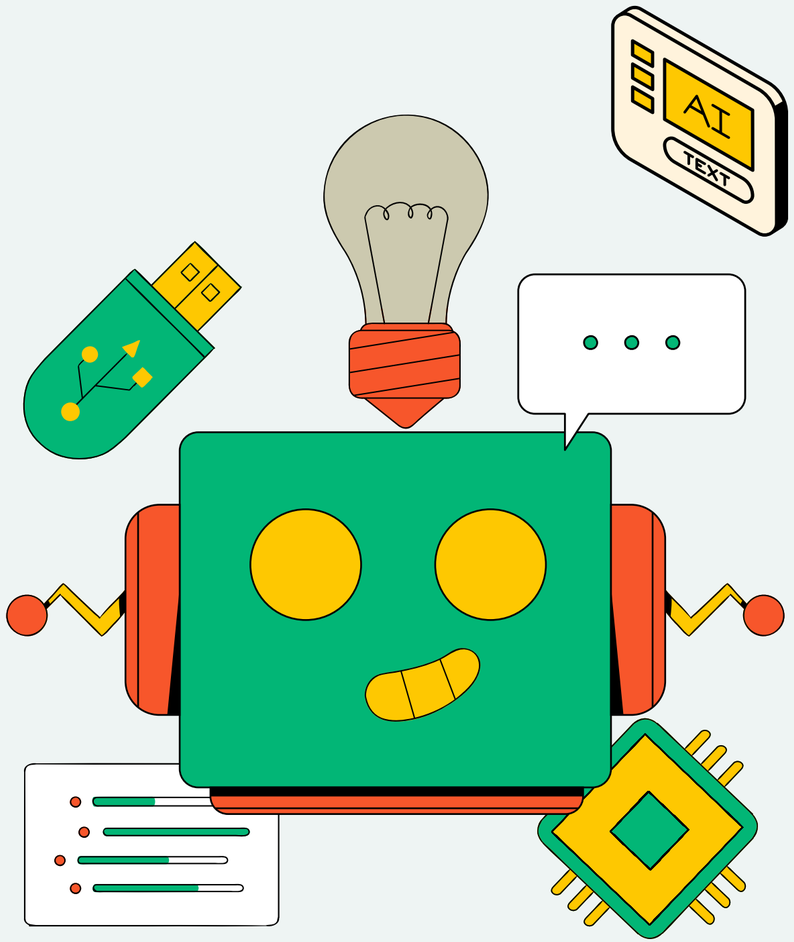
DETECTING AI-ASSISTED INTERVIEW CHEATING THROUGH SPEECH ANALYSIS

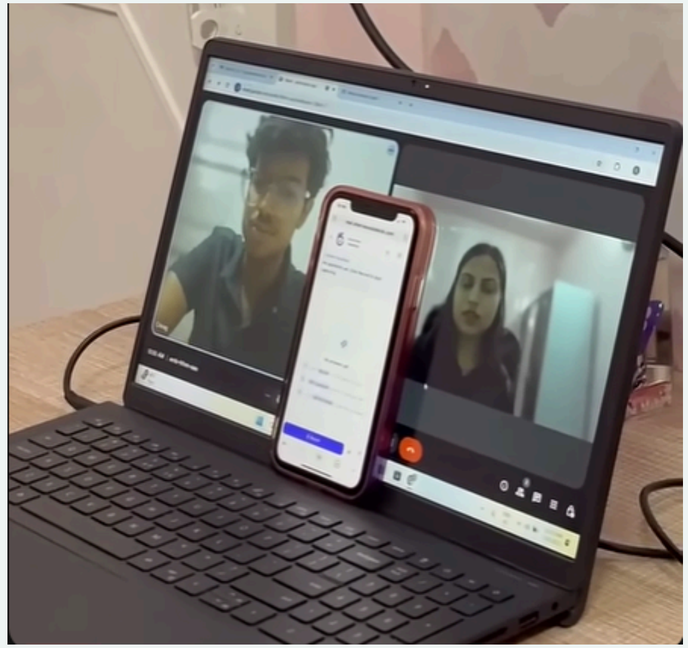


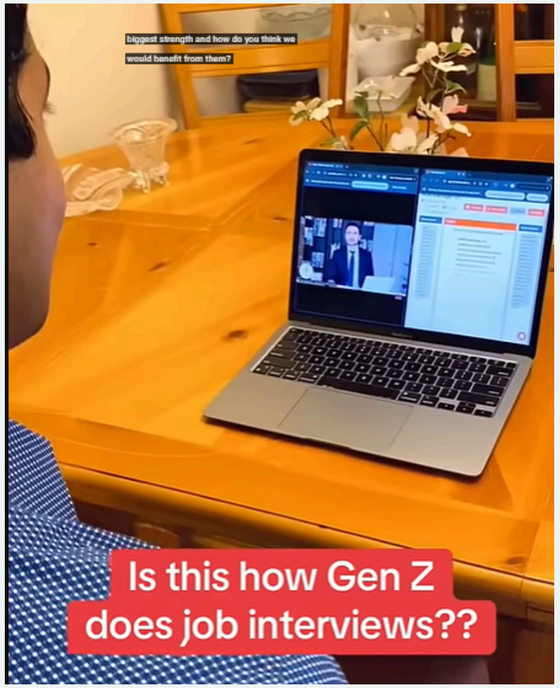
People use AI to cheat! We use AI to CATCH THEM!!!

- AI tools in the market have now entered the interview room, candidates are openly using tools to **generate answers in real time.**

