Vania Abbas

Abstract

Fear of negative evaluation (FNE) and public speaking anxiety (PSA) are widespread problems that have a detrimental impact on people's performance in both their academic and professional lives. This study presents SpeakEasy, an AI-powered speech coaching system that uses coordinated artificial intelligence models to analyze speech content and delivery in order to assist users in improving verbal communication. Based on a thorough need analysis and backed by empirical research, this study offers a clever, scalable, and accessible way to deal with anxiety associated to communication.

SpeakEasy: An AI-driven Speech Coach

Final Year Project

Table of Contents

[**Introduction** 3](#_Toc200811373)

[**Literature Review** 5](#_Toc200811374)

[**Reviewing Existing Applications** 5](#_Toc200811375)

[**Studying Existing Pipelines** 6](#_Toc200811376)

[**Key Technologies** 8](#_Toc200811377)

[**Design** 13](#_Toc200811378)

[**Domain and Users** 13](#_Toc200811379)

[**Justifying Features Based on Users and Domain** 13](#_Toc200811380)

[**System Architecture** 14](#_Toc200811381)

[**Technologies and Methods** 17](#_Toc200811382)

[**User Interface Design** 17](#_Toc200811383)

[**Evaluation Plan** 18](#_Toc200811384)

[**Project Plan** 19](#_Toc200811385)

[**Feature Prototype** 21](#_Toc200811386)

[**Audio Input** 21](#_Toc200811387)

[**Displaying Feedback** 22](#_Toc200811388)

[**User Interface and Experience** 22](#_Toc200811389)

[**Tech Stack Used** 22](#_Toc200811390)

[**Prototype Limitations** 23](#_Toc200811391)

[**Prototype Walkthrough** 23](#_Toc200811392)

[**Significance and Next Steps** 23](#_Toc200811393)

[**References** 24](#_Toc200811394)

[Figure 1 - Framework Architecture as seen in the paper 8](#_Toc200806181)

[Figure 2 - The results of the study as outlined in the paper 10](#_Toc200806182)

[Figure 3 - The results of the study as outlined in the paper 12](#_Toc200806183)

[Figure 4 - Architecture of SpeakEasy 14](#_Toc200806184)

[Figure 5 - Speech to Text Model 15](#_Toc200806185)

[Figure 6 - Language Model 15](#_Toc200806186)

[Figure 7 - Speech Emotion Recognition Model 16](#_Toc200806187)

[Figure 8 - The Front-end Design of the System 17](#_Toc200806188)

[Figure 9 - Gantt Chart Presenting Project Plan 19](#_Toc200806189)

[Figure 10 - User Audio Input Feature 21](#_Toc200806190)

[Figure 11 - Feedback Page 22](#_Toc200806191)

# **Introduction**

Successful academic, social, and professional endeavours require the ability to speak in front of an audience. But it continues to be a major source of tension and concern for a lot of people. Known as public speaking anxiety (PSA), this condition affects a significant percentage of people and shows itself as a variety of physical and mental symptoms, including shaking, trouble speaking, tense muscles, elevated heart rate, and avoidance of eye contact. Those who experience these symptoms may become less confident, perform less well, and be unable to fully engage in professional and academic possibilities. A cognitive disorder called fear of negative evaluation (FNE), which is the dread of receiving an unfavourable judgment from others, frequently lies at the core of this anxiety.

The hallmark of FNE is an overabundance of concern about other people's opinions, which can result in self-monitoring, social avoidance, and even submissiveness. It can cause feelings of inadequacy, diminish student motivation, and have a negative impact on academic achievement, making it especially harmful in educational settings.

Despite this well-established issue, there are still few solutions available, especially for young professionals and students who do not have access to costly or time-consuming in-person coaching. Even if they work for some people, current solutions like therapy, group workshops, and self-help techniques aren't always feasible or scalable. The lack of readily available, customized, and instantaneous coaching tools emphasizes the necessity for a more creative strategy. This study presents *SpeakEasy*, a speech coach powered by artificial intelligence (AI) that helps people overcome their fear of public speaking and boost their confidence when communicating. *SpeakEasy* is a supplemental tool that provides intelligent, flexible, and easily available feedback on the content and delivery of spoken communication; it is not a substitute for therapy or human mentoring.

*SpeakEasy* uses machine learning innovations to give feedback that is comparable to what a human coach might give, but with the advantages of on-demand access, scalability, and privacy. It functions by orchestrating multiple pre-trained AI models, each of which performs a distinct task in the speech analysis pipeline. First, speech-to-text transcription models, like OpenAI Whisper, accurately translate spoken input from users into written text. This transcription serves as the basis for further study. After that, a language model, like GPT or BERT, is used to assess and improve the voice transcription's grammar, sentence construction, and general clarity. It makes recommendations for improving language fluency and rearranging problematic phrases.

The system then evaluates the speech's emotional and tonal qualities using models for sentiment and prosody analysis. In order to gain an understanding of the speaker's emotional state and degree of engagement, these models look at both vocal characteristics (such as pitch, tempo, and volume) and textual emotions (such as confidence, anxiousness, and assertiveness). A feedback creation engine then synthesizes all of this data to provide the user with tailored, helpful suggestions that will help them get better over time. To better grab the audience's attention, this contains ideas for changing vocal tone, rephrasing words, or adding more vocal variation.

What sets *SpeakEasy* apart from conventional voice technologies is the orchestration of these AI models. *SpeakEasy* combines speech recognition, emotion analysis, and grammatical correction into a single coaching system rather than handling them as distinct processes. The end result is a seamless, comprehensive experience that is customized to meet the individual demands of every user. Users receive assistance that changes as they grow, whether they are getting ready for a high-stakes job interview, practicing for a university presentation, or just wanting to become more confident in casual discussions.

*SpeakEasy* was created with accessibility and inclusivity in mind. Users only need a device with a microphone; no further equipment is needed. The system can be implemented as a web platform, mobile application, or incorporated into pre-existing e-learning settings because it is platform-agnostic. An important aspect for those with high FNE is that users can interact with the system without worrying about judgment or shame because feedback is generated automatically and anonymously.

