p-ISSN 2252-7427, e-ISSN 2541-3635
https://ieflj.uniku.ac.id/pub

p-ISSN 2252-7427 e-ISSN 2541-3635
Indonesian EFL Journal
AISEE
The Association of Indonesia Scholars of English Education

THE EFFECT OF VOWEL REDUCTION DUBBING PRACTICE ON WORD ERROR RATE (WER) OF ASR-GENERATED CAPTIONS IN ZOOM PRESENTATIONS BY EFL LEARNERS

Ateng Kurnia*[1], Rudi Hartono[2], Mursid Saleh[3], Sri Wahyuni[4]
[1,2,3,4]*Universitas Negeri Semarang, Indonesia*

Corresponding author's email: atengkurnia70@students.unnes.ac.id

APA Citation: Kurnia, A., Hartono, R., Saleh, M., & Wahyuni, S. (2025). The effect of vowel reduction dubbing practice on Word Error Rate (WER) of ASR-generated captions in Zoom presentations by EFL learners. *Indonesian EFL Journal, 11*(3), 671-682. https://doi.org/10.25134/ieflj.v11i3.68

Received: 10-08-2025 Accepted: 07-10-2025 Published: 30-10-2025

Abstract: This study investigates the effect of vowel-reduction dubbing practice on the Word Error Rate (WER) of automatic speech recognition (ASR)-generated captions in Zoom-based English presentations. Employing a convergent mixed-methods design, the research examines how misarticulation related to vowel reduction—particularly the realization of schwa—shapes ASR performance and caption accuracy. Twenty-seven English teacher candidates completed structured dubbing tasks followed by ASR transcription analysis using Zoom's auto-captioning system. WER was calculated using standard computational formulae, and both quantitative data (percentages, accuracy ratios, and error rates) and qualitative data (error types, substitution patterns, and prosodic deviations) were systematically analyzed. Results indicate that substitution errors remained the most dominant ASR error type, especially when vowel reduction occurred without adequate prosodic support. A substantial improvement was observed following the intervention: the average WER decreased by approximately 52% (from 27.4% to 13.2%), demonstrating a strong effect of targeted dubbing practice on caption intelligibility. Participants who maintained clearer vowel quality and rhythmic timing achieved consistently lower WER scores, while those applying excessive or inconsistent vowel reduction produced higher rates of substitution and deletion errors. The study highlights the pedagogical value of integrating ASR-based feedback with dubbing activities to strengthen learners' pronunciation intelligibility and schwa awareness. It further underscores the importance of TPACK-informed instruction that cultivates learners' sensitivity to both human comprehensibility and machine recognition. Overall, these findings contribute to ongoing developments in pronunciation pedagogy by bridging technology-enhanced learning with EFL teacher education.

Keywords: *vowel reduction; dubbing technique; Zoom ASR caption; Word Error Rate; pronunciation pedagogy.*

INTRODUCTION

In contemporary digital learning environments, automatic speech recognition (ASR) technologies—such as those embedded in Zoom—have become indispensable tools for supporting English as a Foreign Language (EFL) instruction (Bhattacharjee et al., 2025; Ünlü & Doğan, 2024). These real-time captioning systems are designed to transform spoken language into text to enhance accessibility and participation (Gondi & Pratap, 2021; Siamtanidou et al., 2025). Yet in practice, ASR-generated captions frequently display inaccuracies when processing non-native English speech, especially in linguistically diverse EFL contexts where learners often struggle with prosodic control and vowel reduction (Wang et al., 2023; Y. Wang et al., 2025). Such misalignments are pedagogically critical for teacher candidates, who must model intelligible English but may produce speech patterns that obscure meaning and interfere with both human and machine comprehension (Kearns, 2020; Lepage, 2016; X. Li et al., 2024).

These challenges are well explained by the Intelligibility Principle(Levis 2021; Nagle and Huensch 2020), which posits that the primary instructional goal in L2 pronunciation is intelligible—not native-like—speech. According

to this principle, prosodic accuracy, including stress, rhythm, and vowel reduction, plays a more crucial role in intelligibility than segmental precision(Hodgetts 2021; Karunathilake, Kulasingham, and Simon 2023). From the perspective of L2 Perception Models, such as the Speech Learning Model (SLM) and the Perceptual Assimilation Model (PAM-L2), non-native speakers often misperceive and misproduce reduced vowels because they assimilate them to full vowel categories in their L1. This mismatch increases the likelihood of prosodic errors that significantly disrupt ASR systems, which are trained primarily on native-speaker acoustic patterns (Davitti et al., 2024; Thimmaraja Yadava et al., 2024).

Vowel reduction—particularly the centralization of unstressed vowels into schwa /ə/—constitutes a central component of English stress-timed rhythm (Philip & Francis Noyan, 2018; Kearns, 2020). When EFL learners fail to apply appropriate reduction in weak syllables, their speech often becomes overly deliberate or syllable-timed, increasing the likelihood of misrecognition. For instance, pronouncing further with two fully articulated vowels may trigger ASR substitutions such as vibrator or wider, revealing a breakdown in rhythmic and prosodic alignment (Davitti et al., 2024; Holmström, 2019). These errors—especially substitutions—are frequently documented in ASR output and highlight the need for pedagogical approaches targeting prosodic competence(Cheng et al. 2020; X. Wang, Wang, and Yuan 2025).

In response to these challenges, video dubbing has gained prominence as an effective pedagogical technique for improving prosodic accuracy(Bigioi and Corcoran 2023; Fernández-Peña 2023). Dubbing requires learners to replace the original audio track of a video with their own voice, engaging them in precise imitation of model speech patterns, including stress, timing, intonation, and vowel reduction (Bang et al., 2014; Bigioi et al., 2023; Marcet et al., 2022; Park, 2022). Through synchronized speech production, learners develop greater control of articulatory timing and internalize prosodic cues critical for intelligibility (Dahmen et al., 2023) Additionally, the repetitive and multimodal nature of dubbing supports phonological chunking and muscle memory formation for reduced vowels in multisyllabic words (Philip & Francis Noyan, 2018; Pingarrón, 2025). When paired with ASR-caption feedback, dubbing provides learners with immediate, data-driven insights into their pronunciation problems, facilitating self-regulation and reflective learning (Escobar-Álvarez, 2024; Korybski & Davitti, 2024).

Anchored in these theoretical perspectives, the present study investigates how vowel reduction dubbing practice influences the Word Error Rate (WER) of ASR-generated captions in Zoom-based presentations(Purcel 1976; Sheshadri, Vijjini, and Kharbanda 2021). Specifically, this study analyzes speech samples from twenty-seven English teacher candidates following a structured dubbing intervention, focusing on the extent to which mispronunciations of words such as further, wider, set detours, and resiliency produce semantically anomalous substitution patterns in ASR captions (Thu Hien Nguyen et al., 2024). These deviations underline the pedagogical urgency of integrating prosodic instruction into teacher education programs, particularly in Indonesian EFL settings where learners' L1 rhythm patterns often transfer into L2 production.