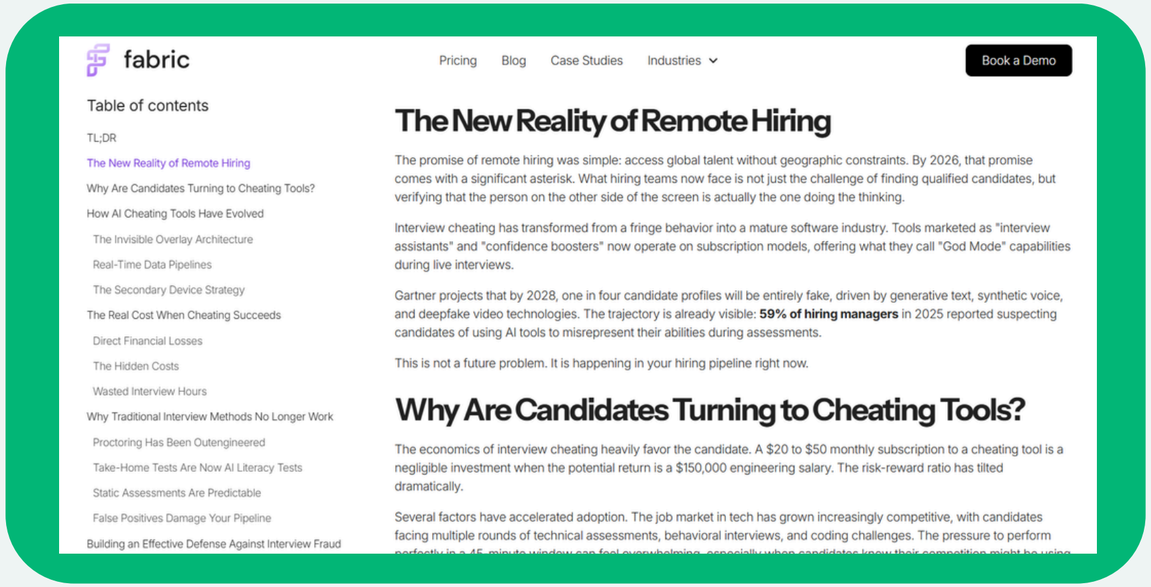
- Platforms like **Interview Copilot** and **Cluely** have emerged specifically to assist candidates during live interviews without detection.

- **Our system fights back by analysing speech patterns, exploiting the natural acoustic difference between how people speak spontaneously versus how they sound when reading AI-generated answers from a screen.**





SIGNICANCE OF OUR PROJECT



- Several studies reveal that **more than 50% of candidates now use AI tools** to assist them during interviews, fundamentally compromising the integrity of remote hiring.

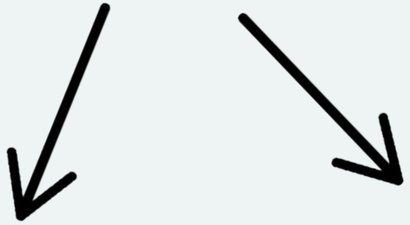
- The situation has grown so severe that **HR leaders** are calling remote hiring 'an **AI literacy test** rather than a skills assessment

- Major companies like **Google and McKinsey** already moving to **in-person interviews**

Ways

Face Expression and
Movements analysis

Speech Analysis



Read

Spontaneous

LITERATURE REVIEW

PAPER 1

CLASSIFICATION OF SPONTANEOUS AND SCRIPTED SPEECH FOR MULTILINGUAL AUDIO

Shahar Elisha1,2, Andrew McDowell1 , Mariano Beguerisse-D´ ıaz1 , Emmanouil Benetos

Overview:

- The paper develops multilingual models to classify scripted vs spontaneous speech from podcast audio.
- It uses a large Spotify dataset covering 15 languages and multiple podcast formats.
- The study compares handcrafted acoustic features with transformer models like Whisper and YAMNet.

Methodology and Key Findings :

- The study uses acoustic features from openSMILE and speaker statistics from Pyannote.
- Whisper-large-v2 achieves the best performance across languages and datasets.
- Whisper reaches about 0.95 AUC, outperforming handcrafted features and YAMNet.
- The models also generalize well on external datasets like CEFC and DIHARD.

Limitations:

- The paper uses mostly clean and professionally recorded podcast audio, which may not generalize well to noisy real-world interview recordings.
- The dataset contains long podcast conversations, which differ significantly from short, high-pressure interview responses seen in actual hiring environments.

PAPER 2

A NOVEL SCHEME TO CLASSIFY READ AND SPONTANEOUS SPEECH SUNIL KUMAR KOPPARAPU1[0000 −0002 −0502 −527 X]

Overview :

- It uses DeepSpeech to convert speech into alphabet sequences for feature extraction.
- The approach focuses on simple and explainable speech features instead of complex deep models.

Methodology and Key Findings :

- Features such as word rate, active alphabet length, and inactive alphabet rate are extracted from DeepSpeech outputs.
- Spontaneous speech contains more pauses and inactive alphabets compared to read speech.
- A lightweight scoring classifier is proposed without requiring model training.
- The system achieved about 88% accuracy on the ALLSSTAR speech dataset.

Limitations

- The approach depends heavily on DeepSpeech-generated alphabet patterns, making performance sensitive to speech recognition errors caused by noisy or low-quality audio.

DATASETS

LIBRI SPEECH–
The Read Speech

- **Overview**: Recordings of English audiobook readings, labeled as scripted/read speech.
- **Relevance to Project**: Provides clean, high quality examples of read speech directly represents how a candidate sounds when reading AI generated answers during an interview.
- **Source of Collection:** Derived from LibriVox audiobook recordings by volunteers, curated by Johns Hopkins University. All recordings are read aloud from existing book texts.
- **Scale**: 28000 samples ,960 hours of worth

BUCKEYE

- **Overview:** 12000–13000 interview samples of 40 speakers.
- **Relevance to Project:** Provides real examples of unscripted conversational speech directly represents how a genuine candidate sounds when answering interview questions from their own knowledge.
- **Source & Collection:** Collected by The Ohio State University through interview-style conversational recordings where speakers discussed everyday topics in natural dialogue settings.
- **Scale:** 12000–13000 samples

DATASET SNIPPETS - LIBRI SPEECH

audio audio · *duration (s)*	text string · *lengths*	speaker_id int64	chapter_id int64	id string · *lengths*
1.29 35	5 576	61 8.56k	960 295k	13 16
▶ 0:00 / 0:03	CONCORD RETURNED TO ITS PLACE…	6,930	75,918	6930-75918-0000
▶ 0:00 / 0:14	THE ENGLISH FORWARDED TO THE…	6,930	75,918	6930-75918-0001
▶ 0:00 / 0:05	CONGRATULATIONS WERE POURED IN…	6,930	75,918	6930-75918-0002
▶ 0:00 / 0:23	FROM THE RESPECT PAID HER ON ALL…	6,930	75,918	6930-75918-0003
▶ 0:00 / 0:11	SHE TAUGHT HER DAUGHTER THEN BY…	6,930	75,918	6930-75918-0004
▶ 0:00 / 0:13	THE COUNT HAD THROWN HIMSELF…	6,930	75,918	6930-75918-0005
▶ 0:00 / 0:05	THIS HAS INDEED BEEN A HARASSING…	6,930	75,918	6930-75918-0006
▶ 0:00 / 0:03	YOU WILL BE FRANK	6,930	75,918	6930-75918-0007

