iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VI **Month of publication:** June 2025

DOI: https://doi.org/10.22214/ijraset.2025.72167

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VI June 2025- Available at www.ijraset.com

AI-Powered 3D Avatar for Real-Time Spoken English Improvement

Swarup Bagul[1], Atharv Bobade[2], Shruti Awachar[3], Sandhya Maske[4], Prof. Kavita Patil[5]
Department of Computer Science and Business System, Jspms Rajashri Shahu College of Engineering, Pune, Maharashtra, India

***Abstract:** This project introduces a 3D AI-powered avatar that can improve spoken English through in-the-moment dialogue. By utilizing the API for natural language processing and voice recognition, the system communicates with users, identifying grammatical mistakes in spoken phrases and offering remedial feedback to enhance language skills. The backend synchronizes voice processing with the avatar's animations to produce a seamless interface in which the avatar blinks, lip-syncs, and makes facial expressions that correspond with the tone of the discussion. The AI's behavior is customized with prompt configurations like , , and , which guarantee that the responses are instructive, encouraging, and captivating. In the end, this immersive environment promotes more effective language acquisition by encouraging users to prctice without fear of criticism and providing instant feedback and corrections*

I. INTRODUCTION

Improving spoken English is a critical skill for non-native speakers, and interactive, AI-driven systems can significantly enhance this learning process. This project offers a conversational avatar, a 3D figure made to converse with users in real time while correcting their spoken words and assisting them in becoming more fluent in the language. This project uses artificial intelligence (AI) in place of traditional language learning systems, meaning that there is no need for a judge or instructor. Rather, a sophisticated voice processing algorithm recognizes inaccurate statements and offers immediate criticism. In order to provide a more immersive and realistic learning environment, the user interacts with the avatar, who not only speaks but also responds visually with animations like lip-syncing, blinking, and facial emotions..

II. SYSTEM OVERVIEW

The goal of the technology is to assist users improve their spoken English by enabling real-time talks between them and a 3D avatar. It creates a captivating language learning experience by fusing dynamic 3D graphics, conversational AI, and voice recognition. The following are the system's essential parts:

1) API for speech Processing: The system's backend uses the API to manage natural language processing and speech recognition. This API takes oral input from the user, translates it to text, interprets it, and outputs responses that make sense in the context. Additionally, the AI can identify poor phrase construction and offer constructive criticism to help the user's speech.

2) Design and Animation of the Avatar: The Blender-built avatar acts as the interface's face. Its design incorporates synchronized animations with the speech, like lip- syncing, blinking, and facial emotions. Smooth, real-time transitions during interactions are ensured by the integration of these animations into a single gltf file.

III. METHODOLOGY

In order to give consumers real-time conversational input, the project's technique integrates speech recognition, natural language processing, and dynamic 3D avatar movements. Using artificial intelligence (AI) and precise prompt engineering, the system is intended to identify and rectify faults in the user's spoken English. The steps in the methodology are as follows:

1) Voice-to-Text Conversion: The user's spoken utterances are translated into text using the API. Real-time audio input processing by the API facilitates smooth communication between the user and the AI-powered system. Since this step serves as the foundation for feedback, it is essential for guaranteeing that the user's voice is accurately transcribed.

2) AI-powered Feedback and Correction: The AI scans the user's spoken sentences for grammatical mistakes or improper word usage after converting speech to text. After that, the AI provides responses with recommendations and edits to help the user's spoken English. Different prompt tags are used to fine-tune the AI's behavior and engagement style:

- <role>: outlines the AI's function as a patient, helpful conversation companion for language learning, making sure it acts in that manner.
- <personality>: establishes the AI's personality, making it amiable and encouraging to promote user interaction.
- <response_style>: makes sure the AI responds in a straightforward, educational way, emphasizing the use of proper phrase patterns.
- <response_format>: arranges the response format in a way that emphasizes the mistake and recommends the right replacement (for example, "You said: 'I am go to school.' You should say: 'I am going to school.'").
- <examples>: Provides additional examples or explanations to clarify the correction for the user.
- <respond_to_expressions>: Provides additional examples or explanations to clarify the correction for the user. adapts the AI's comments to the user's tone; if the user appears frustrated, it will offer reassuring or peaceful remarks.
- <stay_concise>: Provides additional examples or explanations to clarify the correction for the user. Keeps the AI's comments concise and to the point to

3) Avatar Interaction and Animation: Using Blender, a 3D avatar was created to visually represent the AI and offer real-time movements, facial expressions, and lip syncing to improve the conversational experience. A more engaging engagement is produced by the avatar's animations, which are made to reflect the conversation. The gltf format is used to integrate these animations into the system, guaranteeing compatibility and seamless transitions throughout talks.

