From Script to Delivery: An AI-Based Presentation Coaching System

Beom-gyu Choi[1,†], Jeong-min Park[1,†], Tae-i Lee[2,†], Ho-Young Kwak[1], and Joon-Min Gil[1,*]
[1]Dept. of Computer Engineering, [2]Dept. of Artificial Intelligence
Jeju National University, Jeju, Republic of Korea
Email: {fredb, 2022208024, taei0713}@stu.jejunu.ac.kr, {kwak, jmgil}@jejunu.ac.kr

***Abstract*—Presentation skills are essential in academic and professional settings; however, tools that support systematic practice and objective feedback remain limited. Existing approaches often address slide authoring or speech analysis in isolation, with insufficient consideration of the full presentation process from content preparation to delivery. This paper presents an AI-based presentation coaching system that supports the entire workflow from script preparation to delivery practice using document, video, and audio inputs. The system generates slide structures and presentation scripts from uploaded documents through a large language model. During presentation practice, it analyzes non-verbal behaviors via gaze estimation and posture analysis, and identifies and refines speech disfluencies using speech-to-text processing combined with a T5-based model. The outputs of individual modules are organized and visualized through a unified feedback dashboard, enabling structured review of both presentation content and delivery performance. The system is implemented as a mobile application with server-based APIs, and experimental results demonstrate the feasibility of the proposed end-to-end presentation coaching pipeline.**

***Index Terms*—Presentation Coaching, Script Generation, Non-Verbal Behavior Analysis, Disfluency Detection, AI-Based System**

I. Introduction

Presentations have become an essential form of communication in modern society, playing a critical role across academic, business, and professional domains. According to a survey conducted by the Graduate Management Admission Council (GMAC), presentation skills were identified as one of the most important communication competencies valued by employers [1]. Furthermore, approximately 70% of working professionals recognize presentation ability as a core skill for career success [2]. Effective presentations require the integration of multiple factors, including clear content organization, appropriate eye contact, stable posture, and controlled speech delivery. Together, these elements function as key mechanisms through which individuals demonstrate expertise and enhance persuasive impact.

Despite the recognized importance of presentation skills, a large proportion of individuals experience severe anxiety prior to public speaking [3]. Recent studies report that approximately 75% of the population suffers from presentation-related anxiety, and among those affected, 45% exhibit avoidance behaviors such as declining promotion opportunities or refraining from applying for positions that require presentations [4]. This anxiety and its associated avoidance behaviors extend beyond temporary nervousness, acting as significant barriers to career development and preventing individuals from fully realizing their potential. Notably, a study conducted in the United Kingdom revealed that while 77% of employees are required to deliver presentations as part of their job responsibilities, only 23% have received formal presentation training, indicating a substantial imbalance between demand and access to systematic training [5].

Existing presentation training approaches exhibit several limitations, as key components such as script writing, rehearsal, and delivery coaching are often conducted independently, leading to fragmented learning and reliance on subjective self-evaluation without objective or quantitative feedback. While professional coaching can offer high-quality guidance, its accessibility is limited by time and financial constraints, restricting opportunities for continuous practice. Recent studies have further demonstrated through controlled evaluations that multimodal automated feedback systems can lead to measurable improvements in presentation performance even in real classroom settings, although their effects are often complementary rather than substitutive to expert coaching [6]. To address these challenges, this study proposes an AI-based multimodal presentation coaching system that integrates content generation and performance analysis. The system utilizes a large language model (LLM) to generate presentation scripts from user-provided documents and analyzes presentation performance by jointly processing audio and visual information, including eye gaze, posture, and speech disfluencies. The analysis results are presented through an integrated dashboard to support objective evaluation and continuous improvement [7].