The study aims to examine how vowel reduction training delivered through dubbing contributes to lowering WER—particularly substitution errors—and thereby enhances ASR caption accuracy(X. Wang et al. 2025). Consistent with prior research, it is hypothesized that learners who approximate native-like vowel reduction norms will generate clearer, more intelligible speech, resulting in lower WER scores(Kuo and Weismer 2016; Pöchtrager 2024). Empirical evidence from acoustic phonetics further suggests that appropriate vowel reduction increases the contrast between stressed and unstressed syllables, producing a more native-like vowel space that supports both human and machine understanding (Kaori, 2016; Rodd et al., 2021). Thus, the current study situates vowel reduction at the center of pronunciation pedagogy, arguing that mastering this prosodic feature is key to improving intelligibility, reducing recognition errors, and preparing teacher candidates for professional communication.

METHOD

This investigation adopted a convergent mixed-methods design (Creswell & Plano Clark, 2018) to examine how vowel reduction training delivered through video dubbing influences the accuracy of automatic speech recognition (ASR) captions generated in Zoom-mediated presentations(Cheng et al. 2020; Sariani et al. 2024). The design was selected because pronunciation development involves both quantifiable acoustic outcomes—such as reductions in substitution or deletion

errors—and qualitative dimensions related to prosodic control, articulatory timing, and stress patterns. Integrating both datasets allowed a more holistic interpretation of how schwa production and multisyllabic stress influence ASR misrecognition(Bashori et al. 2021).

The methodological framework was informed by the model (Nantha et al. 2024; Valtonen et al. 2020) which supported the integration of technological tools (Zoom live captions, ELAN, Praat), pedagogical strategies (dubbing, guided imitation, flipped instruction), and phonological content knowledge (vowel reduction, weak forms, schwa centralization)(Khoshsaligheh, Pakar, and Ghomi 2024; 王 2023). This alignment reflects the practice-oriented nature of pronunciation training in English teacher education, where learners engage in iterative, reflective, and performance-based tasks.

Twenty-seven pre-service English teachers (coded X1–X27) participated in the study. They were enrolled in an interpretive listening and speaking course within an English teacher education program in West Java, Indonesia. All participants were native speakers of Sundanese or Javanese, representing language backgrounds known for syllable-timed tendencies that influence L2 rhythm acquisition.

English proficiency was classified at the B1–B2 level according to institutional placement measures(Christodoulidou, Nicolaidis, and Stamovlasis 2024; Liu et al. 2021). These classifications were supported by equivalent benchmark ranges—TOEFL ITP scores of 450–500 or IELTS bands of 4.5–5.5—confirming intermediate to upper-intermediate proficiency. None of the participants had previous exposure to dubbing-based training or systematic instruction in vowel reduction, making them appropriate subjects for tracing developmental changes.

Participation was voluntary, with written consent obtained from all students. Ethical clearance was approved by the institutional review board.

The intervention lasted four weeks and targeted vowel reduction, stress patterns, and weak forms. Instruction followed a flipped-classroom model comprising:

Pre-class video lectures introducing schwa centralization, stress-timed rhythm, and JEger/jeGER mnemonic grouping.

Guided Praat demonstrations illustrating reduced-vowel formants and temporal compression.

Deconstruction of native speaker recordings focusing on reduced vowels in multisyllabic words and unstressed syllables.

Participants practiced dubbing a monologue from a short video titled Imagine (87 words). The task required synchronizing articulation with the model's rhythm, intonation, and visual cues. Emphasis was placed on producing reduced vowels consistently, maintaining stress prominence, and avoiding syllable-timed articulation.

Students presented their dubbed audio live on Zoom with automated captioning enabled(Aggarwal and Elembilassery 2023). The ASR output served as the primary data source for WER calculation and qualitative analysis(Sariani et al. 2024).

WER scores were calculated by comparing the ASR-generated captions of each participant's Zoom performance with the original 87-word script using the standard formula:

$$WER = \frac{S+D+I}{N} = X \quad 100$$

where
S = substitutions, D = deletions, I = insertions, N = total reference words.

Raw caption files were exported from Zoom's meeting logs. Transcript alignment was conducted manually using ELAN, with Praat waveform cross-checking to ensure temporal precision(Wang, Li, and Liang 2024). Final counts were tabulated and analysed using Microsoft Excel.

ASR caption outputs were analysed qualitatively to construct a typology of misrecognitions focusing on: (1) Misinterpretation of reduced vowels. (2) Errors linked to mis-stressed syllables. (3) Misalignment of function and content words. (4) Substitution patterns producing semantic anomalies. (5) Insertions associated with disfluency or disrupted timing

Each participant received a five-word diag-nostic summary capturing dominant error tendencies. Codes were organized thema- tically to identify recurring intelligibility issues across the cohort(Hsu 2016; Ngo and Chen 2025).

Analysis followed four sequential stages: (1) WER Scoring. All 27 transcripts were aligned and scored against the reference script. (2) Descriptive Statistics. Means, standard deviations, and score ranges were computed to determine overall ASR accuracy. (3) Qualitative Clustering. Frequent misrecognized items were grouped into thematic categories reflecting vowel reduction failure,

misplaced stress, and improper syllable grouping(Huang, Shi, and Ouyang 2024; Murad, Stuerzlinger, and Munteanu 2019). (4) Interpretive Integration. Mixed methods merging examined how participants with low WER exhibited more consistent vowel reduction and smoother dubbing synchronization(Byers 2017; Verhoeven and Connell 2024). Conversely, high-WER participants displayed over articulated vowels, timing mismatches, or stress displacement(Purcel 1976; Werdiningsih, Marzuki, and Rusdin 2024).

Several procedures enhanced the trustworthiness of findings: (1) Double coding of transcripts by two trained raters, achieving inter-rater agreement above 95%. (2) Triangulation using Zoom logs, ELAN overlays, and Praat spectrograms. (3) ASR benchmarking with two native speaker recordings, yielding WER < 5%, establishing a baseline for comparison.

A methodological limitation concerns the absence of a control group, which restricts strong causal inference. Although the convergent mixed-methods design provides robust triangulation, future studies employing randomized control groups or quasi-experimental designs would more precisely determine the effects of dubbing relative to alternative pronunciation interventions(Baños 2021; Miggiani 2021).

