Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

EmotiCon: Enhancing Public Speaking Training with Context-Aware Feedback on Emotional Delivery

Yash DESAI [a*1], Praneeth KASIRAJU[a+1], Syaamantak DAS[a], Ramkumar RAJENDRAN[a]
[a]*Centre for Educational Technology, IIT Bombay, India*
[*]desaiyash7272@gmail.com, [+]praneethkasiraju@gmail.com

Abstract: Effective public speaking relies on aligning emotional delivery with semantic content. While existing digital coaches analyze paralinguistic cues like tone and pitch, they often neglect the context of what is being said. This paper introduces EmotiCon, a novel system that offers context-aware emotional feedback by analyzing a speaker's audio, transcribed text, and preparation materials. EmotiCon segments speech, detects emotion across modalities, and compares it with relevant contextual passages to assess emotional congruence. It then generates targeted recommendations using a Large Language Model. A user study showed that context-aware feedback was better received than simple feedback. Quantitative evaluation using the extended Technology Acceptance Model (TAM) followed by semi-structured interviews revealed strong scores**.** Our findings suggest that integrating semantic context into feedback generation makes digital speech coaching tools more actionable and effective for learners.

Keywords: Public Speaking, Digital Coach, Emotion Recognition, Semantic Context, Large Language Models, Multimodal Analysis, Automatic Visualisation Feedback, Emotional Speech Training, System Evaluation

1. Introduction

In Technology-Enhanced Language Learning (TELL), developing oral proficiency, particularly public speaking, is a key objective. Public speaking, simply defined as "speaking in front of a community or group of people" (Bilgin, 2022), is central to academic, professional, and social communication. A foundational element of mastering this skill is timely, actionable feedback (Butler & Winne, 1995), which helps learners identify weaknesses and track progress.

Context plays a crucial role in shaping such feedback. It can be defined as the background, topic and setting that clarifies the meaning of the utterance (Rigotti & Rocci, 2006). Effective delivery also requires alignment between tone and the emotional weight or intention of the message (Guyer et al., 2021). In this work, we define *delivery* as the extent to which a speaker's expressed emotion aligns with the semantic content and context of their speech.

Beyond technical accuracy, speakers must convey passion, empathy, or urgency, depending on their message as delivery is influenced by the ability to form an emotional bond (Entong et al., 2024). Prior research on affect analysis of teachers' speech found that monotonous delivery is not well received (T S & Rajendran, 2022). Emotionally resonant speech has historically inspired movements and shaped discourse (Charteris-Black, 2011), underscoring that emotional expression is not merely stylistic but pedagogical. Yet most AI tools flatten emotional nuance by separating vocal expression from semantic context when generating feedback.

[1] *Authors Yash Desai and Praneeth Kasiraju made equal contributions in this paper.*

Recent advances in AI have led to the emergence of improved speech training tools. SpeakAR (Jim et al., 2025) creates immersive audience environments to simulate real-world pressure and manage anxiety. VoiceCoach (Wang et al., 2020) trains modulation using patterns learned from over 2,600 TED Talks. BETTER (Wynn & Wang, 2023) extracts audio tonality and speech emotions, and visualizes their alignment for users to view and infer. However, none fully integrate semantic context into feedback loops for improving emotional delivery.

Tools like **Yoodli**[2] and **Poised**[3] leverage LLMs to assess user delivery in real time while offering suggestions on clarity and confidence. They also factor in the conversational environment or interview setting making feedback more natural. Additionally, recent research has started to explore contextual adaptation in speech training. For example, (Zhang et al., 2025) proposed incorporating environmental context to detect speaker anxiety and offer adaptive recommendations. However, both approaches treat *setting* as context, not the semantic background of the speech. Factoring in the actual content or source material remains underexplored. This gap motivates our Research Question (RQ): **To what extent does integrating semantic context into feedback enhance learner confidence and emotional delivery in public speaking tasks?**

We present **EmotiCon**, a system that integrates semantic context into emotion-based public speaking feedback. EmotiCon accepts a speech audio file and the contextual materials the speaker used for preparation (e.g., essays, articles). It segments the speech using audio-based emotion analysis and transcribes each segment for text emotion analysis. For each segment, it retrieves the most relevant preparation material, analyzes its emotion, and compares alignment across audio, transcript, and context. A Large Language Model then generates actionable recommendations, visualized through an interactive dashboard.