READ (LIBRISPEECH) DATASET

[SNIPPET (First 7 Rows)]

audio_file	speaker_id	chapter_id	duration (s)
libri_198-126831-0017_part000.wav	libri_198	126831	12.94
libri_2092-145709-0004_part000.wav	libri_2092	145709	14.81
libri_1841-159771-0005_part000.wav	libri_1841	159771	15.04
libri_233-155990-0041_part000.wav	libri_233	155990	13.8
libri_226-122538-0030_part000.wav	libri_226	122538	12.84
libri_1081-125237-0038_part000.wav	libri_1081	125237	12.05
libri_103-1240-0000_part000.wav	libri_103	1240	12.09

[DURATION STATS]

	duration (s)
count	6100
mean	12.6271
std	2.4455
min	5.01
25%	12.06
50%	13.14
75%	14.27
max	16.97

DATA SNIPPETS- BUCKEYE

SPONTANEOUS (BUCKEYE/SBC) DATASET

[SNIPPET (First 7 Rows)]

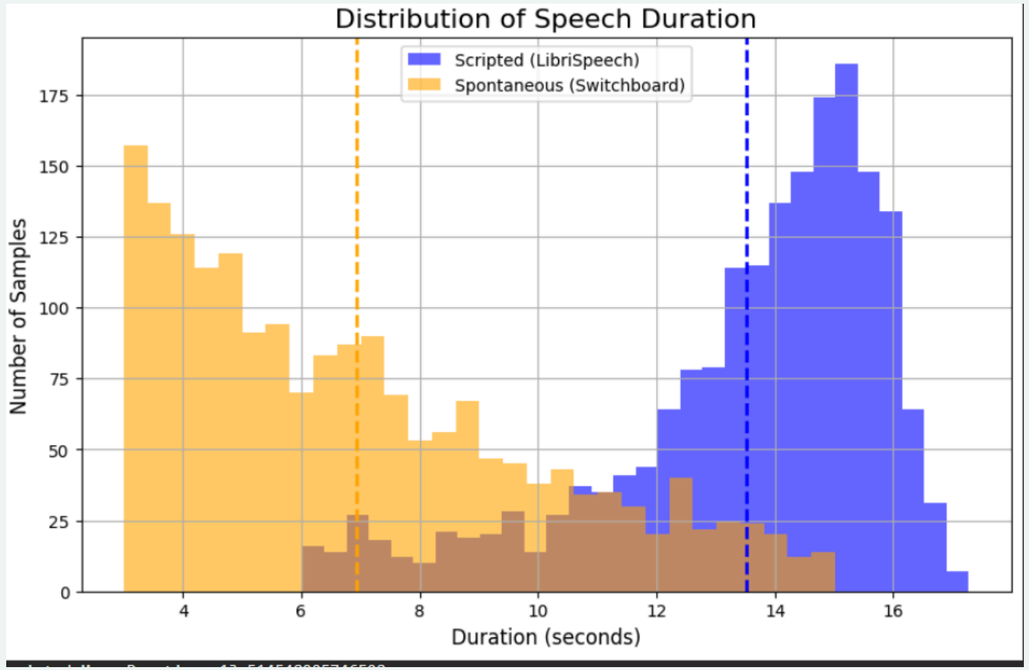
audio_file	speaker_id	chapter_id	duration (s)
s1802a_part023.wav	s1802a	N/A	12.91
s0304b_part007.wav	s0304b	N/A	16.77
s0403b_part006.wav	s0403b	N/A	13.15
s3302b_part025.wav	s3302b	N/A	9.74
s1604a_part015.wav	s1604a	N/A	16.55
s3701a_part004.wav	s3701a	N/A	14.98
s2803a_part007.wav	s2803a	N/A	13.15

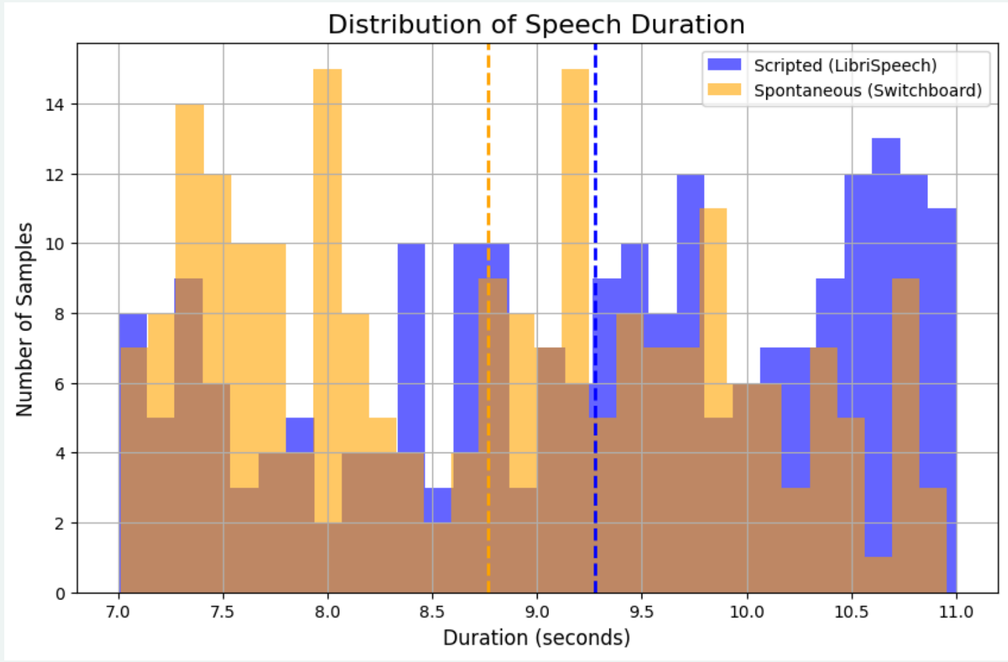
[DURATION STATS]

	duration (s)
count	6120
mean	14.3536
std	1.66634
min	5.26
25%	13.17
50%	14.42
75%	15.6825
max	17