RESULTS AND DISCUSSION

Chart 1 shows that substitution errors overwhelmingly dominate the learners' ASR output, far exceeding deletion and insertion errors (Haupt et al., 1977; Liu et al., 2012; Shiu, 2025). This distribution indicates that vowel reduction, stress assignment, and suprasegmental timing constitute the most persistent barriers to intelligibility for Indonesian EFL learners(Hodgetts 2021; Karunathilake et al. 2023). The pattern directly aligns with predictions of the Speech Learning Model (Flege, 1995; Barbosa, 2024; Zou et al., 2025), which proposes that L2 learners assimilate unfamiliar L2 phonetic elements—such as unstressed English vowels—into the closest existing L1 categories when the contrast is not perceived as meaningful. Recent evidence confirms that Indonesian and other syllable-timed L1 speakers frequently map English schwa onto their L1 full vowels, leading to over-articulated weak syllables that distort lexical identity and trigger ASR substitution errors (Discrimination of Second Language Vowel Contrasts, 2024; Prosodic and Segmental Effects of Pronunciation Training, 2023; Alshangiti & Evans, 2024; Wang & Munro, 2004). For Sundanese and Javanese learners, whose L1 prosodic systems do not employ reduction-based rhythmic distinctions, the tendency to maintain full-vowel clarity in unstressed syllables exacerbates misrecognition, causing ASR to select semantically distant substitutes (Chen & Tian, 2021; Le, 2021; Müller, 1998; Yi, 2012).

Moreover, the predominance of substitution errors reflects a deeper perceptual constraint: learners often fail to detect the acoustic cues that differentiate reduced from non-reduced vowels, resulting in systematic mismatches between target and produced forms. Recent L2 phonological studies also show that inadequate mastery of stress-timing reduces temporal predictability, making L2 speech particularly vulnerable to ASR misalignment in rapid or connected segments(Ünlü and Doğan 2024). These findings echo broader evidence that suprasegmental inaccuracies, rather than segmental errors, constitute the primary source of intelligibility breakdowns for many Asian EFL populations(Nagle and Huensch 2020; O'Neal 2015). As ASR systems are trained predominantly on native English corpora, deviations from expected prosodic patterns are disproportionately penalized, further magnifying substitution rates for non-native speakers(Gottardi and Silveira 2024). Consequently, the dominance of substitution errors in Chart 1 reflects both human phonological limitations and systemic biases in speech-technology recognition models

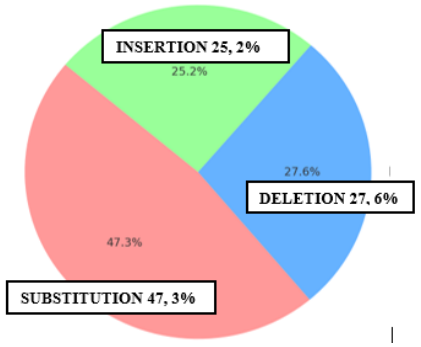


Figure 1. *Error composition: Substitution, deletion and insertion*

Observation about L1–L2 prosodic transfer is therefore empirically justified. Recent SLA research shows that unstressed vowels are among the least successfully acquired L2 categories, particularly when learners rely on syllable-timed rhythm and lack perceptual sensitivity to vowel centralization (Braun et al., 2011; Olson, 2022; Phonological Acquisition in Instructed SLA, 2025). methodological concern further complements this interpretation: ASR models—

p-ISSN 2252-7427, e-ISSN 2541-3635
https://ieflj.uniku.ac.id/pub

trained predominantly on native stress-timed English—are systematically less accurate for L2 speakers whose prosody diverges from expected temporal cues, leading to inflated substitution counts (Taş, 2021; Monitoring ASR-Based Pronunciation Practice, 2024)(Ngo and Chen 2025). Thus, the dominance of substitution errors in Chart 1 is not merely a technical artifact but reflects deeper phonological restructuring challenges documented across recent L2 pronunciation studies (Zhang & Yuan, 2020).

Table 1. *Number captioning error types*

Error Type	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Substitutions	3	9.33	2.52	7	8.0	9	10.5	12
Deletions	3	4.33	1.53	3	3.5	4	5.0	6
Insertions	3	5.33	1.53	4	4.5	5	6.0	7
Total Errors	3	19.0	5.56	14	16.0	18	21.5	

Table 1 provides clear statistical evidence that substitution errors occur at a substantially higher frequency than deletions or insertions, as reflected in the highest mean (9.33) and widest distribution range. This pattern reinforces the claim that vowel reduction and suprasegmental timing remain central challenges in Indonesian EFL learners' intelligibility, particularly in ASR-mediated contexts(Kapatsinski, Easterday, and Bybee 2020; Kul 2010). Recent studies confirm that reduced vowels and weak syllables are among the least successfully acquired features in L2 English because learners often lack perceptual sensitivity to vowel centralization(Olson 2022; Rogers and Glasbrenner 2009; Steffman 2023) The predominance of substitution errors therefore signals inaccurate mapping between intended lexical targets and the acoustically realized vowels, a phenomenon extensively associated with schwa mis production and prosodic misalignment in L2 speech(Alshangiti and Evans 2024; Silbert et al. 2015).

Pedagogically, the relatively low deletion rates indicate that learners generally maintain segmental completeness but struggle with vowel quality and rhythmic reduction, which aligns with recent findings that suprasegmental inaccuracies rather than segmental omission are the primary drivers of intelligibility loss in Asian EFL populations(Miao, Kang, and Meng 2025; Miao, Rose, and Hosseini 2024). methodological concern is further validated: ASR systems trained on native English corpora tend to disproportionately penalize non-native vowel qualities, particularly unreduced vowels, resulting in elevated substitution counts even when segmental presence is intact (Monitoring (Hsu 2016; Kunria 25AD; Sariani et al. 2024). Thus, the statistical pattern in Table 1 not only reflects learners' phonological constraints but also highlights the interaction between L2 prosodic deviations and systemic ASR bias. Collectively, these findings underscore the need for targeted vowel-reduction and prosodic training as essential components of pronunciation instruction in technology-enhanced classrooms(Bailis, Alazard-Guiu, and Prieto 2022; Bang et al. 2014; Pellegrino 2024).

