MLPR Final Presentation

Expression Intensity Alignment With Dialogue Context

Quantifying the Gap Between Script Intent & Physical Delivery

Abhishek, Manan Singla, Shivika Dhawan

Problem Statement

Why does emotional alignment matter?

In acting, emotional delivery depends on both emotion type and intensity, which should align with the meaning and intent of the dialogue.

Existing models generally asks: "What emotion is this?"

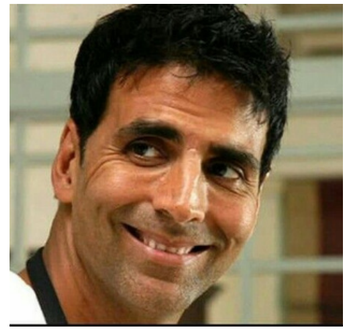
But what also matters is: "Does the delivery match what the script demands?"

Under-Expressive

Script demands anger,
Actor delivers a
flat, monotone whisper

Over-Expressive

Script has mild surprise
Actor screams and
overacts dramatically



Correctly Aligned

Physical delivery
matches the emotional
gravity of the script

Potential Applications

- **Acting and performance feedback**: assessing whether emotional delivery matches dialogue intent.
- **Human computer interaction:** enabling AI systems to better interpret human emotions in context.
- **Media analysis**: automated analysis of emotional expression in **Acting Industry** and may act as a benchmark.
- **Emotion aware AI systems:** improving virtual assistants, social robots, and conversational agents.

Imapact

- Advances research in multimodal emotion analysis.
- Helps AI understand **emotions with context**, instead of just predicting emotion labels.
- Enables development of more emotionally intelligent and context-aware AI systems.

Literature Review & Gaps

What's been done and what we uniquely solve?

Literature Review

S. No.	Title	Year	Journal / Conference	Location	Methods	Accuracy	Author	Citation
1	MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations	2019	ACL Conference	Florence, Italy	Multimodal emotion recognition using audio, video, and text features	~65%	Soujanya Poria and others.	Poria et al., 2019
2	Multimodal Sentiment and Emotion Recognition using Deep Learning	2017	EMNLP	Copenhagen, Denmark	Deep neural networks combining speech, facial expressions, and language features	~70%	Amir Zadeh and others.	Zadeh et al., 2017
3	CMU-MOSEI (Zadeh et al.) Multimodal Sentiment	2018	ACL(Association for Computational Linguistics) conference	Pennsylvania, USA	Multimodal sentiment analysis	~80%	Paul Pu Liang, Soujanya Poria, Erik Cambria, Amir Ali Bagher and Louis-Philippe Morency	Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018).
4	Speech Emotion Recognition using Deep Neural Networks	2018	IEEE Transactions on Affective Computing	USA	Acoustic feature extraction (MFCC, pitch, energy) with deep neural networks	~72%	Björn Schuller and others.	Schuller et al., 2018
5	Multimodal Emotion Intensity Estimation using Audiovisual Signals	2016	ACM International Conference on Multimodal Interaction	Tokyo, Japan	Fusion of facial expressions, body gestures, and speech features	~74%	H. Ranganathan and others.	Ranganathan et al., 2016

Existing Solutions & their Limitations

1. Speech Emotion Recognition (SER)

These systems detect emotions only from audio signals.

Limitation: They only use voice information and ignore facial expressions and dialogue context.

2. Facial Expression Recognition (FER)

These systems detect emotions from facial expressions in images or videos.

Limitation: They only use visual information and ignore speech and dialogue meaning.

3. Multimodal Emotion Recognition (MER)

These systems combine multiple modalities such as: audio, facial expressions, text

Limitation: They mainly classify emotions (happy, sad, angry) and do not evaluate whether the emotional expression matches the dialogue meaning.

4. Emotion Intensity Estimation

Some research tries to estimate how strong an emotion is, rather than just the emotion type.

Limitation: These systems estimate intensity but do not relate it to dialogue context.

Research Gaps

No Continuous Intensity

- Existing models give a discrete label.
- They can't say how intensely angry someone is on a 0.0–1.0 scale.

→

Our Extension

Knowledge distillation from RoBERTa + EmoRoBERTa to generate continuous Expected Intensity scores.

