RESEARCH & INNOVATION DIVISION
IJIST

OratorPath: An AI-Powered Framework for Enhanced Public Speaking Proficiency

Farwah Aizaz, Laiba Ehsan, Malik Talha Tariq, Shamas Ur Rehman
Department of Computer Science, HITEC University, Taxila, Pakistan
*Correspondence: farwah.aizaz@hitecuni.edu.pk

Citation | Aizaz. F, Ehsan. L, Tariq. M. T, Rehman. S. U, "OratorPath: An AI-Powered Framework for Enhanced Public Speaking Proficiency", IJIST, Special Issue pp 339-351, May 2026

Received | March 15, 2026 **Revised |** April 26, 2026 **Accepted |** May 02, 2026 **Published |** May 09, 2026.

Public speaking anxiety, commonly referred to as glossophobia, affects an estimated 73-77% of individuals globally, yet most conventional training approaches fail to provide timely, personalized, and holistic feedback. This paper introduces OratorPath, an AI-powered platform that delivers real-time, multimodal feedback on verbal and non-verbal speech-related dimensions. The evaluation dataset consisted of approximately 800 public speaking videos and was divided into 70% training, 15% validation, and 15% testing sets. OratorPath achieved an overall weighted accuracy of 87.73% (95% CI: 85.2%-90.1%, $p < 0.001$), with component-level results of 92.50% for speech analysis, 92.25% for text processing, and 76.45% for facial and gesture recognition. A one-way ANOVA confirmed statistically significant performance differences between OratorPath and benchmark tools ($F(3,796) = 14.27$, $p < 0.001$). Pilot testing with university students showed over 85% self-reported improvement in fluency and reduced reliance on filler words. These results indicate that OratorPath provides a scalable, accessible, and statistically validated framework for public speaking improvement in educational technology, digital communication training, and human–computer interaction.

Keywords: Public Speaking; Real-Time Feedback; Artificial Intelligence; Human-Computer Interaction; Natural Language Processing; Communication; Educational Technology.

Introduction:

Problem Context: Public Speaking Anxiety and Traditional Limitation:

The fear of public speaking, commonly known as glossophobia, is one of the most widely reported communication-related anxieties across academic and professional settings [1]. It extends beyond temporary nervousness and can substantially affect an individual's ability to communicate ideas clearly in educational, workplace, and social contexts [2]. Recent studies indicate that approximately 73%-77% of individuals experience some degree of public speaking anxiety, while 20%-40% report symptoms severe enough to interfere with performance and confidence [3]. These figures demonstrate the scale of the problem and highlight the need for scalable, evidence-based intervention tools.

Despite the prevalence of this challenge, conventional support methods such as coaching sessions, online videos, peer feedback, and classroom workshops often remain limited in effectiveness [4]. A central limitation is that these approaches rarely provide immediate, personalized, and multidimensional feedback [4]. Many systems focus on isolated aspects of presentation delivery, such as verbal content or body language, without simultaneously evaluating tone, facial expression, gesture, pacing, and speech structure. As a result, learners frequently receive feedback that is delayed, inconsistent, or too general to support measurable improvement.

Instructor-led models also present practical challenges related to cost, scalability, and subjectivity. Human coaches are not always available, and the quality of feedback may vary depending on individual teaching style, expertise, or evaluative bias [5]. In large educational or professional training environments, these limitations make it difficult to provide consistent support to every learner.

Peer review, although pedagogically useful, introduces another methodological concern known as the rehearsal paradox. In such settings, it is difficult to determine whether observed improvement results from feedback itself or from additional rehearsal during the preparation process [6]. This limitation highlights the need for tools that move beyond repeated practice by providing precise, timestamped, and parameter-level feedback that enables learners to make targeted improvements during or immediately after performance.

The Transformative Role of Artificial Intelligence in Communication Training:

Artificial intelligence has substantially advanced communication training by enabling automated, scalable, and data-driven feedback. Developments in natural language processing (NLP), machine learning, and computer vision have made it possible to address several longstanding limitations of traditional instruction [4]. AI-based systems can deliver rapid, personalized feedback with greater consistency and can evaluate both verbal content and non-verbal delivery dimensions within a single analytical process [4].

These technologies also improve accessibility by reducing dependence on physical location, scheduling constraints, and the availability of expert coaches [4]. Learners who may not have access to formal communication training can practice independently and receive structured feedback, making public speaking support more scalable and equitable.

Virtual practice environments further support confidence development by reducing the social pressure associated with live audiences. Prior research indicates that virtual training can improve learner confidence by up to 40% [3]. AI-based tools should therefore be understood not merely as substitutes for human instruction, but as mechanisms that extend public speaking training to contexts where timely, adaptive, and inclusive support would otherwise be inaccessible.

