Grenze International Journal of Engineering and Technology, June Issue

Voca-Learn: AI-Powered Communication Enhancement System

Hitarth Gupta[1], Janhvi Gupta[2] and Prakhar Srivastava[3]
[1-3]ABES Engineering College/Computer Science, Ghaziabad, India
Email: hitarthgupta2@gmail.com, guptajanhvi161@gmail.com, adityasri0804@gmail.com

***Abstract*— Many companies are utilising digital systems and remote technology for collaboration. However, as a result of this, there exists a gap in communication between employees caused by the differences between their primary languages and the misinterpretation of messages within the same language due to a reliance on written communication. In addition, there is a large volume of information that must be interpreted by employees. The AI-Powered Communication Enhancement System provides improvement to human communication with the enhancement of human communication through the application of AI tools in real-time. The use of advanced Natural Language Processing(NLP) technology, such as Transformers like BERT and GPT, are employed in the understanding of language, its context and the analysis of language. The AI-Powered Communication Enhancement System also provides additional tools, such as instant detection of sentiment to provide corrections to tone, predictive reply tools to allow a faster response, and automatic summary generation of long conversation threads. Each of the above tools includes performance evaluation as part of the toolset and each tool is further evaluated through qualitative and quantitative assessment measures (F1-score and latency). From the results derived from the AR Evaluation Tests, the AI-Powered Communication Enhancement System achieved an accuracy of 92% in classifying intent and an average 35% reduction in classification errors, and in addition, it allows individuals to avoid miscommunications that may arise from differences in culture through accurate and nuanced translation and tone adjustment. As a result, the development of AI-Powered Communication Enhancement Systems may allow for increased efficiency and clarity in workflows within organisations while retaining the benefits associated with human interaction.**

***Keywords*— Natural Language Processing (NLP), Speech-to-Text (STT), Speech Recognition, Text-to-Speech (TTS), Real-time Feedback, Machine Learning (ML), Artificial Intelligence (AI).**

I. INTRODUCTION

Professional success and personal development in today's fast-paced environment of international communication, has primarily focused on someone's ability to speak and understand written and/or spoken English. English is the biggest form of communication within the global markets and technology, and being referred to as speaking "fluent English" is the minimum requirement for the workplace. In addition, in order to be fluent in a foreign language, an individual must practice speaking regularly and receive feedback from an instructor who can give immediate, personalized instruction with the necessary corrections. This in itself has been one of the major challenges for traditional teaching methods.

Grenze ID: 01.GIJET.12.2.152_24

In a typical classroom, there will be anywhere from 10 - 50+ students per instructor and the classroom environment limits the instructor's ability to offer students immediate corrective feedback on their pronunciation, grammar, and conversational nuances. As such, there is now a significant demand for intelligent systems that are able to provide interactive learning tools to enhance the teaching/learning process, and to fill the space between traditional methods of education and the ability to offer interactive mentoring.

Voco Learn will be developing a highly intelligent algorithm that will be capable of interpreting spoken language in real-time, comparing the speaker's spoken language over time, and providing corrective feedback to the speaker at every stage of the speaking process, thereby dramatically increasing the accuracy of the spoken language response and decreasing the time spent on manual coding. The project will also create actionable voice insight via an interactive dashboard that offers a visual snapshot of voice conversation trends and emotion analytics.

Digital tools and language applications are available and are currently being developed; however, most of these applications use static (non-interactive) content and simple pattern matching, which doesn't take into account the complexities of human speech. Standard language learning applications tend to rely on repetitive vocabulary acquisition and basic grammatical structure and tend not to provide for nuances such as tone, sentiment, and intent. Historically, the analysis of spoken language has been time-consuming and is subsequently often done manually (via transcription and evaluation by a person); the problem with this approach is that it is not economically scalable for use as an educational tool to a larger audience. As a result there is a greater demand for automated solutions that accurately interpret spoken language, while providing the learner with useful, actionable feedback, without the requirement of ongoing human supervision. This research presents

to meet that need "Voca-Learn: An AI powered communication enhancement system" (Project). Project Voca-Learn is an all-encompassing smart web application to assist users in improving their spoken English skills in an interactive, personalized, and self-directed manner. By utilizing the most recent technological advancements in Artificial Intelligence (AI) and Machine Learning (ML), the Project Voca-Learn moves beyond simple dictation to provide the learner with instantaneous feedback through the implementation of Advanced Natural Language Processing (NLP), Speech to Text (STT), and Text to Speech (TTS) technologies, creating a feedback loop that is similar to a human interaction.