*SpeakEasy* has significant societal ramifications in addition to personal advantages. Improved employment, greater civic involvement, and greater participation in scholarly conversations can all result from the system's ability to enable more people to speak fluently and coherently. *SpeakEasy* can be a scalable addition to speech curriculum for educational institutions, providing students with quantifiable and regular feedback without adding to the workload of teachers. The application may help firms with employee training by giving staff members a risk-free setting in which to hone their presentation and negotiating skills.

In conclusion, it has been demonstrated that both academic and professional achievement are impacted by the ongoing problem of public speaking anxiety and the dread of receiving a poor grade. Even if they work for some people, traditional coaching options are not available to everyone. *SpeakEasy* fills this gap by fusing cutting-edge AI technology to produce a customized, expandable, and intuitive voice coaching platform. *SpeakEasy* is a relevant and useful invention that supports confident communication in a judgment-free setting. It is based on a strong machine learning architecture and psychology research.

# **Literature Review**

The emergence of AI-powered platforms for communication and education has sparked the creation of a number of resources targeted at enhancing English language competency, especially in speech and grammar. It is crucial to critically assess the current research and methods in this field because *SpeakEasy* is intended to improve spoken English and boost user confidence by providing feedback on vocabulary, emotional tone, and fluency. In order to determine the advantages, disadvantages, and gaps in the current solutions, this literature review examines the models for correcting grammatical errors, evaluation methodologies, and speech assessment platforms.

Rich front-end features are a common component of commercial apps, but the underlying technologies and development processes are frequently opaque. Academic research that reveals the technical frameworks and procedures utilized to develop these systems is therefore given particular attention. To help with the technological viability and pedagogical foundation of *SpeakEasy*, research on the accuracy of speech-to-text transcription, emotion detection from vocal cues, and grammatical correction using big language models are also reviewed. This review not only serves as a basis for defending the suggested features, but it also aids in placing *SpeakEasy* in relation to other voice coaching and language learning tools.

## **Reviewing Existing Applications**

**Speeko AI** [1]

Speeko is an application committed to helping people become confident in their communication. It enables users to become confident, clear and empathetic in public speaking. The features that make up this application include its ability to track voice and speech patterns in real-time, providing friendly alerts to keep one speaking their best and lastly it provides live coaching sessions which need to be booked in advance.

The user interface consists of three modules: Coaching: this harbours training practices and some daily activities, Record: this module has 4 main options including Presentation, Meeting, Practice Interview and Random Prompt. It simulates each of those situations and provides real-time feedback and lastly Progress: This module provides informative learning analytics and statistics based on the feedback data it has stored from the other two modules. However, whilst this application provides benefits, it comes at some costs, quite literally. The beneficial features that promote confidence in public speaking such as the ability to provide personalised feedback and real-time guidance. The free aspect of this application only provides basic insights as feedback, a few training exercises as opposed to the paid version that includes over a 1000. Moreover, this application is currently available for Mac, iPhone and iPad users thus not being accessible to the Android market as of yet. Furthermore, its ability to provide friendly alerts can cause interruptions whilst the user is speaking for a presentation or any other important event. This eventually causes hinderance in the flow of speech and make the user forget where they were before being alerted.

**Poised AI** [2]

Poised AI is another application which helps users speak confidently. It is an AI powered application that offers real-time feedback to help speak confidently during calls. Users can get personalised suggestions and actionable insights on their speeches and communication skills. It automatically brings up meeting notes and crosses items off as they are covered in the meeting so as to help the user keep track of what has been covered and what is yet to be covered. It provides immediate actionable feedback on what the user could do better based on goals of each specific meeting.

However, despite the benefits that this application offers, it does not prepare one for a meeting/event beforehand. Instead, it provides feedback while the user is engaged in a live meeting or event. This can thus lead to interruptions while the user is speaking and can make the user lose their train of thoughts. Moreover, Poised AI focuses on tone, filler words and pacing and does not analyse emotional variations in speech.

**Relevance to *SpeakEasy*:**

These applications serve as a baseline for what is readily available to consumers for a speech coach. *SpeakEasy* aims to provide more in-depth feedback of speech by analysing the emotions, clarity and tone. The limitations of the aforementioned applications highlight the gap for a more sophisticated, AI-driven speech coach focused on helping its users speak with clarity and confidence.

**Gap Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Speeko** | **Poised** | ***SpeakEasy*** |
| **Provides In-depth Feedback** | O | O | O |
| **Checks Grammar and Provides Alternative text suggestions** | X | X | O |
| **Provides Emotion and Tone Analysis** | X | X | O |

## **Studying Existing Pipelines**

Applications currently in use emphasize front-end functionality while disclosing minimal information about their technological underpinnings. Reading up on development tools, architectures, and data processing was crucial to the successful creation of *SpeakEasy*. This behind-the-scenes knowledge guarantees that the app's design adheres to industry best practices and is technically possible given the limitations and scale of the project.

**Speech Coach: A framework to evaluate and improve speech delivery** [3]

Through methodical speech evaluation, researchers Adhish Deshpande, Rohit Pandharkar, and Subodh Deolekar offer a web-based platform designed to assist users in analyzing, comprehending, and improving their elocution abilities. The user uploads an audio recording as the first step in their framework's free and user-friendly methodology. Google's Automatic Speech Recognition (ASR), which the authors claim has a 95% accuracy rate for English speech, is then used to transcribe the voice. Acoustic characteristics like intensity, speed, pitch, intonation, pauses, and filler words are processed with the recovered transcript. In order to give feedback on the user's delivery, these features are then displayed through user-friendly graphs.

The framework uses a variety of tools, such as Aeneas to synchronize audio with the transcript, Flask for the web interface, and the my-voice-analysis API for feature extraction. Although the platform places a strong emphasis on user-friendliness, its reliance on self-assessment—in which users compare transcripts and interpret graphical outputs—can limit the objectivity of feedback in the absence of expert evaluation and add subjective bias. Additionally, the approach seems to be adapted to Indian English pronunciation standards, which would limit its use in other linguistic or cultural situations without further modification. Even though the application is meant to be user-friendly, less tech-savvy users may still find it difficult to use because it requires technical knowledge to grasp data visualizations and vocal metrics. Significantly, the framework has not yet been subjected to extensive user testing or empirical review to determine how well it improves speech delivery. This is a significant drawback when thinking about its application in more extensive coaching or instructional settings.