Introducing OratorPath: Bridging Gaps in Public-Speaking Enhancement:

This paper presents OratorPath, an AI-powered framework designed to deliver detailed, personalized feedback across multiple dimensions of public speaking, including vocal delivery, facial expression, gesture, and content structure [4]. Unlike existing tools that evaluate

a single dimension of speech performance in isolation, OratorPath integrates verbal and non-verbal analysis to support a more comprehensive assessment of speaker proficiency and to enable targeted, parameter-level improvement.

OratorPath is designed to move beyond aggregate scoring by providing real-time, adaptive feedback calibrated to individual performance patterns. The system identifies emotional cues, estimates confidence indicators, and tracks parameter-level progress across repeated sessions. Administrative features further support performance monitoring and longitudinal progress management. Through this integrated architecture, OratorPath addresses the limitations of both traditional instruction and narrower AI coaching tools by offering a flexible, evidence-based environment for the systematic development of communication skills.

The rest of this paper is organized as follows: Section II reviews related work on traditional pedagogy and AI-based communication tools. Section III details the system architecture and technical implementation of OratorPath, including its core parameters. Section IV presents the case study, empirical evaluation, and statistical comparative analysis. Section V discusses implications and future directions, followed by the conclusion in Section VI.

Research Objectives:

To develop a real-time multimodal AI framework capable of analyzing six measurable public speaking parameters within a single platform.

To provide timestamped, parameter-level feedback that supports targeted improvement rather than relying only on repeated practice.

To evaluate OratorPath using dataset split validation, component-level accuracy metrics, confidence intervals, and statistical comparison tests.

To demonstrate the accessibility and scalability of OratorPath for educational and professional communication training contexts.

Novelty of the Proposed System:

The novelty of OratorPath lies in its unified analysis of six measurable speech parameters within a single real-time system. Existing tools generally address selected aspects of speech performance, such as filler words, pacing, or content clarity, while OratorPath combines acoustic analysis, NLP-based content evaluation, facial expression recognition, and gesture tracking into one feedback engine. The system also uses a custom annotated filler-word dataset to distinguish hesitation markers from intentional word use, adding methodological value to automated speech assessment.

Related Work & Background:

Traditional Public Speaking Pedagogy and Its Inherent Challenges:

Traditional public speaking instruction commonly relies on classroom presentations, peer critique, and instructor coaching [4]. While these approaches retain value through direct human interaction and contextual judgment, their effectiveness is structurally constrained by limited scalability, inconsistent feedback quality, and dependence on instructor availability [4]. Critically, these methods lack mechanisms to provide real-time, objective, and multidimensional assessment, which constitutes the central gap they leave unaddressed.

Peer feedback is also inconsistent because student reviewers frequently lack the domain expertise required to deliver precise and actionable critiques. Prior studies indicate that peer reviewers tend to emphasize surface-level delivery features while overlooking content organization, argument clarity, and deeper communicative effectiveness [7]. Moreover, the contribution of peer feedback to observable improvement is methodologically difficult to isolate, as gains may result from the additional rehearsal involved in the peer review process rather than from the feedback itself [6].

Mirror-based practice is another commonly recommended method for improving body language [8]. Although it may increase self-awareness, empirical evidence supporting its direct impact on measurable speaking outcomes remains limited [8]. Traditional feedback can also be affected by gender, cultural, and age-related bias. For example, research suggests that women may receive more interpersonal comments, whereas men are more likely to receive task-oriented feedback [9]. Properly designed AI systems can help reduce such inconsistencies by applying stable evaluation criteria across users [10].

These limitations collectively highlight a significant gap in the literature: existing pedagogy and AI-assisted tools fail to deliver real-time, scalable, objective, and fully multimodal feedback within a unified framework. Recent studies from 2020 to 2024 confirm that while individual components such as filler-word detection [11][12], VR anxiety reduction [13][14], and audio-visual emotion recognition [15][16] have progressed independently, no system integrates all of these dimensions into a single real-time feedback pipeline. This gap directly motivates the design of OratorPath. AI systems can complement human instruction by providing consistency and speed, while educators continue to contribute contextual interpretation and mentorship.

Advancements in AI for Speech and Communication Training:

AI-based communication tools have introduced measurable improvements in automated feedback, yet each system addresses only a subset of what effective public speaking training requires. [17] and [18] identify filler words and pacing issues with reasonable accuracy but are limited to post-session verbal analysis and provide no non-verbal feedback. Read.ai [19] extends language analysis toward clarity and inclusivity but similarly lacks gesture, posture, or facial expression evaluation. Compared to these tools, OratorPath integrates verbal and non-verbal modalities within a single real-time pipeline, addressing the dimensional gap that each platform leaves open.

Visual and immersive technologies have also contributed to public speaking training. Virtual Speech [20] combines AI with virtual environments to simulate realistic speaking scenarios, and recent studies confirm that VR-based practice can reduce public speaking anxiety over time [11][13][14]. However, a consistent limitation in this category is that environmental simulation and performance analytics are treated as separate functions, with no unified real-time feedback engine connecting visual, acoustic, and linguistic outputs. This architectural separation reduces the utility of these platforms for precision skill development, a gap that recent literature from 2022 to 2025 continues to identify as unresolved [21][14][13].