No Decoupling of Text vs. Physical

- Models mash text and AV together.
- If script says 'I am furious' but actor smiles, models average and get confused.

→

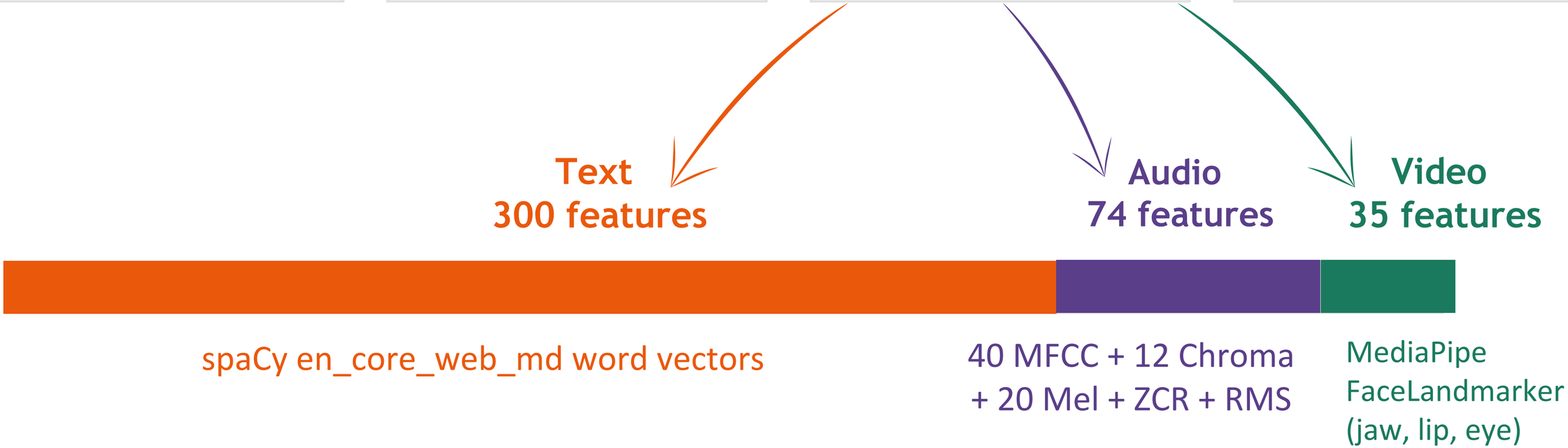
Separate BiLSTM brains for Text and Audio/Video , independently measuring Expected vs. Expressed.

Dataset: CMU-MOSEI

A massive, widely used dataset created by **Carnegie Mellon University** for analyzing human sentiment and emotion in online videos

CMU-MOSEI Overview

23,000+	15,347	409	50
Raw videos in dataset	Synced data points used	Features per time-step	Time-steps per sequence



Why CMU-MOSEI?

- In-the-wild YouTube monologue videos (real background noise, varied lighting, diverse accents)
- Forces model to be robust, not a clean lab recording
- Pre-extracted SDK .pkl files available (allow GPU-efficient training).

Data & Feature Preprocessing

The **ANNOTATION** journey & pipeline

The Missing Feature: Knowledge Distillation

CMU-MOSEI had **NO 'Expected Intensity' label**

16,000 Transcripts

CMU-MOSEI **raw text**

→

RoBERTa distilroberta-base

General emotion probability dist.

→

EmoRoBERTa j-hartmann model

High-arousal emotion probabilities

→

Expected Intensity 0.0 - 1.0 Score

roberta_score + emoberta_score = **CSV**

- RoBERTa & EmoRoBERTa are specialized NLP transformers fine-tuned for emotion.
- They output mathematical probability distributions over high-arousal emotions (Anger, Fear, Joy).
- We mathematically collapse these into a 0–1 continuous intensity score.

Pre-processing Pipeline

Temporal Interpolation

- Interpolated variable-length videos (2-10s) to exactly 50 time-steps via OpenCV.
- Final tensor shape: (Batch, 50, 409)

Nearest-Neighbor Sync

- Synchronized the text and video data by their timestamps, allowing for up to 2.5 seconds of delay.

Feature Standardization

- Audio/Video features normalized using Z-score: (X − Mean) / Std

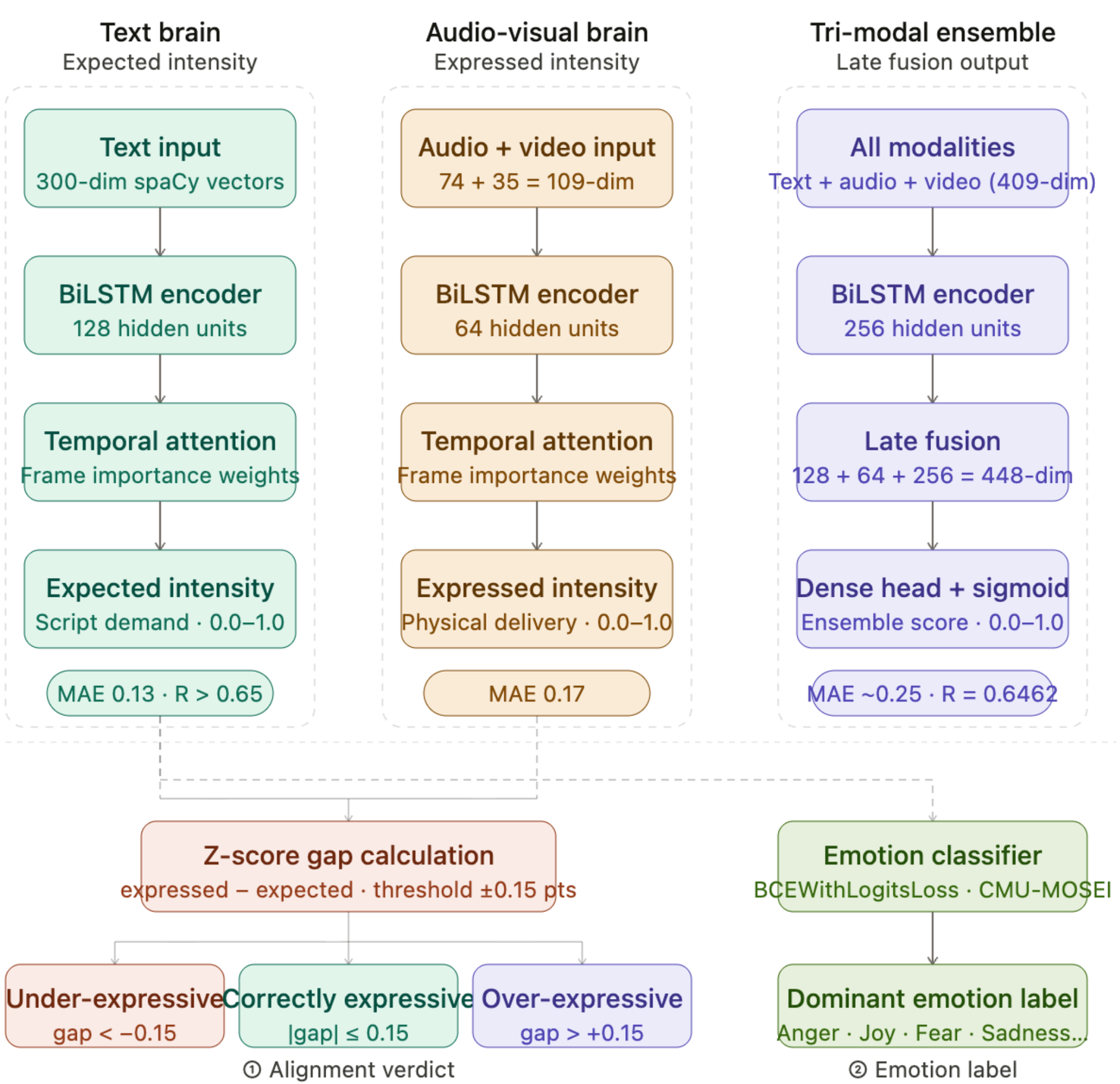
Train/Val/Test Split

- 80/10/10 stratified split ensuring balanced distribution of intensity labels.