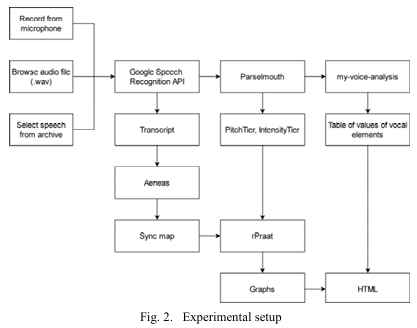


Figure - Framework Architecture as seen in the paper

## **Key Technologies**

1. **Speech to Text**

**Assessing Speech-to-Text Translation Quality: An Overview of Key Metrics** [4]

Labied, Belangour, and Banane offer a thorough summary of evaluation criteria that are crucial for determining how well speech-to-text translation systems perform, especially in multilingual and real-time applications. The assessment requirements for end-to-end models that integrate both processes and cascade models, which comprise distinct stages of machine translation (MT) and automatic speech recognition (ASR), are distinguished in the study. Commonly used metrics like BLEU, Character Error Rate (CER), and Word Error Rate (WER) are examined for their advantages and disadvantages in terms of accurately capturing transcription and translation. For a more linguistically nuanced assessment of MT outputs, the authors point out that upstream ASR faults in cascade systems might spread throughout the pipeline, requiring metrics like METEOR and Translation Edit Rate (TER). Additionally, they emphasize how crucial pipeline latency is as a practical factor, particularly in real-time or time-sensitive applications. Additional metrics such as Joint Error Rate (JER), ROUGE, COMET, and LEPOR are added for end-to-end models, providing more comprehensive evaluations that take into account both processing latency and semantic accuracy. Notably, in order to take into consideration both deeper contextual fidelity and surface-level precision, the authors recommend combining various measurements. The study does not, however, empirically benchmark the measures across actual datasets or systems; instead, it focuses on providing a theoretical evaluation. Furthermore, despite the metrics' good categorization, user experience and the ways in which these assessment techniques result in noticeable enhancements for end users receive scant attention. Nevertheless, the framework provides useful advice for developers looking to maximize voice translation models, particularly in fields where it's important to strike a balance between accuracy, latency, and semantic integrity. When choosing models for speech to text conversion, the metrics indicated in the report will be useful in determining which metrics to use as the foundation for my experiments. The best model for transcribing user-inputted audio files will be identified with the aid of these measures.

**Comparing Speech-to-Text Algorithms for Transcribing Voice Data from Surveys** [5]

Landesvatter, Behnert, and Bauer compare various automatic speech recognition (ASR) systems for recording survey voice responses. This technique is being employed more and more to improve respondent engagement and data richness in smartphone-based research. Their study benchmarks the performance of four major ASR frameworks on 99 randomly selected spoken responses from a diverse U.S. survey sample. These frameworks are Google Cloud Speech-to-Text, Meta's wav2vec 2.0 (base and large), NVIDIA NeMo (xlarge), and OpenAI's Whisper (medium and large). Word Error Rate (WER) is the main indicator, and the authors observe significant heterogeneity in model performance when comparing the model to human-generated transcripts.

The lowest WER was 4.7% for Whisper (big), 5.6% for Whisper (medium), and 7.2% for NeMo (xlarge). Wav2vec (big and base) trailed with WERs of 10.8% and 21.1%, respectively, while Google Cloud's model fared even worse with a WER of 13.3%. These results call into question Google's ASR's dominant position in social science research and support the wider use of more recent, open-source models.

The inclusion of different ASR configurations without fine-tuning, which matches normal usage scenarios, the study's transparent approach, and its real-world, noisy data environment are among its strong points. However, generalizability can be hindered by its small sample size and absence of qualitative error analysis. Furthermore, although the article provides recommendations for researchers utilizing ASR in survey settings, it ignores model performance across different demographic accents or languages, which are essential for inclusive survey design. Nevertheless, the authors include useful implementation code and support a repeatable assessment approach, which makes this an invaluable resource for academics looking to successfully incorporate voice-based replies into survey tools.

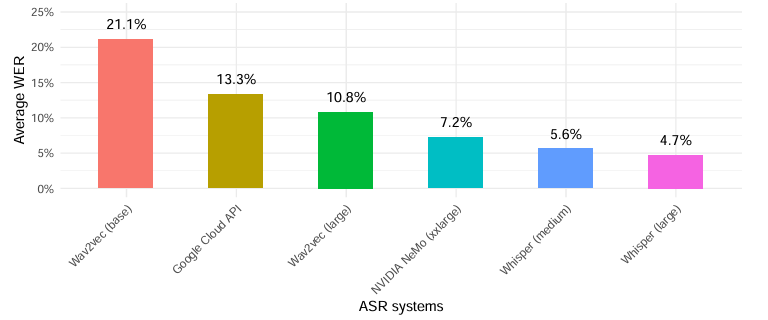


Figure - The results of the study as outlined in the paper

This study helps in the selection of speech-to-text models for *SpeakEasy*. The test results in this study assist in selecting the model with a low Word Error Rate thereby producing accurate transcribed text.

1. **Grammar Correction**

**Ground Truth for Grammatical Error Correction Metrics** [6]

Napoles et al. (2015) challenge the reliability of popular automatic metrics like MaxMatch (M2), BLEU, and I-measure in order to fill a significant gap in the assessment of Grammatical Error Correction (GEC) systems. Despite their popularity, these measurements are not based on human judgment, which is the very standard they are meant to emulate. In order to address this, the authors carried out the first extensive human assessment of the CoNLL-2014 Shared Task system outputs. They developed a human-derived system rating to act as ground truth using more than 28,000 pairwise comparisons gathered using the Appraise tool and scored using the TrueSkill model.

According to the study, there is a poor correlation between current metrics and human appraisal. While BLEU and I-measure shown little to no connection, occasionally even negative, M2, the standard in shared tasks, only moderately matched human assessment. The authors present GLEU, a modified BLEU variation that penalizes grammatical errors and promotes accurate modifications. Spearman's ρ = 0.555 indicates that GLEU0 (without penalizing false negatives) had the strongest correlation with human ranks, indicating that it is more in line with the judgments of real language users.