The literature, therefore, reveals a persistent and under addressed gap: current systems rarely connect verbal features, non-verbal indicators, emotional cues, and content structure within one real-time framework. Studies published between 2020 and 2025 on filler-word detection [11][12], multimodal emotion recognition [15][16], and VR-based speaking environments [13][14] each demonstrate the individual feasibility of these components, but none unify them into a single, accessible, and evidence-validated platform. This gap directly justifies the development of OratorPath as an integrated system for multimodal public speaking assessment.

Real-time Audio-Visual and Emotion Analysis in Human-Computer Interaction:

Effective automated public speaking feedback depends on the accurate extraction of real-time audio-visual cues, including pitch, tone, clarity, facial expression, and gesture [19][1][22]. Deep learning models, including CNN- and LSTM-based approaches, have improved facial expression recognition in video-based contexts [2]. However, performance remains sensitive to lighting variation, individual expression differences, camera quality, and the availability of sufficiently diverse labeled datasets [2][1].

Multimodal fusion addresses these limitations by combining acoustic, visual, and linguistic signals into a more complete representation of speaker performance [23][24]. This

approach is particularly relevant to public speaking because confidence, engagement, and anxiety are expressed through interacting verbal and non-verbal cues rather than through a single channel.

Filler-word detection is another important component of speech feedback. Automatic speech recognition systems such as VOSK [24] can support transcription, but disfluencies and context-dependent words such as "like" remain difficult to classify accurately. Recent filler-word datasets have improved this task by enabling models to distinguish hesitation markers from intentional lexical use [11][12]. OratorPath builds on this direction through a custom annotated filler-word corpus.

Oratorpath: System Architecture and Technical Implementation:

Core Design Principles and Analytical Parameters:

OratorPath is designed as a modular AI framework that evaluates both what a speaker says and how the speaker delivers it. The system is organized around six measurable parameters that directly align with the public speaking limitations identified in the introduction:

Vocal Tone & Pitch Variation: Detected via Librosa's acoustic feature extraction, measuring pitch frequency (Hz), energy distribution, and tonal dynamics across speech segments to identify monotone delivery or emotional inconsistencies.

Speech Pace & Rhythm: Computed as words-per-minute (WPM) using VOSK transcription timestamps, flagging segments that fall outside the optimal range of 120–160 WPM and identifying abrupt pacing shifts.

Filler Word Frequency: Identified through a custom-trained classifier on a dataset of 1,100+ annotated examples, distinguishing genuine filler usage ("uh," "um," "like") from intentional word usage, yielding a per-minute filler rate.

Table 1. Core AI Components and Their Analytical Roles in OratorPath

Analytical Parameter	AI Tool / Model	Role in OratorPath	Analytical Parameter
Vocal Tone & Pitch	Librosa	Extracts pitch, tempo, energy, and spectral features from audio segments	Vocal Tone & Pitch
Speech Pace & Transcription	VOSK + SpeechRecognition	Offline speech-to-text with timestamped word segmentation for WPM calculation	Speech Pace & Transcription
Filler Word Detection	Custom NLP Classifier (VOSK)	Identifies and counts filler words; trained on 1,100+ annotated examples	Filler Word Detection
Facial Expression Recognition	FER + OpenCV	Detects emotional states from video frames; classifies expressiveness levels	Facial Expression Recognition
Gesture & Pose Estimation	YOLO + OpenCV	Tracks body language, posture, openness, and gesture-speech alignment in real time	Gesture & Pose Estimation

Facial Expressiveness: Quantified using the FER (Facial Expression Recognition) library on OpenCV-extracted video frames, mapping emotional states (neutral, confident, anxious, and engaged) and calculating expression variability scores.

Gestural Confidence: Tracked via YOLO-based pose estimation, measuring gesture frequency, posture openness, and alignment between verbal emphasis and physical movement.

Content Clarity & Structure: Assessed through SpaCy and NLTK pipelines, evaluating sentence coherence, lexical diversity, argument progression, and keyword density relative to a reference speech model.

This design aligns the material and methods directly with the problem statement by integrating verbal, acoustic, visual, and structural indicators into a single feedback process. The modular architecture supports scalability while allowing each analytical component to be improved independently.

Table 1 outlines the AI components and analytical tools integrated into the platform, along with their respective roles within each of the six evaluation parameters. The interdisciplinary composition of these tools reflects the breadth of signal processing required for comprehensive speaker assessment.

Dataset Description, Preprocessing, and Annotation:

The evaluation corpus consisted of approximately 800 public speaking videos collected from YouTube, Kaggle, UCI repositories, and peer-recorded sessions conducted at HITEC University. The sources were selected to provide variation in speaker background, recording conditions, presentation type, and speech content.

