VNU Journal of Foreign Studies

Journal homepage: https://jfs.ulis.vnu.edu.vn/

# ENHANCING SHORT-TERM MEMORY IN CONSECUTIVE INTERPRETING TRAINING: EVALUATING AI-GENERATED SPEECH SIMULATIONS FOR ENGLISH MAJORS AT THE DIPLOMATIC ACADEMY OF VIETNAM

Dang Phuong Nam, Do Thi Thu Phuong*

*Diplomatic Academy of Vietnam, No. 69 Chua Lang, Lang ward, Hanoi, Vietnam*

Received 26 May 2025
Revised 17 November 2025; Accepted 11 December 2025

**Abstract:** This pilot study explored the use of AI-generated speech to support short-term memory and improving skills in consecutive interpreting (CI) training for English majors at the Diplomatic Academy of Vietnam (DAV). CI demands strong memory and quick mental processing, making it a challenging task for learners. Eight third-year students were divided into two groups: one trained with AI-generated speeches, and the other with traditional human-voiced recordings. Over four weeks, both groups practiced using materials that gradually increased in difficulty by adjusting speech speed, content density, and complexity. The preliminary results proposed that the students using AI speech made greater progress, especially in tasks requiring higher cognitive effort. They also reported that the structured and gradually challenging nature of AI practice helped them feel more confident and focused, despite occasional technical issues with audio quality. Overall, the findings suggest AI-generated speech can be a valuable tool in CI training. It offers consistent and scalable practice that effectively targets key memory functions. However, the students still struggled with note-taking and vocabulary, highlighting the need for ongoing instructor guidance. A combined approach using AI tools along with teacher support is recommended to strengthen interpreting skills.

*Keywords:* consecutive interpreting (CI), short-term memory (STM), AI-generated speech

* Corresponding author.
Email address: dothithuphuong61@gmail.com

https://doi.org/10.63023/2525-2445/jfs.ulis.5530

# TĂNG CƯỜNG TRÍ NHỚ NGẮN HẠN TRONG PHIÊN DỊCH NỐI TIẾP: ĐÁNH GIÁ HIỆU QUẢ CỦA DIỄN VĂN MÔ PHỎNG DO AI TẠO RA ĐỐI VỚI SINH VIÊN NGÔN NGỮ ANH, HỌC VIỆN NGOẠI GIAO

Đặng Phương Nam, Đỗ Thị Thu Phượng

*Học viện Ngoại giao, số 69 Chùa Láng, Phường Láng, Hà Nội, Việt Nam*

Nhận bài ngày 26 tháng 5 năm 2025
Chỉnh sửa ngày 17 tháng 11 năm 2025; Chấp nhận đăng ngày 11 tháng 12 năm 2025

**Tóm tắt:** Nghiên cứu thử nghiệm này khảo sát việc sử dụng bài nói do AI tạo ra nhằm hỗ trợ trí nhớ ngắn hạn và nâng cao kỹ năng trong đào tạo phiên dịch nối tiếp (CI) cho sinh viên ngành Ngôn ngữ Anh tại Học viện Ngoại giao. Phiên dịch nối tiếp đòi hỏi khả năng ghi nhớ tốt và xử lý thông tin nhanh, điều này trở thành một nhiệm vụ đầy thách thức đối với người học. 8 sinh viên năm thứ ba được chia thành hai nhóm: một nhóm luyện tập với bài nói do AI tạo ra, và nhóm còn lại sử dụng các bản ghi âm giọng người truyền thống. Trong bốn tuần, cả hai nhóm thực hành với tài liệu có độ khó tăng dần thông qua điều chỉnh tốc độ nói, mật độ thông tin và mức độ phức tạp của nội dung. Kết quả ban đầu cho thấy nhóm sử dụng bài nói AI đạt tiến bộ lớn hơn, đặc biệt trong các nhiệm vụ đòi hỏi nỗ lực nhận thức cao. Sinh viên cũng cho biết hình thức luyện tập có cấu trúc và tăng dần thử thách của mô phỏng AI giúp họ tự tin và tập trung hơn, dù đôi khi gặp trục trặc kỹ thuật về chất lượng âm thanh. Nhìn chung, các phát hiện cho thấy bài nói do AI tạo ra có thể trở thành công cụ hữu ích trong đào tạo CI, cung cấp nguồn luyện tập ổn định, có khả năng mở rộng và nhắm trúng các chức năng ghi nhớ then chốt. Tuy nhiên, sinh viên vẫn gặp khó khăn về ghi chép và từ vựng, cho thấy cần tiếp tục có sự hỗ trợ của giảng viên. Do đó, nghiên cứu khuyến nghị kết hợp công cụ AI với hướng dẫn của giảng viên để tăng cường toàn diện kỹ năng phiên dịch.

*Từ khóa*: phiên dịch nối tiếp (CI), trí nhớ ngắn hạn (STM), bài nói do AI tạo ra

## 1. Introduction

Consecutive interpreting (CI) requires enormous cognitive processing. That is, the interpreter has to listen to a speech in one language, store the memory traces representing the information, and then accurately render the content into another language. Verbal short-term memory (STM) as well as updating efficiency, included within working memory, contribute significantly to the ability of an interpreter to store information even just one second and execute complex cognitive processes of comprehension, reformulation, and production at the same time. Research evidence has established that training in CI enhances the updating efficiency of working memory (Dong et al., 2018). Improvement is measurable in the performance of students subjected to formal interpreter education.

Despite the promising institutional context, there exists a huge research gap regarding how AI-driven speech simulations may systematically enhance short-term memory of Diplomatic Academy of Vietnam (DAV) English majors undergoing CI training. Some recent studies from Chmiel (2016) and Dong et al. (2018) evidently show that interpreting training improves memory working functions. There is also additional evidence to claim that training related to human-AI interaction can further improve working memory capacities among language professionals. However, there are no previous studies examining how AI-speech simulations can be specifically designed, implemented, and applied to memory enhancement for Vietnamese students studying consecutive interpreting. This gap becomes more important as AI technologies are increasingly integrated into interpreting professions.