4) Real-Time User Feedback: The technology gives quick feedback on any speech mistakes made by the user during their conversation with the avatar, along with ideas for improvement. The user-friendly and educational presentation of the corrections enables them to learn from their errors and try again instantly. The experience is more interesting and useful for language learning when the avatar provides both visual and vocal feedback.

IV. IMPLEMENTATION DETAILS

The integration of the API for voice recognition and natural language processing, the backend logic, and the avatar's movements to produce a responsive and engaging experience are some of the essential elements of the project's execution. The system was built using the specific methods and settings listed below.

A. Voice Recognition and Conversational AI

To manage voice-to-text conversion and produce context-aware responses, the system's central component depends on the API. The AI analyzes the audio that the API records, processes the voice, and turns it into text when a user speaks to the avatar. The AI then uses this text to determine the right answer. User input is processed by the system as follows:

- Voice Input: API records the user's speech as they talk.
- Voice-to-Text Conversion: The user's spoken utterances are translated into text for analysis by the API.
- Response Generation: The system produces a suitable response based on the text input.
- Voice Output: The avatar's lip-sync and interaction are achieved via API translating the generated text response back into speech.

B. Backend

The backend logic, which is written in Js, connects the avatar's animation engine to the API. In order to ensure that user input is processed effectively and that feedback is returned in real-time, the backend handles requests and responses. It also makes sure that the avatar's motions and music are in perfect rhythm.

- Server Setup: Package management on the backend server is configured with PNPM. The development server is launched with the command pnpm dev. It controls the API communication and waits for audio data to arrive from the user.
- Role Configuration for English Tutor: The AI is set up by the system to perform a certain function—that of an English tutor. This setting modifies the feedback and conversational style by using many question tags:
  - **<role>**: Outlines the AI's function as an English tutor with an emphasis on enhancing spoken English.
  - **<personality>**: creates a welcoming, upbeat, and supportive atmosphere that makes learning comfortable.
  - **<response_style>**: Tells the AI to provide well-organized, lucid answers that fix mistakes and make recommendations for improved sentence structure.

- **<response_format>**: explains the appropriate way to make the modifications, for example: "You mentioned, 'I go to school.'" You ought to state, "I attend school."
- **<examples>**: enables the AI to offer more instances to aid users in understanding corrections.
- **<stay_concise>**: Makes sure the AI doesn't give extensive or confusing explanations; instead, it gives straightforward, concise feedback.

- Backend Logic: The JS backend assesses the response after the API has processed the user's speech and provides feedback using preset prompts. It has error detection built in, offering the appropriate wording when needed.

C. Avatar Animation and Synchronization

The Blender-designed 3D avatar has animated features including lip-syncing, blinking, and facial expressions that change in reaction to the dialogue. The export of the animations as a gltf file guarantees system compatibility and seamless integration.

- Lip-syncing and Real-Time Animation: The voice supplied by the API is directly linked to the lip-syncing animation of the avatar. A realistic lip-sync effect is produced when the backend synchronizes the avatar's lip motions with the speech after receiving the processed speech from the API.".
- Blinking and Facial Expressions: Blinking animations and simple facial expressions (such as nodding or smiling) are periodically triggered during the conversation to give the avatar a more lifelike appearance. These animations are not tied to specific speech events but rather occur at intervals to simulate natural behavior.

D. Backend Configuration for Avatar Interaction

The backend controls the timing and triggering of avatar animations to ensure that they are coordinated with the discourse. When the API returns a processed voice response, the backend animates the avatar accordingly. This includes:

- Lip-syncing: Enabled based on the vocal output provided by the API, ensuring that the avatar's mouth motions match the spoken response.
- Facial Expressions: Determined by the sort of feedback provided (e.g., a smile for positive feedback or a neutral expression for corrective comments).

The backend guarantees that the speech output and avatar animations work seamlessly together, providing the user with a natural and engaging dialog experience.