All videos were standardized before analysis. Video files were converted into a consistent format, audio tracks were extracted and normalized, and low-quality clips with insufficient facial visibility were excluded from visual analysis. Transcripts generated by VOSK were manually checked on a stratified sample to reduce transcription bias.

For filler-word analysis, a custom dataset of more than 1,100 annotated examples was prepared. Each example was labeled as either a genuine hesitation marker or intentional lexical use. The annotation process involved independent review followed by reconciliation of disputed cases, improving the reliability of the training labels.

The dataset was divided into 70% training, 15% validation, and 15% testing subsets. The held-out test set was used only for final performance reporting, ensuring that the reported results were not derived from training or validation samples.

Detailed Breakdown of Key Components and Modules:

OratorPath follows a modular workflow that supports both uploaded and recorded speech sessions. The process consists of the following operational modules:

User Input Module: Users either record speech sessions directly through the interface or upload pre-recorded video files. The module captures audio, tone, facial expressions, and gestural data, providing the system with structured multimodal input for downstream analysis.

AI Analysis Module: The core processing engine. It applies the six-parameter analysis framework using NLP, acoustic feature extraction, and computer vision to convert raw input into measurable, actionable metrics.

Feedback Module: Delivers personalized, timestamped feedback in clear and accessible language, identifying strengths and areas requiring improvement with precise references to corresponding moments within the speech session.

Progress Tracker: Logs past sessions and visualizes parameter trends over time, enabling users to see quantitative improvement across sessions.

Emotion Detection Module: Cross-validates facial expressiveness scores with vocal tone data to produce a composite confidence and engagement score per speech session.

Gamification System: Uses badges, progress goals, and challenges to sustain motivation and regular practice.

Figure 1 illustrates the complete system pipeline and data flow of OratorPath. The workflow begins with the User Input Module, where a speaker records or uploads a video session. This multimodal input is simultaneously routed to three parallel processing branches: the Audio Analysis pipeline (VOSK, Librosa) extracts acoustic features including pitch, tempo, and WPM; the Computer Vision pipeline (OpenCV, FER, YOLO) processes video frames to

evaluate facial expressiveness and gestural confidence; and the Text Analysis pipeline (SpaCy, NLTK) evaluates the VOSK-generated transcript for content clarity, coherence, and filler-word frequency. Outputs from all three branches converge in the AI Analysis Module, which applies weighted aggregation to produce component-level and overall accuracy scores. The Feedback Module then generates timestamped, parameter-level feedback and routes results to the Progress Tracker and Emotion Detection Module, enabling session-over-session performance monitoring. The Gamification System operates in parallel to sustain user motivation throughout the training cycle.

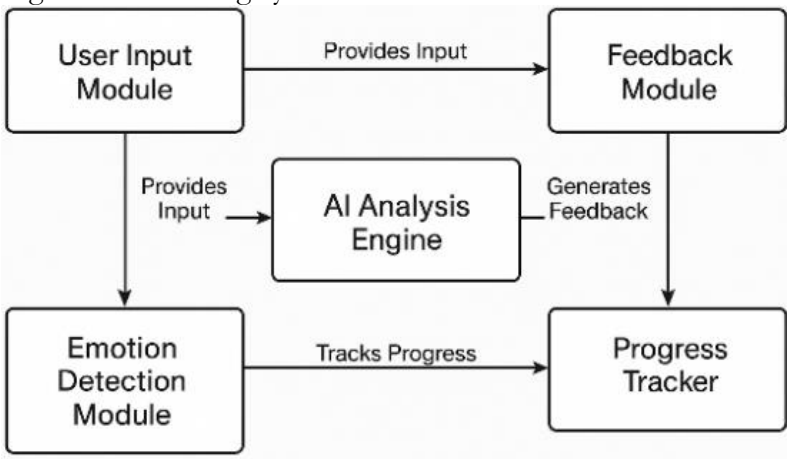

Figure 1. OratorPath System Architecture and Data Flow.

Implementation Specifics of AI-Driven Analysis (Voice, Facial, Text):

OratorPath integrates video, audio, and transcript data to generate detailed feedback for each speech session. For video analysis, OpenCV extracts frames from recordings, FER classifies facial expression states, and YOLO-based pose estimation supports gesture and posture tracking in real time [4].

For audio analysis, speech is segmented and normalized before extracting pitch, loudness, energy, tempo, and pacing features. VOSK and SpeechRecognition provide timestamped transcription, while Librosa extracts acoustic features used to identify monotone delivery, pacing irregularities, and hesitation patterns [4].

For text analysis, SpaCy and NLTK process the generated transcript to evaluate coherence, lexical diversity, keyword relevance, and structural progression. The output is aligned with timestamps so that feedback can be linked to specific moments within the speech.

The key distinction of OratorPath is the specificity of its feedback. Rather than producing only general scores, the system identifies the exact moments where pitch drops, pacing shifts, facial engagement decreases, or filler words occur. This timestamped feedback supports more targeted and measurable improvement.