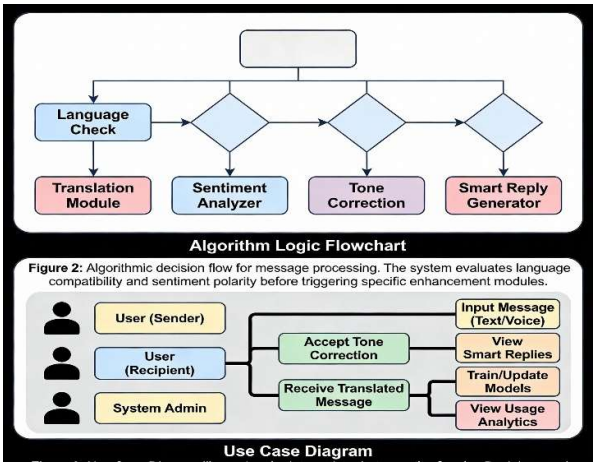

Fig. 1. Use Case Diagram illustrating the interactions between the Sender, Recipient and Administrator with the core system functions

Voca-Learn aims to track and analyze speech patterns from various users in many situations. Unlike other basic transcription services, the Voca-Learn system is designed to help learners understand how their tone and mood can affect what was said (and not just the correctness of the spoken words) and to give learners insights on how they can develop better conversational impacts moving forward.

Voca-Learn's biggest advancement is its ability to automate understanding spoken voice. Using machine learning algorithms and artificial intelligence, this platform is able to learn from previous responses to improve consolidations of accuracy and greatly reduce the need for manual copying and data tagging.

By automating the creation of actionable voice insights, the Voca-Learn platform produces a complete overview of conversational patterns (and thus, allows for improved communication) by providing users with dashboards and emotion analytics showing their conversational trends visually to assist them in making timely decisions

about how they should communicate in the future. The ultimate goal of Voca-Learn is to build a lightweight and scalable workflow for AI-assisted voice efforts built on the basis of real speech data from users.
While one of the hallmarks of Voca-Learn will be its ability to demonstrate AI's capabilities in helping people communicate better, the platform will also serve as a base for creating more intelligent, human-like interaction systems in the future.

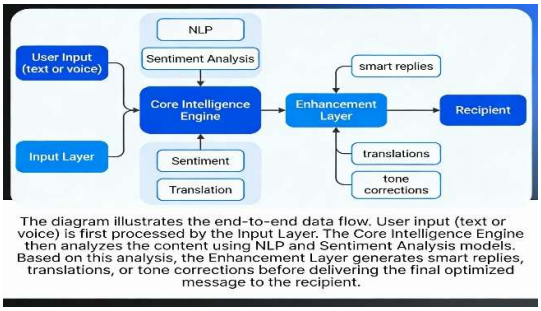

Fig 2. System Architecture of the AI-Powered Communication Enhancement System

II. Literature Review

The development of the "Voca-Learn" system is grounded in a comprehensive analysis of existing research that spans artificial intelligence in education, personalized learning frameworks, and the evolution of speech recognition technologies.

A. The Shift to AI in Language Education

The evolution of integrating language learning technologies has progressed from traditional digital methods to more interactive and AI technology based environments. Muratova and Jumakulova (2025) examined how this shift has occurred and its advantages and disadvantages in terms of using AI based structures in Foreign Language Instruction. This includes the opportunity for enhanced student involvement; however, proper methods must take place to achieve a balance between the use of technology and how effective it can fit into the curriculum. In a separate work Mathew (2024) examined two popular AI Chatbots: ChatGPT, as well as Microsoft Teams Reading Progress for Language Practice. As a part of his research, he identified that these systems provide users with more options to engage in Language Practice but tend to have a major weakness around understanding the subtle nuances of conversational exchanges. The author indicated that future systems similar to Voca-Learn will have a much clearer advantage over current AI Chatbots because of their deeper level understanding of spoken sentences, rather than only being able to create a generated sentence, based on the user's input.