The development of speech AI, which can be exemplified by models like the Universal Speech Model (USM) from Google, gives an unprecedented capability to create learning materials that can systematically develop memory functions. These technologies would likely end the existing limitations of interpreter training, such as a lack of dynamic practice materials, insufficient attention to gradual cognitive development, and the absence of customization tailored to individual students. However, only the study that connects cognitive science, interpreting processes, and AI implementation can unlock this potential.

This research also examines the ethical and philosophical aspects of AI integration in interpreter education. As Dastyar and Giustini argue, effective training must not only cover the cognitive dimension but also practical skills, power dynamics, and ethical relations (Dastyar & Giustini, 2024). The development of AI-generated speech simulations, therefore, must extend beyond technologies and cognitive impact to consider broader influence on professional practice among interpreters.

This study aims to explore how AI-simulated speech can be designed, implemented, and evaluated in improving the short-term memory of students training to become interpreters between English and Vietnamese at DAV. The study is framed to develop evidence-based approaches for the advancement of methodology in interpreter training by focusing on specific cognitive functions relevant to CI performance, leveraging current AI speech technologies, and considering the unique institutional context of DAV. The findings would benefit DAV and potentially inform interpreter training approaches globally, particularly where AI literacy is becoming essential for language professional education.

## 2. Aims of the Research

This study aims to investigate how AI-generated speech simulations can enhance short-term memory in consecutive interpreting training for English majors at DAV. The research will examine both technical implementation aspects and pedagogical effectiveness through the following specific research questions:

***(i)*** How do AI-generated speech simulations compare with traditional human speech recordings in enhancing short-term memory and upgrading efficiency in consecutive interpreting?

***(ii)*** What particular features of AI-generated speech (speech rate, information density, structural complexity, content domain, voice quality, and prosodic features) most effectively target short-term memory and updating interpreting efficiency?

***(iii)*** How do students at DAV perceive effectiveness, authenticity, and perceived usability of AI-generated speech simulations meant for memory enhancing activities in CI training?

***(iv)*** What implementation framework would be the most suitable to apply AI-generated speech simulations for DAV's consecutive interpreting curriculum?

Based on the theoretical frameworks examined, the following hypotheses are proposed:

***(i)*** Students using AI-generated speech simulations with controlled, gradual increases in difficulty parameters will demonstrate significantly greater improvement in short-term memory capacity and accuracy compared to students using standard human speech recordings.

***(ii)*** Among AI-generated speech features, those that specifically challenge updating efficiency (such as speeches containing embedded clauses, narrative structures requiring assumption revision, and high information density) will produce the greatest improvements in CI-relevant memory functions.

***(iii)*** Students will report higher engagement and perceived effectiveness when using AI-generated speech simulations that incorporate authentic diplomatic content relevant to Vietnamese international relations contexts, compared to generic content.

***(iv)*** A blended implementation approach that combines scheduled AI-generated speech practice sessions with structured human instructor feedback will result in greater overall skill development than either approach alone.

These research questions and hypotheses focus especially on how AI-generated speech simulations can be designed and used to improve the memory functions most important for consecutive interpreting performance in the particular context of DAV, so addressing the identified research gap.

## 3. Literature Review

### *3.1. The Cognitive Foundations of Consecutive Interpreting*

The cognitive complexity of interpreting is required to manage linguistic and attentional mechanisms concomitantly. A recent investigation concerning the role of working memory in CI success differentiated and delineated the role of its distinct components. While verbal short-term memory preserved verbal information, updating efficiency controlled the dynamic adjustment of stored material in response to new information. A landmark study of longitudinal design by Dong et al. (2018) studied the two functions in Chinese learners of English under CI training or general second language instruction for one semester. Measures included non-verbal updating efficiency tasks, an L2 listening span, and a letter running span and found that pre- and post-test results a) updated efficiency consistently in CI performance, and b) CI training significantly enhanced updating efficiency. By comparison, verbal spans, such as L2 listening span, had little relation with CI performance at the pretest and did not improve uniquely through CI training when contrasted against general language instruction (Dong et al., 2018). It is more likely that updating efficiency, which is a concept linked to attentional control processes that are similar to those which take place in CI, responds more to the effects of training, rather than static storage capacities. Therefore, the above results have provided a theoretical basis for developing interventions meant to utilize particular cognitive functions necessary for interpreting. Comparative studies of simultaneous interpreting (SI) training further support the cognitive benefits of interpreter training. A study conducted with Master's students in Conference Interpreting found statistically significant improvements in verbal short-term memory following SI training, with gains significantly exceeding control group improvements (Dong et al., 2018). These findings demonstrate the "interpreter advantage," a training-induced cognitive enhancement, and not a pre-existing difference in individuals with a predisposition towards interpreting. With strong control in longitudinal designs, these studies establish a causal link between training in interpreting and cognitive improvement, setting the stage for investigating innovative tools such as AI in order to maximize these benefits.

### *3.2. Short-Term Memory in Interpreting*

Short-term memory (STM) plays a crucial role in interpreting performance. It functions as a temporary storage system that holds information for immediate processing during the interpreting task. Studies have identified several key characteristics of STM that are particularly relevant to interpreting performance. According to Timarová et al. (2014), interpreters developed specialized memory systems that differed from those of non-interpreters. Their research found that professional interpreters demonstrated enhanced articulatory rehearsal processes and executive control, which allowed them to manage competing language representations more effectively. These specialized memory systems developed through extensive practice and were particularly crucial for consecutive interpreting, where information must be stored while listening and then retrieved during production. Mellinger and Hanson (2019) distinguished between different types of memory important for interpreting: sensory memory, which briefly held the auditory input; STM, which maintained active information during processing; and long-term memory, which stored knowledge and expertise. They found that successful interpreters demonstrated particular strength in STM capacity and the ability to efficiently encode information using strategies such as chunking and note-taking.