We evaluated EmotiCon in a user study with eight higher-education students. Each submitted a short speech and preparation materials, then received context-aware feedback through the dashboard before taking part in semi-structured interviews. Participants found the feedback meaningful, helping them identify tone-content mismatches and boosting confidence for future presentations. Complementing the qualitative findings, participants completed a TAM questionnaire, which showed strong ratings for Perceived Usefulness (5.8/7), Attitude Toward Use (6.37/7), and a dedicated RQ dimension with questions to capture the research question, namely whether EmotiCon enhanced learners' confidence and delivery (6.37/7).

2. System Architecture

EmotiCon is designed to provide visual feedback of the user's varying emotions throughout the speech and give context-based recommendations for the same. EmotiCon implements a multimodal emotion analysis pipeline for audio data, which combines speech, text, and context modalities to produce granular visualisations of the segments.

As shown in Figure 1, EmotiCon first loads the user's speech audio and divides it into non-overlapping emotion segments. Each audio segment is processed using a pretrained "HuBERT"[4]-based transformer model ("superb/hubert-large-superb-er"[5]) for Speech Emotion Recognition (SER). The HuBERT model outputs a probability distribution over four emotion classes (happy, sad, angry, neutral) and selects the highest as the segment's dominant emotion with its prediction confidence[6]. Consecutive audio segments having the same

[2] Yoodli - AI powered speech coaching tool.
[3] Poised - Real-time communication feedback tool.
[4] HuBERT - Self-supervised speech representation model.
[5] Superb-Hubert - Benchmark suite version of Hubert for speech tasks.
[6] Confidence scores of the emotions detected by the model are referred to here as prediction confidence

dominant emotion are merged together. Following this, each audio segment is transcribed into text using Assembly AI's[7] external Automatic Speech Recognition (ASR) service. The transcribed text for each audio segment is analyzed using a transformer-based model, specifically the "j-hartmann/emotion-english-distilroberta-base"[8] (a fine-tuned DistilRoBERTa[9]) for emotion classification. The model outputs a probability distribution over emotion classes for each segment and records the highest-probability class as the segment's predicted emotion along with its prediction confidence. Labels from different sources are normalized to canonical forms.

The system extracts text from the context PDF and divides it into chunks. Each chunk is encoded into a dense vector using a pretrained sentence-transformer model. The embedding of each audio transcript is compared to these chunk embeddings, and the most relevant chunk is retrieved as contextual background. Each retrieved context chunk is analyzed with a text emotion classifier, and its dominant emotion and prediction confidence are recorded. Recommendations are generated by comparing the emotional alignment of the audio with the transcribed text and the context, and phrased by Gemini.