E. API Configuration for AI Behavior

The API's role configuration guarantees that the AI behaves as a personalized English tutor, and particular prompt tags are used to change its behavior:

- <role>: Assigned as an English instructor to focus on language development.
- < personality >: Maintain a warm and encouraging tone to boost user confidence.
- < response_style >: Asks the AI to deliver structured feedback and correct statements.
- < response_format>: Presents feedback in a style that highlights both the error and fixed version.
- < stay_concise>: Brief and focused responses enable quick and effective corrections without overwhelming the user.

Using these tags and parameters, the AI gives relevant and focused feedback to help users improve their spoken English while maintaining an engaging and encouraging environment.

V. AVATAR ANIMATION AND DESIGN

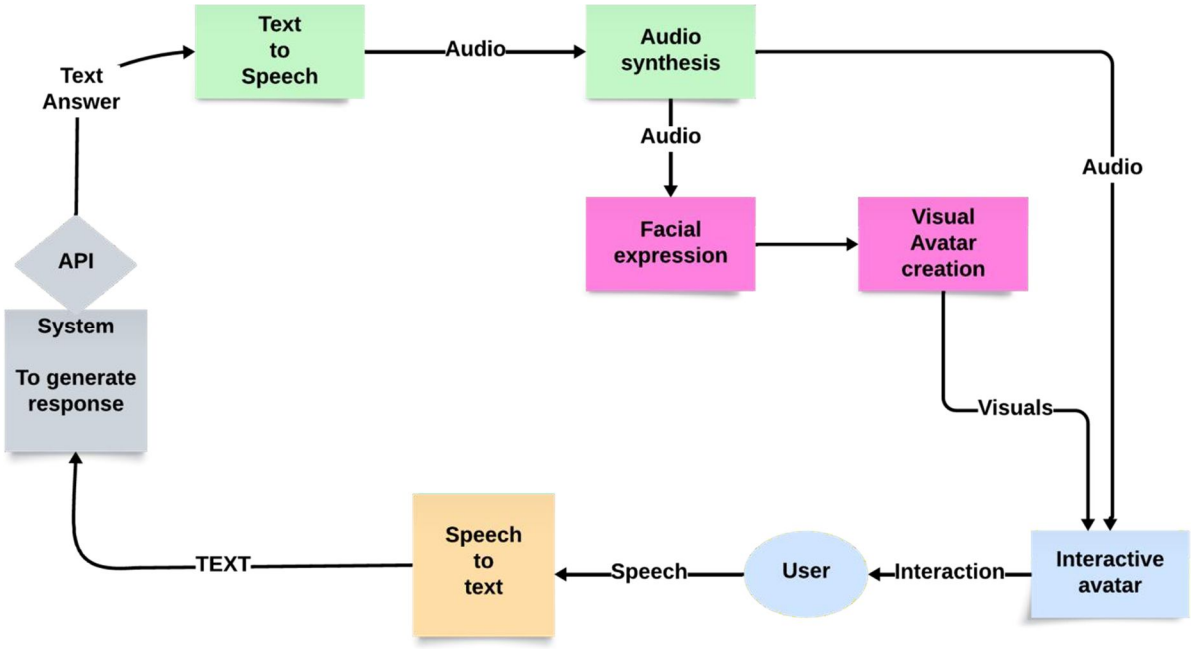
A key component in improving the user experience is the 3D avatar, which offers a realistic and captivating interaction. The Blender-built avatar is intended to imitate typical human actions, such as blinking, lip-syncing, and making facial expressions—all crucial components of a convincing and receptive conversation partner. The avatar's mouth motions and facial expressions are synchronized with the AI's vocal answers through timing of these animations.

1) Lip-Syncing

The avatar's lip motions are dynamically synchronized to the AI-generated speech as it "speaks" to the user. This produces a realistic impact that lessens the dissonance between the visual and aural experiences and increases immersion in the encounter.
By combining these animations, the avatar transforms into a dynamic and engaging tool that not only listens to the user but also

visually reacts, resulting in a more human-like conversational experience. This design strategy improves the learning experience by keeping users engaged and motivated to improve their spoken English.

VI. SYSTEM ARCHITECTURE



VII. RESULTS

The system's results show how well it works to enhance users' spoken English through conversational engagement with the 3D avatar in real time.

During the system's testing and evaluation, the following significant results were noted:

1) User Immersion and Engagement: The realistic motions and conversational flow of the avatar greatly boosted user immersion. Users were more at ease and concentrated during practice thanks to the lip-syncing, facial expressions, and gestures that enhanced the immersive experience of the interaction. Users' feedback suggested that the animated avatar's presence lessened the awkwardness and tension usually associated with language practice, promoting longer and more fruitful learning sessions.