DATA CLEANING &PREPROCESSING

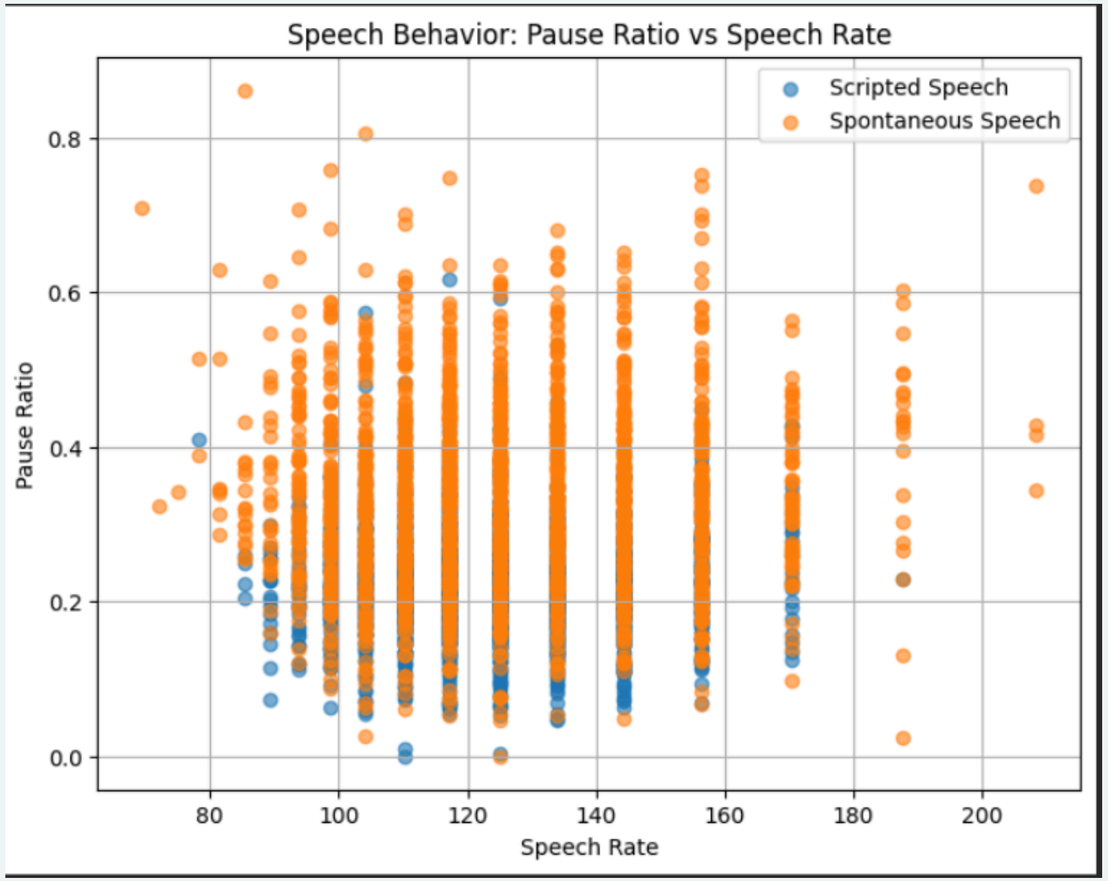
1.Removed clips under 7 seconds and over 11 seconds , to match duration of audio clips distributions between datasets

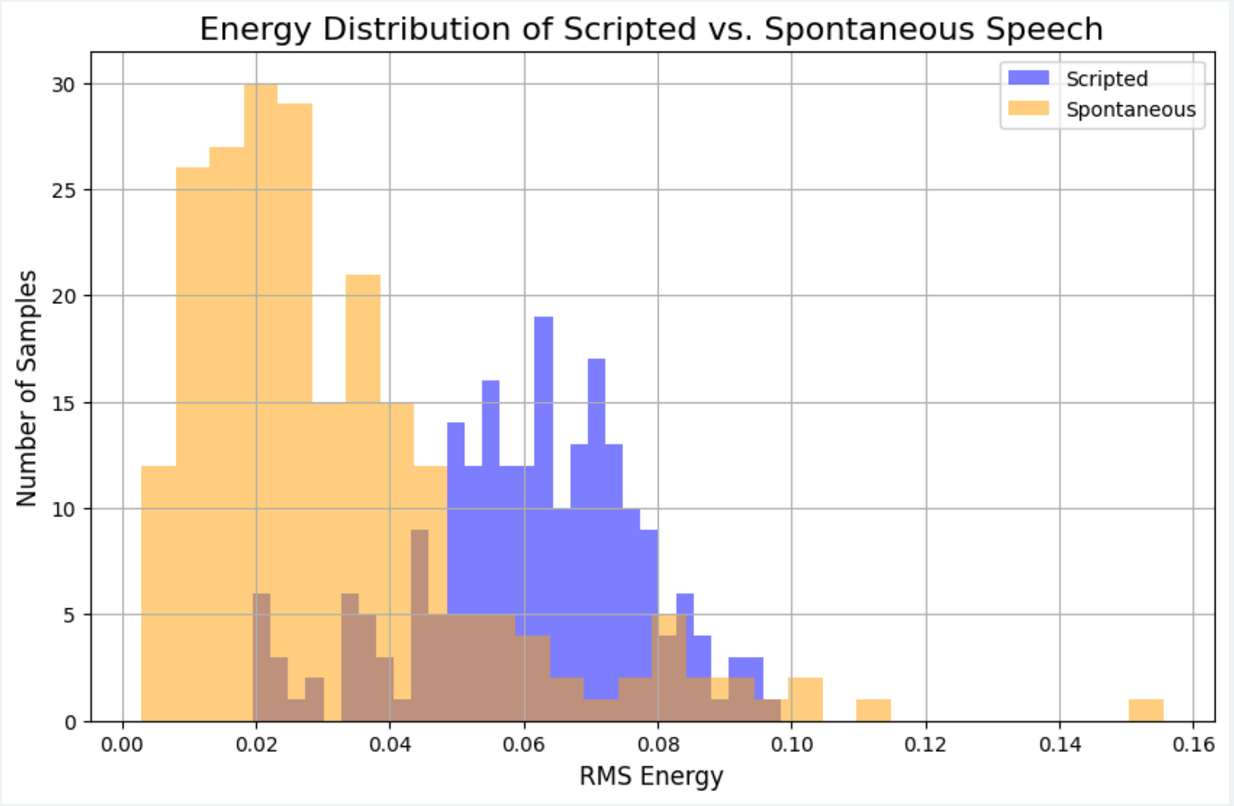




- Using an RMS (Root Mean Square) threshold which measures average audio energy to filter out near-silent clips that contain no meaningful speech signal.