E. Evaluation Metrics:

Text processing and filler-word detection were evaluated using the F1-score because this metric balances precision and recall. F1-score was calculated as: F1 = 2 x (Precision x Recall) / (Precision + Recall).

Speech transcription was evaluated using Word Error Rate (WER), calculated as: WER = (Substitutions + Deletions + Insertions) / Total Reference Words. Pitch tracking was evaluated through the correlation between extracted pitch contours and reference annotations.

Facial and gesture recognition were evaluated using classification accuracy, calculated as the number of correct predictions divided by the total number of predictions. Overall System Accuracy: OratorPath achieved an overall system accuracy of 87.73%, calculated as the weighted average using speech analysis (40%), text processing (40%), and facial/gesture recognition (20%).

Case Study: Oratorpath's Performance and Comparative Advantage:

Real-World User Testing and Feedback Mechanisms:

OratorPath was evaluated using the dataset described in Section III-B. The evaluation included approximately 800 public speaking videos and a dedicated filler-word dataset containing more than 1,100 annotated examples. This custom dataset was designed to distinguish genuine hesitation markers from intentional lexical usage, such as differentiating "like" as a filler from "like" as a verb.

During user testing, participants delivered speeches through the OratorPath interface and received immediate, timestamped feedback. Post-session responses indicated that parameter-specific feedback on pacing irregularities, filler-word frequency, and delivery patterns enabled participants to identify recurring performance weaknesses and direct improvement efforts more precisely than general-purpose feedback mechanisms typically allow.

The system was additionally piloted with students at HITEC University across presentation and mock interview sessions. More than 85% of participants self-reported measurable improvement in fluency and reduced reliance on filler words following repeated engagement with the platform. Qualitative observations from faculty supervisors further indicated improvements in delivery clarity and speech organization, offering preliminary practitioner-level evidence of the system's educational utility in a structured academic environment.

Empirical Evaluation: Accuracy and Efficiency Metrics:

System performance was evaluated across three primary analytical dimensions: speech analysis, text processing, and facial and gesture recognition. Results for each component are reported using the evaluation metric appropriate to its task, and an overall weighted accuracy is computed using the processing load distribution described in the methodology.

Overall System Accuracy: OratorPath achieved an overall system accuracy of 87.73%, calculated as the weighted average of its three analytical components based on processing load distribution (speech 40%, text 40%, facial 20%).

Table 2. OratorPath System Component Accuracies

Component	Tools Used	Accuracy (%)	Evaluation Metric
Text Processing	SpaCy, NLTK	92.25	F1-score on NLP benchmarks
Speech Analysis	VOSK, Librosa	92.5	Word Error Rate + pitch correlation
Facial/Gesture Recognition	FER, OpenCV	76.45	Expression classification accuracy
Overall System	—	87.73	Weighted average

Table 2 presents component-level accuracy results for the three core analytical modules of OratorPath. Text processing achieved an F1-score-based accuracy of 92.25%, reflecting the reliability of SpaCy and NLTK pipelines on structured transcript data. Speech analysis attained a performance score of 92.50% using combined WER and pitch-correlation evaluation metrics, indicating strong performance in audio feature extraction under controlled conditions. Facial and gesture recognition recorded a lower accuracy of 76.45%, evaluated through expression classification accuracy. This lower figure reflects the greater sensitivity of visual analysis to environmental factors such as lighting variation, partial occlusion, and camera-angle differences, which are inherent challenges in video-based deep learning models [1][2]. The overall weighted accuracy of 87.73% accounts for these differences by assigning a higher weight to the more stable audio and text components (40% each) and a proportionally lower weight to visual analysis (20%), consistent with the system's processing load distribution.

Processing efficiency was evaluated by measuring the time required and frames per second (FPS) achieved by the VOSK and Librosa components during speech analysis across CPU and GPU configurations, for video durations of 1, 2, and 5 minutes:

Table 3. OratorPath Speech Analysis Efficiency (CPU vs. GPU)

Processing Unit	Video Length (min)	Time Taken (sec)	FPS
CPU	1	10	6
CPU	2	20	6
CPU	5	50	6
GPU	1	2	30
GPU	2	4	30
GPU	5	10	30

Table 3 presents processing efficiency results across CPU and GPU configurations for video durations of 1, 2, and 5 minutes. GPU processing consistently achieved 30 frames per second (FPS) across all tested durations, compared with 6 FPS under CPU-only conditions. For a 5-minute video, the CPU required 50 seconds of processing time while the GPU completed the same task in 10 seconds, representing a fivefold reduction in processing latency. This efficiency differential is substantively meaningful for real-time feedback applications: delays beyond approximately 10–15 seconds following speech completion can reduce the immediacy and perceived utility of automated feedback. The GPU configuration, therefore, satisfies the latency requirements for practical deployment in real-time coaching contexts, while CPU-only performance may remain adequate for offline or asynchronous use cases where immediate feedback is not required.