The remainder of this paper is organized as follows. Section II reviews related research on automated content generation and multimodal behavioral analysis for presentations. Section III presents the architecture of the proposed AI-based presentation coaching system, detailing the LLM-based script generation, computer vision-based non-verbal analysis, and speech disfluency detection modules. Section IV evaluates each module's performance using experimental results. Finally, Section V and VI discusses the implications of this study and

† These authors contributed equally to this work (Co-first authors).
* Corresponding author: Joon-Min Gil (jmgil@jejunu.ac.kr)

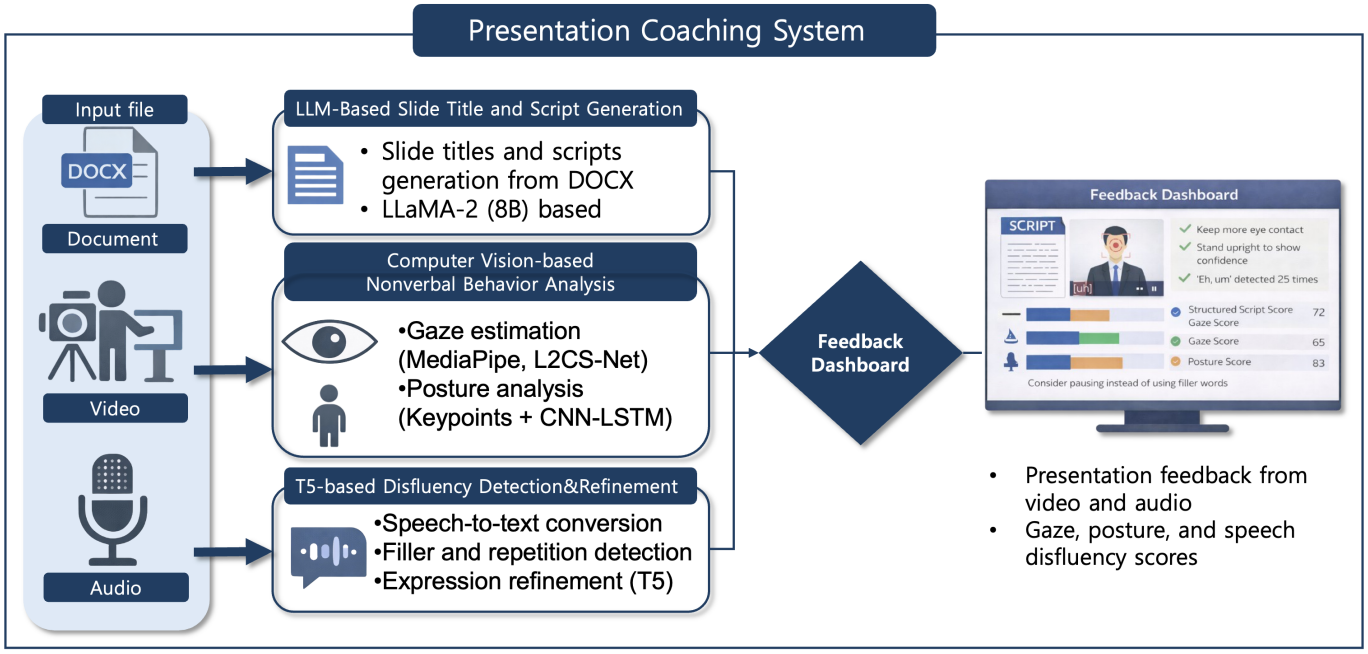

Fig. 1. Overall architecture of the proposed AI-based presentation coaching system

concludes the paper by suggesting future work.

II. Related Work

The field of AI-based presentation coaching has evolved around three core pillars: automated content generation, nonverbal behavior analysis, and disfluency detection.

Aggarwal et al. [8] introduced the 'PASS' pipeline, designed to automate the conversion of general Word documents into structured slide layouts, presentation scripts, and AI-synthesized narrations. While previous efforts in automated slide generation were largely restricted to academic manuscripts, this work broadens the utility to business and educational contexts. By intelligently partitioning textual and visual elements, the PASS architecture enhances the clarity and visual impact of the generated content, marking a significant step toward end-to-end presentation automation.

Wang et al. [9] proposed the Region Attention Network (RAN) to ensure robust facial expression recognition under challenging conditions, such as facial occlusions or significant pose variations. Rather than processing the face as a single global image, the RAN framework decomposes facial features into multiple regions and utilizes a self-attention mechanism to prioritize salient areas like the eyes and mouth. This localized approach allows the model to maintain high accuracy even within the visual complexities often encountered in real-world presentation settings.

Bhat et al. [10] explored an adversarially-trained sequence tagging model for disfluency correction in low-resource linguistic environments. By leveraging a strategic combination of large-scale synthetic data and limited real-world utterances, their model maximizes learning efficiency for detecting stuttering and other speech irregularities. Their findings demonstrated superior performance across multiple languages and speech impairments, providing a robust foundation for automated fluency enhancement.