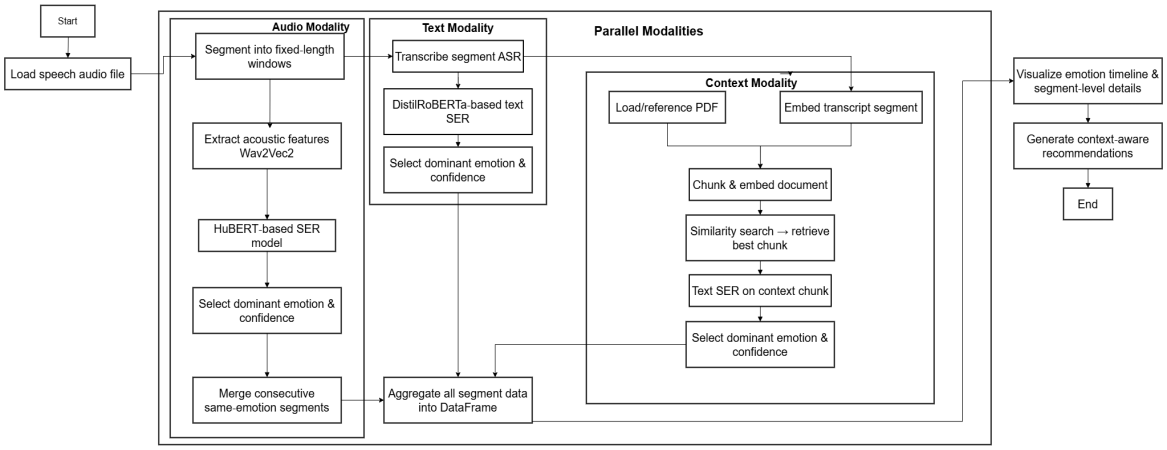

Figure 1: Emoticon's analysis and feedback generation pipeline integrating audio, text, and context modalities and generating recommendations.

3. Visual Feedback Dashboard

Figure 2 shows an interactive dashboard displaying each speech segment horizontally with the emotion concordance of the three sources (audio, text, context) revealing per-segment emotion details on hover. The color coding used for the segments in Figure 2 is:

- 🟩 **Green:** All three emotion sources match.
- 🟨 **Yellow:** Any two match.
- 🟥 **Red:** All differ.

A sidebar lists the segments by timestamp to navigate the speech. Selecting a segment updates the main panel to show a detailed analysis for that time frame, including the detected audio emotion, the transcribed text, the relevant context chunk, and the emotion scores for each modality.

[7] Assembly AI - Speech-to-text and audio intelligence API.
[8] Hartmann's DistilRoBERTa - Fine-tuned variant of DistilRoberta for NLP tasks.
[9] DistilRoBERTa - Fine-tuned variant of DistilRoberta for NLP tasks.

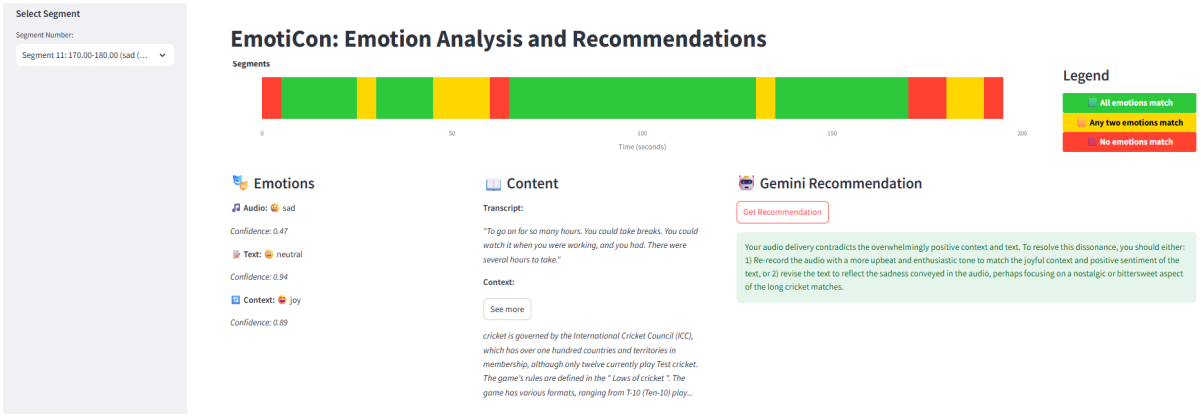

Figure 2: EmotiCon's feedback dashboard with emotion alignment visualization and recommendations.

Furthermore, the user can generate a specific recommendation for the selected segment. The system passes the segment's emotions, context, transcribed text, and neighboring transcripts into Google's Gemini 1.5 (temperature set to 0.2 to reduce hallucinations) to produce a response (Woollaston et al., 2024).