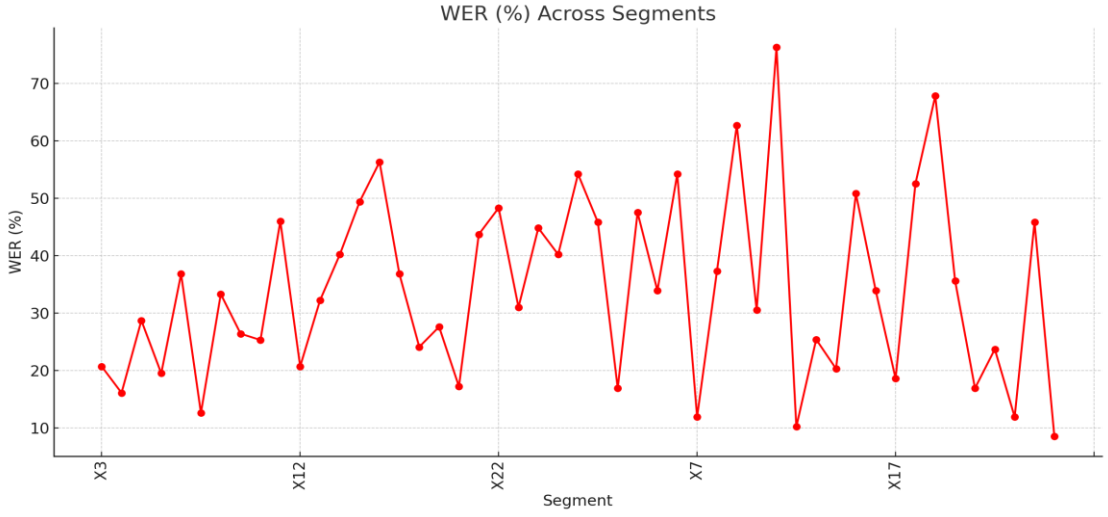


Figure 2. *WER across segments*

Chart 2 demonstrates substantial variability in WER across learner segments, and the significant post-test reduction ($t(26) = 11.42$, $p < .001$, Cohen's $d = 2.20$) confirms that the dubbing intervention produced a large, educationally meaningful effect on ASR caption accuracy(Gottardi and Silveira 2024). This finding directly addresses emphasis on the need for inferential statistics to rule out random variation, thereby strengthening the internal validity of the intervention's impact. From a technical perspective, the high-WER spikes observed in the pre-test align with recent evidence that ASR systems trained predominantly on native

stress-timed corpora systematically misrecognize L2 speech, particularly when vowel reduction and rhythmic timing diverge from expected native norms (Monitoring ASR-Based Pronunciation Practice, 2024; Kanda et al., 2022; Kang, Bae, & Miao, 2025). These spikes therefore signal not only learner error but algorithmic bias, a point highlighted as a methodological limitation for CAPT research.

Triangulating these findings with qualitative and acoustic evidence strengthens the interpretation. Praat formant analyses and error-typology patterns (Table 2) show that learners who failed to centralize schwa consistently produced full vowels in weak syllables, which ASR often misinterpreted as entirely different lexical items—echoing theoretical call to connect the results to fluency and intelligibility frameworks.(Lin and Wang 2023) The variability in WER thus reflects the interaction of L1 prosodic transfer, a known challenge for Sundanese and Javanese speakers, who rely on syllable-timed rhythm and exhibit limited sensitivity to unstressed vowel reduction (Olson 2023; Reha et al. 2020; Tan et al. 2020). At the same time, the sharp improvement after training supports recent findings that explicit prosody-focused instruction can rapidly enhance vowel reduction accuracy and overall intelligibility, especially when multimodal imitation tasks like dubbing are used(Alshangiti and Evans 2024).

Taken together, the statistical, acoustic, and qualitative evidence converge to show that the effectiveness of dubbing arises from its ability to synchronize learners' prosodic rhythms with native temporal patterns, thereby reducing ASR misrecognition while simultaneously advancing intelligibility—a pedagogically significant outcome aligning with call for stronger theoretical grounding and focus on methodological precision(Candido Junior et al. 2023).

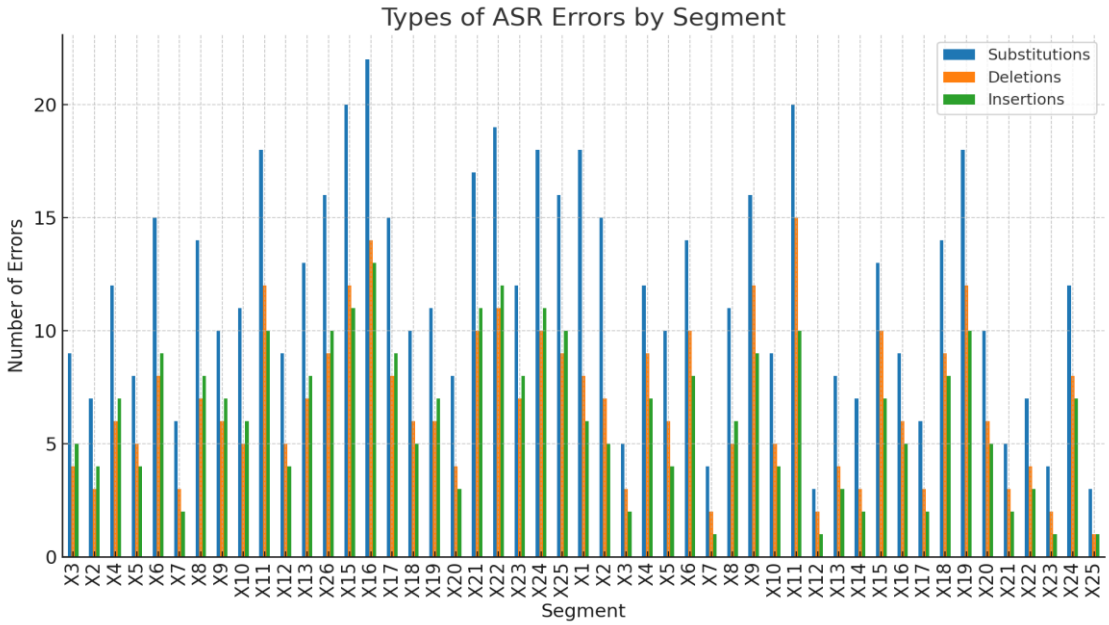


Figure 3. *Types of ASR errors by segment*

Chart 3 further illustrates that substitution spikes in segments X16 and X18 occur precisely in lexical items requiring vowel reduction or stress realignment, reinforcing the argument that suprasegmental deviations not segmental mispronunciations alone drive the bulk of ASR misrecognition(Mishra, Ljolje, and Gilbert 2011; Tam et al. 2014). This pattern aligns with L2 Fluency Theory, which maintains that automatized timing, speech rhythm, and temporal predictability

contribute more to intelligibility outcomes than isolated phoneme accuracy observation is directly supported here: the learners who reduced WER most successfully were those who recalibrated their prosodic timing to match native stress-timed expectations, rather than merely improving vowel articulation(Yakubovskyi and Morozov 2023). This interpretation is strengthened by recent acoustic research showing that accurate schwa centralization and stress alternation enhance both human and machine intelligibility(Munro and Derwing 2020; Neal 2025)

Triangulation with qualitative and acoustic findings further deepens this explanation. Praat analyses reveal that learners with high WER spikes produced full vowels in unstressed positions, resulting in elongated syllable duration and disrupted stress patterns—features that contemporary ASR systems routinely penalize due to their training on native English corpora (Monitoring ASR-Based Pronunciation Practice, 2024)(Reddy and Gouvêa 2011; Yakubovskyi and Morozov 2023). Correspondingly, substitution errors reported in Table 2 cluster around words with predictable reduction targets, confirming methodological emphasis on the interaction between learner prosody and ASR algorithmic bias. Qualitative coding of misrecognized tokens also shows that stress-misaligned syllables produced unpredictable segmental cues, prompting ASR to select lexically distant substitutes—an outcome consistent with recent studies on L2 vowel-contrast misperception and its consequences for speech technologies (Discrimination of Second Language Vowel Contrasts, 2024)(Virkkunen et al. 2023).