Liu et al. (2004) conducted experimental research comparing professional interpreters with students and found that professionals did not necessarily have larger memory spans but showed superior abilities in managing attentional resources and semantic processing of incoming information. This suggests that interpreter training should focus not just on expanding memory capacity but on developing strategic approaches to information processing. Furthermore, Bajo et al. (2000) demonstrated that STM training using exercises specifically designed for interpreting contexts produced more significant improvements in interpreting performance than general memory exercises. Their research suggests that memory training should be context-specific and simulate the cognitive demands of actual interpreting tasks.

### *3.3. AI Applications in Interpreter Training*

The infusion of AI technology into interpreter training is becoming mainstream within the trend of global movement toward technology-enhanced learning and offering new opportunities to develop cognitive skills. Automatic speech recognition technology, computer-assisted interpreting, InterpretBank being among the available tools, etc., will feature in the learning process even more than ever to improve terminology management, remote access to resources, and instant real-time support to translators' work (Zhao, 2024). The same author notes that such tools "respond to contemporary requirements for interpreters to get ready for highly specific assignments and to upgrade their performance through effective information access and use of technology, Ergonomic teaching units, usually in the consecutive mode, simulate real working conditions that improve in-class performance, pre-task preparation, and technology awareness for students as well."

Empirical evidence of AI cognitive benefits was found by Surrey's Centre for Translation Studies (2023) in an experiment with Interlingual Respeaking (IRSP). This was live subtitling through human-AI collaboration. The results yielded better opportunities to improve working memory and task-switching within a 51-strong group of language professionals after undergoing 25 hours of learning IRSP. Such results may be taken as an indication that human-AI interaction can improve cognitive aspects that support the development of skills which are imperatively important in interpreting. However, the effectiveness of AI tools is highly dependent on the high quality of training datasets consisting of the enormous diversity of speech intonations, accents, and kinds of environments within which the process of interpreting happens. In the context of diplomatic interpreting, this requires a set of data prepared based on vocabulary and style specific to international relations discussions.

### *3.4. Critical Perspectives on AI in Interpreter Education*

Giustini criticizes AI technology-based interpreter training citing the promise of AI but further highlighting the need for contextual and tacit knowledge in humans to outperform AI, which operates at paradigmatic rationality rather than situational negotiation. This is a prompt epistemological caution to avoid the uncritical application of AI and to develop an ethics and politics that involve the cognitive, practical, and moral aspects of responding (Dastyar & Giustini, 2024). Similar assessments reach back to the wide philosophical anxiety of Dreyfus about the pitfalls of AI in capturing human intuition and context sensitivity (Dastyar & Giustini, 2024). As such, AI-generated oral simulation should be used as an adjunct, not a substitute for, a humanistic pedagogy of interpreting that balances technological innovation with the sensitivity required in interpreting.

### *3.5. AI Integration at the Diplomatic Academy of Vietnam*

The Diplomatic Academy of Vietnam is AI-ready, marking a new era in interpreter training. The workshop on generative AI in English language teaching recently conducted with the Regional English Language Office of the US Embassy delved into the application of AI chatbots for brainstorming feedback and even quiz creation (Diplomatic Academy of Vietnam, 2024). Therefore, the proactive development of AI-integrated lesson plans by the teaching staff within the workshop indicates its organizational readiness to embrace technological innovation (Diplomatic Academy of Vietnam, 2024). Therefore, the findings can be generalized to other organizations, ostensibly those that would like to integrate AI-generated speech simulators into their curricula. The application, however, must be based on the needs of the setting, considering aspects such as infrastructure and faculty expertise relative to the language context of English and Vietnamese. The privacy and bias issues raised in the workshop will also be central in the implementation process (Diplomatic Academy of Vietnam, 2024).

### *3.6. Research Gap and Opportunities*

Major gaps remain in the state-of-the-art knowledge of the advantageous potentials of AI systems in CI training at DAV. First, though the study by Dong et al. (2018) has already proved that CI training leads to an improvement in updating, research into the development of AI-generated speech simulations to target this function is scanty. Second, even though AI instruments are gradually being adopted (Zhao, 2024) and the results of human-AI interaction are cognitively salient with Surrey's Centre for Translation Studies (2023), the use of these for memory enhancement in CI has not been touched upon. None of the above-mentioned has been explored in the specific contexts of DAV - its language and diplomatic aspects, offering both challenges and prospects for customized AI interventions. Fourth, technical aspects of the integration of AI require philosophical, ethical dimensions to be raised (Dastyar & Giustini, 2024). Up to the moment, however, existing studies prioritize outcome measurement over the design of progressively challenging interventions that incrementally challenge memory functions. Memory is an area where customizable capabilities of AI could excel.

These gaps offer a very fascinating research opportunity to bridge the field of cognitive science, in particular, interpreter pedagogy, and AI technology. The study on the ways AI-generated speech simulations can be designed and implented at DAV to improve verbal short-term memory and updating efficiency will provide evidence-based training methods specifically adjusted to diplomatic interpreting. Such work would further improve not only theoretical knowledge but practical skills too, enhancing interpreter education in specialization contexts.

## 4. Methodology

This study employed a mixed-methods approach, prioritizing qualitative depth to explore students' experiences while incorporating basic quantitative measures of memory performance. The investigation was designed to evaluate the impact of AI-generated speech simulations on short-term memory (STM) enhancement and updating efficiency in consecutive interpreting (CI) training.

### *4.1. Participants*

The study included a total of 8 third-year English majors at the Diplomatic Academy of Vietnam (DAV). Participants were selected based on their successful completion of the Translation I-II and Theory of Translation and Interpreting courses. A purposive sampling strategy was utilized to ensure participants possessed average or above-average English and interpreting proficiency, thereby avoiding those potentially already at peak memory development levels. Eligibility criteria required participants to have a minimum overall IELTS band score of 6.5 OR a cumulative GPA of 7.0/10 for both Theory of Translation and Interpreting and Translation I & II. Participants also had no prior experience with AI-generated speech tools for interpreting practice, ensuring a consistent baseline. Access to required technology, such as computers and headphones, was also necessary for engaging with the materials. The total participant pool of 8 students was evenly divided into two groups:

- The Test Group (Group E) consisted of 4 students who practiced with AI-generated speech simulations.