Despite these advances, significant challenges remain in integrating individual technologies into a cohesive presentation coaching system suitable for practical use. Existing LLM-based automation often lacks fine-grained contextual control, which can result in redundant or overlapping content across slides. In addition, vision-based analysis methods frequently incur high computational costs, limiting their applicability in time-sensitive feedback scenarios. Speech disfluency analysis also relies heavily on synthetic datasets, which do not fully capture the irregular and context-dependent speech patterns observed in real presentation settings, particularly under conditions of anxiety.

To address these limitations, this study proposes an AI-based presentation coaching architecture that organizes document, video, and audio analysis modules within a unified system. The proposed approach processes document, video, and audio inputs through independent analysis modules and organizes their outputs into a unified feedback representation. This modular design enables efficient computation while supporting both presentation preparation and delivery practice within a single system.

III. Proposed Method

The proposed presentation coaching system is designed to support presentation preparation and practice through document, video, and audio analysis. As shown in Fig. 1, the system follows a modular pipeline in which document, video, and

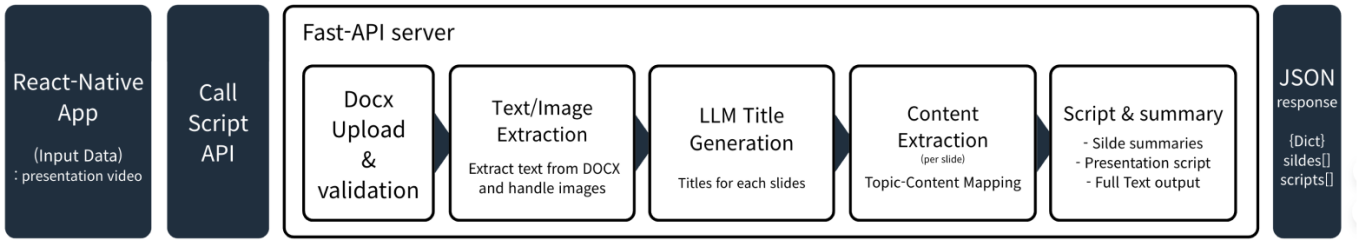

Fig. 2. Workflow of the proposed slide and script generation system, showing the transformation of DOCX documents into slide titles, summaries, and presentation scripts.

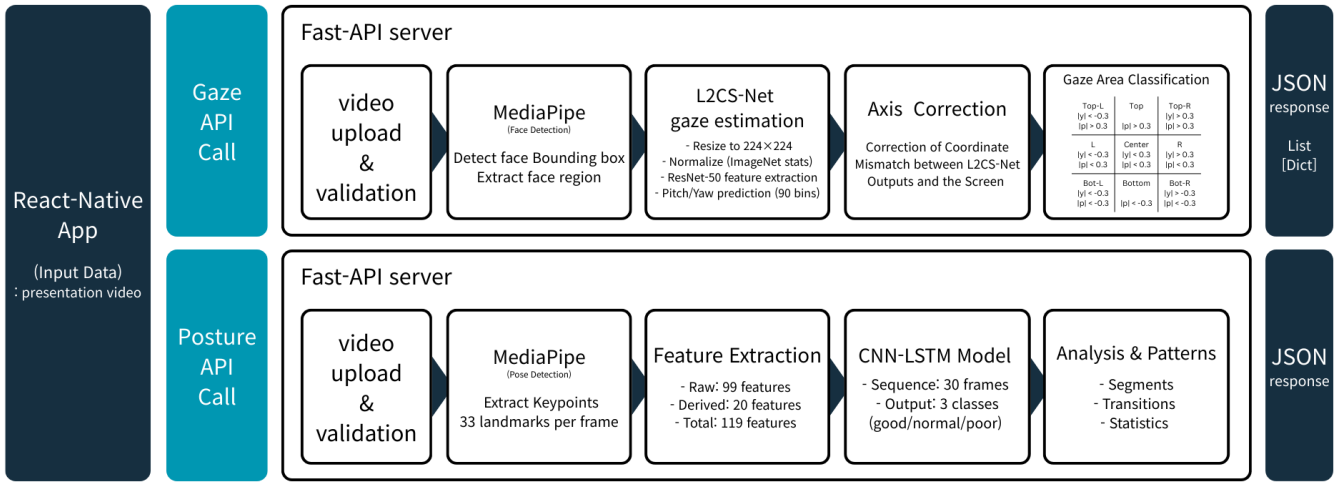

Fig. 3. Workflow of Nonverbal behavior analysis system. (Top) Gaze estimation: MediaPipe → L2CS-Net → 9-region classification. (Bottom) Posture analysis: MediaPipe Pose → Feature extraction → CNN-LSTM → 3-class classification.

audio inputs are processed by independent analysis modules, each focusing on a specific aspect of presentation performance.