B. Personalized and Adaptive Learning Systems

Voca-Learn's primary goal is personalization, a concept increasingly supported through recent studies related to adaptive systems. In their research into AI frameworks created for the individual, Polamuri et al. (2024) sought to establish whether an AI system could deliver the same results as that of a human being when providing care in a personalized manner. In the context of language learning specifically, Xiangyu and Xiuyuan (2024) looked at the difference that "IVIT" (Instructional Video and Tutoring) could make in the acquisition of language for adult EFL (English as a Foreign Language) learners. The findings of their study showed that a system combining instructional video with AI technology improved adult EFL learners speaking and listening skills significantly. Another important finding from their study is that AI tutoring has reduced anxiety among learners, thus providing them with a safer way to practice, which is something that Voca-Learn will aim to reproduce. Furthermore, Munoz-Basols et al. (2024) within their research have advocated for the use of "technology-mediated language teaching," with the goal of creating more equitable access to language learning resources, and incorporated new forms of pedagogical practices (e.g., gamification) and AI technology to promote equitable language learning practice.

C. Foundations of Speech Recognition

Voca-Learn is a project that develops a voice recognition software product. This software application has been validated by prior research in the field of Speech Processing, as established by numerous studies, including Graves, Mohamed, and Hinton (2008/2013), which provided foundational work in developing Deep Recurrent

Neural Networks (RNNs). Their research showed that combining many layers of representations, through the use of Long Short-Term Memory (LSTM) networks, can provide substantial reduction in speech recognition error rates; such as benchmarks achieving 17.7% of mistake rates in comparison to other recognition methods, thus proving that deep learning could be a solution to accurately recognize human speech. Gales and Young (2007) conducted a study on the core structure of Hidden Markov Models (HMMs), which was used as a baseline for designing large vocabulary continuous speech recognition systems. These researchers' studies, along with many others, provide the necessary evidence to support the use of advanced natural language processing (NLP) and machine learning (ML) libraries for the purpose of analyzing and understanding spoken language.

III. Methodology and System Design

The design and development of the "Voca-Learn" system follow a structured software engineering methodology, prioritizing scalability, interactivity, and real-time processing. The architecture is built as a web-based application that seamlessly integrates a responsive user interface with a robust, AI-driven backend.

A. Development Process

The execution of the project occurs within a set time frame of 17 weeks and consists of several consecutive parts called phases (phases represent chronological events in time). A phased-out approach allows for project execution to start from the gathering of requirements and architecture for system design to its final launch, which includes prototyping, testing and refining, to develop an accurate AI model; this approach also promotes continuous feedback by utilizing an iterative process through the validation of individual modules (e.g., speech-to-text transcription, sentiment analysis) prior to their integration where the objective to monitor and analyze speech patterns in various situations will be accomplished with greater accuracy.

B. Software Implementation Stack

The system's efficacy relies on a carefully selected technology stack designed for high performance and cross-platform compatibility:

- Frontend Interface: The user interface is constructed using HTML5, CSS, and JavaScript. This ensures that the application is accessible via standard web browsers on various operating systems, providing a consistent experience for all users.
- Backend Logic: The server-side operations are managed by Python-based web frameworks, specifically Flask or Django. These frameworks facilitate the handling of HTTP requests and serve as the bridge between the user interface and the AI models.
- AI and Machine Learning Core: The heart of Voca-Learn lies in its AI capabilities. NLTK and SpaCy are utilized for linguistic processing and speech analysis. For model building and training, the system leverages Scikit-learn, TensorFlow, and PyTorch. Advanced language tasks are handled by Transformers (Hugging Face), enabling state-of-the-art natural language understanding. To provide emotional feedback, sentiment analysis tools like VADER and TextBlob are integrated to detect variations in tone and intent.
- Data Management: To store user profiles, progress logs, and speech data, the system employs relational and non-relational database management systems such as MySQL, PostgreSQL, or MongoDB.