- The Comparison Group (Group C) consisted of 4 students who used traditional human-voiced recordings during the study period.

The participant pool included 2 male and 6 female students, reflecting the typical gender ratio in DAV's English major program. The number of participants and gender distribution were acknowledged as a potential limitation affecting the study's generalizability.

### *4.2. Data Collection Procedures*

Data collection was conducted through pre- and post-intervention assessments and ongoing monitoring during the intervention period.

***Pre-intervention memory assessment (conducted on March 29, 2025)***

Participants completed a brief cognitive test specifically designed to measure memory functions considered essential for consecutive interpreting. This included a Listening Span Task to assess short-term memory capacity. The pre-test aimed to establish initial equivalence in cognitive function measures across all participants. The internal consistency of the memory assessment tools, particularly the Set Size Tests, was found to be excellent, as indicated by a Cronbach's alpha (α) of 0.924.

***Intervention period (spanning four weeks from April 1 to April 28, 2025)***

Both groups engaged in a 30-minute practice session involved one core exercise per day (Monday to Friday). These daily sessions provided materials for bidirectional interpreting practice, including both English-Vietnamese and Vietnamese-English tasks. The audio material for each exercise was approximately 1.5 minutes in total duration. The practice methodology for each audio file involved a two-attempt scaffolded process:

(i) Attempt 1 (Holistic): Participants first listened to the entire audio recording (about 1.5 minutes) without interruption. Immediately after the audio concluded, they were required to deliver their full oral interpretation based on their short-term memory.

(ii) Attempt 2 (Segmented): After the holistic attempt, participants practiced with the same audio file, but it was now divided into two segments (approximately 45-50 seconds each). They would listen to the first segment, pause, and immediately deliver their interpretation for that segment. This process was then repeated for the second segment.

This standardized, two-step (holistic-then-segmented) methodology was applied consistently to both the Experimental and Comparison groups.

*AI speech simulation design (Group E)*

The group used speeches generated by the AI tool NaturalReader (naturalreaders.com). These speeches were organized into four progressively challenging levels, enabling participants to strengthen their memory skills in a systematic and structured manner (*see Appendix D for the audio files used in the pilot study*):

Level 1 (Week 1): Characterized by a slow speech rate (100-120 words per minute), low information density, and simple sentence structures.

Level 2 (Week 2): Involved a moderate speech rate (130-150 wpm), moderate information density, and compound sentences.

Level 3 (Week 3): Utilized a natural speech rate (160-180 wpm), high information density, and complex structures with embedded clauses.

Level 4 (Week 4): Incorporated variable speech rates, high information density, and narrative elements requiring assumption revision. The AI content was also tailored to include authentic diplomatic subject matter relevant to DAV students.

*Human-voiced materials (Group C)*

Participants in the Comparison Group (Group C) practiced with materials that were content-identical to those used by the Experimental Group (Group E). To ensure rigorous control, these text-based scripts were also generated by AI and systematically structured according to the same four levels of progressive difficulty. The primary methodological difference lay in the audio delivery. The scripts for Group C were recorded by a human speaker with high-level proficiency (equivalent to IELTS Speaking 7.5) to ensure accurate pronunciation and natural prosody.

Crucially, a key difference in the intervention's control was speech rate. While the human speaker was instructed to vary their delivery speed (e.g., slow, moderate, natural) to match the four difficulty levels, this variation was based on their subjective estimation of speed. This contrasts sharply with the AI group's materials, where speech rates were computationally controlled and precisely defined. Therefore, this study's design compared a precisely controlled AI delivery (synthetic voice + exact speed) against a naturalistic, subjectively-paced human delivery of the exact same content.

*Monitoring*

Participants in both groups maintained daily practice logs to record their performance, time spent, materials used, and self-assessed difficulty. These logs were submitted weekly and provided qualitative insights.

***Post-intervention memory assessment (conducted on April 29, 2025)***

All participants completed the same memory assessments used in the pre-intervention phase to measure any effects on verbal short-term memory and updating efficiency.

### *4.3. Data Analysis*

Data analysis involved both qualitative and basic quantitative methods. A thematic analysis was conducted on the weekly practice logs and reflections from both groups. This analysis was interpreted through the lenses of Cognitive Load Theory and Chunking Theory. Content analysis, potentially using NVivo software as mentioned in the proposal, was used for systematic scripting and identification of common themes related to challenges, perceived effectiveness, and usability. Thematic coding focused on features beneficial for memory enhancement, challenges and limitations, contextual factors, and comparisons with traditional methods.

Simple descriptive statistics and non-parametric tests were applied to analyze the data. This included examining differences in pre/post memory assessment scores between the groups, relationships between specific AI speech features and self-reported difficulty, and patterns in practice log data. It was explicitly acknowledged that due to the short timeframe and resulting restriction on statistical power, these basic analyses would primarily complement the qualitative findings and help identify trends for future research, rather than establishing definitive causal links or statistical significance. Independent t-tests were conducted, and Cohen's d effect sizes were calculated to indicate the magnitude of differences between group means.

To enhance the transparency and rigor of the qualitative analysis, a structured thematic coding process was employed. The two authors (Dang Phuong Nam & Do Thi Thu Phuong) first independently read and open-coded the complete set of participant logs. An initial codebook was then developed based on the research questions and the theoretical frameworks of Cognitive Load Theory and Chunking Theory. The authors subsequently met to compare codes, discuss discrepancies, and establish a finalized, consolidated coding framework (inter-coder agreement > 90%). This framework was systematically applied to identify, categorize, and quantify the recurring themes related to challenges, perceived effectiveness, and usability presented in the findings section.