FEATURE PLOTS





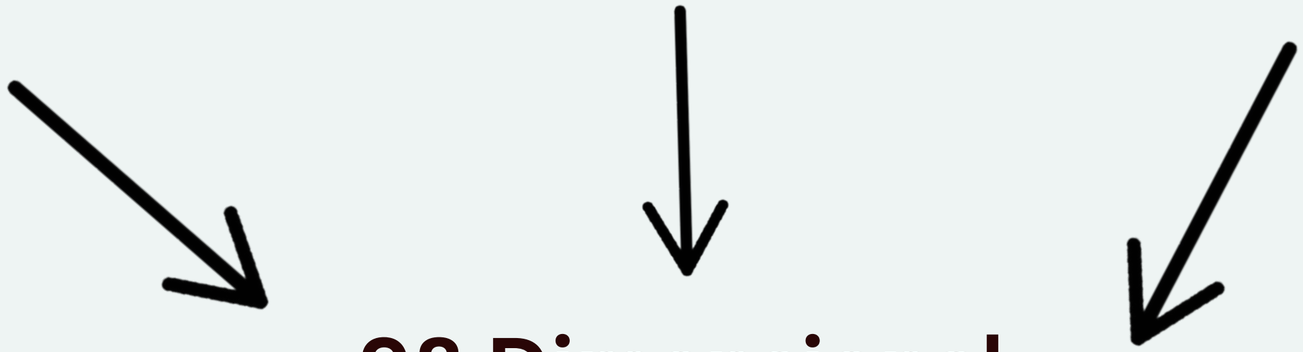
PART 1

FEATURE EXTRACTION

Open Smile

eGeMAPSV02

Supplemental Statistics



98 Dimensional Fature Vector

PCA → Training

SVM Classifier

But The model Cheated!!

Classification Report Table

Class	Precision	Recall	F1-Score	Support
Spontaneous	1.0000	0.9950	0.9975	200
Read	0.9950	1.0000	0.9975	200
Accuracy			0.9975	400

The Real Problem!!! – Audio Quality

Dataset 1 – Read

Audio Preview

Audio Properties

Format	Duration (s)	Bitrate (bps)	Sample Rate (Hz)
wav	14.96	256023	16000
Channels	**Codec**	**LUFS**	**True Peak (dB)**
1	pcm_s16le	-33.36	-12.08
Loudness Range	**Max Volume (dB)**	**RMS**	**Probe Score**
4	-12.1	-33.3	99

Dataset 2 – Spontaneous

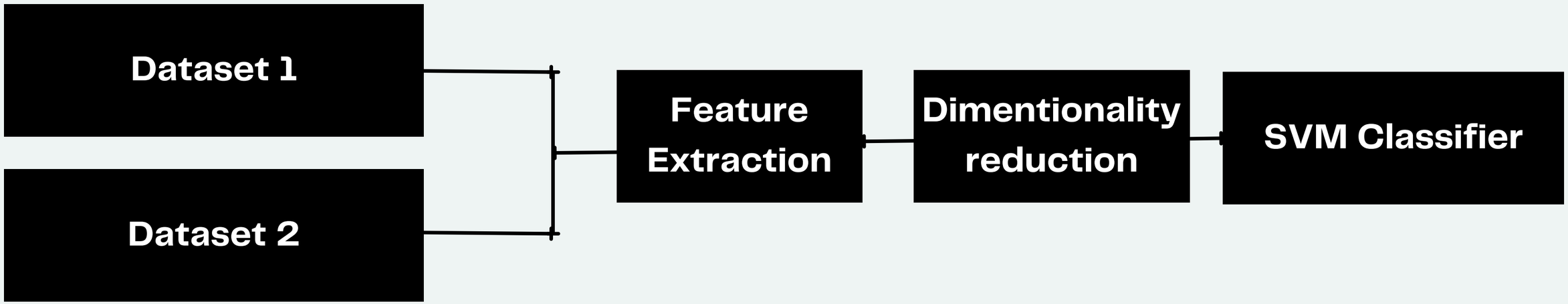
Audio Preview

Audio Properties

Format	Duration (s)	Bitrate (bps)	Sample Rate (Hz)
wav	10.05	256035	16000
Channels	**Codec**	**LUFS**	**True Peak (dB)**
1	pcm_s16le	-28.23	-8.76
Loudness Range	**Max Volume (dB)**	**RMS**	**Probe Score**
2.5	-8.8	-27.8	99

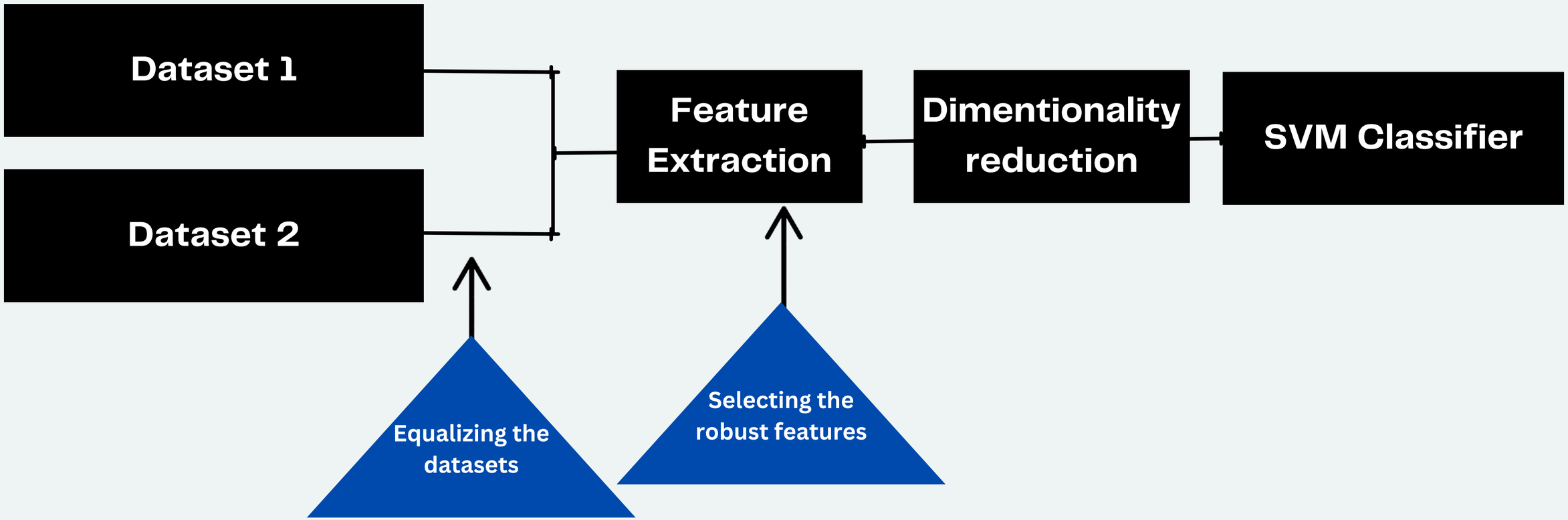
Now this was no more an Ml problem but an engineering problem!!