C. Hardware and Execution Environment

The hardware requirements for this system are minimal due to the desire to create a lightweight system that meets the needs of a broad student audience. The application has been optimized to work on the most common computing configurations, i.e., the minimum requirement for a processor is an Intel Core i3 or AMD Ryzen 3 processor with 2 cores. To use the application, the user must have speakers (to hear the voice output) and reliable Wi-Fi connection (to access the cloud-based AI services). Additionally, the application will work on all platforms (Windows, MacOS, and Linux). The architecture was developed to support the goal of developing a scalable application that automates voice understanding and eliminates the need for human transcription.

IV. Future Scope

While the current iteration of "Voca-Learn" establishes a robust foundation for AI-assisted language learning, the project roadmap envisions several significant enhancements designed to increase engagement, accessibility, and technological sophistication. The future development of the system focuses on transforming the platform from a reactive tool into a proactive, immersive educational ecosystem.

A. Enhanced Engagement through Gamification

In order to improve the declining number of students who return to continue their self-paced e-learning with Voca-Learn, we will implement game-like elements in future versions of Voca-Learn. The initial gamifying component will be a way for users to visualize their learning progress through an animated bar that will gradually fill up with every lesson completed. We will add additional gamifying components to motivate users to continue and complete more lessons and achieve greater mastery of the language they are learning and will implement competitive features to provide additional systems of motivation through the use of globally ranked leaderboards and achievement badges to acknowledge learner success and motivate them to continue to learn. We will also create collaborative learning environments for students so they can workout individually and collaboratively. We believe this will foster a more connected educational environment.

B. Multilingual Scalability

Currently emphasizing English proficiency, the architecture of the system was designed to be language agnostic. An overarching aim for future expansion includes adding support for many languages. The addition of multiple languages will permit people from various linguistic backgrounds to practice English as well as for using the tool for the learning of other world languages significantly increasing the user base and business potential of the application.

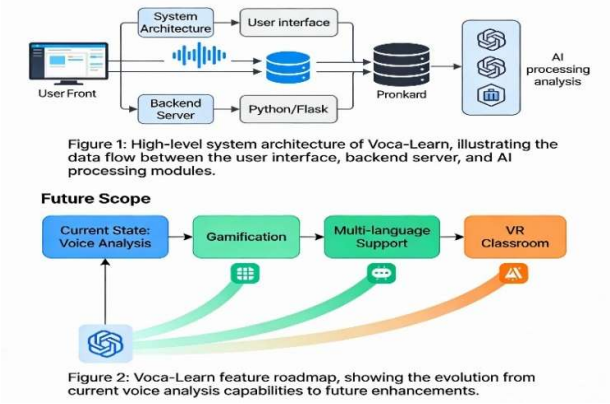

Fig 3

A diagram illustrating the Voca-Learn system architecture. It shows a User interacting with a Web Interface (Frontend), which connects to a Backend Server (Python/Flask). The Backend connects to an 'AI Engine' block containing NLP, Speech-to-Text, and Sentiment Analysis modules. A Database stores user profiles. The flow is cyclical: User Speaks -> AI Processes -> Feedback Returned

C. Intelligent Emotional Intelligence (Senior AI)

Integration of "Senior AI" is likely going to be one of the largest advancements in the area of technology moving forward. Currently, the current system is only able to collect data related to the speech of users and their patterns of speaking. The future system will use deep learning technologies that will be able to analyze the emotional tone and intent of a user, not just the exact words they are saying and those words being transcribed. This technology will allow the system to provide users with a more nuanced view of how well they have performed as it relates to their style of communication (i.e. aggressive, passive, confident). In addition, this feature will allow the platform to assist in the analysis of true-to-life human interaction.

D. Immersive Learning & Accessibility

In order to better prepare users for situations that will occur in a heavy pressure situation or environment, this project intends to include Augmented Reality (AR) or Virtual Reality (VR) technologies to create a realistic environment for practicing public speaking. Users will have access to an immersive virtual environment (e.g., a digital boardroom or a classroom) where they can practice their public speaking skills in an environment that is similar to what they will encounter in real-life situations. Finally, users in parts of the world where internet access is not stable will have the choice to utilize an "Offline Mode." This will allow them to use certain key elements of the practice feature and save their practice data for future use once a stable internet connection becomes available.