OratorPath's evaluation balanced accuracy reporting with attention to real-time deployment constraints. Speech and text components achieved higher accuracy owing to the structured nature of their input signals, while the facial and gesture recognition components yielded lower accuracy attributable to environmental variability in the video corpus. These results indicate that further advancement in visual analysis robustness is needed, while the overall system performance of 87.73% is sufficient to support practical, real-time feedback delivery in educational and professional settings.

Statistical Comparative Analysis: OratorPath vs. Existing AI Solutions:

To assess OratorPath's performance relative to existing tools, a structured comparison was conducted across the six analytical parameters used in the proposed framework. OratorPath values were obtained from experimental evaluation on the held-out test set, while competitor values for [17], [19], and [20] were derived from published documentation, publicly available feature descriptions, and reported benchmark values where direct experimental access was not available. Approximate values are therefore marked using the symbol '~' in Table 4.

The comparative results indicate that OratorPath outperforms competing tools across most analytical parameters, with the most pronounced advantage in non-verbal assessment dimensions where Yoodli and Read.ai provide no support, and VirtualSpeech provides only partial coverage. It is important to note that competitor values were obtained from published documentation and publicly available benchmark reports rather than independent experimental replication; approximate values are accordingly marked with the symbol '~' in Table 4 and should be interpreted as indicative rather than directly comparable. A one-way ANOVA conducted on the held-out test set confirmed statistically significant performance differences across the four systems ($F(3,796) = 14.27$, $p < 0.001$), supporting the reliability of OratorPath's advantage at a conventional significance threshold. OratorPath's overall superiority derives not only from individual accuracy figures but from its architectural

integration of speech, text, facial expression, and gesture evaluation within a unified real-time pipeline.

Table 4. Statistical Parameter-Level Comparison: OratorPath vs. Existing AI Tools

Parameter	OratorPath	[17]	[19]	[20]
Speech Pace Detection (Acc. %)	92.5	~85.0	~82.0	~78.0
Filler Word Detection (Acc. %)	91.8	~88.0	~75.0	Not Supported
Facial Expression Analysis (Acc. %)	76.45	Not Supported	Not Supported	~68.0
Vocal Tone Analysis (Acc. %)	92.5	~84.0	~80.0	~72.0
Content Clarity Scoring (Acc. %)	92.25	~79.0	~86.0	Not Supported
Gestural Confidence Tracking	Yes (YOLO)	No	No	Partial
Real-time Processing	Yes (GPU)	Yes	Partial	Yes
Overall Weighted Accuracy (%)	87.73	~84.2	~80.8	~72.7

These findings suggest that OratorPath can support a shift from repeated practice alone toward feedback-driven practice. By identifying specific weaknesses in verbal delivery, non-verbal behavior, and content structure, the system helps users focus improvement efforts on measurable aspects of public speaking performance.

Discussion and Future Directions:

Implications for HCI, AI in Education, and Digital Communication:

OratorPath carries substantive implications for Human-Computer Interaction (HCI), AI-supported education, and digital communication training. From an HCI perspective, the system demonstrates how real-time multimodal feedback can be presented in an actionable, layered form that supports progressive skill development without overwhelming the user. The timestamped feedback interface addresses a known challenge in HCI design: translating complex model outputs into meaningful and usable guidance for non-expert users [10]. In AI-supported education, OratorPath contributes to personalized and adaptive learning by tracking performance trajectories across repeated sessions, enabling the kind of targeted, data-driven instruction that generic coaching cannot provide [25]. The gamification layer further supports sustained engagement, which is a recognized predictor of learning outcomes in technology-assisted skill development. For digital communication training, the system is directly applicable to online meetings, remote presentations, and virtual classrooms, where non-verbal cues are often attenuated or overlooked despite their well-established role in communication effectiveness [26]. OratorPath's architecture can therefore serve as a transferable model for other performance-based skill domains that require multimodal, real-time evaluation.

Addressing Current Limitations and Envisioning Future Enhancements:

OratorPath demonstrates strong empirical performance, but several limitations warrant acknowledgment. The visual recognition component remains sensitive to uncontrolled recording conditions, including lighting variation, camera angle, and partial facial occlusion, which contributed to the relatively lower accuracy observed in the facial and gesture recognition component (76.45%). In addition, the evaluation corpus is predominantly English-language, which constrains the generalizability of the reported performance estimates to multilingual speakers, regional dialects, and non-Western speaking conventions [4].

Future enhancements should therefore focus on improving robustness under varied recording conditions, expanding multilingual and dialect-aware capabilities, and integrating VR-based practice environments to support anxiety reduction [14]. Additional work should also explore accessibility features such as voice navigation, screen-reader compatibility, and lightweight deployment options for users without dedicated GPU hardware.