The paper's main merit is its emphasis on the subjectivity of GEC, where not all fixes are equally important and where strict edit-based metrics could ignore fluency or contextual appropriateness. The results' generalizability may be limited by the use of a particular dataset, and even GLEU's enhanced correlation is not as good as human agreement.

By offering a more human-aligned option in GLEU and establishing a human-based baseline for GEC evaluation, the paper contributes significantly overall. It promotes human-centric evaluation as a top priority for future measure development, particularly for jobs where readability and linguistic nuance are crucial.

**Grammar Correction: A Comparison of T5, LLAMA 2, and ChatGPT** [7]

Nurhasanah et al. (2025) address the persistent drop in English competence in Indonesia and promote better language learning outcomes by conducting a systematic comparison of four AI-driven Grammatical Error Correction (GEC) models: T5 Mini, T5 Tiny, LLAMA 2, and ChatGPT 3.5-turbo.

Utilizing the JFLEG dataset and the GLEU metric, the research assesses each model's capacity to rectify grammatical errors while maintaining meaning and fluency. With the highest GLEU score of 0.565, LLAMA 2 outscored the others, according to the data, demonstrating its superior ability to match human corrections, especially for formal grammar problems. With a GLEU score of 0.524, T5 Mini came in second, providing a good mix between accuracy and efficiency. T5 Tiny, on the other hand, had a slightly lower score of 0.518, making it appropriate for usage in settings with constrained computational resources.

With a score of 0.491, ChatGPT 3.5-turbo came in last in automatic scoring but distinguished itself in qualitative expert ratings for producing input that was conversationally natural, coherent, and contextually rich. The study emphasizes that although ChatGPT may score poorly on rigorous grammatical tests, it is excellent at generating natural-sounding, conversational responses, which is useful in interactive, real-world learning situations.

The study's limitations, which may impact its generalizability, are its limited dataset scope (JFLEG only) and its absence of evaluation for multilingual or casual speech use situations, despite its sound methodology. Furthermore, even though LLAMA 2 received the highest score, it occasionally used less exact language (such as "more better") and had some redundancy problems. All things considered, the study provides useful advice for developers and educators who want to incorporate AI-based grammar feedback into teaching resources. It recommends that model selection should be informed by application context in addition to accuracy, whether for formal assessments or dynamic conversational learning.

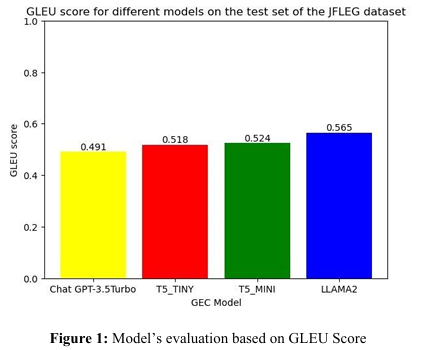


Figure - The results of the study as outlined in the paper

This study helps in selecting a language model that accurately corrects the grammar of the speech transcribed. It assists in the consideration of a language model for *SpeakEasy* to correct grammar in the speech uploaded by the user.

# **Design**

## **Domain and Users**

*SpeakEasy* works in the domain of Education, with a specific focus on enhancing communication in English and soft skills. The system is designed to build users’ confidence and ease in spoken English by providing comprehensive feedback on both verbal communication and non-verbal cues, such as tone of voice.

*SpeakEasy*’s target audience includes:

* Individuals seeking to improve their English-speaking abilities.
* Students aiming to develop essential communication-related soft skills.
* Job seekers and individuals preparing for interviews, presentations, or public speaking events.
* Non-native English speakers striving for greater fluency and self-assurance in English conversations.

*SpeakEasy* empowers users to become more confident, articulate, and effective communicators in a variety of real-world contexts.

## **Justifying Features Based on Users and Domain**

*SpeakEasy*'s features have been carefully chosen to serve the demands of non-native speakers, students, job seekers, and public speakers, given its domain in education—specifically, improving English communication and soft skills. *SpeakEasy* has a more targeted approach, whereas existing programs on the market (as evaluated in the Literature Review) provide a broad range of capabilities. It prioritizes depth and usefulness over mimicking the scope of existing systems, making sure that every feature directly advances the objective of enhancing user confidence and fluency. Realistic time and resource restrictions also inform this approach.

The proposed features are:

1. **Audio Upload for Analysis** – In order to facilitate flexible practice in a variety of real-life speaking scenarios, including interviews, class presentations, and spontaneous responses, users can upload recordings of themselves speaking.
2. **Speech-to-Text Transcription** – transforms the speech that has been uploaded into text, which serves as the foundation for linguistic analysis. Finding trends in vocabulary usage, grammar, and fluency requires this.
3. **Fluency Feedback** – detects speaking pace, filler words, and awkward pauses. This helps presenters who are apprehensive or worried by pointing out areas that could interfere with a smooth delivery and allowing for focused improvement.
4. **Grammar and Vocabulary Suggestions** – helps non-native speakers by providing corrections and recommendations that enhance word choice, sentence structure, and intelligibility. It facilitates more accurate and self-assured English expression.
5. **Emotion Detection via Speech** – employs a speech-to-emotion model to deduce emotional states like anxiety, hesitancy, or excitement from vocal. This knowledge enables users to consider how others might interpret their tone, which is important for soft skills like professionalism, audience engagement, and persuasion.

With an emphasis on speech content and vocal delivery, each feature is based on the application's primary goal of assisting users in becoming more confident, expressive, and eloquent English communicators.

