CHAPTER II
LITERATURE REVIEW AND CONCEPTUAL FRAMEWORK

A. Literature Review

1. Theoretical Review

a. Speaking Fluency in EFL Contexts

According to David Nunan (1995), speaking is a productive aural and oral skill that requires learners to produce systematic verbal utterances in order to communicate meaning in real time. Because the speaker must formulate ideas, select vocabulary, and construct grammatical structures as the communication unfolds, speaking becomes one of the most cognitively demanding language skills. Nation (2011) describes fluency as the ability to speak at a natural and steady rate with minimal hesitation, and he argues that fluent performance depends on efficient lexical retrieval and rapid formulation of utterances so that communication can proceed without unnecessary interruptions. Peter Skehan (2009) identifies fluency as one of the three principal dimensions of oral performance, alongside complexity and accuracy. In his framework, fluency represents the temporal aspect of speech production and reflects the speaker's capacity to manage processing demands while maintaining continuity of discourse. In this view, fluency is not simply fast speech. Instead, it is the speaker's ability to sustain delivery through efficient attention management and through the automatization of linguistic resources.

Norman Segalowitz (2010) expands the discussion of fluency by distinguishing between cognitive fluency, utterance fluency, and perceived fluency. Cognitive fluency refers to the speed and efficiency of internal mental processes involved in language production. Utterance fluency refers to the measurable temporal characteristics of speech, such as pauses, repetitions, and speech rate. Perceived fluency refers to the listener's overall impression of how smoothly the speech is delivered. Segalowitz explains that disfluency in the form of long pauses, filled pauses, false starts, or uneven timing often signals difficulties in cognitive processing or an overload in attentional resources. These disruptions occur because many L2 learners have not yet automatized essential components of language production, which forces them to allocate substantial attention to lexical retrieval

or syntactic planning. When attention is heavily consumed by these processes, the temporal flow of speech becomes irregular.

An additional implication of this framework is that fluency development is closely tied to practice conditions that promote automatization. Repeated exposure to meaningful communication, time pressure, and opportunities for spontaneous production are necessary for reducing cognitive load during speech. Without such conditions, learners tend to rely on controlled processing rather than automatic processing, which slows down production and increases hesitation. This is particularly relevant in EFL contexts, where input and interaction are often limited compared to second language environments. In EFL settings, where opportunities for extended speaking practice are limited and where accuracy-oriented instruction and affective pressures frequently increase cognitive demands, learners often struggle to internalize the automatic processing required for fluent, sustained speech.

b. Technology and AI in Language Learning

Technology-Enhanced Language Learning (TELL) emphasizes how digital tools can extend learners' opportunities to practice language skills in flexible, authentic, and interactive ways (Chapelle, 2003). Within this framework, technology is not merely a medium for content delivery but a facilitator of meaningful interaction, allowing learners to engage with language beyond the constraints of traditional classroom environments. The integration of multimedia resources, online communication platforms, and adaptive systems has enabled learners to access input, produce output, and receive feedback in increasingly individualized ways.

Among recent advancements, Artificial Intelligence (AI) has played a central role in enabling systems capable of understanding and generating human-like language. AI in language learning typically involves machine learning algorithms that analyze linguistic input and produce contextually appropriate responses or feedback. Applications such as virtual tutors, speech recognition systems, and chatbots have been shown to enhance pronunciation, vocabulary acquisition, and communicative competence by providing immediate, scalable, and

often personalized interaction (Kukulska-Hulme & Shield, 2008). These systems reduce dependency on human instructors for routine practice, thereby increasing the frequency and accessibility of language exposure.

AI-driven speech recognition tools, in particular, have contributed significantly to speaking practice. These systems analyze learners' spoken input and provide automatic feedback on pronunciation, fluency, and accuracy, allowing learners to identify and correct errors in real time. A meta-analysis by Ngo, Chen, and Lai (2023) indicates that such technologies can lead to measurable improvements in pronunciation and speaking performance, especially when feedback is immediate and iterative. This aligns with established theories of language learning that emphasize the importance of timely corrective feedback in skill acquisition.