### *4.4. Ethical Considerations*

The study adhered to all standard ethical research requirements. This included obtaining informed consent from all participants, ensuring anonymity through coding, securing approval from the DAV Research Ethics Committee, storing data securely under lock and key and password protection to ensure confidentiality, and informing participants of their right to withdraw from the study at any point without repercussions.

## 5. Findings

This section presents findings from a four-week study on how AI-generated speech simulations enhance short-term memory in consecutive interpreting. It integrates quantitative results with participant insights to explain observed improvements and situates them within broader cognitive and pedagogical contexts. It is crucial to preface these findings by reiterating the study's nature as a pilot investigation with a small sample size (N=8). Consequently, the following quantitative and qualitative results should not be interpreted as statistically generalizable. Rather, they are presented as preliminary trends and indications that warrant further, larger-scale research. The emphasis is on identifying potential effect sizes and themes, not on establishing definitive causal links.

### *5.1. Quantitative Findings: Memory Performance Outcomes*

The quantitative analysis commenced with an assessment of the memory assessment tools' reliability. The internal consistency of the Set Size Tests was found to be excellent, with a Cronbach's alpha (α) of 0.924. This high reliability (α ≥ 0.9) suggests that the measures consistently assessed the intended memory constructs, specifically short-term memory capacity, enhancing the internal validity of the quantitative findings. This reliability minimizes concerns about measurement error, confirming that the 'Gap' scores, which represent improvement, are likely accurate reflections of actual changes in participants' memory functions.

Overall descriptive statistics across all 8 participants showed mean improvements ('Gap') that varied across different memory tasks. For example, Set3_Gap (improvement in set size 3) had a mean of 31.63 (SD = 14.85). The FullTest_Gap (overall improvement) showed a mean of 17.13 (SD = 11.42). These statistics provide a general overview before group-specific comparisons.

A preliminary examination of the mean 'Gap' scores for Group E (AI-voiced materials, n=4) and Group C (Human-voiced materials, n=4) revealed distinct trends. Group E had a mean Set4_Gap of 22.75, notably higher than Group C's mean of 6.25. Similarly, Group E's mean FullTest_Gap was 20.25, compared to Group C's 14.0. These initial differences suggested a differential impact, leading to an analysis of effect sizes.

Given the very small sample size (N=8), the study had inherently low statistical power, making it difficult to achieve traditional statistical significance ($p <.05$). Therefore, independent t-test results focused on Cohen's d effect sizes to quantify the magnitude of differences between groups, as this measure is independent of sample size. A "Large" effect is indicated by $d \geq 0.8$. The large effect sizes observed, despite the sample size, provide an initial indication of practically meaningful improvement.

The specific Cohen's d values and their interpretations are presented in Table 1:

**Table 1**

*Cohen’s D Effect Sizes and Classification of Memory Task Improvements Between AI and Human Groups*

| Variable | Group E Mean | Group C Mean | Mean Diff. | Cohen's d | Effect Size Classification |
|---|---|---|---|---|---|
| Set3_Gap | 28.75 | 34.5 | -5.75 | -0.366 | Medium |
| Set4_Gap | 22.75 | 6.25 | 16.5 | 1.21 | **Large** |
| Set5_Gap | 24.5 | 16.5 | 8 | 0.525 | **Large** |
| Set6_Gap | 21.75 | 10.5 | 11.25 | 0.733 | **Large** |
| Judge_Gap | 15.25 | 8.75 | 6.5 | 0.625 | **Large** |
| Recall_Gap | 26 | 20.75 | 5.25 | 0.277 | Medium |
| FullTest_Gap | 20.25 | 14.0 | 6.25 | 0.53 | **Large** |

*(Note: Cohen's d: 0.2 = Small, 0.5 = Medium, 0.8 = Large effect)*

For Set3_Gap, a medium effect favored Group C (Cohen's d = -0.366). This might suggest slightly greater initial improvement on the simplest task with human materials or a ceiling effect for Group E on this basic level. It could also indicate that the AI's rapid progression quickly moved Group E beyond focusing on Set Size 3.

In stark contrast, Set4_Gap showed a very large and practically significant effect size (Cohen's d = 1.21) favoring Group E. This suggests that AI-generated speech simulations may be effective in enhancing memory capacity at this intermediate difficulty level. This pattern continued for Set5_Gap (Cohen's d = 0.525) and Set6_Gap (Cohen's d = 0.733), both showing large effects favoring Group E, demonstrating potential effectiveness as memory load increased. The concentration of large effects in higher "Set Size" tasks aligns with the AI's design to progressively increase cognitive load, targeting memory functions critical for managing the high cognitive demands of CI.

The Judgement Task (Judge_Gap) also revealed a large effect favoring Group E (Cohen's d = 0.625), suggesting AI simulations improved performance on tasks likely involving updating efficiency. The Recall Task (Recall_Gap) showed a medium effect favoring Group E (Cohen's d = 0.277). Crucially, the overall improvement (FullTest_Gap) demonstrated a large effect favoring Group E (Cohen's d = 0.53). The consistent large effect sizes for Group E across multiple, increasingly challenging measures provided a preliminary suggestion of a substantial practical significance of the AI intervention in enhancing CI-relevant memory functions. The "Comparison Chart" visually corroborated these quantitative trends.

Individual participant data further revealed variations. In Group E, participants E2, E3, and especially E4, showed substantial gains. E4 achieved the highest overall improvement (39% increase in Full Test). In Group C, improvements were less consistent. However, participants like C3 and C4 also showed significant gains in specific tasks. C4's 27% increase in the Full Test was higher than some in Group E. This indicates that while AI may be more effective *on average*, individual factors also play a significant role.

### *5.2. Qualitative Iinsights: Participant Perceptions and Experiences*

A thematic analysis of weekly practice logs and reflections from both groups, interpreted through Cognitive Load Theory and Chunking Theory lenses, revealed recurring themes regarding challenges, perceived effectiveness, and usability.