## **System Architecture**

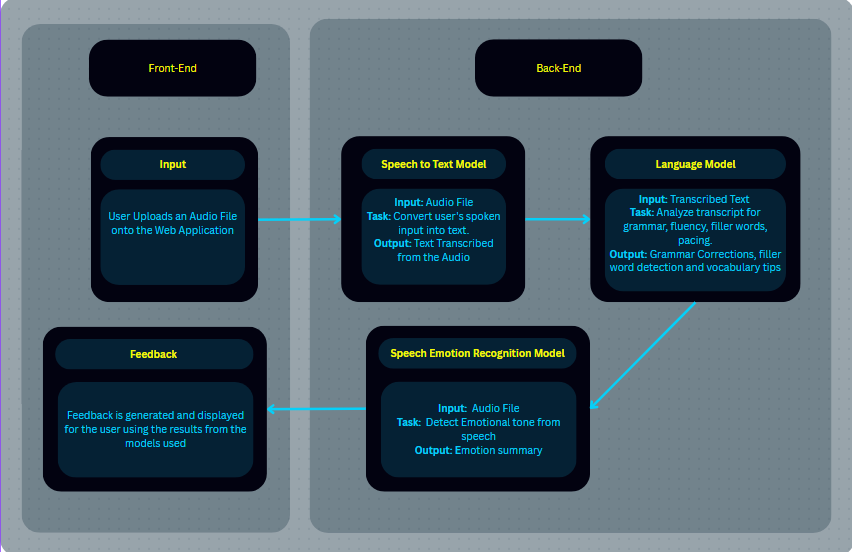


Figure - Architecture of SpeakEasy

*SpeakEasy* employs a variety of models, including emotion detection, language processing, and speech-to-text, to produce detailed, tailored feedback. Input, Speech-to-Text Conversion, Language Analysis, Emotion Recognition, and Feedback Generation are the five consecutive steps of the pipeline. Every step is essential to the platform's ability to provide precise, useful information.

**1. Input**

* **Action**: The user uploads an audio file containing a sample of their speech onto the web application.
* This serves as the initial entry point for the analysis process.
* The uploaded file is expected to be a spoken response, such as a speech, interview answer, or general spoken communication.

**2. Speech-to-Text Model**

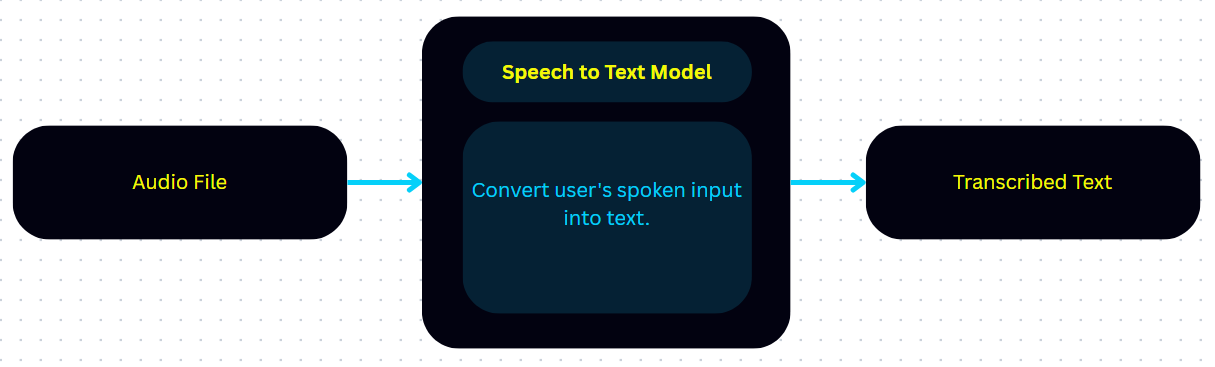
****

Figure - Speech to Text Model

|  |  |
| --- | --- |
| **Input** | The uploaded audio file. |
| **Task** | The system uses a speech-to-text model to convert spoken language into written text. |
| **Output** | A text transcription that accurately reflects what the user said in the recording. |

This is a critical step because most of the downstream analysis—grammar checking, vocabulary enhancement, and fluency evaluation—requires a text version of the speech.

**3. Language Model**

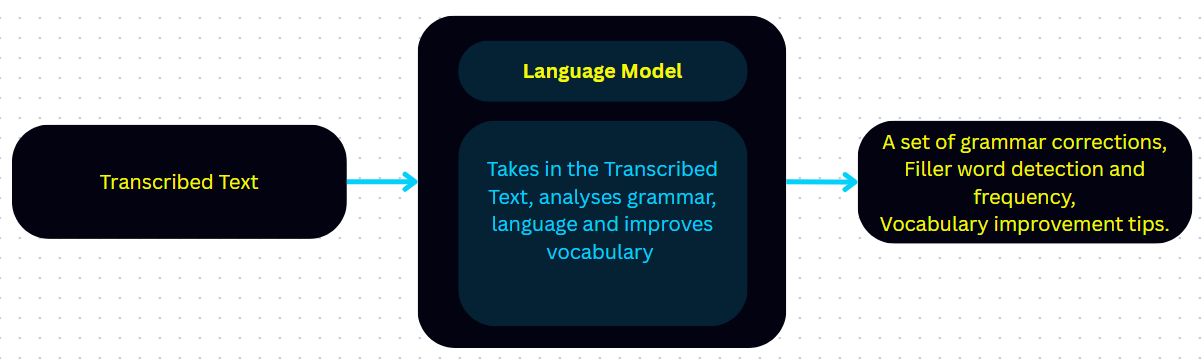
****

Figure - Language Model

|  |  |
| --- | --- |
| **Input** | The transcribed text from the stage prior |
| **Task** | The language model examines the transcript to provide feedback on:   * Grammar: Identifying grammatical errors and suggesting corrections. * Fluency: Detecting filler words (e.g., “um,” “like,” “you know”), awkward pauses, and irregular pacing. * Vocabulary: Highlighting basic or repetitive word choices and offering suggestions for more precise or sophisticated alternatives. |
| **Output** | A set of grammar corrections,  Filler word detection and frequency,  Vocabulary improvement tips. |

**4. Speech Emotion Recognition Model**

****

Figure - Speech Emotion Recognition Model

|  |  |
| --- | --- |
| **Input** | The original audio file |
| **Task** | Analyses the audio for paralinguistic cues—non-verbal elements like tone, pitch, intonation, and voice modulation—to detect emotional states such as:   * Nervousness * Confidence * Enthusiasm * Calmness * Uncertainty |
| **Output** | An emotion summary that reflects the emotional tone conveyed by the speaker. |

**5. Feedback**

In the last step, the output from all of the earlier models—the language analysis, emotion detection, and speech-to-text transcription—is combined to produce thorough, approachable feedback. Along with comments and corrections for vocabulary, grammar, and fluency, this feedback also provides information about the user's voice's emotional tone. Through the integration of expressive delivery analysis and linguistic accuracy, the system offers comprehensive recommendations customized to each user's communication preferences. The web program then presents the comments in an easy-to-understand, structured manner, assisting users in identifying their areas of strength and growth in terms of both content and emotional expression.