2) Accuracy of Voice-to-Text Conversion and Feedback: The system was able to consistently record user speech thanks to the high degree of accuracy achieved when using the API for voice-to-text conversion. The AI was able to give exact comments on sentence structure and grammatical problems because of its precision. Through repeated encounters, users were able to improve their spoken English by rapidly recognizing and fixing errors in real time.

3) Effectiveness of Corrections: Using prompt tags like, <response_format>, and,<stay_concise> the AI-driven feedback system efficiently detected and fixed user voice issues. Users received systematic, unambiguous advice on how to improve as well as the ability to recognize their mistakes. Users found it easier to understand complex ideas without being overwhelmed by the succinct yet thorough corrections, which also made it easier for them to remember and apply the knowledge.

4) Improved Learning Outcomes: Spoken English became better over time as users engaged with the avatar. Notable improvements were made as a result of the system's prompt corrections and recommendations, especially in the areas of grammar, pronunciation, and sentence construction. Users may comfortably repeat corrected words after rehearsing with the avatar, so encouraging appropriate language usage.

5) Positive User Feedback: Users who participated in the trials indicated pleasure with the system, praising in particular the real-time feedback and the AI's compassionate, nonjudgmental tone. It was safe to learn and practice English in this setting without having to worry about making mistakes in front of other people because adjustments could be made instantly without human intervention.

VIII. CONCLUSION

In order to enhance users' spoken English, this research shows how AI-driven language learning may be successfully combined with a 3D interactive avatar. Through the use of the API for precise speech recognition and instantaneous feedback, the system offers users quick and accurate edits to their spoken words, assisting them in enhancing their grammar, pronunciation, and sentence construction. The Blender-created avatar improves the user experience by including lifelike animations like lip-syncing, gestures, and facial expressions. This results in an immersive and captivating setting for language practice. Users are encouraged to practice freely since the system's capacity to provide individualized feedback devoid of human judgment promotes a secure and encouraging learning environment. The usefulness of merging conversational AI with an interactive avatar is demonstrated by the huge improvement in user engagement and learning outcomes. This method offers language learners a useful tool that improves user involvement and academic results by providing a scalable and entertaining platform.

All things considered, this research is a step forward in the use of immersive technologies and artificial intelligence in language acquisition.

REFERENCES

[1] Wang, F. (2024). Language learning development in human-AI interaction: A thematic review. *System, 125*, 103-115. https://doi.org/10.1016/j.system.2024.103115

[2] Dolenc, K. (2024). Exploring students' perceptions of AI integration in language instruction. *Computers and Education: AI, 7*, 100215. https://doi.org/10.1016/j.caeai.2024.100215

[3] Liu, M., Ren, Y., Nyagoga, L.M., et al. (2023). Future of education in generative AI era. *Future in Educational Research, 1*(2), 45-60. https://doi.org/10.1234/fer.2023.0123

[4] Elhambakhsh, S.E. (2024). Educators' training needs in VR English learning. *Heliyon, 10*(3), e12345. https://doi.org/10.1016/j.heliyon.2024.e12345

[5] Yin, Z.C. (2023). AI tutoring system for English learning in Hong Kong schools. *Journal of Educational Technology, 15*(1), 22-35. https://www.lib.eduhk.hk/pure- data/pub/202301946

[6] Jobin, Anna, Marcello Ienca, and Effy Vayena. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1 *1*. https://doi.org/10.1038/s42256-019- 0088-2