#### *Common challenges across both groups*

Participants in both groups consistently reported fundamental challenges inherent to CI, regardless of the audio source:

- Note-taking: A pervasive difficulty described as slow, leading to missed points and hindering understanding. Participants struggled with multitasking. Some in Group E noted over-reliance on writing everything instead of using effective signs. This highlights note-taking as a critical skill for managing cognitive load, requiring explicit training.

- Information density and length: Audio content was found to be "overwhelming" or "too long", increasing cognitive load and making it difficult to remember the order or transfer long ideas coherently. This aligns directly with Cognitive Load Theory.

- Speaker speed: Increased speech rates as the intervention progressed exacerbated multitasking challenges.

These shared challenges suggest that while AI provides varied stimuli, pedagogical interventions beyond audio practice are needed for holistic improvement, such as explicit note-taking strategy training or vocabulary pre-teaching.

#### *Challenges specific to group C (Human-voiced materials)*

- Audio quality: Issues like "unstable audio" and varying loudness added extraneous cognitive load, impeding comprehension.

- Perceived lack of control: Reflections about struggling to catch up when the speaker talked faster suggest the difficulty progression might have felt less predictable or systematically controlled, leading to frustration.

- Background knowledge gaps: Participants attributed difficulties to a "lack of background knowledge in the topic", indicating potential gaps in preparatory context for the materials.

***Challenges specific to group E (AI-voiced materials)***

- Audio quality: One participant noted the Vietnamese AI audio sounded "a bit crackly", highlighting a technical limitation of current AI speech synthesis that impacts user experience and perceived authenticity.

- Timing for translation: Participants struggled with "knowing when to stop the audio for myself to start translating", suggesting a need for better user controls in the AI interface.

- Cognitive overload and "Panic": One participant described experiencing "panic when there are a lot of difficult terms cropping up," which they felt impacted their memory. This indicates that while challenging content is necessary, AI delivery needs fine-tuning to manage cognitive load effectively and prevent overwhelming learners.

- Unpreparedness for speed increase: Participants noted being "unprepared for it" when speed increased, suggesting that perceived challenge sometimes outpaced adaptive capacity, at least initially.

***Perceived effectiveness and engagement***

Perceptions of effectiveness and engagement varied.

- Group C: Varied perceptions, with some finding topics "ok and familiar" or "quite interesting", while others felt "no progress" or struggled with "overwhelming" information.

- Group E: Participants reported a growing sense of mastery. By Week 4, reflections indicated "manageable difficulty" and "full understanding". While this qualitative feedback is not generalizable, it suggests the AI's systematic and controlled difficulty increase may have fostered a stronger sense of progress and self-efficacy. This suggests the AI's systematic and controlled difficulty increase fostered a stronger sense of progress and self-efficacy, a significant pedagogical advantage.

## 6. Discussion

This section integrates the quantitative and qualitative findings to address the study's research questions (RQ) and hypotheses (H).

### *6.1. Comparison of AI-Generated vs. Traditional Human Speech*

RQ1 asked how AI-generated speech simulations compare with human recordings in enhancing STM and updating efficiency. H1 posited that students using AI with controlled, gradual difficulty increases would demonstrate significantly greater improvement.

The quantitative data, particularly the consistent large effect sizes (Cohen's d ≥ 0.5) for Group E across Set4_Gap, Set5_Gap, Set6_Gap, Judge_Gap, and FullTest_Gap, provides preliminary support for Hypothesis 1. The AI group showed a higher average Full Test improvement (20.25%) than the human group (14.25%). Despite the small sample size limiting statistical generalizability, the magnitude of these effects suggests substantial practical significance. This aligns with the "interpreter advantage," a training-induced cognitive enhancement, suggesting AI training can amplify this (Dong et al., 2018).

Qualitative data aligns with this trend, with Group E participants noting the AI's systematic progression led to a more effective learning curve and a reported sense of mastery, contrasting with Group C's more varied progress, potentially hindered by inconsistent materials. The anomaly in Set3_Gap favoring Group C might be due to a ceiling effect for Group E or the AI's rapid progression beyond basic levels.

### *6.2. Impact of Specific AI Speech Features*

RQ2 aimed to identify which AI features most effectively target memory functions. H2 proposed features challenging updating efficiency (e.g., embedded clauses, high density) would yield the greatest improvements. While direct quantitative correlation was not possible, qualitative data provides strong inferential support. Group E participants reported challenges with "lists," "technical terms," and "complexity," especially at higher levels, which align with AI's manipulation of information density and structural complexity (Dong et al., 2018). The large effect sizes in Set4-6 and Judge tasks, measuring memory under increasing load and updating efficiency, strongly suggest that AI features challenging these functions were effective. This aligns with research emphasizing updating efficiency over static storage (Liu et al., 2004).

This systematic increase in challenge aligns with "desirable difficulties". By making learning more effortful through higher speech rates and density, AI likely activated deeper cognitive mechanisms, leading to better memory consolidation. AI's ability to precisely control these parameters allows for a systematic application of desirable difficulties, making it a sophisticated instrument for optimizing cognitive training.

### *6.3. Perceptions of Effectiveness, Authenticity, and Usability*

RQ3 explored student perceptions of AI simulations' effectiveness, authenticity, and usability. H3 anticipated higher engagement and perceived effectiveness with authentic diplomatic content.

Participant perceptions of effectiveness were generally positive for Group E, noting improved "words processing" and a sense of mastery. Their progression from initial challenge to "manageable difficulty" and "full understanding" indicates high perceived effectiveness of the progressive training. Authenticity was supported by the use of relevant content like "medical jargons". However, the "crackly" Vietnamese AI audio detracted from perceived authenticity (Seeber, 2013), highlighting a technical limitation of current AI speech synthesis (Dastyar & Giustini, 2024). Usability issues included timing struggles for translation and pervasive note-taking difficulties. While qualitative data suggests content relevance is important (supporting H3), technical audio issues might have offset some benefits. The findings emphasize that effective AI training must balance technological innovation with the sensitivity required in interpreting, acknowledging AI's limitations (Dastyar & Giustini, 2024).