## **Technologies and Methods**

*SpeakEasy* is developed using a lightweight and efficient tech stack suited for deploying AI-driven web applications. The backend is built using Flask, a Python microframework that handles routing, file uploads, and integration with machine learning models. The frontend is developed using HTML and CSS, providing users with a clean and responsive interface for uploading speech samples and viewing feedback.

For natural language processing tasks such as grammar correction, filler word detection, and vocabulary enhancement, the application leverages pre-trained transformer models accessed via the Hugging Face Transformers library. These models, such as T5 or BERT-based variants, are integrated into the Flask server to process and analyse transcribed speech.

Speech-to-text conversion is handled using a suitable speech recognition model (e.g., Whisper), while speech emotion recognition is performed using models that classify vocal features into emotional states like nervousness or confidence. Audio features may be extracted using libraries like Librosa to support emotion classification.

All results—text transcription, language corrections, and emotional insights—are synthesized and returned to the user as interactive feedback through the web interface.

## **User Interface Design**

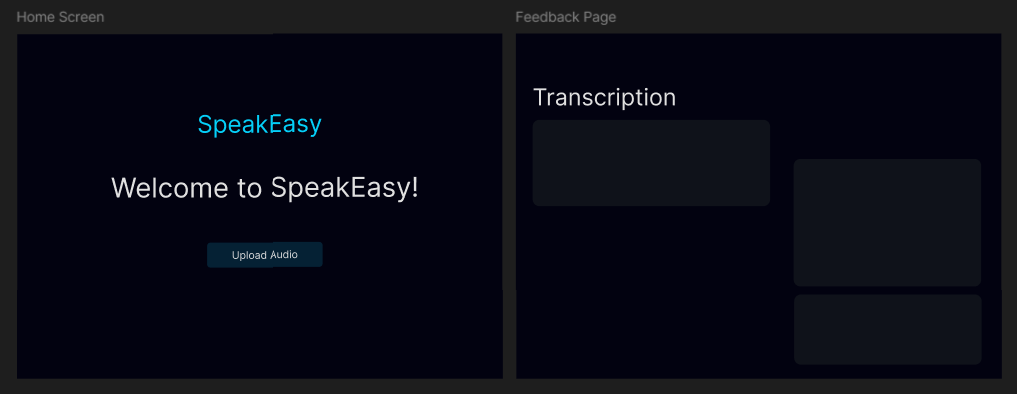


Figure - The Front-end Design of the System

The frontend of the system consists of two webpages. The landing page – referred to as the Home Screen in the figure - will be that of one where the user uploads the audio file. After that, it is the job of the backend to perform its analysis and the results will be displayed in the form of feedback on the feedback page.

## **Evaluation Plan**

An evaluation plan for *SpeakEasy* aims to methodically determine whether the platform has successfully met its objective of improving users' communication and speech delivery abilities. This entails using a mix of quantitative and qualitative techniques to measure system performance as well as user outcomes. After using the application, the evaluation will ascertain whether users exhibit quantifiable improvements in critical speech characteristics like confidence, clarity, fluency, and pronunciation. The plan will also include an analysis of usability, user happiness, and how useful and actionable users find *SpeakEasy's* comments. Crucially, the strategy will take into account the application's flexibility and inclusivity, evaluating if it can support a range of English dialects and speech situations.

**Quantitative Evaluation**

* **Accuracy of Speech to text Model**

Since the speech-to-text model forms the basis for later models in the pipeline, its performance is essential to the system's overall success. We employ the Word Error Rate (WER) metric to assess the precision of the text that is transcribed from audio input. WER calculates the number of substitutions, deletions, and insertions needed to match the reference in order to quantify the differences between the transcribed output and the actual (reference) text. Higher transcription accuracy is indicated by a lower WER, which shows that the model is more successful at turning speech into text.

* **Language Analysis, Grammar Correction:**

In order to guarantee grammatical correctness, fluency, and clarity, the transcribed text must be refined by the grammar correction model. Automated metrics can be used to assess this model's performance:

* **1. Grammar Correction Metrics:**
  + **GLEU (Grammar Error Correction BLEU):** An adaptation of the BLEU score, GLEU compares the model’s corrected output to a set of human-corrected references. It balances precision and recall to measure both correctness and completeness.
* **2. Error Reduction Rate:**
  + Calculate the percentage of grammatical errors corrected in the output compared to the original input.
  + Formula:

A combination of these methods provides a comprehensive assessment of how well the grammar correction model improves language quality.

* **Emotion Detection Accuracy Using Speech Audios**

A test dataset of pre-labelled emotional speech samples will be used to assess the Speech Emotion Recognition (SER) model's performance. To evaluate accuracy, the model's predictions will be contrasted with labels of the human-annotated ground truth. Evaluation metrics for each emotion class (e.g., anxiousness, confidence, enthusiasm) will be computed, including precision, recall, and F1-score. To gauge robustness, more testing will be done with different accents, speech lengths, and background noise levels. To evaluate the perceived accuracy and use of emotion-based insights in the context of speech enhancement, user input will also be gathered.

**Qualitative Evaluation**

* **User Feedback via surveys**

This will evaluate how well the user interacts with the program overall and how much it affects their confidence when speaking. Prior to and following a predetermined amount of time spent using *SpeakEasy*, participants will be asked to score their own level of confidence in public speaking. To gather both qualitative and quantitative data, these surveys will contain reflection prompts, open-ended answers, and Likert-scale questions. Key areas of emphasis will be the features' relevancy, convenience of use, feedback's clarity and utility, and encouragement to get better. To determine whether the app has a noticeable impact on users' self-efficacy in speaking situations, confidence scores before and after use will be compared.

Furthermore, the surveys will investigate user happiness, recommendations for improvement, and willingness to promote *SpeakEasy* to others. Feedback from a variety of demographic groups (such as age, language proficiency, and use case) will also be examined to make sure the platform caters to a wide range of consumers. The objective is to use these insights to improve the user experience and make sure *SpeakEasy* stays a user-centred tool for confidence-building and voice coaching, in addition to validating the application's impact.