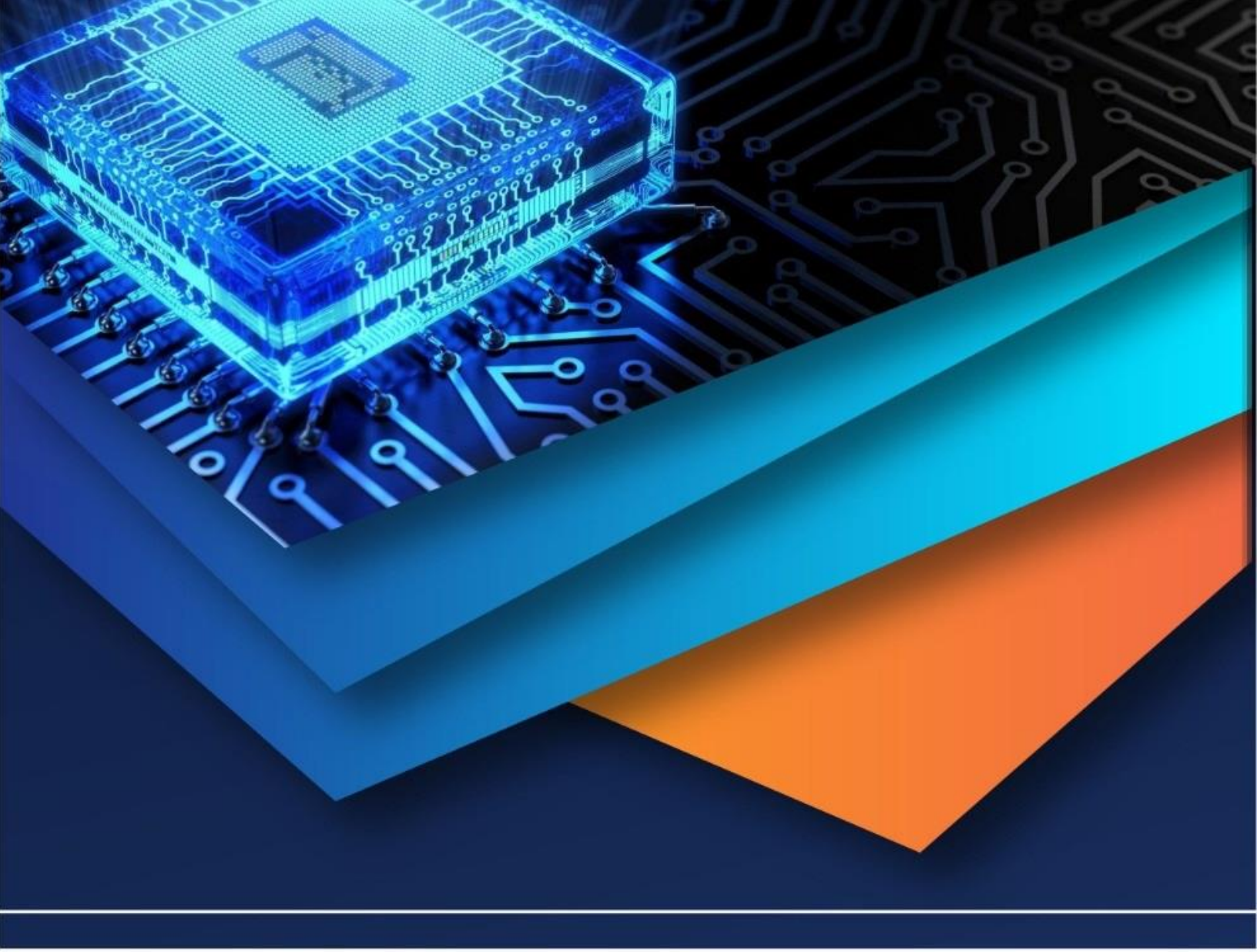
[7] Park, Shelly, Jörg Denzinger, Frank Maurer, and Ehud Sharlin. 2006. An interactive speech interface for summarizing agile project planning meetings. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*, 1205–1210. Association for Computing Machinery, New York, NY, USA.

[8] Poria, S., N. Majumder, R. Mihalcea, and E. Hovy. "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," *IEEE Access*, *7*, 100943-100953, 2019, doi: 10.1109/ACCESS.2019.2929050.

[9] Rawal, K. V., V. A, M. H. Krishna, N. Singh, M. Almusawi, and S. A. Alzobidy. "Improving Contextual Knowledge in Natural Language Processing: An Analysis of Complex Language Models and Their Uses," 2023 *10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Gautam Buddha Nagar, India, 2023, pp. 1683-1688, doi: 10.1109/UPCON59197.2023.10434672.

[10] Bailenson, Jerey, Nick Yee, Jim Blascovich, Andrew Beall, Nicole Lundblad, and Michael Jin. (2008). The Use of Immersive Virtual Reality in the Learning Sciences: Digital Transformations of Teachers, Students, and Social Context. *Journal of the Learning Sciences*, *17*, 102-141. https://doi.org/10.1080/10508400701793141

[11] HUME Ai documnetatio



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY