# THE IMPACT OF AI-POWERED SPEECH RECOGNITION TOOLS ON PRONUNCIATION ACCURACY IN EFL UNIVERSITY LEARNERS

***Gulrukh Fayziyeva Shoniyozovna***
*Assistant of Department of Philology and teaching languages,*
*International school of Finance and Technology Samarkand branch*
*+998972858111*
*gulrukhfayziyeva@gmail.com*

**Abstract:** This study investigates the effects of integrating AI-powered speech recognition (ASR) tools into university-level English as a Foreign Language (EFL) speaking instruction. Using a mixed-methods design, data were collected from 118 undergraduate students across two intact groups at a state university in Uzbekistan over one academic semester (16 weeks). The experimental group (n = 59) used ASR-based speaking practice sessions three times per week as a supplement to regular instruction, while the control group (n = 59) followed a conventional teacher-led speaking curriculum. Pre- and post-test scores on the English Pronunciation Scale (EPS) and the Communicative Speaking Assessment Rubric (CSAR) were used to measure pronunciation accuracy and overall oral proficiency. Findings revealed a statistically significant improvement in vowel articulation, word stress, and connected speech in the ASR group ($p < .01$), while improvements in the control group were modest and largely limited to segmental features. Qualitative analysis of learner interviews and speaking journals identified reduced anxiety, increased self-monitoring behavior, and a stronger sense of speaking autonomy as key psychological benefits. The study also revealed several constraints: overreliance on the tool's feedback, lower motivation among intermediate proficiency learners, and inconsistencies in ASR recognition of regionally accented English. Pedagogical implications are discussed, including teacher readiness, tool selection criteria, and a proposed integration framework for EFL speaking classrooms.

**Keywords:** AI-powered speech recognition, EFL speaking instruction, pronunciation accuracy, learner autonomy, higher education, Central Asia

## Introduction

Walk into any English-language speaking class in a Central Asian university and you will likely witness a familiar scene: students hesitating before they speak, producing rehearsed phrases in near-silence, and waiting anxiously for the teacher's correction. For all the advances in communicative language teaching over the past three decades, the spoken word remains a source of persistent anxiety and uneven progress for EFL learners — especially those whose first language phonological systems differ

substantially from English.

The emergence of AI-powered speech recognition (ASR) technologies has introduced a new variable into this picture. Tools such as Google Speech-to-Text, ELSA Speak, Speeko, and integrated voice engines in platforms like Duolingo and Babbel now claim to offer immediate, individualized phonetic feedback — something that a single teacher managing thirty students cannot realistically provide. The theoretical appeal is obvious. If learners can receive real-time pronunciation analysis outside the classroom, practice volume increases, feedback loops shorten, and the teacher's role can shift from error-corrector to facilitator of meaning-making.

Yet enthusiasm should not precede evidence. The research base on ASR in language learning, while growing, has several gaps. Much of the existing work focuses on learners in North American or European contexts (Liakin, Cardoso, & Liakina, 2015; Golonka et al., 2012), uses small samples over short intervention periods, or conflates pronunciation improvement with overall speaking gains. Studies involving EFL learners in post-Soviet educational systems — where learner autonomy, technology literacy, and the phonological gap between L1 (Uzbek or Russian) and English are all contextually specific — are especially sparse.

This paper reports the findings of a 16-week classroom-based study conducted at a state university in Uzbekistan, involving 118 undergraduate students and targeting the use of ASR-assisted speaking practice as a pedagogical supplement. Our research questions were:

RQ1. To what extent does supplementary ASR-assisted speaking practice improve pronunciation accuracy in EFL university learners compared to conventional instruction alone?

RQ2. What specific phonological features (segmental vs. suprasegmental) are most affected by ASR-based practice?

RQ3. How do learners perceive the role of ASR tools in their speaking development, and what psychological factors are involved?

By engaging with these questions in a specific sociolinguistic and institutional context, we aim to contribute data that moves beyond generalized claims and toward pedagogically actionable conclusions.

ASR technology and language learning: a brief overview: automatic speech recognition has been applied to language education since the late 1990s, initially in computer-assisted pronunciation training (CAPT) systems that provided visual waveform feedback to learners (Neri, Cucchiarini, & Strik, 2002). Early systems were limited by narrow acoustic models, high error rates with non-native accents, and the need for specialized hardware. The past decade has witnessed a dramatic transformation. Deep neural network architectures have made modern ASR systems far more accurate even with accented input, and the proliferation of mobile applications

has democratized access to the technology (Nassaji, 2020).

Research on CAPT broadly supports its potential for improving segmental accuracy — the production of individual phonemes — particularly in controlled laboratory settings (Liakin et al., 2015). Improvements in suprasegmental features such as stress, rhythm, and intonation are documented less consistently, partly because these features are harder to model acoustically and partly because learners receive less explicit feedback on them from ASR systems (Derwing & Munro, 2015).

One of the more consistent findings in the ASR literature concerns the affective dimension of computer-mediated speaking practice. Multiple studies have found that learners report lower anxiety when speaking to a machine compared to a human interlocutor, particularly in high-stakes classroom settings (Chapelle, 2010; Xu & Ye, 2014). This finding has particular relevance in educational cultures where speaking mistakes carry social consequences — a characteristic noted in Confucian-heritage contexts but also well-documented in post-Soviet classrooms (Karpinska-Musiał, 2015).

Autonomy is a related construct. Holec's (1981) foundational work defined learner autonomy as the capacity to take charge of one's own learning, and subsequent work has linked it to deeper processing and more durable skill gains. ASR-assisted practice, because it decouples pronunciation feedback from teacher presence, has potential as an autonomy-building tool — but only if learners develop metacognitive awareness of how to interpret and respond to the feedback (Little, 2007).

Despite promising findings, several limitations in the existing literature warrant attention. First, most studies involving ASR in EFL contexts report data from East Asian learners (Liakin et al., 2015; Wang & Young, 2015), leaving learners from Central Asia, the Middle East, and sub-Saharan Africa largely absent from the empirical record. The phonological challenges facing an Uzbek-speaking EFL learner — including the production of interdental fricatives, vowel reduction in unstressed syllables, and English word stress patterns — are markedly different from those facing a Mandarin or Korean speaker.

Second, most studies last fewer than eight weeks. Speaking fluency and pronunciation accuracy are habitual motor skills; meaningful change typically requires sustained practice over months rather than days (DeKeyser, 2015). Third, very few studies examine what happens when ASR is used as a supplement within a regular speaking class rather than as a standalone replacement — which is the more ecologically valid scenario for most institutional contexts.

This study employed a mixed-methods quasi-experimental design. A quasi-experimental rather than fully randomized design was used because students were organized into pre-existing intact classes. Quantitative data came from pre- and post-tests measuring pronunciation accuracy and communicative speaking ability.

Qualitative data were gathered through semi-structured interviews (conducted at weeks 4, 10, and 16) and reflective speaking journals maintained by all participants.

One hundred and eighteen second-year undergraduate students (62 female, 56 male) enrolled in a compulsory English for Academic Purposes course participated in the study. Their ages ranged from 18 to 22 (M = 19.4). All students were Uzbek-English bilinguals with varying degrees of Russian proficiency. Based on a pre-study placement test using the Oxford Quick Placement Test (OQPT), participants were distributed across B1 (n = 74) and B2 (n = 44) proficiency levels according to the CEFR framework. No student exceeded C1 or fell below A2.

Students were assigned to the experimental group (EG, n = 59) and control group (CG, n = 59) based on their class sections. An independent samples t-test confirmed no statistically significant difference between the two groups' pre-test mean scores (EG: M = 41.3, SD = 7.2; CG: M = 40.8, SD = 6.9; t(116) = 0.38, p = .70), indicating baseline equivalence.

Both groups attended three 90-minute English speaking classes per week for 16 weeks, taught by the same instructor following the same communicative syllabus. The experimental group additionally completed three 20-minute ASR-assisted practice sessions per week using ELSA Speak (AI English Language Speech Assistant), accessed individually via smartphones.

ELSA Speak was selected based on four criteria: (1) documented phoneme-level feedback capability; (2) compatibility with both Android and iOS; (3) a pedagogically sequenced curriculum covering consonants, vowels, stress, and connected speech; and (4) affordability — students used the university-subsidized institutional license. Weekly guided practice tasks were aligned with the speaking topics in the main syllabus (e.g., presentations, discussions, debates), ensuring content coherence between in-class and out-of-class practice.

Students in the control group received no ASR-based supplement. Their out-of-class speaking practice consisted of the standard homework tasks assigned to both groups: recording a 2-minute monologue on each week's topic and submitting it for teacher feedback.