However, earlier AI-based systems were limited in several respects. Many operated on rule-based or narrowly trained models, restricting interactions to scripted dialogues and predefined responses. As a result, learners were often exposed to predictable conversational patterns that lacked the variability and unpredictability of real-world communication. Additionally, these systems typically focused on controlled pronunciation tasks rather than spontaneous speech production, limiting their effectiveness in developing communicative fluency. Emotional engagement was also minimal, as interactions were often mechanical and lacked the natural flow of human conversation. These limitations highlight the gap between earlier AI applications and more recent developments in large language models, which aim to provide more dynamic and context-sensitive interaction.

c. ChatGPT Voice Mode in Speaking Practice

To contextualize the role of ChatGPT Voice Mode, it is necessary to first understand the underlying technology. Recent developments in artificial intelligence have led to the emergence of Large Language Models (LLMs), which are computational models trained on large-scale text corpora to generate human-like language. One prominent example is ChatGPT, which is built upon the Generative Pre-trained Transformer (GPT) architecture. Models such as GPT-4

operate by predicting the most probable sequence of words based on contextual input, allowing them to produce coherent and contextually appropriate responses across a wide range of tasks (OpenAI, 2023). Unlike earlier rule-based systems or narrow AI applications, LLMs rely on deep neural networks and probabilistic modeling, enabling them to generalize across topics and maintain extended interactions. This shift from deterministic systems to generative models marks a significant advancement in how machines process and produce language, particularly in conversational settings.

In addition, LLMs are designed to handle contextual dependencies over extended discourse, meaning they can maintain coherence across multiple turns of interaction. This capability is particularly relevant for language learning, where sustained dialogue and contextual understanding are essential for developing communicative competence. The scalability of LLMs also allows them to be continuously improved through training on diverse datasets, resulting in increasingly natural and flexible language use. However, their outputs are still probabilistic rather than rule-governed, which means responses may vary and are not always guaranteed to be accurate. This characteristic should be considered when integrating such tools into educational settings.

Within educational contexts, LLMs have increasingly been recognized as tools that can support personalized and interactive learning. Kasneci et al. (2023) argue that systems like ChatGPT can facilitate adaptive dialogue, immediate feedback, and extended learner engagement, all of which are critical components of effective language learning environments. Unlike traditional computer-assisted language learning (CALL) tools that often rely on fixed exercises, LLM-based systems allow for dynamic and context-sensitive interaction, enabling learners to engage in more authentic communicative practices. Furthermore, the accessibility of such systems lowers barriers to practice, as learners can interact with AI at any time without the constraints of classroom schedules or availability of human interlocutors.

ChatGPT Voice Mode represents a significant innovation in AI-assisted language learning. It enables learners to engage in real-time spoken dialogue with

a natural conversational flow, simulating authentic human interaction. Unlike traditional language applications that focus primarily on pronunciation drills, ChatGPT Voice supports spontaneous communication, helping learners to process language in real time and respond to unpredictable input. This aligns with the Interaction Hypothesis (Long, 1996) and Output Hypothesis (Swain, 1985), which highlight the role of meaningful interaction and language production in language development. The tool also provides a low-pressure environment that reduces anxiety and promotes confidence, which are conditions conducive to second language acquisition as explained by Krashen's (1982) Affective Filter Hypothesis. Although ChatGPT Voice does not yet offer detailed phonetic scoring like specialized ASR systems, its capacity for natural interaction offers valuable opportunities for fluency-oriented practice.

d. Supporting Theories

Several theoretical frameworks support the use of ChatGPT Voice Mode to enhance speaking fluency. Long's (1996) Interaction Hypothesis suggests that language acquisition occurs through negotiation of meaning during communication. Swain's (1985) Output Hypothesis posits that producing language encourages learners to notice linguistic gaps and refine their output, promoting fluency. Segalowitz's (2010) Automaticity Theory emphasizes repeated, meaningful practice as the basis for fluent performance. Together, these theories imply that continuous verbal interaction with ChatGPT Voice can facilitate automatization and fluency improvement. In addition, Krashen's (1982) Affective Filter Hypothesis highlights the importance of reducing anxiety and promoting motivation, which are two affective factors addressed by the supportive and judgment-free nature of AI-based voice interaction.

2. Previous Studies

Han (2020), in a controlled experimental study examining the effects of a voice-based AI chatbot on Korean EFL learners, found that ten weeks of structured voice-chat sessions yielded measurable gains across several speaking components, including pronunciation, language use, and overall speaking test performance. Importantly for the present research, Han's analysis indicated a statistically significant improvement in

fluency-related indicators within the speaking rubric, suggesting that sustained spoken interaction with an AI agent can create conditions conducive to more automatic speech production. Safitri, Hidayati, and Ciptaningrum (2025), in a systematic review synthesizing empirical studies on AI chatbots in EFL contexts, also identified recurring evidence of increased learner engagement, willingness to communicate, and reduced speaking anxiety. However, their review emphasized that despite positive effects on general speaking proficiency, most included studies operationalised "speaking improvement" broadly and rarely incorporated standardized fluency metrics such as speech rate or pause analysis. This gap in methodological rigor raises concerns about the extent to which reported gains can be attributed specifically to enhanced fluency rather than to holistic performance improvements. Wardat and Akour (2025), investigating AI-supported tools in Jordanian university settings, further demonstrated that while AI integration can alleviate academic pressure and foster comfort during speaking tasks, reductions in fear of negative evaluation were inconsistent. Their findings imply that improvements in affective variables do not always translate into more fluent or spontaneous oral production, underscoring the need for studies that measure fluency directly rather than infer it from affective changes alone.

Salsabil, Rakhmawati, and Febtwenesty (2025), focusing on Indonesian upper-secondary EFL students using ChatGPT in voice-chat mode, reported learner-perceived gains in fluency, vocabulary, grammar, and confidence, attributing these results to increased opportunities for extended oral turns and reduced interpersonal pressure when speaking to a non-evaluative AI interlocutor. Nevertheless, they noted that cognitive load and technological dependency posed challenges for some learners, which may influence fluency performance in unpredictable ways. Pituxcoosuvarn, Tanimura, Murakami, and White (2025), through a comparative, game-based experimental design, examined how AI versus human speaking partners affected learners' fluency development during word-guessing tasks. Their findings indicated that AI partners can support incremental fluency gains under structured conditions because they provide consistent pacing, predictable turn-taking, and immediate feedback. However, their study also highlighted that such improvements may be task-specific and not necessarily generalizable to broader communicative contexts. Pratiwi et al. (2024), in a conceptual review of ChatGPT's voice-conversation features, similarly noted strong pedagogical potential for enhancing speaking

practice but warned that evidence on fluency remains preliminary, fragmented, and often based on learner perceptions rather than on quantitative fluency measures. Shahab et al. (2024) also contributed relevant evidence by showing that AI-powered pronunciation tools can improve aspects of speaking performance; yet their findings concentrate on segmental and suprasegmental accuracy rather than fluency, further illustrating the disproportionate emphasis on pronunciation in current AI-speaking research. In conclusion, although existing studies collectively demonstrate that AI-mediated spoken interaction can enhance certain components of EFL speaking empirical research providing robust, quantitative evidence of fluency improvement remains limited, particularly confidence, pronunciation, and willingness to communicate. Few studies employ standardized fluency metrics, longitudinal designs, or voice-mode AI systems capable of natural spoken interaction, thereby establishing a clear research gap for examining the fluency-specific impact of tools such as ChatGPT Voice Mode.

B. Conceptual Framework

The conceptual framework of this study is built upon the idea that regular and meaningful spoken interaction can significantly improve learners' fluency in a foreign language. This assumption is grounded in theories such as Long's (1996) Interaction Hypothesis, which posits that language acquisition occurs through active communication and negotiation of meaning, and Swain's (1985) Output Hypothesis, which emphasizes that producing language output helps learners internalize linguistic forms. In addition, Segalowitz's (2010) Automaticity Theory suggests that repeated and meaningful language use enables learners to process speech more efficiently, ultimately leading to fluency. When learners engage with ChatGPT Voice Mode, they participate in an environment that naturally supports these theoretical principles

In this framework, the independent variable (IV) is the use of ChatGPT Voice Mode as an AI-based speaking practice tool. It allows students to interact in real-time conversations that resemble authentic communication, giving them opportunities to produce spontaneous speech, receive immediate responses, and practice language beyond the classroom setting. ChatGPT Voice provides a safe, low-pressure environment that can lower learners' affective filters (Krashen, 1982) and encourage more frequent speaking practice.

The dependent variable (DV) is speaking fluency, which in this study refers to learners' ability to produce speech that is smooth, coherent, and continuous, with minimal hesitation or

unnatural pauses. Speaking fluency can be measured through several indicators, including speech rate, the number and length of pauses, and overall flow and coherence during oral performance. These measurable elements reflect how efficiently learners retrieve and produce language in real time.

The relationship between the two variables is mediated by affective and motivational factors. Regular practice with ChatGPT Voice Mode not only increases the quantity of spoken output but also reduces speaking anxiety and promotes confidence. As learners become more comfortable expressing themselves, their attention shifts from form to meaning, facilitating the development of automaticity. Consequently, more frequent and confident practice is expected to lead to measurable improvements in fluency.

C. Hypothesis Formulation

H_0: There is no significant difference in students' speaking fluency before and after using ChatGPT Voice Mode.

H_1: There is a significant improvement in students' speaking fluency after using ChatGPT Voice Mode.