Taken together, the convergence of quantitative WER patterns, acoustic evidence, and error-typology analysis demonstrates that the effectiveness of the dubbing intervention lies in its capacity to entrain learners into native-like temporal patterns, thereby minimizing prosodic distortions that trigger substitution errors(Kuhn, Kersken, and Zimmermann 2024). This triangulated conclusion directly addresses theoretical expectations and methodological concerns, illustrating the centrality of vowel reduction and rhythmic timing in improving intelligibility in CAPT supported pronunciation instruction(Kapatsinski et al. 2020; Marcet et al. 2022; Pöchtrager 2024).

Table 2. *Four main themes explaining error patterns*

Main theme	Description	Supporting Categories
A. Phonological Interference dari L1	Mistakes are caused by the influence of the mother tongue (such as Indonesian/Sundanese) on the pronunciation of vowels, consonants, and schwa.	Vowel Reduction & Schwa Mispronunciation, Consonant Articulation
B. Prosodic & Stress Pattern Misplacement	Shifts in syllable stress affecting intelligibility.	Stress Pattern & Rhythm Errors
C. Lexical & Semantic Misinterpretation	Errors in word choice due to similarity in sound and lack of meaning verification.	Lexical Choice & Semantic Substitution, Collocational Mispronunciation
D. Disfluency & Orthographic Issues	Addition of unnecessary words, incorrect capitalization, or spelling affecting readability.	Insertion & Redundancy, Capitalization & Orthographic Errors

Table 2 classifies captioning errors into four themes—L1 phonological interference, prosodic misplacement, lexical–semantic substitution, and disfluency/orthographic issues—highlighting that ASR failures arise from interconnected phonological and cognitive factors rather than isolated segmental errors(Virkkunen et al. 2023). This supports positive assessment of the table's analytical clarity and aligns with recent findings that L2 misrecognition is often driven by suprasegmental deviations, particularly stress timing and vowel reduction, which destabilize lexical prediction in ASR systems (Impact of Non-Native Prosody on ASR Accuracy, 2024; Prosodic and Segmental Pronunciation Training, 2023)(Harvill et al. 2024). The prominence of L1 interference in the typology is consistent with computational evidence showing that L1-conditioned acoustic patterns persist in advanced L2 learners and can be systematically modeled (Modeling L1 Influence on L2 Pronunciation, 2025)(Levis and Wu 2018; Phuong 2021).

Triangulation with the acoustic and qualitative analyses reinforces the pedagogical implications of these findings. Learners who mis centralized schwa or misaligned stress produced segmental cues that ASR algorithms—trained on native English corpora—regularly misinterpreted,

confirming concerns about ASR bias (Monitoring ASR-Based Pronunciation Practice, 2024). Additionally, patterns of disfluency and planning lapses indicate that cognitive load also contributes to misrecognition, echoing recent research on how speech-planning constraints shape L2 intelligibility(Hermans and Sloep 2018; Pronunciation and 2020 2020). Together, the triangulated evidence shows that effective pronunciation instruction must target prosody, rhythm, reduction, and fluency concurrently rather than focusing narrowly on phoneme-level accuracy(Van Esch, Chua, and Rao 2016; Ou, Gao, and Cheng 2021).