This is our existing pipeline



Now this was no more an Ml problem but an engineering problem!!

This is our existing pipeline, We added 2 more stages in our pipeline to eliminate the overfitting of model,



PART 2

1.Datasets Equalization

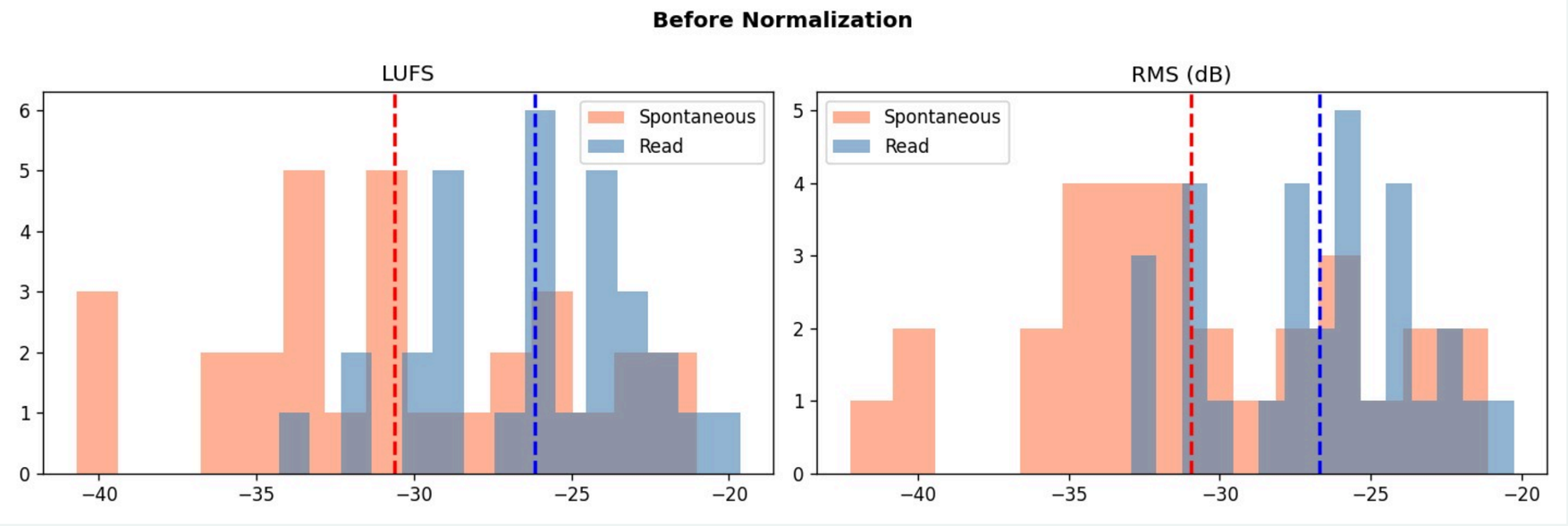
To reduce the recording-quality mismatch between the clean LibriSpeech corpus and real-world spontaneous speech recordings, we introduced a four-stage audio normalization and cleaning pipeline. This preprocessing ensured that the model learned speech-style characteristics rather than dataset-specific recording artifacts.

- **Reduced excessive high–frequency detail present in studio recordings.**
- **Matched the spectral bandwidth characteristics of conversational audio**

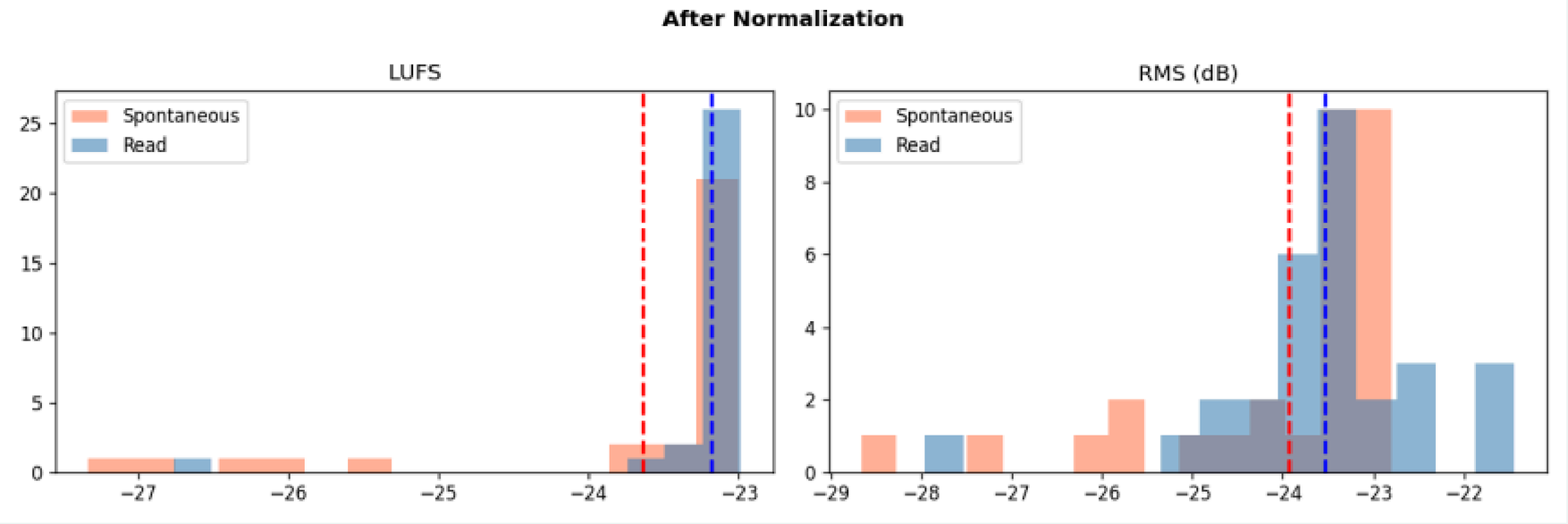
- **Standardized perceived loudness across both datasets.**
- **Prevented the model from using volume differences as a shortcut feature.**