4. Methodology of Study Design

Figure 3: Study design flowchart.

A pilot study was conducted with eight higher-education students (7 male, 1 female) aged 18-35, from diverse academic and linguistic backgrounds. All had prior experience with academic presentations in English, which was chosen as the delivery language to ensure comparability. Participants selected one topic from five context-demanding options, chosen to require critical thinking, synthesis of sources, and avoidance of personal anecdotes. They were given 30 minutes to prepare a speech and were required to document all referenced textual material (URLs, sources, copied text). Audio or video sources were not permitted to maintain parity in semantic context processing, and participants were instructed to avoid unverifiable statements. Each participant then delivered a 3-6 minute speech before an audience, recorded in audio and video. The collected context material was compiled into a PDF and, along with the audio recording, served as input to EmotiCon. After this, participants explored their own speech analysis through the interactive dashboard, followed by a semi-structured interview to gather feedback on the context-based recommendations.

5. Results and Discussions:

RQ: To what extent does integrating semantic context into feedback enhance learner's confidence and emotional delivery in public speaking tasks?

To answer our RQ, we first conducted a qualitative analysis of the semi-structured interviews. Participant responses were reviewed and summarized, with illustrative quotes highlighting recurring insights or specific comments. The interviews showed that context-based feedback increased participants' awareness of their speaking style and conveyed emotions. One participant noted it enabled them to "set the narrative to pursue

their audience". The "clean, temporal segmentation of the speech" was described as informative, and using context as a modality was considered "novel for digital coaches". Several participants said the segment-level emotion feedback and alignment visualization made them "more confident, as they did not think about it before". Recommendations were perceived as engaging and insightful, aligning with prior work on the role of context in emotional delivery. Participants also noted that the UI "displays too much at once" and suggested condensing recommendations into shorter points. Another suggestion was an "audio modulator" feature to play back segments aligned with the intended emotional tone.

To quantitatively evaluate the system, participants completed a Technology Acceptance Model - TAM (Granić & Marangunić, 2019) based questionnaire rating 17 items on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). We assessed 6 parameters in the questionnaire, namely - Perceived Usefulness (PU): The specificity and actionable quality of the context-aware recommendations. Perceived Ease of Use (PEOU): The clarity and usability of the dashboard. Perceived Enjoyment (PE): Engagement with interactive feedback. Attitude Toward Use (ATT) & Intention to Use (INT): Overall disposition towards the AI coach and likelihood of future use. In addition to the standard TAM constructs, we included an "RQ" dimension aligned with our research question, asking participants whether EmotiCon gave participants actionable feedback to improve their confidence and delivery.

Table 1. Results of the mean and standard deviation of the 6 parameters of the TAM questionnaire and the RQ dimension.

	PU	**PEOU**	**PE**	**ATT**	**INT**	**RQ**
Mean	5.8	5.43	5.41	6.37	5.31	6.37
Standard Deviation	0.55	0.96	1.76	0.41	1.83	0.74

The TAM and RQ results from Table 1, reinforce the qualitative findings. Perceived Usefulness scored highly (M = 5.8), indicating that participants found the recommendations actionable, consistent with their remarks about identifying tone–content mismatches. Attitude toward Use was the highest-rated dimension (M = 6.37), reflecting strong positivity and aligning with feedback that the system felt "novel" and "engaging." The added RQ dimension, measuring perceived improvements in confidence and delivery, also received a high score (M = 6.37), supporting interview claims that participants felt more confident and aware of their speaking style. By contrast, Intention to Use showed more variation (M = 5.31, SD = 1.83), suggesting the dashboard being overwhelming. These results suggest that EmotiCon was well received, with participants finding the system's feedback both actionable and useful.

6. Conclusion

We introduced EmotiCon, a novel system designed to enhance public speaking training by providing context-aware feedback on emotional delivery. The results of the pilot study validate and reaffirm the importance of integrating semantic context in feedback with multimodal analysis (audio, text and source material), which generates specific and effective recommendations for a speech. Such systems could benefit non native English speakers to practice their delivery and help them improve their communication skills. EmotiCon could potentially contribute to treating social anxiety or stage fright by providing a safe space to learn and practice emotion and tone in speech delivery.