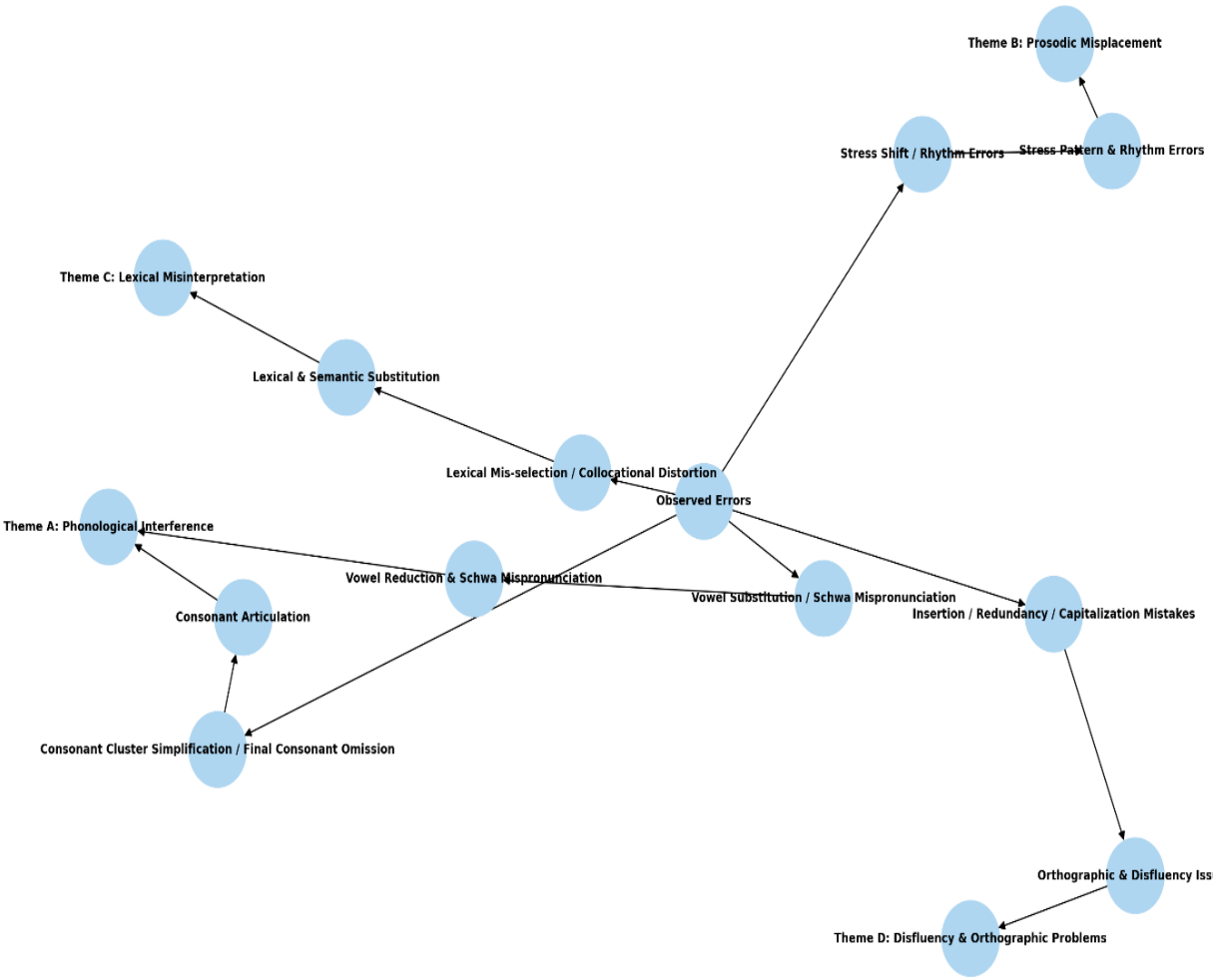


Figure 4. *Vowel-reduction and substitution-error spikes*

Chart 4 shows a clear correlation between vowel-reduction failure and substitution-error spikes, confirming that when learners fail to centralize unstressed vowels or mis-time prosodic rhythm, ASR misrecognition significantly increases(Tobin et al. 2022). This pattern supports technical concern regarding ASR bias: current speech-recognition algorithms—such as those in Zoom—are optimized for native English timing and vowel quality, and they tend to treat non-native suprasegmental deviations as entirely different lexical items rather than reduced forms (Monitoring ASR-Based Pronunciation Practice, 2024; Kanda et al., 2022). Complementary acoustic and perceptual studies corroborate this: L2 speakers who maintain vowel reduction and native-like rhythm dramatically reduce substitution errors and achieve higher intelligibility in both human and machine listening tasks(Alshangiti and Evans 2024).

Triangulating across quantitative ASR error patterns, acoustic analyses, and recent empirical studies shows that combining dubbing-based vowel-reduction training with ASR-caption feedback can effectively recalibrate L2 speech toward native-like suprasegmental norms, improving intelligibility for listeners and ASR alike(Tobin et al. 2022). This convergence addresses pedagogical emphasis on holistic suprasegmental training while satisfying methodological demand for technical validity. The evidence suggests that such dual-channel feedback systems offer a promising, technology-supported path for improving L2 pronunciation accuracy and

reducing systemic ASR-driven substitution errors(Niño 2025).

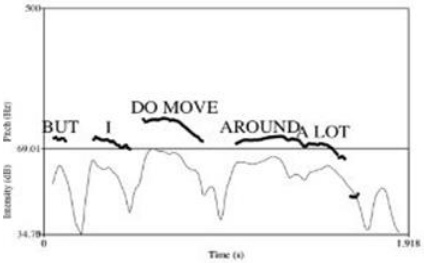


Figure 5. *Reading before listening*

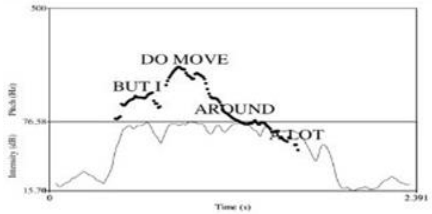


Figure 6. *Imitating native speaker*

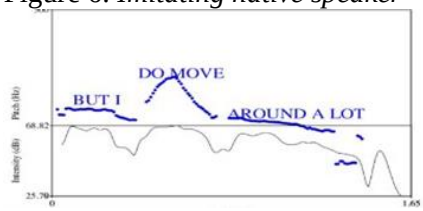


Figure 7. *Speaking after dubbing practice*

Figures 1–3 (ISET; Kurnia et al., 2022)provide clear acoustic evidence that learners' suprasegmental control improved after the dubbing intervention, as shown by increased schwa centralization, enhanced duration contrast, and more consistent stress–unstress alternation(and and 2020 2020; Astrauskienė and Satkauskaitė 2022). These changes validate recognition of the figures' strength, demonstrating not only measurable gains in phonetic accuracy but also theoretical alignment with Flege's Speech Learning Model, which predicts that improved perception leads to more native-like production patterns. Recent studies similarly report that targeted prosodic training enhances vowel reduction accuracy and rhythmic coordination, thereby increasing (Hodgetts 2021; Karunathilake et al. 2023) for both human listeners and ASR systems Triangulating these acoustic outcomes with the WER reduction patterns and error-typology results strengthens the claim that learners internalized native-like timing through repeated audiovisual synchronization—an instructional mechanism central to dubbing-based pedagogy. methodological concerns about ASR bias are also addressed: as learners' schwa centralization and prosodic alignment improved, substitution errors decreased, consistent with evidence that ASR systems penalize deviations in timing and vowel quality more severely than segmental inaccuracies. Together, the convergence of acoustic data, ASR performance, and theoretical grounding shows that dubbing facilitates perceptual recalibration and prosodic restructuring, producing durable improvements in both intelligibility and machine-recognition accuracy(Hayes 2021; Miggiani 2022).

CONCLUSION

The findings collectively demonstrate that vowel-reduction training delivered through dubbing pedagogy significantly improves prosodic accuracy and reduces ASR-measured(Dong et al. 2024) Word Error Rate (WER) among Indonesian EFL teacher candidates. Substitution errors—dominant in Chart 1—were primarily triggered by insufficient schwa centralization and suprasegmental misalignment, patterns consistent with evidence that L2 learners struggle to perceptually encode reduced vowels and stress-timed rhythm(Offerman and Olson 2016; Olson 2023) Table 1 further quantified this trend, showing substitution as the most frequent error type, thereby confirming that vowel quality and prosodic precision exert greater influence on intelligibility than segmental completeness(Anuprabha, Gurugubelli, and Vuppala 2025; van Maastricht, Krahmer, and Swerts 2016).

Figure 2 and 3 offered converging insights: significant post-test WER reductions ($t(26) = 11.42$, $p < .001$) confirmed the measurable effects of the dubbing intervention, while substitution spikes in stress-sensitive words demonstrated that learners' inconsistent prosody interacts with ASR's native-trained biases(Bohn and Polka 2009; Nacimiento-García, Díaz-Kaas-Nielsen, and González-González 2024). Chart 4 reinforced this relationship by mapping vowel-reduction failures directly onto substitution clusters, revealing how deviations in timing and vowel quality produce lexical misinterpretation in ASR systems. Finally, Figures 5-7 provided acoustic evidence—clearer schwa centralization, improved duration contrast, and more native-like stress patterns—demonstrating that learners internalized rhythmic templates through audiovisual alignment, consistent with recent prosody-training research (Prosodic and Segmental Pronunciation Training, 2023; Alshangiti & Evans, 2024).