- **Equalized maximum signal amplitude.**
- **Prevented clipping and inconsistent peak**

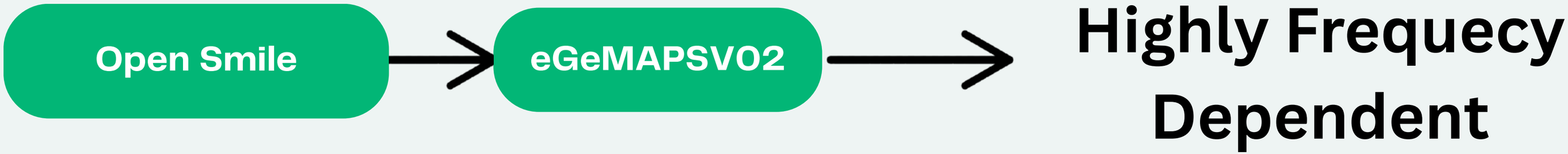
- **Applied only to the cleaner LibriSpeech recordings.**
- **Simulated realistic ambient conversational conditions.**
- **Matched the background noise floor of spontaneous recordings.**



Normalization minimizes variance between speech types, standardizing LUFS and RMS distributions.



2.Feature selection



- Extracted 98 handcrafted acoustic and prosodic features using OpenSMILE, including MFCCs, pitch, energy, spectral, and temporal descriptors.

- Many features were highly correlated with recording quality (noise, loudness, microphone artifacts), causing the model to learn dataset differences instead of speech style.

- Through iterative feature ablation and leakage testing, we removed unstable features and retained only robust conversational cues such as pauses, rhythm, articulation dynamics, and speaking variability.

98 Features

↓

58-Feature Set (Partial Leakage)

Removing MFCC Features to Reduce Dataset Bias

Why did we remove MFCCs?

- **MFCC features dominated the 98-feature model.**
- **Leakage analysis showed MFCCs were strongly correlated with:**
  - **microphone characteristics**
  - **recording environment**
  - **channel quality**
- **The model was learning dataset identity instead of speech behavior.**

98 Features

↓

Removed 40 MFCC-based Features

↓

58 Remaining Features

MODEL CHEATED AGAIN!!

Classification Report Table

Class	Precision	Recall	F1-Score	Support
Spontaneous	0.9949	0.9850	0.9899	200
Read	0.9851	0.9950	0.9900	200
Accuracy			**0.9900**	400

“Removing MFCCs alone was insufficient to eliminate dataset leakage.”

DOMAIN KNOWLEDGE ABLATION

IDENTIFY WHICH FEATURES ARE GENUINELY BEHAVIORAL AND WHICH FEATURES ARE LEAKING DATASET IDENTITY

Removed 107 standard acoustic features after leakage analysis showed strong dataset dependence. They were inherently biased by the microphone quality.

We engineered **8 highly-targeted behavioral features** using a **Dynamic Voice Activity Detector (VAD)**. The VAD mathematically isolates the unique background noise of **every single audio file** (calculating the 10th percentile of RMS energy) and sets a dynamic threshold 10dB above it to perfectly identify when a user is pausing versus talking, completely ignoring background static.

The Final 8 Features: **`speech_ratio`**, **`silence_ratio`**, **`mean_speech_dur`**, **`std_speech_dur`**, **`mean_silence_dur`**, **`std_silence_dur`**, **`n_speech_segs`**, **`n_silence_segs`**.

DYNAMIC VOICE ACTIVITY DETECTOR (VAD)

“Dynamic VAD adaptively separates speech and silence regions based on the background noise profile of each audio file.”

ADAPTIVE BEHAVIORAL FEATURE EXTRACTION PIPELINE

Audio Signal

↓

Frame Windowing (25 ms)