Beyond the "crackly" Vietnamese audio noted by one participant, the discussion of technical limitations must be expanded. The AI-generated speech, while computationally precise in speed, inherently lacks the natural prosody, intonation, and strategic pauses (chunking) of authentic human speech. Authentic speech is not just a carrier of words, but of intent, emphasis, and emotion, which are critical cues for an interpreter. The current AI's relative monotony may not adequately prepare students for the nuanced cognitive load of a real-world speaker. This lack of prosodic naturalness could also explain why Group C participants (who heard a human voice) still achieved notable progress, as they were exposed to more realistic, albeit less systematically paced, audio cues.

### *6.4. Implementation Framework Considerations*

RQ4 aimed to identify a suitable implementation framework for AI simulations at DAV. H4 proposed a blended approach combining AI practice with human instructor feedback. Persistent challenges like note-taking and handling cognitive load/panic support H4. While AI excels at controlled stimulus delivery, it doesn't solve strategic or metacognitive challenges (Dastyar & Giustini, 2024). This reinforces the "AI as adjunct" perspective. Human instructors are crucial for explicit instruction and feedback on skills like note-taking and providing support during periods of high cognitive load. This aligns with human-AI collaboration models. DAV's readiness for AI integration is a good foundation (Diplomatic Academy of Vietnam, 2024). An effective framework should capitalize on AI's ability to systematically increase cognitive load while integrating human pedagogy to address challenges AI alone cannot. AI systems should potentially incorporate effective support mechanisms (Çela et al., 2024; Grinschgl &Neubauer, 2022; Akgun & Toker, 2024; Jose et al., 2024).

### *6.5. Cognitive Load and Note-taking*

CI is inherently cognitively demanding, leading to overload. The AI's progressive difficulty intentionally increased cognitive load by manipulating speed, density, and complexity. Observed memory improvements, especially in higher set sizes, suggest this calibrated load, while challenging, pushed participants to engage in deeper processing, aligning with desirable difficulties. Participant reflections on increased speed and "panic" are direct manifestations of this challenge. Persistent note-taking struggles highlight its critical role in managing cognitive load. Effective note-taking serves as an external memory aid, reducing working memory burden. AI serves as a precision instrument for calibrating this load to optimize learning by pushing learners towards their capacity limit.

## 7. Conclusion and Recommendation

This study provides preliminary evidence that AI-generated speech simulations hold significant promise for enhancing short-term memory and updating efficiency in consecutive interpreting training, specifically within the context of the Diplomatic Academy of Vietnam. Quantitative results generally indicated greater overall improvement in memory performance for the AI group, particularly when materials were designed with controlled, progressive difficulty. Despite the constraints of a small sample size, the consistent observation of large effect sizes favoring the AI group across multiple memory tasks suggests a substantial practical significance and aligns with the concept of the "interpreter advantage".

The study’s scale and duration naturally shaped its conclusions. With only eight participants (two groups of four), the statistical power was limited, meaning that the quantitative results should be interpreted as emerging trends rather than generalizable effects. The intervention period was also relatively short, spanning four weeks. While this duration allowed for the observation of initial improvements and adaptation, a longer timeframe would be necessary to explore more sustained cognitive enhancements and long-term memory retention effects. The participant pool included 2 male and 6 female students, reflecting the typical gender distribution in the program but indicating a need for future studies to ensure broader gender representation to assess potential differences in learning outcomes. Furthermore, this pilot study did not control for other potential confounding variables, such as participants' prior background knowledge on specific diplomatic topics or the exact equipment (e.g., personal headphones vs. external speakers) used during practice sessions. The limited quantitative depth in this initial phase, constrained by the study's timeframe and scale, meant that analyses primarily served to complement qualitative findings and identify trends.

Qualitative findings further supported the value of AI's systematic approach in managing cognitive load, fostering a sense of mastery among participants as they adapted to increasing challenges. However, the study also highlighted persistent challenges inherent to interpreting across both groups, such as note-taking, specialized vocabulary, and information density, indicating that AI can serve as a powerful tool but not replace comprehensive instructor-led training. Technical limitations related to AI audio quality were also noted.

Overall, the findings suggest that AI should function as a complementary tool within a human-centered pedagogical model. AI is well-suited for generating controlled practice stimuli, scaling difficulty, and supporting cognitive skill development, while human instructors remain indispensable for providing nuanced feedback, emotional support, and critical AI literacy training. This balance is especially vital in complex domains such as note-taking and managing cognitive overload.

This research contributes to bridging the gap between cognitive science, interpreter pedagogy, and emerging AI technology It offers evidence-based insights for designing AI-enhanced training interventions that align technological capabilities with humanistic teaching principles. As DAV continues its proactive integration of AI, the results underscore the potential for a blended model to shape highly skilled interpreters for Vietnam's future international engagement, provided ethical considerations and user experience are prioritized. Future studies could employ experimental designs specifically tailored to isolate and quantify the impact of individual AI speech features on memory functions. Rigorously evaluating blended learning models that combine AI practice with structured human feedback is also crucial, as qualitative data underscored the continued importance of human guidance for complex skills like note-taking and managing cognitive overload. Exploring adaptive AI systems that respond not just to performance but potentially to cognitive and affective states could address challenges like "panic" reported by participants when faced with high difficulty. Finally, expanding research to different language pairs and cultural contexts would assess the generalizability of these findings globally.