For presentation content preparation, the system takes a DOCX document as input and generates slide titles, contextual structures, summaries, and presentation scripts using a large language model. To evaluate non-verbal behaviors during presentation practice, presentation videos are analyzed to estimate gaze direction and body posture based on computer vision techniques. In addition, presentation audio is transcribed through speech-to-text processing, after which disfluency detection and refinement are applied to identify filler words, repetitions, and disfluent expressions.

Rather than combining analysis results at the model level, the proposed system organizes the outputs from each module and presents them through a feedback dashboard. The dashboard provides quantitative scores and qualitative feedback derived from presentation practice videos and audio, enabling users to review different aspects of their presentation performance. The following subsections describe each component of the proposed system in detail.

A. LLM-Based Slide Title and Script Generation

Fig. 2 illustrates the workflow of the slide title and script generation module. This module is implemented as the first stage of the proposed pipeline and provides the textual basis for subsequent feedback. When a user uploads a DOCX document through the mobile application, the file is transmitted to a FastAPI-based server, where textual contents are extracted and organized into structured inputs for generation.

Based on the extracted document content, slide titles and slide-level summaries are generated using a large language model. To produce presentation scripts suitable for actual delivery, the system employs task-specific prompts that reflect presentation conditions, such as slide structure, target audience level, and narrative flow. These prompts guide the model to generate scripts that align with the logical progression of slides rather than producing isolated textual outputs.

The generation process is performed using a LLaMA-2 8B language model deployed on a GPU server. The resulting slide titles, summaries, and presentation scripts are returned in a structured JSON format, which is then used by the feedback dashboard to present script-related guidance alongside other analysis results.

B. Computer Vision-based Nonverbal Behavior Analysis

The computer vision analysis, as shown in Fig. 3, is composed of two modules designed to assess visual engage-

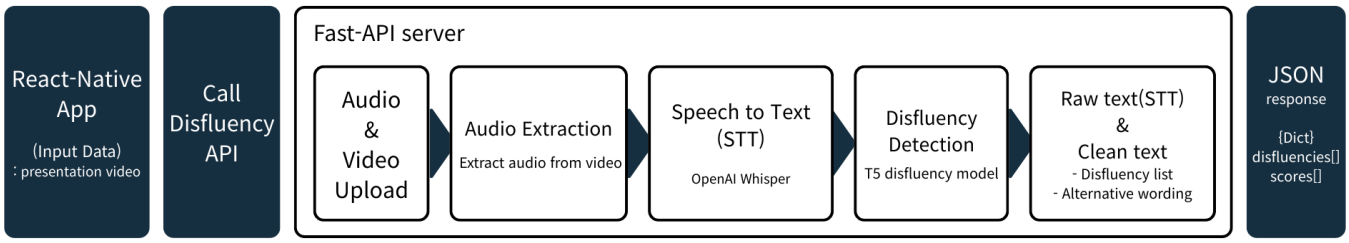

Fig. 4. Workflow of the proposed disfluency detection and refinement system. It illustrates the transition from raw audio to structured feedback through Whisper and T5 models.

ment and physical expressiveness: gaze estimation and pose analysis.

The gaze estimation module is implemented as a two-stage pipeline to quantitatively analyze the presenter's audience-facing (camera-facing) gaze behavior. In the first stage, MediaPipe Face Mesh is used to detect the facial region in each frame and extract 468 three-dimensional facial landmarks. In the second stage, gaze direction is estimated using L2CS-Net (Look-To-Camera System Network). L2CS-Net is a pre-trained appearance-based model that takes a facial image as input and outputs gaze direction in terms of yaw (horizontal) and pitch (vertical) angles. Over the entire presentation, the system computes quantitative metrics including camera gaze ratio, gaze dispersion, and gaze aversion intervals. The analysis results are returned in JSON format via a FastAPI server and visualized on a dashboard using timeline graphs and heatmaps.