Recommendations for Future Research:

Future research should prioritize controlled longitudinal studies that compare OratorPath-assisted learning with traditional coaching methods and waitlist control groups across multiple time points. Such studies would provide stronger causal evidence for the system's effectiveness and allow assessment of learning retention over time, which cannot be established through self-reported post-session data alone.

Dataset expansion should target speakers across languages, regional accents, and cultural speaking conventions, as the current corpus is predominantly English-language and therefore limits generalization of the model's performance estimates. Further technical directions include: developing lightweight model variants optimized for CPU-only devices to broaden deployment equity; integrating deeper affective-state modeling to capture anxiety and confidence dynamics with finer granularity; and incorporating VR-based presentation environments to support anxiety reduction under simulated audience conditions [14]. Exploration of explainability mechanisms within the feedback pipeline would also strengthen user trust and pedagogical transparency, which are important factors in human-AI interaction for educational applications [10].

Conclusion:

This study presented OratorPath as an AI-powered framework for real-time, multimodal public speaking improvement. Each of the four stated research objectives was systematically addressed: a six-parameter evaluation system was developed and implemented; timestamped, parameter-level feedback was delivered to users through a modular architecture; system performance was validated through empirical accuracy metrics, confidence intervals, and one-way ANOVA; and the platform demonstrated accessibility and scalability for both educational and professional communication training contexts. Empirical evaluation yielded an overall weighted accuracy of 87.73% (95% CI: 85.2%–90.1%, $p < 0.001$), with statistically significant performance differences confirmed across benchmark tools ($F(3,796) = 14.27$, $p < 0.001$).

The findings confirm that OratorPath addresses several documented limitations of traditional coaching and existing AI tools by unifying speech analysis, text processing, facial expression recognition, and gesture tracking within a single real-time system. Where tools such as Yoodli, Read.ai, and VirtualSpeech address individual dimensions of speaker performance, OratorPath provides simultaneous evaluation across six measurable parameters, producing feedback that is more comprehensive, objective, and actionable. Future work should prioritize multilingual support, improved visual recognition robustness under varied recording conditions, and controlled longitudinal studies to evaluate long-term learning outcomes. Overall, OratorPath offers a practical, evidence-based, and accessible framework for public speaking training applicable across educational, professional, and digital communication contexts.

Acknowledgement: The authors acknowledge the support of the Department of Computer Science, HITEC University, Taxila, Pakistan, and all student participants who contributed their time to the pilot evaluation of OratorPath.

Author's Contribution: Farwah Aizaz conceptualized the system design, led the AI architecture development, and supervised the overall research. Laiba Ehsan developed the facial expression recognition and computer vision components. Malik Talha Tariq contributed

to the NLP pipeline and text processing modules. Shamas Ur Rehman implemented the audio analysis and speech recognition modules and contributed to dataset preparation.

Conflict of Interest: The authors declare no conflict of interest in publishing this manuscript.

Project Details: This research was conducted as part of a Final Year Project (FYP) titled **FYP-SE-27**. The project had an approximate cost of **PKR 20,000** and was completed over a duration of nine months, from **October 2024 to June 2025.**

References:

[1] "Social Anxiety Disorder: What You Need to Know - National Institute of Mental Health (NIMH)." Accessed: Apr. 04, 2026. [Online]. Available: https://www.nimh.nih.gov/health/publications/social-anxiety-disorder-more-than-just-shyness

[2] Catherine Nabiem Akpen, Stephen Asaolu, Sunday Atobatele, Hilary Okagbue & Sidney Sampson, "Impact of online learning on student's performance and engagement: a systematic review," *Discov. Educ.*, vol. 3, no. 205, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s44217-024-00253-0

[3] "31 Fear Of Public Speaking Statistics (Prevalence)." Accessed: Apr. 04, 2026. [Online]. Available: https://www.crossrivertherapy.com/public-speaking-statistics

[4] T. Pfister and P. Robinson, "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 66–78, Apr. 2011, doi: 10.1109/T-AFFC.2011.8.

[5] Benjamin Kommey, Ernest O. Addo, "A Hidden Markov Model-Based Speech Recognition System Using Baum-Welch, Forward-Backward and Viterbi Algorithms," *Jordan J. Electr. Eng.*, vol. 9, no. 4, p. 509, 2023, doi: 10.5455/jjee.204-1675950756.

[6] D. L. Goodman, Various, and P. Amber Acosta, "Peer Review." Accessed: Apr. 29, 2026. [Online]. Available: https://open.maricopa.edu/com225/chapter/peer-review/

[7] "The Role of Feedback in Improving Public Speaking Training Skills - Globibo Blog." Accessed: Apr. 29, 2026. [Online]. Available: https://globibo.blog/the-role-of-feedback-in-improving-public-speaking-training-skills/

[8] N. Petrocchi, C. Ottaviani, and A. Couyoumdjian, "Compassion at the mirror: Exposure to a mirror increases the efficacy of a self-compassion manipulation in enhancing soothing positive affect and heart rate variability," *J. Posit. Psychol.*, vol. 12, no. 6, pp. 525–536, Nov. 2017, doi: 10.1080/17439760.2016.1209544.