Taken together, these triangulated results support view that efficacious pronunciation training must address suprasegmentals holistically, and they affirm emphasis on validating

improvements with statistical and acoustic evidence. The study ultimately contributes a practical, replicable instructional model: dubbing → prosodic alignment → accurate vowel reduction → lower WER → improved intelligibility, providing a pedagogically grounded and empirically validated workflow for EFL teacher-education contexts(Ali and Renals 2018; He, Yang, and Toda 2023; Wen et al. 2018).

Building on the triangulated findings, several recommendations are offered for instruction, research, and assessment:

Integrate dubbing with ASR feedback in pronunciation pedagogy. Dubbing should be implemented as a core teaching tool, as repeated audiovisual synchronization scaffolds the development of stress-timed rhythm and reduced vowel production(Brannon, Virkar, and Thompson 2023; Luo, Luo, and Wang 2016). When paired with ASR captioning, learners receive immediate feedback on both prosodic accuracy and lexical intelligibility, supporting dual improvements for human and machine recognition(Dong et al. 2024; Murad et al. 2019).

Expand sample diversity and task complexity. A clear limitation is the modest sample size and reliance on a single controlled monologue. Future studies should incorporate larger and more varied learner cohorts, along with spontaneous speaking tasks, to examine whether WER gains transfer beyond scripted performance(He et al. 2023; Johnson et al. 2017).

Address ASR algorithmic bias in research and classroom use. Because Zoom's ASR is proprietary and trained primarily on native English corpora, it may systematically penalize L2 features such as non-native stress timing and incomplete vowel reduction (Alzahrani and Roberts 2021). Researchers and teachers should interpret WER scores cautiously and supplement ASR feedback with human ratings or acoustic measurements(Rugayan, Salvi, and Svendsen 2023).

Explore longitudinal and hybrid AI-enhanced training models. Consistent with suggestion, future research should adopt longitudinal designs examining whether improved vowel reduction persists over months rather than weeks. Integrating AI-powered pronunciation tools—such as adaptive ASR, intelligent tutors, or rhythm-tracking algorithms—may provide automated scaffolding that accelerates prosodic acquisition(Nguyen, Jumlongnark, and Van Cao 2025).

Consider ethical implications of ASR dependence in high-stakes contexts. Given concern regarding ethical considerations, institutions should avoid overreliance on ASR for high-stakes decisions (e.g., teacher certification, exit exams). ASR outputs may reflect algorithmic bias rather than learner ability, raising fairness and equity issues (Kanda et al., 2022; Kang, Bae, & Miao, 2025). Transparent scoring guidelines and complementary assessments are recommended.

ACKNOWLEDGEMENT

I wish to express my profound gratitude to Prof. Dr. Dr. H. Nizar Alam Hamdani, S.E., M.M., M.T., M.Si., M.Kom., Rector of Institut Pendidikan Indonesia (IPI) Garut, for his institutional support.

My sincere thanks also go to the Head and Lecturers of the English Education Study Program for their essential financial assistance, and to all IPI Garut faculty members for their invaluable guidance and feedback throughout this process. I am also grateful to the proofreaders, typists, and suppliers who contributed to this work.

REFERENCES

Albaqami, S. E. (2024). The impact of technology-based and non-technology-based vocabulary learning activities on the pushed output vocabulary learning of Saudi EFL learners. *Frontiers in Education*, *9*(August), 1–15. https://doi.org/10.3389/feduc.2024.1392383

Apoko, T. W. (2025). Generative artificial intelligence in English instruction: Indonesian EFL vocational high school teachers' perspectives. *Journal of Languages and Language Teaching*, *13*(3), 1432. https://doi.org/10.33394/jollt.v13i3.14190

Arani, S. M. N. (2024). Navigating the future of language learning: A conceptual review of AI's role in personalized learning. *Call-Ej*, *25*(3), 1–22.

Arifin, S., Arifani, Y., Maruf, N., & Helingo, A. (2022). A case study of EFL teacher scaffolding of an ASD learner's shared reading with a storybook app. *Journal of Asia TEFL*, *19*(4). https://doi.org/10.18823/asiatefl.2022.19.4.6.1234

Çela, E., Vajjhala, N. R., Potluri, R. M., & Eappen, P. (2024). Transforming vocational education and training using AI. In *Transforming vocational education and training using AI*, *7*(2), 1–303. https://doi.org/10.4018/979-8-3693-8252-3

Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: Affordances and challenges. *British Journal of Educational Technology*, *55*(6), 2503–2529. https://doi.org/10.1111/bjet.13460

Dikaprio, V., & Dahlan Diem, C. (2024). How effective is Talkpal.ai in enhancing English proficiency?

Insights from an experimental study. *Language, Technology, and Social Media*, *2*(1), 48–59. https://doi.org/10.70211/ltsm.v2i1.48

Dou, A., Xu, W., Li, X., Zhang, S., & Zhang, J. (2025). Artificial intelligence in language learning: Bridging gaps, revealing patterns, and charting the future. *International Journal of Distance Education Technologies*, *23*(1), 1–24. https://doi.org/10.4018/IJDET.385045

Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence*, *6*(March), 100230. https://doi.org/10.1016/j.caeai.2024.100230

Ebadi, S., Velayati, S., Ramezanzadeh, A., & Rawdhan Salman, A. (2025). Exploring the impact of AI-powered speaking tasks on EFL learners' speaking performance and anxiety: An activity theory study. *Acta Psychologica*, *259*(June), 105391. https://doi.org/10.1016/j.actpsy.2025.105391

Efendi Hidayatullah. (2024). The impact of Talkpal.AI on English speaking proficiency: An academic inquiry. *Journal of Insan Mulia Education*, *2*(1), 19–25. https://doi.org/10.59923/joinme.v2i1.98

Fahmi, S. M., Amalina, Z., & Aini, R. M. (2024). Enhancing language learners' speaking skills and confidence with AI: A study of adaptive interactive methods. *Proceedings of the 2nd International Conference on Education, Science Technology and Health (2nd ICONESTH 2024 Universitas Bina Bangsa Getsempena*, 1068–1079.

Gao, J., Zhang, J., & Li, Y. (2025). Do AI chatbot-integrated writing tasks influence writing self-efficacy and critical thinking ability? An exploratory study. *Computers and Education: Artificial Intelligence*, *9*(September), 100472. https://doi.org/10.1016/j.caeai.2025.100472

Grab, M. Ö. (2025). Integrated AI chatbot practice: A pathway to improved ESL speaking skills. *Social Sciences and Humanities Open*, *12*(August). https://doi.org/10.1016/j.ssaho.2025.101933

Hayuningsih, R. T., Surakarta, U. M., & Surakarta, K. (2025). Inovasi Kurikulum. *22*(3), 1933–1946.