The pose analysis module employs a CNN–LSTM hybrid architecture to analyze the presenter's body movements with temporal context. In the first stage, MediaPipe Pose is used to extract 33 body keypoints (e.g., shoulders, elbows, and wrists) from each frame. In the second stage, a custom-designed CNN–LSTM network is used for posture analysis. The CNN layers learn spatial relationships between body joints, while the LSTM layers capture temporal patterns of posture changes. The model is trained to classify posture into stable posture, excessive movement, rigid posture, and effective gestures, using a custom dataset of real presentation videos with data augmentation to improve generalization.

Finally, the system computes quantitative indicators such as posture stability score, average movement range, gesture usage frequency, and unnecessary movement intervals. The analysis results are transmitted via a FastAPI server and visualized on a dashboard, providing objective feedback to the presenter.

C. T5-based Disfluency Detection & Refinement

Fig. 4 illustrates the workflow of the proposed disfluency detection and refinement system. The module is designed as a three-stage pipeline to identify and refine disfluent patterns in presentation speech. Initially, OpenAI's Whisper model transcribes audio into a "Raw Text" that captures all filler words and repetitions. This text is then processed by a T5 (Text-to-Text Transfer Transformer) model using a specific "disfluency correction" prompt to remove linguistic noise and generate a grammatically precise "Clean Text."

Finally, the system performs a token-level comparative analysis between the Raw and Clean texts to identify the specific types and frequencies of disfluencies. These analytical results are returned via a FastAPI-based server in a structured JSON format, which is then visualized on a dashboard to facilitate user self-correction.

IV. Experimental Result

The experimental results of the proposed system are reported in this section. The three core modules—slide and script generation, non-verbal behavior analysis, and disfluency detection and refinement—were trained and evaluated on a workstation equipped with an NVIDIA A6000 GPU. Each module was implemented and tested independently in accordance with its specific task characteristics.

The experiments aim to assess both the performance and practical feasibility of each module within the overall presentation coaching pipeline. Instead of relying on a single end-to-end evaluation metric, the analysis focuses on module-wise results, reflecting the modular design of the proposed system. The following subsections present the experimental setup and evaluation outcomes for each module in detail.

A. LLM-Based Slide Title and Script Generation

The Slide Title and Script Generation module takes a DOCX document as input and performs slide construction and script generation as the first stage of the proposed pipeline. The input document enters the system directly without manual preprocessing and proceeds through the document handling process.

For evaluation, the module processes a structured English background document describing Jeju Island. The document covers multiple thematic domains, including geological characteristics, climate and ecosystems, natural heritage, cultural traditions, economic transformation, and environmental and social issues. As the document does not contain predefined slide boundaries or presentation-oriented annotations, the module determines appropriate slide segmentation based on semantic structure and topic transitions.

After extracting textual content, the module partitions the document into presentation units and generates slide titles and slide-level summaries using a LLaMA-2 8B language model deployed on a GPU server. To ensure presentation-oriented outputs, the module applies task-specific prompts that

emphasize topic coherence, contextual continuity across slides, and spoken-style narration. Rather than producing all outputs in a single step, the module sequentially generates slide titles, contextual summaries, and presentation scripts according to the constructed slide structure.

Using the generated slide structure, the module produces presentation scripts that reflect the logical emphasis and narrative flow of each slide. The system stores all generation results in a structured JSON format, which supports subsequent visualization and feedback presentation within the proposed platform.

Representative examples of the generated slide titles, summaries, and presentation scripts are presented in Table 1.