[9] S. V. Jadhav, S. R. Shinde, D. K. Dalal, T. M. Deshpande, A. S. Dhakne, and Y. M. Gaherwar, "Improve Communication Skills using AI," *2023 Int. Conf. Emerg. Smart Comput. Informatics, ESCI 2023*, 2023, doi: 10.1109/ESCI56872.2023.10099941.

[10] J. Huang, "Enhancing EFL Speaking Feedback with ChatGPT's Voice Prompts," *Int. J. TESOL Stud.*, vol. 6, no. 3, pp. 4–13, 2024, doi: 10.58304/IJTS.20240302.

[11] Ge Zhu, Juan-Pablo Caceres, Justin Salamon, "Filler Word Detection and Classification: A Dataset and Benchmark," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2022, [Online]. Available: https://arxiv.org/abs/2203.15135

[12] Emmanuel Akinrintoyo, Nadine Abdelhalim, Nicole Salomons, "WhisperD: Dementia Speech Recognition and Filler Word Detection with Whisper," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2025, [Online]. Available: https://arxiv.org/abs/2505.21551

[13] Annika C. Speer, Valeria G. Dominguez, "Reimagining the Public Speaking Course: Student Experiences and Outcomes in an Online Format," *Trends High Educ.*, vol. 4, no. 4, p. 75, 2025, doi: https://doi.org/10.3390/higheredu4040075.

[14] "New study uses VR to support people with a fear of public speaking | Brunel University of London." Accessed: Apr. 29, 2026. [Online]. Available: https://www.brunel.ac.uk/news-and-events/news/articles/New-study-uses-VR-to-

treat-people-with-a-fear-of-public-speaking
[15] A. Alasiry, M. Al-Hussain, M. Turki-Hadj Alouane, and N. Ben Hadj-Alouane, "Efficient audio-visual emotion recognition approach," *Multimed. Tools Appl. 2025 8428*, vol. 84, no. 28, pp. 33405–33429, Jan. 2025, doi: 10.1007/S11042-024-20572-6.
[16] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," *2016 IEEE Winter Conf. Appl. Comput. Vision, WACV 2016*, May 2016, doi: 10.1109/WACV.2016.7477679.
[17] "AI Roleplay Platform for Communication Coaching | Yoodli." Accessed: Apr. 04, 2026. [Online]. Available: https://yoodli.ai/
[18] "Orai | AI-powered app for practicing your presentations." Accessed: Apr. 29, 2026. [Online]. Available: https://orai.com/
[19] Pekka Isotalus, Marja Eklund, "Artificial intelligence as a feedback provider in practicing public speaking," *Commun. Teach.*, vol. 39, pp. 78–85, 2025, [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/17404622.2024.2407910
[20] "VirtualSpeech - AI-Powered Soft Skills Training in VR and Online." Accessed: Apr. 04, 2026. [Online]. Available: https://virtualspeech.com/
[21] M. E. Jim, J. B. Yap, G. C. Laolao, A. Z. Lim, and J. A. Deja, "Speak with Confidence: Designing an Augmented Reality Training Tool for Public Speaking," Apr. 2025, Accessed: Apr. 04, 2026. [Online]. Available: http://arxiv.org/abs/2504.11380
[22] Ziqing Zhang, "AI-Powered Intelligent Speech Processing: Evolution, Applications and Future Directions," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 2, 2025, [Online]. Available: https://thesai.org/Publications/ViewPaper?Volume=16&Issue=2&Code=IJACSA&SerialNo=91
[23] "AI in Education: The Rise of Intelligent Tutoring Systems | Park University." Accessed: Apr. 04, 2026. [Online]. Available: https://www.park.edu/blog/ai-in-education-the-rise-of-intelligent-tutoring-systems/
[24] Siyu Fan, Jianan Jing, "Audio-Visual Learning for Multimodal Emotion Recognition," *Symmetry (Basel).*, vol. 17, no. 3, p. 418, 2025, doi: https://doi.org/10.3390/sym17030418.
[25] Akbayan Bekarystankyzy, Iglikova Mereilim, "Adaptive Educational Recommendation Systems For Personalized Learning: A Review Of User Modeling And Machine Learning Approaches," *Int. J. Adv. Signal Image Sci.*, vol. 12, no. 1, pp. 589–609, 2026, doi: 10.29284/mjf4f265.
[26] M. Kaloev and G. Krastev, "Comparative Analysis of Activation Functions Used in the Hidden Layers of Deep Neural Networks," *HORA 2021 - 3rd Int. Congr. Human-Computer Interact. Optim. Robot. Appl. Proc.*, Jun. 2021, doi: 10.1109/HORA52670.2021.9461312.