## **Project Plan**

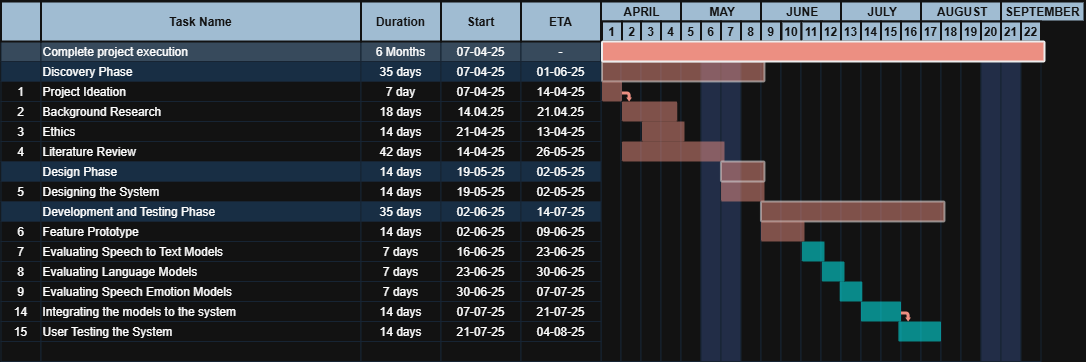


Figure - Gantt Chart Presenting Project Plan

This Gantt chart is a presentation of the project journey including the tasks that have been successfully completed (red bars) and future tasks that are yet to be completed (blue bars).

Portraying the Information in terms of a Sprint Backlog:

Sprint 1 –

|  |  |
| --- | --- |
| **Tasks Assigned to Sprint** | **Sprint Duration** |
| Evaluating Speech to Text Models  Evaluating Language Models  Evaluating Speech Emotion Models | 3 weeks |

To ensure the selection of accurate and efficient models for *SpeakEasy,* it is deemed necessary for them to be evaluated first before their integration into the system.

Sprint 2 –

|  |  |
| --- | --- |
| **Tasks Assigned to Sprint** | **Sprint Duration** |
| Integration of Models in the System | 2 weeks |

The evaluated models are then ready for system integration

Sprint 3 –

|  |  |
| --- | --- |
| **Tasks Assigned to Sprint** | **Sprint Duration** |
| User Testing of the System | 2 weeks |

After the system is complete, user testing is done to evaluate the entire system as a whole.

# **Feature Prototype**

The main goal of creating the *SpeakEasy* feature prototype is to model the system's essential operations before complete AI integration. In addition to helping stakeholders envision how the system will function in its final configuration, this early-stage implementation attempts to evaluate the fundamental user interaction flow, from speech input to feedback delivery. In order to facilitate iterative testing of the user interface and guarantee a seamless transition to the incorporation of actual AI models in subsequent stages of development, the prototype uses dummy data to mimic feedback.

## **Audio Input**

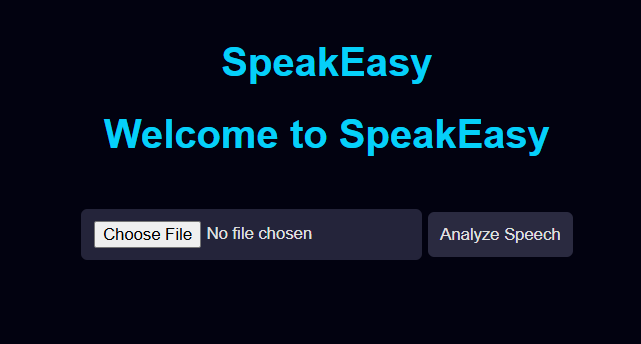


Figure - User Audio Input Feature

Users can submit their recorded voice for analysis using an easy-to-use interface thanks to the audio input capability. Currently, the system allows users to submit files in widely used file types as MP3 and MP4. The landing page prompts users to upload a speech file via a prominently displayed file selection window. This part is essential to *SpeakEasy*'s intended functionality since it simulates the first stage of the entire analysis pipeline, which comprises the transcription and analysis of user speech using AI models.

