Educ.*, vol. 1, no. 1, pp. 31–38, 2024.

[34] Y. Wang, "Reducing anxiety, promoting enjoyment and enhancing overall English proficiency: The impact of AI-assisted language learning in Chinese EFL contexts," *Brit. Educ. Res. J.*, to be published.

[35] B. T. Sayed, Z. B. Bani Younes, A. Alkhayyat, I. Adhamova, and H. Teferi, "To be with artificial intelligence in oral test or not to be: A probe into the traces of success in speaking skill, psychological well-being, autonomy, and academic buoyancy," *Lang. Test. Asia*, vol. 14, no. 1, p. 49, Oct. 2024.

[36] M. Muthmainnah, "AI-CiciBot as conversational partners in EFL education, focusing on intelligent technology adoption (ITA) to mollify speaking anxiety," *J. English Lang. Teaching Appl. Linguistics*, vol. 6, no. 4, pp. 76–85, Oct. 2024.

[37] G. Pons and D. Masip, "Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4764–4771, Jun. 2022.

[38] Y. Fu, S. Okada, L. Wang, L. Guo, Y. Song, J. Liu, and J. Dang, "Context- and knowledge-aware graph convolutional network for multimodal emotion recognition," *IEEE MultimediaMag.*, vol. 29, no. 3, pp. 91–100, Jul. 2022.

[39] C.-T. Lu, Y.-C. Liu, and Y.-C. Pan, "An intelligent playback control system adapted by body movements and facial expressions recognized by OpenPose and CNN," *Multimedia Tools Appl.*, vol. 83, no. 10, pp. 31139–31160, Sep. 2023.

[40] C.-T. Lu, C.-W. Su, H.-L. Jiang, and Y.-Y. Lu, "An interactive greeting system using convolutional neural networks for emotion recognition," *Entertainment Comput.*, vol. 40, Jan. 2022, Art. no. 100452.

[41] T. Mittal, A. Bera, and D. Manocha, "Multimodal and context-aware emotion perception model with multiplicative fusion," *IEEE MultimediaMag.*, vol. 28, no. 2, pp. 67–75, Apr. 2021.

[42] J.-Q. Yang, I.-T. Chu, and C.-T. Lu, "Emotion recognition using self-attention convolutional neural network for interactive learning system," in *Proc. IET Int. Conf. Eng. Technol. Appl.*, 2024, pp. 1–16.

[43] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–8.

[44] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2001, pp. 511–518.

CHING-TA LU (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from the National Taiwan University of Science and Technology, Taipei, in 1991 and 1995, respectively, and the Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2006. He had been with the Department of Electronic Engineering, Ming Chi University of Technology, Miaoli, Taiwan, from August 1995 to January 2008. He had been with the Department of Information Communication, Asia University, Taichung, Taiwan, from February 2008 to January 2023. He has been a Full Professor with the Department of Communications Engineering, Feng Chia University, since February 2023. His research interests include artificial intelligence applications, speech enhancement, image recognition, image denoising, speech coding, and speech signal processing.

YEN-YU LU received the B.S. degree from the Department of Information Communication, Asia University, Taiwan, and the M.S. degree from the Department of Computer Science, National Chengchi University, Taiwan. Currently, she is pursuing the Ph.D. degree with the Department of Engineering Science, National Cheng Kung University, Taiwan. Her research interests include the applications of artificial intelligence and image signal processing.

YI-RU LU is currently pursuing the bachelor's degree with the Department of Electronic Engineering, National Taipei University of Technology, Taiwan. Her research interest includes artificial intelligence applications.

YING-CHEN PAN received the B.S. degree from the Department of Information Communication, Asia University, Taiwan. She is currently pursuing the degree with the Department of Electronic Engineering, National Taiwan Normal University, Taiwan. Her research interests include the applications of artificial intelligence and image signal processing.

YU-CHUN LIU received the B.S. degree from the Department of Information Communication, Asia University, Taiwan. She is currently pursuing the degree with the Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan. Her research interests include the applications of artificial intelligence and speech signal processing.