TABLE I
EXAMPLES OF GENERATED SLIDE TITLES, SUMMARIES, AND PRESENTATION SCRIPTS

ID	Slide Title	Slide Summary	Script Excerpt
S1	Geological Formation and Landscapes of Jeju Island	Jeju Island is a volcanic island formed through repeated eruptions over the past 1.2 million years. The island is composed largely of basaltic lava, with Hallasan rising at its center.	Jeju Island, located at the southernmost edge of the Korean Peninsula, has a fascinating geological history. Over millions of years, volcanic activity shaped its unique landscapes, dominated by basaltic formations and Hallasan, the highest mountain in Korea.
S2	Climate and Ecosystems of Jeju Island	Jeju experiences a mild oceanic climate with high humidity, strong winds, and higher-than-average precipitation, leading to diverse ecosystems influenced by Hallasan.	Jeju Island's climate is characterized by warm summers, mild winters, and frequent winds from the surrounding seas. These conditions have created diverse ecosystems, ranging from subtropical vegetation to unique high-altitude habitats.
S3	Jeju's Role in Renewable Energy and Smart City Initiatives	Jeju plays a leading role in renewable energy and smart-city initiatives, with high wind and solar penetration rates and widespread electric vehicle adoption.	Jeju Island is at the forefront of Korea's renewable energy transition. With extensive wind and solar infrastructure and active smart-city pilot projects, the island serves as a living testbed for low-carbon development.

B. Computer Vision-based Nonverbal Behavior Analysis

The proposed gaze estimation system achieved a precision of 87.3%, a recall of 84.6%, and an overall accuracy of 86.1%. Analysis of gaze distribution revealed that presenters, on average, directed approximately 40% of their gaze toward the central region of the screen. Furthermore, effective presenters tended to distribute their gaze more evenly across the nine predefined screen regions. The system automatically identifies time intervals where gaze is excessively concentrated in a limited region and provides feedback emphasizing the need for improved gaze dispersion. The proposed CNN–LSTM hybrid model achieved a classification accuracy of 78.3%. The class-wise F1-scores were 0.82 for "good," 0.76 for "normal," and 0.77 for "poor." Considering the limited size of the training dataset (50 presentation videos) and the inherently subjective nature of posture evaluation, this performance is regarded as reasonable. Analysis of the confusion matrix indicates occasional misclassification between the "normal" and "poor" classes, which can be attributed to the ambiguous boundary between these posture categories. In terms of computational performance, the gaze estimation module operated at an average speed of 28 FPS (36 ms per frame), while the pose analysis module achieved 24 FPS (42 ms per frame). These processing speeds are sufficient for practical post-processing feedback in presentation training scenarios.

TABLE II
PERFORMANCE EVALUATION OF NONVERBAL BEHAVIOR ANALYSIS MODULES

Module	Metric	Performance
Gaze Estimation	Precision	87.3%
	Recall	84.6%
	Accuracy	86.1%
	Processing Speed	28 FPS
Pose Analysis	Accuracy	78.3%
	F1-Score (Good)	0.82
	F1-Score (Normal)	0.76
	F1-Score (Poor)	0.77
	Processing Speed	24 FPS

C. Performance of Disfluency Detection & Refinement

The proposed T5-based disfluency detection system demonstrated high refinement performance and real-time efficiency using actual presentation data. The integrated Whisper-T5 pipeline intelligently removed various disfluencies—including fillers, stutters, and repetitions—achieving an identification accuracy of over 90% while significantly enhancing readability and delivery. Performance evaluations on an NVIDIA RTX A6000 server showed an average latency of less than one second for one minute of audio, confirming its capability for immediate feedback. Furthermore, token-level comparative analysis provided structured data on disfluency patterns, offering presenters effective, objective metrics for linguistic habit self-correction.

V. DISCUSSION

The experimental results demonstrate that the proposed system is feasible as a modular presentation coaching pipeline supporting both preparation and rehearsal feedback. Since

TABLE III
EXAMPLES OF DISFLUENCY REFINEMENT AND DETECTION RESULTS

Category	Raw Transcript	Refined Script	Detected
Filler	"Um, today's talk is, uh..."	"Today's talk is..."	um, uh
Repetition	"This is... this is the point."	"This is the point."	this is
Stutter	"The sy... sy... system."	"The system."	sy... sy...
Redundancy	"I actually, like, enjoyed it."	"I really enjoyed it."	actually, like

each module operates independently, the discussion focuses on module-level behavior rather than a unified end-to-end metric.

The LLM-based slide and script generation module serves as the foundation of the system by transforming unstructured documents into presentation-oriented content. While the generation quality is not evaluated using explicit quantitative metrics, the use of task-specific prompts enables consistent slide structure and narrative flow. However, the quality of generated scripts remains dependent on the organization of the input document, indicating that lightweight user review is still necessary.