## **Displaying Feedback**

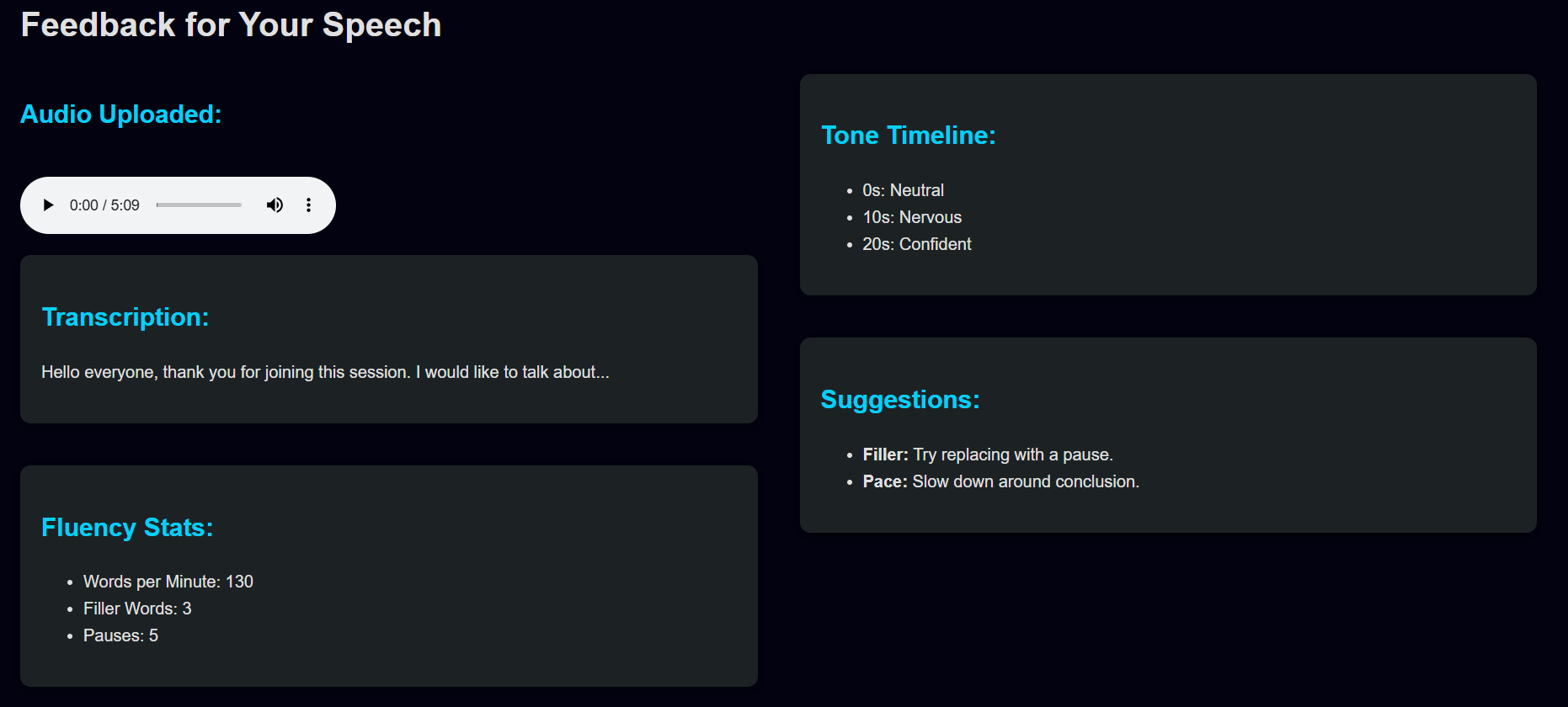


Figure - Feedback Page

The current version of *SpeakEasy* simulates analytical feedback using mock JSON data because the integration of AI models is a later development milestone. Predefined recommendations and scores pertaining to grammar, language usage, tone, and confidence level are included in this fake data. Even without real-time AI analysis, this method aims to duplicate the end-user experience by providing feedback in an organized and insightful manner. The information is useful for improving feedback display tactics and confirming how users may perceive and respond to feedback.

## **User Interface and Experience**

The front-end of the prototype was created with user comfort and clarity in mind. The interface, which was created with HTML and CSS, has a clear layout with enough padding, buttons that are easy to use, and text that is readable. Users are taken to a feedback page after uploading an audio file, where analysis findings are shown under various categories like "Confidence Score," "Filler Words," "Grammar Suggestions," and "Tone Assessment." The results are fuelled by simulated data. Users may quickly navigate through the material and comprehend how various components of their speech are assessed because to the feedback pieces' separation.

## **Tech Stack Used**

To enable quick development and testing, the prototype makes use of an effective and lightweight tech stack:

1. Backend: File uploads and routing are managed using Flask, a Python program.
2. Frontend: HTML and CSS are used to style and organize pages.
3. Data Simulation: Structured feedback material in the JSON mock data format.

In order to facilitate rapid prototyping and maintain scalability for future integration with AI services, these technologies were chosen.

## **Prototype Limitations**

Real-time AI analysis is not yet included in this feature prototype. The current implementation simulates dynamic model-generated feedback using static mock data. Furthermore, no speech sentiment analysis, grammar correction, or transcribing is actually done. The layout, user interaction, and feedback presentation are the only things being tested in this release. As a result, feedback's precision and adaptability do not represent the finished output.

## **Prototype Walkthrough**

The prototype's user experience is organized as follows:

1. Landing Page: A button titled "Upload Your Speech" asks the user to upload an audio file.
2. File Upload: From their device, the user chooses an audio file.
3. Feedback Generation (Simulated): The user is redirected to the feedback page by the system after submission.
4. Feedback Page: The user can examine their simulated performance by viewing mock feedback that is presented in many categories, including grammar fixes, tone insights, and a confidence score.

This straightforward yet useful walkthrough offers a clear picture of the main system flow meant for the finished *SpeakEasy* version.

## **Significance and Next Steps**

This prototype is fundamental to the creation of *SpeakEasy*. It makes it possible to test system architecture, interface design, and usability early on without needing the full AI backend. Before model integration starts, layout, interaction patterns, and user comprehension will be improved based on feedback from this prototype phase. The following are the upcoming development milestones:

1. Transcribing uploaded audio involves integrating a Speech-to-Text model.
2. implementation of a language model to offer recommendations in real time.
3. Vocal emotions and delivery are evaluated through the speech emotion models
4. Switch to real-time, adaptive feedback production based on real user voice from fake data.

This prototype guarantees a more seamless transition to the fully functional AI-driven speech coaching platform that *SpeakEasy* aspires to become by verifying the basic user experience early in the process.

# **References**

[1] Speeko - AI Speech Coach for Public Speaking. *Speeko*. Retrieved May 5, 2025 from https://www.speeko.co

[2] Poised: AI-Powered Communication Coach. Retrieved May 3, 2025 from https://www.poised.com/

[3] Adhish Deshpande, Rohit Pandharkar, and Subodh Deolekar. 2020. Speech Coach: A framework to evaluate and improve speech delivery. In *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, September 2020. 1–5. https://doi.org/10.1109/ICCCSP49186.2020.9315267

[4] Maria Labied, Abdessamad Belangour, and Mouad Banane. 2024. Assessing Speech-to-Text Translation Quality: An Overview of Key Metrics. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, December 11, 2024. IEEE, Manama, Bahrain, 1–6. https://doi.org/10.1109/DASA63652.2024.10836447

[5] Camille Landesvatter, Jan Behnert, and Paul C. Bauer. 2023. Comparing Speech-to-Text Algorithms for Transcribing Voice Data from Surveys. https://doi.org/10.31235/osf.io/vk6wj

[6] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground Truth for Grammaticality Correction Metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015. Association for Computational Linguistics, Beijing, China, 588–593. https://doi.org/10.3115/v1/P15-2097

[7] Jihan Apriliani Nurhasanah, Healty Susantiningdyah, Iwan Saputra, Ilham Rahmaddani Adhie, and Muchammad Chandra Cahyo Utomo. Grammar Correction: A Comparison of T5, LLAMA 2, and ChatGPT.