Despite the positive outcomes, there exist limitations to be acknowledged. First, the system in its current state has no method to detect intentionally conflicting emotions, such as sarcasm, which makes it ineffective for satirical or comedic speeches. Additionally, our study only included participants over 18, so the impact on younger audiences remains unexplored.

In future work, we would like to address the shortcomings of the system and focus on simplifying the user dashboard for easier interpretation of scores and reducing the volume of data on-screen for a smoother experience. Additionally, detecting and supporting ironic styles or figures of speech and conducting broader studies across age groups and use cases is also planned. Addressing factors such as speech impairments and language nativity while expanding to facial recognition will help enhance the recommendation system making it more inclusive, reliable, and personalized.

7. Acknowledgment

The authors would like to thank all the participants and anonymous reviewers of this article. This work is supported by the SBI Foundation Hub for Data Science and Analytics under grant ID SBIF002-007.

References

Bilgin, R. (2022). A review of public speaking and its components. *Canadian Journal of Educational and Social Studies, 2*(3), 37–49. https://doi.org/10.53103/cjess.v2i3.39

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Rigotti, E., & Rocci, A. (2006). Towards a definition of context. *Studies in Communication Sciences, 6/2,* 155-180. https://doi.org/10.5169/SEALS-791116

Guyer, J. J., Briñol, P., Vaughan-Johnston, T. I., Fabrigar, L. R., Moreno, L., & Petty, R. E. (2021). Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments. *Journal of Nonverbal Behavior*, *45*(4), 479–504. https://doi.org/10.1007/s10919-021-00374-2

Entong, M. B. M., Garil, B. A., Muarip, V. C., & Chavez, J. V. (2024). Language delivery styles in academic trainings: Analysis of speaker's emotional connection to audience for lasting learning. *Forum for Linguistic Studies*, *6*(3), 326–342. https://doi.org/10.30564/fls.v6i3.6533

T S, A., & Rajendran, R. (2022). Audio feature based monotone detection and affect analysis for teachers. *2022 IEEE Region 10 Symposium (TENSYMP)* (pp. 1–6). IEEE. https://doi.org/10.1109/TENSYMP54529.2022.9864393

Charteris-Black, J. (2011). *Politicians and rhetoric: The persuasive power of metaphor*. Springer.

Jim, M. E., Yap, J. B., Laolao, G. C., Lim, A. Z., & Deja, J. A. (2025). *Speak with confidence: Designing an augmented reality training tool for public speaking.* arXiv. https://doi.org/10.48550/arXiv.2504.11380

Wang, X., Zeng, H., Wang, Y., Wu, A., Sun, Z., Ma, X., & Qu, H. (2020). VoiceCoach: Interactive evidence-based training for voice modulation skills in public speaking. *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). ACM. https://doi.org/10.1145/3313831.3376726

Wynn, A., & Wang, J. (2023). BETTER: An Automatic feedBack systEm for supporTing emoTional spEech tRaining. *Lecture Notes in Computer Science,* 746–752. https://doi.org/10.1007/978-3-031-36272-9_66

Zhang, T., Tan, J., Xiong, Z., Wu, B., & Zheng, C. (2025). Causality-guided context-aware multimodal public speaking anxiety detection for out-of-distribution generalization. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICASSP49660.2025.10888426

Woollaston, S., Flanagan, B., Ocheja, P., Dai, Y., & Ogata, H. (2024). TAMMY: Supporting EFL translation practice with an LLM-powered chatbot. In *Proceedings of the 32nd International Conference on Computers in Education (ICCE 2024)*. APSCE. https://doi.org/10.58459/icce.2024.4901

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology, 50*(5), 2572–2593. https://doi.org/10.1111/bjet.12864