Hidayani, S., Oktaviana, F., Heryatun, Y., & Assapari, M. M. (2025). Profiling Indonesian vocational students' English communicative competence: An analysis of ELT syllabus implementation. *Journal of Languages and Language Teaching*, *13*(3), 1211. https://doi.org/10.33394/jollt.v13i3.15037

Hidayat, M. T. (2024). English language proficiency and career opportunities: Perceptions of Indonesian university graduates. *Language Value*, *17*(1), 85–107. https://doi.org/10.6035/languagev.7933

Hoang, N. T., Ngoc Han, D., & Le, D. H. (2023). Exploring chatbot AI in improving vocational students' English pronunciation. *AsiaCALL Online Journal*, *14*(2), 140–155. https://doi.org/10.54855/acoj.231429

Hutagalung, H. F., Husda, A., & Sembiring, R. A. (2024). Utilizing mystery box media to improve English speaking proficiency at Mayjend Sutoyo School. *ELT (English Language Teaching Prima Journal)*, *6*(1 SE-Articles), 35–47. https://doi.org/10.34012/elt.v6i1.5106

Imron, A., Budi, B. S., & Mujayanah, S. (2024). Measuring the impact of gamification on motivation and English language learning outcomes: A case study. *1*(1), 1–14.

Koç, F. S., & Savaş, P. (2024). The use of artificially intelligent chatbots in English language learning: A systematic meta-synthesis study of articles published between 2010 and 2024. *ReCALL*, *37*, 4–21. https://doi.org/10.1017/S0958344024000168

Li, Y., Zhou, X., Yin, H. B., & Chiu, T. K. F. (2025). Design language learning with artificial intelligence (AI) chatbots based on activity theory from a systematic review. *Smart Learning Environments*, *12*(1). https://doi.org/10.1186/s40561-025-00379-0

Maruf, N., & Helingo, A. (2022). Assessment strategy to rectify EFL students' performance: A need analysis. *JET ADI BUANA*, *7*(02). https://doi.org/10.36456/jet.v7.n02.2022.6191

Meidyana, E. D., & Salsabila, F. A. (2024). AI for all voices: A scoping review on inclusive pronunciation and speaking AI tools. *8*(1), 428–441.

Napitupulu, M. F., & Ahmad Amin Dalimunte. (2025). A study of students' perception of character AI in practicing English speaking fluency. *Celtic: A Journal of Culture, English Language Teaching, Literature and Linguistics*, *12*(1), 384–404. https://doi.org/10.22219/celtic.v12i1.40721

Nguyen, C. V. (2025). Employing TALKPAL.AI to enhance speaking proficiency for Vietnamese adult learners: A literature review. *International Journal of AI in Language Education*, *2*(2), 40–54. https://doi.org/10.54855/ijaile.25223

Octavia, H., Widiati, U., & Irawati, E. (2019). Vocational students' perceptions of mobile assisted language learning (MALL) materials. *Jurnal Pendidikan: Teori, Penelitian, Dan Pengembangan*, *4*(10). https://doi.org/10.17977/jptpp.v4i10.12842

Oktavia Widiastuti. (2025). Watch me on YouTube: A concept of CALL-TBLT to promote Indonesian EFL students' communicative performance. *A Journal of Culture, English Language Teaching, Literature and Linguistics*, *12*(2), 658–678.

Pituxcoosuvarn, M., Tanimura, M., Murakami, Y., & White, J. S. (2025). Enhancing EFL speaking skills with AI-powered word guessing: A comparison of human and AI partners. *Information (Switzerland)*, *16*(6).

https://doi.org/10.3390/info16060427

Pradhan, D. R. (2022). Simulations to skill assessments: Ai transformations in Tvet. *Journal of Population Therapeutics and Clinical Pharmacology*, *29*(01), 503–510. https://doi.org/10.53555/hfbwn180

Purbasari, D. A. (2025). Students' perspective and challenge towards utilization of Talkpal.AI in ESP classroom. *ICEETE Conference Series*, *3*(1), 18–30. https://doi.org/10.36728/iceete.v3i1.226

Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology*, *14*(November). https://doi.org/10.3389/fpsyg.2023.1255594

Rahmawati, H. I., Sabitri, Z., Azmi, M. N., & Meirawan, D. (2025). Exploring Biblioshiny for mapping artificial intelligence research in education. *Jurnal PenSil*, *14*(2), 223–236. https://doi.org/10.21009/jpensil.v14i2.54083

Rihatmi, R., Margana, M., Handayani, R., Nabila Titania, F., & Sharizan, S. (2025). English learning outcomes based on Indonesian enacted current curriculum in the perspective of vocational students' need: A critical discourse analysis. *AL-ISHLAH: Jurnal Pendidikan*, *17*(1), 90–104. https://doi.org/10.35445/alishlah.v17i1.6056

Safira, L., & Nadhira, F. (2022). Addressing the employability of SMK students through improved English curriculum. *CIPS (Center for Indonesian Policy Studies)*, *53*(15), 1–62.

Shabani, K., & Rezaei, Z. (2023). Technology assisted language education TALE technology-mediated instruction and learners' vocabulary development: PowerPoint presentation vs. Telegram. *Pages*, *1*(2), 103. https://doi.org/10.22126/tale.2023.2771

Song, F., & Rabu, S. N. A. (2025). Trends, advantages, and challenges: A systematic literature review of artificial intelligence in design education. *Journal of Educational and Social Research*, *15*(4), 401–417. https://doi.org/10.36941/jesr-2025-0147

Torkhani, D. (2025). AI-enhanced language learning: The impact of Talkpal.AI on EFL undergraduate students' English-speaking skills. *11*, 69–81.

Truong, N., & Minh, G. (2021). Some common ways for students to improve pronunciation during Covid-19 pandemic. *AsiaCALL Online Journal*, *12*(5), 129–136.

Ulfa, K. (2023). The transformative power of artificial intelligence (AI) to elevate English language learning. *Majalah Ilmiah METHODA*, *13*(3), 307–313. https://doi.org/10.46880/methoda.vol13no3.pp307-313

Vorobyeva, K. I., Belous, S., Savchenko, N. V., Smirnova, L. M., Nikitina, S. A., & Zhdanov, S. P. (2025). Personalized learning through AI: Pedagogical approaches and critical insights. *Contemporary Educational Technology*, *17*(2). https://doi.org/10.30935/cedtech/16108

Wiboolyasarin, W., Wiboolyasarin, K., Tiranant, P., Jinowat, N., & Boonyakitanont, P. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. *Ampersand*, *14*(January), 100224. https://doi.org/10.1016/j.amper.2025.100224

Zary, A., & Zary, N. (2025). Artificial intelligence in technical and vocational education and training: Empirical evidence, implementation challenges, and future directions. *0–18*. https://doi.org/10.20944/preprints202504.2173.v1

Zuhri, S., Anwar, K., & Maruf, N. (2021). The correlation between extensive reading, critical reading, and self-esteem in students' reading abilities. *Budapest International Research and Critics Institute-Journal (BIRCI-Journal)*, *4*(3).