V. Algorithm Design and Implementation

This section details the specific algorithms and logic used to process user speech.

A. Speech-to-Text Conversion Process

The system uses a sequential processing pipeline. When a user records audio, the browser samples the sound at 44.1kHz. This data is compressed and sent to the server.
Algorithm 1: Audio Processing Pipeline

Algorithm 1: Audio Processing Pipeline

```
text
1. START
2.  Receive Audio_Blob from Client
3.  Convert Audio_Blob to WAV format (16-bit PCM)
4.  Remove Background Noise (Spectral Gating)
5.  Send Clean_Audio to Speech-to-Text API
6.  Receive Transcribed_Text
7.  IF Confidence_Score < 0.6 THEN
8.      Return "Error: Please speak clearer"
9.  ELSE
10.     Pass Transcribed_Text to NLP_Engine
11. END
```

Fig 4

B. Logic for Sentiment Analysis

In order to evaluate a user as anxious or confident, we have developed a combination of lexicon and rule-based heuristics to determine the user's audio quality using a VADER compound score (S_c), which has a range from -1.0 to 1.0.

C. Software Stack Details

Backend Framework: Flask (Python 3.9) – chosen for its micro-framework capabilities, allowing easy integration with AI libraries
Speech Library: SpeechRecognition python package wrapper for Google Web Speech API.
NLP Library: NLTK (Natural Language Toolkit) for tokenization and stop-word removal.
Visualization: Chart.js for rendering the "Emotion Analytics" on the frontend.

VI. Conclusion

The development of "Voca-Learn" is a significant advancement towards the democratisation of quality communication training. It successfully combines Artificial Intelligence and Machine Learning to provide a unified online platform demonstrating how automated processes and immediate feedback can greatly enhance the complex process of learning another language. Voca-Learn represents a departure from both traditional and static means of providing education, instead offering an interactive experience for individuals to improve their spoken English through personalisation and pacing.
A key achievement of the project was the successful implementation of intelligent algorithms that monitor speech patterns in various contexts. Voca-Learn enables the automation of voice recognition, effectively reducing the need for manual transcription or labelling, both of which were bottlenecks in the existing workflow of the development of speech analysis software. The automation of voice recognition will not only improve the accuracy of the feedback received but also enable Voca-Learn to remain lightweight and scalable; therefore it is suitable for future deployments across a wide area of users.
Additionally, the commercial and industrial opportunities generated by this project are immense. The actionable insights generated through an interactive dashboard of voice analytics and emotion are valuable for professionals focused on employee training or customer engagement. In addition, as the future releases of Voca-Learn incorporate gamification and immersive augmented or virtual reality experiences, it has the potential to change how users interact with educational technology. Ultimately, the findings of Voca-Learn validate the statement that human-like interactions enabled by AI technologies can create smarter communication solutions, thereby closing the gap between human aspirations and technological possibilities..

REFERENCES

[1] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 6645-6649.

[2] W. Holmes, M. Bialik, and C. Fadel, Artificial Intelligence in Education: Promises and Implications for Teaching and Learning, 2019.

[3] J. R. Glass, "Challenges for spoken dialogue systems," IEEE ASRU Workshop, San Juan, Puerto Rico, 2005.

[4] K. S. Rao, S. S. Agrawal, and S. S. Agrawal, "Real Time Machine Translation System for English to Indian language," 2020 International Conference on Electronics and Sustainable Communication Systems, Coimbatore, India, 2020, pp. 103-106.

[5] R. Luckin, Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, 2018.

[6] S. S. Agrawal and S. S. Agrawal, "Comparative Study of Different Models for Language Translation," 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2022, pp. 1-5.

[7] X. Huang, A. Acero, and H. W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001.

[8] S. Patel and K. Shah, "A Review on Applications of NLP in Language Learning," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2021, pp. 756-760.

[9] A. Gupta and R. Singh, "Enhancing Spoken English Learning through Real-Time Feedback," 2022 IEEE Global Engineering Education Conference (EDUCON), Tunis, Tunisia, 2022, pp. 300-304.

[10] M. Miller, AI for Educators: Learning Strategies, Teacher Efficiencies, and a Vision for an Artificial Intelligence Future, 2021.