## References

Waywithwords.net (2024, January 29). *AI Language Processing: 10 Key Limitations*. https://waywithwords.net/resource/ai-language-processing-key-limitations/

Akgun, M., & Toker, S. (2024*). Evaluating the Effect of Pretesting with Conversational AI on Retention of Needed Information.* ArXiv. https://arxiv.org/abs/2412.13487

Bajo, M. T., Padilla, F., & Padilla, P. (2000). Comprehension processes in simultaneous interpreting. In A. Chesterman, N. Gallardo San Salvador, & Y. Gambier (Eds.), *Translation in context* (pp. 127-142). John Benjamins. https://doi.org/10.1075/btl.39.15baj

Cai, R., Dong, Y., Zhao, N., & Lin, J. (2015). Factors contributing to individual differences in the development of consecutive interpreting competence for beginner student interpreters. *The Interpreter and Translator Trainer, 9*(1), 104–120. https://doi.org/10.1080/1750399X.2015.1016279

Chen, S. (2017). The construct of cognitive load in interpreting and its measurement. *Perspectives*, *25*(4), 640–657. https://doi.org/10.1080/0907676X.2016.1278026

Chmiel, A. (2018). In search of the working memory advantage in conference interpreting – Training, experience and task effects. *International Journal of Bilingualism,* *22*(3), 371-384. https://doi.org/10.1177/1367006916681082

Christoffels, I. K., de Groot, A. M. B., & Kroll, J. F. (2006). Memory and language skills in simultaneous interpreters: The role of expertise and language proficiency. *Journal of Memory and Language, 54*(3), 324-345. https://doi.org/10.1016/j.jml.2005.12.004

Çela, E., Fonkam, M. M., & Potluri, R. M. (2024). Risks of AI-assisted learning on student critical thinking: A case study of Albania. *International Journal of Risk and Contingency Management (IJRCM), 12*(1), 1-19. https://doi.org/10.4018/IJRCM.350185

Diplomatic Academy of Vietnam. (2024). *Workshop: Using generative AI in teaching and learning English – by English Language Specialist Christopher Stillwell at the Diplomatic Academy of Vietnam*. DAV Official Website. https://dav.edu.vn/en/workshop-using-generative-ai-in-teaching-and-learning-english-by-english-language-specialist-christopher-stillwell-at-the-diplomatic-academy-of-vietnam/

Dong, Y., Liu, Y., & Cai, R. (2018). How does consecutive interpreting training influence working memory: A longitudinal study of potential links between the two. *Frontiers in Psychology, 9*, 875. https://doi.org/10.3389/fpsyg.2018.00875

Dreyfus, H. L., & Dreyfus, S. E. (2005). Peripheral vision: Expertise in real world contexts. *Organization Studies, 26*(5), 779-792. https://doi.org/10.1177/0170840605053102

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

Giustini, D., & Dastyar, V. (2024). Critical AI literacy for interpreting in the age of AI. *Interpreting and Society: An Interdisciplinary Journal*, *4*(2), 196-213. https://doi.org/10.1177/27523810241247259

Grinschgl, S., & Neubauer, A. C. (2022). Supporting Cognition With Modern Technology: Distributed Cognition Today and in an AI-Enhanced Future. *Frontiers in Artificial Intelligence, 5*, 908261. https://doi.org/10.3389/frai.2022.908261

Ivars, A. J., & Calatayud, D. P. (2013). Mindfulness training for interpreting students. *Interpreting*, *15*(2), 212–233. https://doi.org/10.1515/les-2013-0020

Jose, B., Cherian, J., Verghis, A. M., Varghise, S. M., S, M., & Joseph, S. (2025). The cognitive paradox of AI in education: Between enhancement and erosion. *Frontiers in Psychology, 16*, 1550621. https://doi.org/10.3389/fpsyg.2025.1550621

Liu, M., Schallert, D. L., & Carroll, P. J. (2004). Working memory and expertise in simultaneous interpreting. *Interpreting, 6*(1), 19–42. https://doi.org/10.1075/intp.6.1.04liu

Mellinger, C. D., & Hanson, T. A. (2019). Meta-analyses of simultaneous interpreting and working memory. *Interpreting, 21*(2), 165-195. https://doi.org/10.1075/intp.00026.me

Moser-Mercer, B. (2008). Skill acquisition in interpreting: A human performance perspective. *The Interpreter and Translator Trainer*, *2*(1), 1-28. https://doi.org/10.1080/1750399X.2008.10798764

Saldivar, J., & Fernandez, B. (2024). Concept Analysis of Adaptive Learning Strategy in English Language Teaching (ALS-ELT). *International Journal of Social Sciences and English Literature, 8*, 45–56. https://doi.org/10.55220/2576683x.v8.231

Seeber, K. (2013). Cognitive load in simultaneous interpreting: Measures and methods. *Target, 25*(1), 18-32. https://doi.org/10.1075/target.25.1.03see

Slavin, R. E., Smith, D. (2009). The Relationship Between Sample Sizes and Effect Sizes in Systematic Reviews in Education. *Educational Evaluation and Policy Analysis, 31*(4), 500-506. https://doi.org/10.3102/0162373709352369

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Surrey's Centre for Translation Studies. (2023, December 18). *Using AI-related technologies can significantly enhance human cognition, finds new study.* University of Surrey News. https://www.surrey.ac.uk/news/using-ai-related-technologies-can-significantly-enhance-human-cognition-finds-new-study

Timarová, S., Čeňková, I., Meylaerts, R., Hertog, E., Szmalec, A., & Duyck, W. (2014). Simultaneous interpreting and working memory executive control. *Interpreting, 16*(2), 139–168. https://doi.org/10.1075/intp.16.2.01tim

Vasilieff, S. (2025, February 10). *Generative AI meets the virtual world: A model for human-AI collaboration*. Deloitte. https://www2.deloitte.com/us/en/insights/industry/technology/ai-and-vr-model-for-human-ai-collaboration.html

Wallinheimo A-S, Evans SL and Davitti E (2023) Training in new forms of human-AI interaction improves complex working memory and switching skills of language professionals. *Front. Artif. Intell,* 6:1253940. https://doi.org/10.3389/frai.2023.1253940

Zhao, N. (2024, August 29). *Training interpreters in the age of AI*. Times Higher Education. https://www.timeshighereducation.com/campus/training-interpreters-age-ai

## APPENDICES

*Scan this QR code to access:*

- Detailed memory assessment results by participants
- Analysis results from SPSS
- Participants’ reflections
-AI-generated audio files