Project Synopsis: SpeakWise - Intelligent Speech Quality & Style Analyzer

1. Project Title

SpeakWise: Intelligent Speech Quality & Style Analyzer. The title reflects the project's core capability of using artificial intelligence to analyze speech for quality and stylistic elements, providing a clear insight into the work to be undertaken.

2. Problem Statement

Effective communication is crucial in professional and personal settings. Many individuals struggle to assess and improve their speaking skills due to a lack of objective, readily available feedback. They may be unaware of their use of filler words, their speaking pace, or the perceived confidence and formality of their tone. There is a need for an accessible tool that can provide data-driven, insightful analysis to help users enhance their oral communication abilities.

3. Objective and Scope of the Project

Objective : The primary objective of SpeakWise is to develop a web application that helps users improve their speech quality and style through AI-driven feedback. The project aims to deliver a platform that can accurately transcribe audio, identify speaking habits, and provide actionable insights. The core objectives include:

- **Speech-to-Text Conversion**: To accurately transcribe spoken audio to text.
- **Filler Word Detection**: To automatically identify disfluencies like "uh" and "um" to measure fluency.
- **Speech Quality Classification**: To classify a speech sample as "Good" or "Needs Improvement" based on linguistic features.
- **Formality & Confidence Analysis**: To analyze and classify the tone (Formal/Informal) and confidence (Confident/Hesitant) of the speech.
- **Interactive Feedback**: To present the analysis to the user through a visual and interactive dashboard with targeted suggestions.
- **Emotion Detection** : Analyzing acoustic and linguistic cues to detect emotions such as nervousness or excitement.

Scope

The scope of the Minimum Viable Product (MVP) will focus on core functionalities. The system will allow users to upload an audio file and receive a detailed analysis report. This report will visualize key metrics such as filler word count,emotions, speech rate, and confidence and formality. The project includes the entire process from data preprocessing and feature

engineering to training classification models and exposing the service via an API. Advanced features like multi-language support, and personalized coaching are considered out of scope for the MVP but are planned for future iterations.

4. **Methodology (Process Description)**

The proposed system will follow a sequential process to analyze speech from input to feedback generation. The data and control flow are described below:

1. **Speech Ingestion & Transcription** : The user uploads an audio file. The system ingests this audio and uses a high-accuracy Automatic Speech Recognition (ASR) model, such as Whisper, to convert the speech into a textual transcript. This transcript is enriched with timestamps for further analysis

2. **Feature Engineering & Analysis** : The generated transcript undergoes preprocessing to normalize and clean the text. The system then computes various speech-level features, including:

    - **Filler Word Detection** : Identification of disfluencies ("um", "uh", etc.).
    - **Linguistic Features**: Calculation of speech rate, pause density, and lexical diversity.
    - **Embedding Generation**: Creation of text embeddings to detect tone and style.

3. **ML-Based Classification**: The engineered features are fed into pre-trained machine learning models:
    - **Speech Quality Classification** : A classifier predicts if the speech is "Good" or "Needs Improvement".
    - **Formality & Confidence Analysis** : NLP models classify the tone as "Formal/Informal" and the confidence level as "Confident/Hesitant". Models like Random Forest, SVM, or BERT will be trained and evaluated for these tasks.

4. **Interactive Feedback Dashboard** : The final analysis is presented to the user on an interactive dashboard. This dashboard provides:
    - A preview of the transcript.
    - Visualizations of fluency metrics, filler words, confidence, and tone scores.
    - A detailed report with actionable suggestions for improvement.

5. Hardware & Software to be used

- **Software Requirements**:

    - **Programming Language**: Python 3.8+
    - **Audio Processing Libraries**: speech_recognition, whisper, pydub
    - **NLP Libraries**: spaCy, NLTK

    - **ML/Modeling Libraries**: scikit-learn, xgboost
    - **Development Tools**: Jupyter/Colab, HuggingFace models
    - **API Framework**: FastAPI or Flask, Azure Speech API

- **Hardware Requirements**:
    - A standard development machine (PC/Laptop) for coding and initial model training.
    - A server with GPU capabilities for efficient training of deep learning models (like Whisper and BERT) and for deploying the web application and API.

6. Future Work of this Project

Future iterations of the project will focus on expanding its capabilities and enhancing the user experience. The planned enhancements include:

- **Multi-Language & Accessibility** : Integrating multilingual models to support various languages and adding accessibility features like audio-read reports.

- **Personalized Coaching** : Tracking user progress over time to provide personalized improvement tips and gamifying the experience with badges.

- **Speech Summarization** : Using Large Language Models (LLMs) to generate summaries of lengthy speeches.

- **Gamified Learning Mode** : Introducing interactive challenges and milestones to engage users.

- **Model Serving & Scheduling**: Automating model retraining and feature updates using tools like Airflow.

7. The Schedule of the project

The project will be developed in phases based on the priority of the requirements.

Phase	Duration (Est.)	Key Tasks
Phase 1: Core Functionality (MVP)	8 Weeks	- Setup technology stack - Develop speech-to-text pipeline - Implement feature engineering and filler word detection. - Initial training of models for quality, formality, and confidence - Build basic interactive dashboard and API.
Phase 2: Refinement & Validation	4 Weeks	- Conduct Exploratory Data Analysis (EDA) on results . Perform human validation to ensure model accuracy .Refine models and feedback suggestions based on validation.
Phase 3: Future Enhancements	TBD	- Begin development of medium-to-low priority features such as multi-language support and personalized coaching - Implement automated model retraining and serving .

8. Conclusion

The "SpeakWise" project presents a robust and well-conceived vision for an intelligent speech analysis platform. By leveraging state-of-the-art ASR and NLP technologies, it offers an innovative solution to the common challenge of improving public speaking skills. The project stands out by providing users with personalized, data-driven feedback that is not easily obtainable otherwise. While the initial MVP is clearly defined, successful implementation will depend on a strong data strategy for training accurate and reliable models for classifying

abstract concepts like "good" speech and "confidence".

9. References and Bibliography

[1] Radford, A., et al. "Robust Speech Recognition via Large-Scale Weak Supervision." *arXiv preprint arXiv:2212.04356*, 2022.

[2] Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830.

[3] Bird, S., Klein, E., and Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[4] Honnibal, M., and Montani, I. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." *To appear*, 2017.

[5] S. Burnwal, "Speech Emotion Recognition," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition

[6] E. J. Lok, "CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/ejlok1/cremad

[7] UWRF Kaggler, "RAVDESS Emotional speech audio," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

[8] Microsoft, "Azure AI Speech," Microsoft Azure. [Online]. Available: https://azure.microsoft.com/en-us/products/ai-services/ai-speech