Pronunciation accuracy was measured using the English Pronunciation Scale (EPS; adapted from Munro & Derwing, 2006), a 50-point diagnostic instrument assessing segmental accuracy (25 points: consonants, vowel quality, minimal pair distinctions) and suprasegmental accuracy (25 points: word stress, sentence stress, intonation contours, connected speech features). Each participant produced three speech samples — a read-aloud passage, a semi-spontaneous picture description, and a spontaneous opinion response — which were rated by two trained raters blind to group membership. Inter-rater reliability was established at ICC = .88.

Overall oral proficiency was assessed using the Communicative Speaking

Assessment Rubric (CSAR), a 6-band descriptive rubric covering fluency, pronunciation, vocabulary, grammar, and interactive competence. The CSAR was administered through a structured speaking task completed in pairs.

Qualitative data consisted of 36 semi-structured interviews (18 from each group, purposively sampled for proficiency level diversity) and 118 reflective journals submitted every two weeks throughout the semester.

Table 1 presents the pre- and post-test mean scores on the EPS for both groups, disaggregated by subscale.

**Table 1. Pre- and Post-test English Pronunciation Scale (EPS) Scores by Group**

| Subscale | EG Pre | EG Post | CG Pre | CG Post | t | p |
|---|---|---|---|---|---|---|
| Segmental (25 pts) | 13.1 | 17.9 | 12.9 | 14.8 | 4.21 | < .001 |
| Suprasegmental (25 pts) | 10.7 | 16.4 | 10.4 | 11.6 | 6.83 | < .001 |
| **Total EPS (50 pts)** | **23.8** | **34.3** | **23.3** | **26.4** | **8.14** | **< .001** |

Note. t values reflect independent samples comparisons of post-test scores; df = 116 for all comparisons.

The experimental group demonstrated significantly greater improvement across both subscales. The suprasegmental gain (EG: +5.7 points; CG: +1.2 points) is particularly noteworthy, as suprasegmental features are traditionally considered harder to develop through implicit practice. A within-group paired samples t-test confirmed that both the EG and CG showed significant pre-to-post gains, but Cohen's d effect sizes were substantially larger for the EG (d = 1.24) than the CG (d = 0.42), indicating a large vs. small practical effect respectively.

## Conclusion

Language teaching has always involved a negotiation between the human and the technological, from the writing tablet to the tape recorder to the language lab to the smartphone. AI-powered speech recognition is the latest entrant in this long conversation, and like its predecessors, it neither solves everything nor destroys what came before.

What this study adds to the conversation is a specific, contextualized answer to a practical question: if EFL students at a Central Asian university use an ASR-based speaking app three times a week for one semester, does it help them speak more accurately? The answer, on the available evidence, is a cautious yes — particularly for pronunciation patterns that conventional classroom instruction tends to underaddress, such as word stress and connected speech features.

But the more important finding may be qualitative rather than quantitative:

students who used the tool regularly reported thinking differently about their own voices. They began to hear themselves, to notice mismatch between intention and production, and to feel that improvement was within their own power. That shift — from passive recipient of correction to active analyst of one's own speech — is precisely what language educators have been trying to engineer for decades.

AI did not make that happen on its own. It happened because a thoughtful teacher embedded the technology in a coherent curriculum, responded to what the data showed, and kept the goal firmly in view: not perfect pronunciation, but confident, intelligible communication. That, ultimately, is what the technology is for.

**References:**

1. Bajorek, J. P. (2019). Voice recognition still has significant race and gender biases. Harvard Business Review. Retrieved from https://hbr.org
2. Brazil, D. (1994). Pronunciation for advanced learners of English. Cambridge University Press.
3. Chapelle, C. A. (2010). The spread of computer-assisted language learning. Language Teaching, 43(1), 66–74.
4. DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), Theories in second language acquisition (2nd ed., pp. 94–112). Routledge.
5. Derwing, T. M., & Munro, M. J. (2015). Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research. John Benjamins.
6. Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners. Computers in Human Behavior, 75, 461–468.
7. Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2012). Technologies for foreign language learning: A review of technology types and their effectiveness. Computer Assisted Language Learning, 27(1), 70–105.
8. Holec, H. (1981). Autonomy and foreign language learning. Pergamon Press.
9. Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. Modern Language Journal, 70(2), 125–132.
10. Karpinska-Musiał, B. (2015). Intercultural education in higher education contexts: Challenges in post-Soviet spaces. Journal of Intercultural Communication, 38, 1–14.