For nonverbal behavior analysis, the gaze estimation module achieved stable performance at practical processing speeds, providing interpretable indicators such as gaze dispersion and central fixation. The posture analysis module showed reasonable accuracy given the limited dataset size, though ambiguity between posture categories suggests the need for clearer evaluation criteria and larger datasets.

The disfluency detection and refinement module effectively converted raw speech into refined scripts while providing structured disfluency statistics. Nevertheless, detection reliability may vary with recording conditions, highlighting the importance of presenting refinement results as suggestions rather than automatic replacements.

A key design choice of the proposed system is the absence of model-level multimodal fusion. Instead, modality-specific results are organized through a unified dashboard, improving interpretability and extensibility while limiting direct modeling of cross-modal interactions.

VI. CONCLUSION

This study presents an AI-based presentation coaching system that supports the full presentation workflow, including preparation, practice, and feedback. The proposed system integrates slide and script generation with multimodal analysis of gaze, posture, and speech to provide accessible and comprehensive presentation training. By enabling repeated self-directed practice without professional coaching, the system offers a practical solution to presentation anxiety and limited training opportunities.

Despite its practical applicability, several limitations remain, including restricted training data for vision models, sensitivity to environmental conditions, dependency on document quality for script generation, and limited multilingual support. Future work will focus on improving generalization and robustness through large-scale data augmentation for vision models, fine-tuning LLMs with domain-specific presentation data, and expanding multilingual capabilities. Additionally, incorporating context-aware and culturally adaptive analysis is expected to further enhance feedback quality and real-world applicability.

ACKNOWLEDGMENTS

Following are results of a study on the "Convergence and Open Sharing System" Project, supported by the Ministry of Education and National Research Foundation of Korea. This research was also supported by the Basic Science Research Program to the Research Institute for Basic Sciences (RIBS) of Jeju National University through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2019-NR040080).

REFERENCES

[1] Graduate Management Admission Council, "Corporate recruiters survey 2024," 2024, [Online]. Accessed: Dec. 2024. [Online]. Available: https://www.gmac.com/-/media/files/gmac/research/employment-outlook/2024-corporate-recruiters-survey/2024_gmac_research_crs_deansummary.pdf

[2] Insivia, "70% say presentation skills are critical for career success," 2024, [Online]. Accessed: Dec. 2024. [Online]. Available: https://www.insivia.com/quoter/70-say-presentation-skills-are-critical-for-career-success/

[3] N. Fourati, A. Barkar, M. Dragée, L. Danthon-Lefebvre, and M. Chollet, "Probing experts' perspectives on ai-assisted public speaking training," arXiv preprint arXiv:2507.07930, 2025, [Online]. Accessed: Jul. 2025. [Online]. Available: https://arxiv.org/abs/2507.07930

[4] Novoresume, "60+ eye-opening public speaking statistics you should know," 2025, [Online]. Accessed: Dec. 2024. [Online]. Available: https://novoresume.com/career-blog/public-speaking-statistics

[5] Buffalo 7, "Uk employees and presentations survey," 2021, [Online]. Accessed: Dec. 2024. [Online]. Available: https://buffalo7.co.uk/blog/presentations-skills-survey/

[6] X. Ochoa and F. Dominguez, "Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1615–1630, 2020, [Online]. Accessed: Dec. 2024. [Online]. Available: https://doi.org/10.1111/bjet.12987

[7] E. N. Kimani, P. Murali, A. Shamekhi, D. Parmar, S. Munikoti, and T. Bickmore, "Multimodal assessment of oral presentations using hmms," in *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20)*. Virtual Event, Netherlands: Association for Computing Machinery, 2020, pp. 650–654, [Online]. Accessed: Dec. 2024. [Online]. Available: https://doi.org/10.1145/3382507.3418888

[8] T. Aggarwal and A. Bhand, "Pass: Presentation automation for slide generation and speech," arXiv preprint arXiv:2501.06497, 2025, [Online]. Accessed: Dec. 2024. [Online]. Available: https://arxiv.org/abs/2501.06497

[9] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020, [Online]. Accessed: Dec. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/8851398

[10] V. Bhat, P. Jyothi, and P. Bhattacharyya, "Adversarial training for low-resource disfluency correction," arXiv preprint arXiv:2306.06384, 2023, [Online]. Accessed: Dec. 2024. [Online]. Available: https://arxiv.org/abs/2306.06384