↓

RMS Energy Calculation

↓

Dynamic Noise Floor Estimation

↓

Adaptive Thresholding

↓

Speech / Silence Mask

↓

Behavioral Statistics

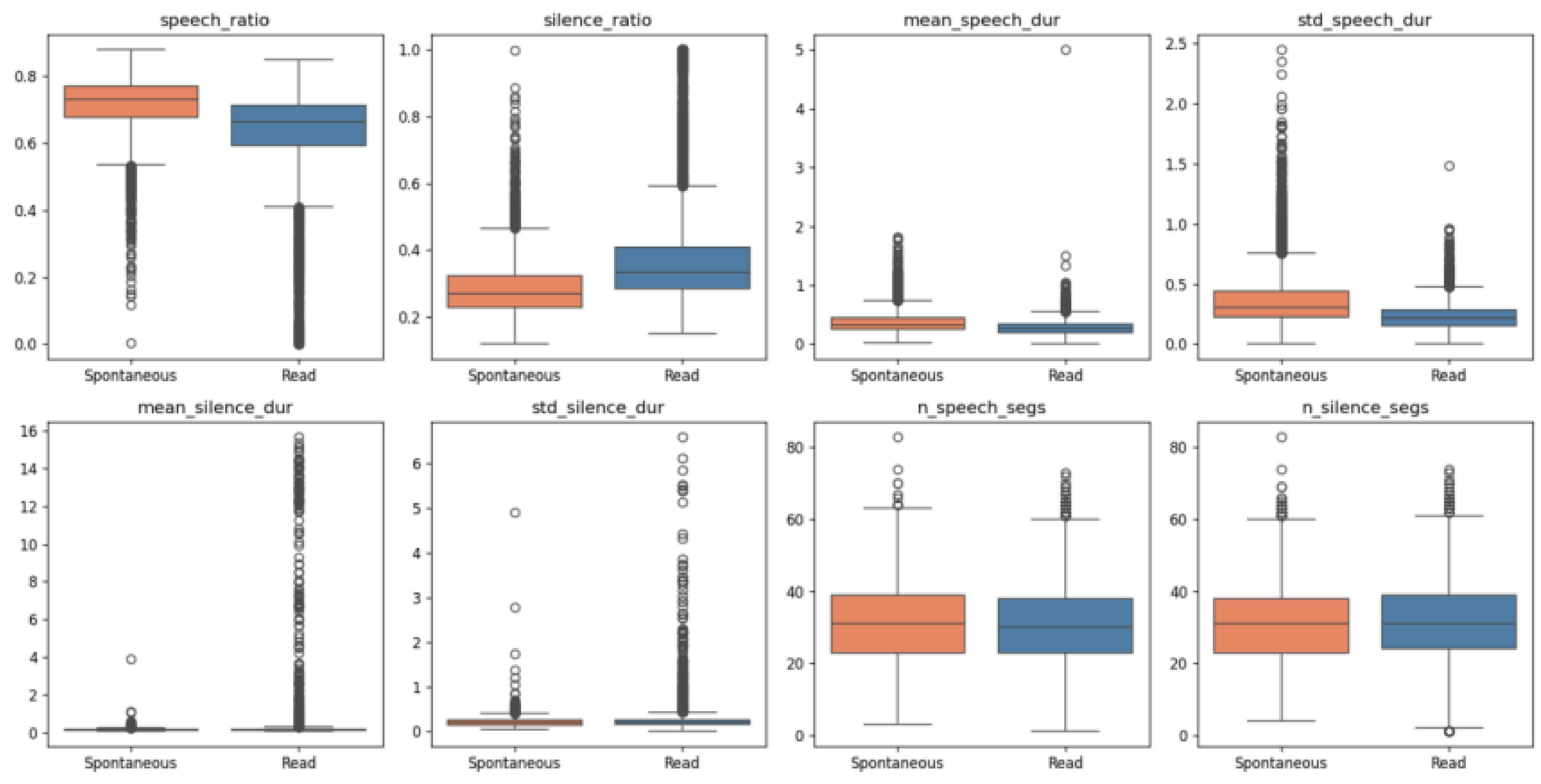
↓

Final 8 Features

Fixed-threshold VAD failed due to large noise differences between datasets.

Threshold=Noise Floor+10dB

FEATURE DISTRIBUTION



MODEL USED

Why Support Vector Machine (SVM)?

- Effective for small-to-medium feature spaces
- Works well with non-linear decision boundaries
- Robust against overfitting after feature reduction
- Performs strongly on behavioral statistical features

Audio Input

↓

Dynamic VAD

↓

8 Behavioral Features

↓

Feature Standardization

↓

SVM Classifier (RBF Kernel)

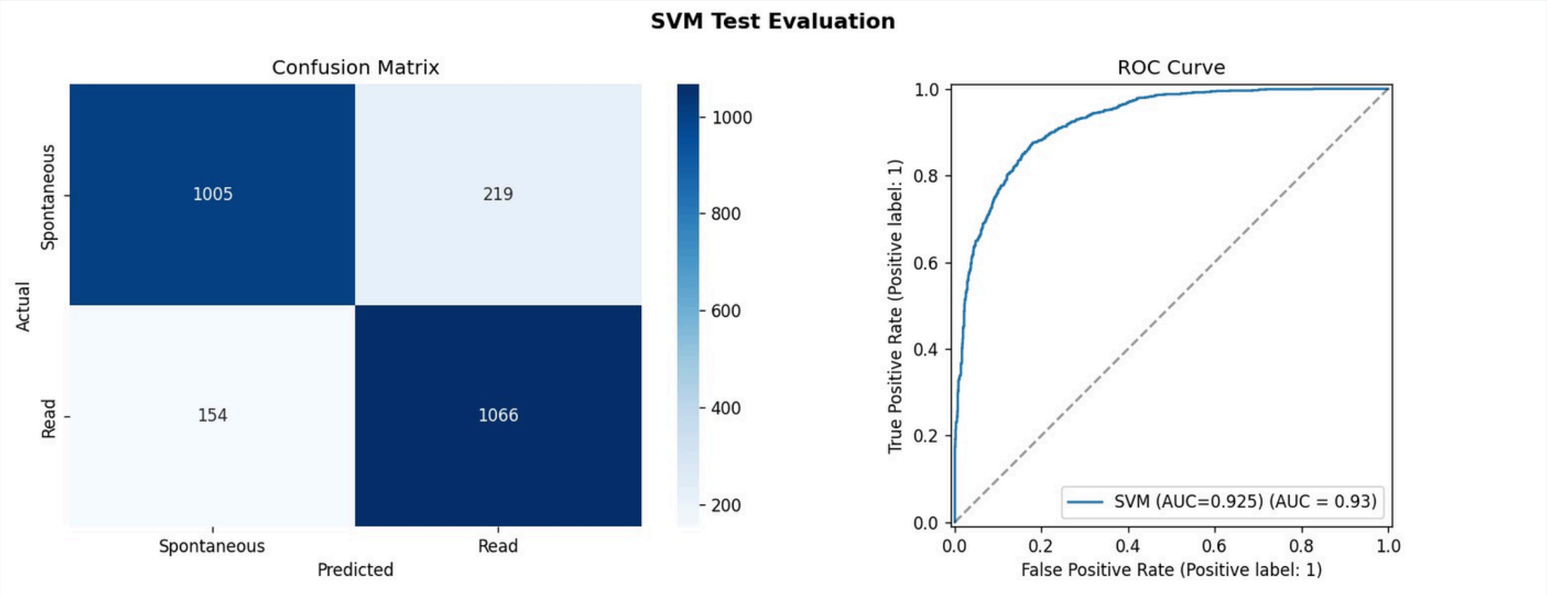
↓

Prediction:
Read vs Spontaneous

RESULTS

Behavior-Focused Classification Performance

Class	Precision	Recall	F1-Score	Support
Spontaneous	0.87	0.84	0.85	1224
Read	0.84	0.87	0.86	1220
Accuracy			**0.86**	2444
Macro Avg	0.86	0.86	0.86	2444



DEPLOYMENT AT PLAKSHA

HOW IT COULD BE DEPLOYED

Integration: Run as a lightweight backend service connected to Plaksha's existing online interview or proctoring platform.

Post-call processing: Each recorded interview session is processed after the call — audio is cleaned, features extracted, and an SVM score returned.

Output: Confidence estimate and per-segment breakdown flagging likely AI-assisted responses.

Infrastructure: Near-instant inference (~98-dim features, SVM) on a single CPU instance with sub-second latency per file — no GPU required.

SCALING CHALLENGES

Multilingual generalization: The model was trained on English speech; it may not transfer reliably to interviews conducted in Hindi or other languages spoken on campus.

Domain/audio shift: If recording conditions differ significantly from training data, model performance may degrade without re-calibration.

Adversarial adaptation: Candidates who learn that reading-style signals trigger flags may consciously mimic spontaneous speech prosody while still reading.

Threshold tuning: A fixed 0.5 decision threshold may require adjustment to control false-positive rates fairly across diverse speakers.

Privacy & ethics: Automated flagging systems require transparent policies and human review to avoid unjust disqualification.

THANK YOU