ML Methodology

BiLSTM + Temporal Attention + Late Fusion

ML Pipeline


Text brain
Expected intensity
Text input
300-dim spaCy vectors
BiLSTM encoder
128 hidden units
Temporal attention
Frame importance weights
Expected intensity
Script demand · 0.0–1.0
MAE 0.13 · R > 0.65
Audio-visual brain
Expressed intensity
Audio + video input
74 + 35 = 109-dim
BiLSTM encoder
64 hidden units
Temporal attention
Frame importance weights
Expressed intensity
Physical delivery · 0.0–1.0
MAE 0.17
Tri-modal ensemble
Late fusion output
All modalities
Text + audio + video (409-dim)
BiLSTM encoder
256 hidden units
Late fusion
128 + 64 + 256 = 448-dim
Dense head + sigmoid
Ensemble score · 0.0–1.0
MAE ~0.25 · R = 0.6462
Z-score gap calculation
expressed – expected · threshold ±0.15 pts
Under-expressive
gap < −0.15
Correctly expressive
|gap| ≤ 0.15
Over-expressive
gap > +0.15
① Alignment verdict
Emotion classifier
BCEWithLogitsLoss · CMU-MOSEI
Dominant emotion label
Anger · Joy · Fear · Sadness...
② Emotion label

BiLSTM + Temporal Attention

Why BiLSTM?

- Emotion is sequential so the build-up matters.
- BiLSTMs read all 50 time-steps forwards AND backwards, maintaining contextual memory of the full sequence.
- This captures build-up, peak, and emotional resolution.

Why Temporal Attention?

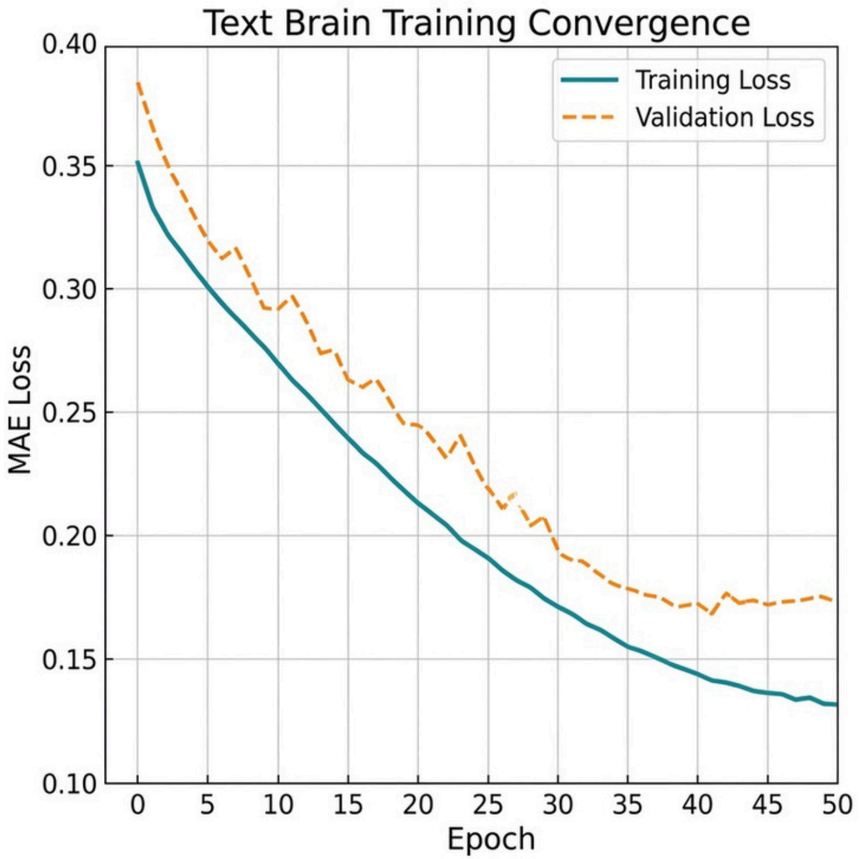
- 45 frames of silence + 5 frames of screaming!
- A plain LSTM would average the screaming out.
- Temporal Attention computes importance weights (0–1) for every frame, dynamically highlighting exactly where critical emotional behavior occurs.

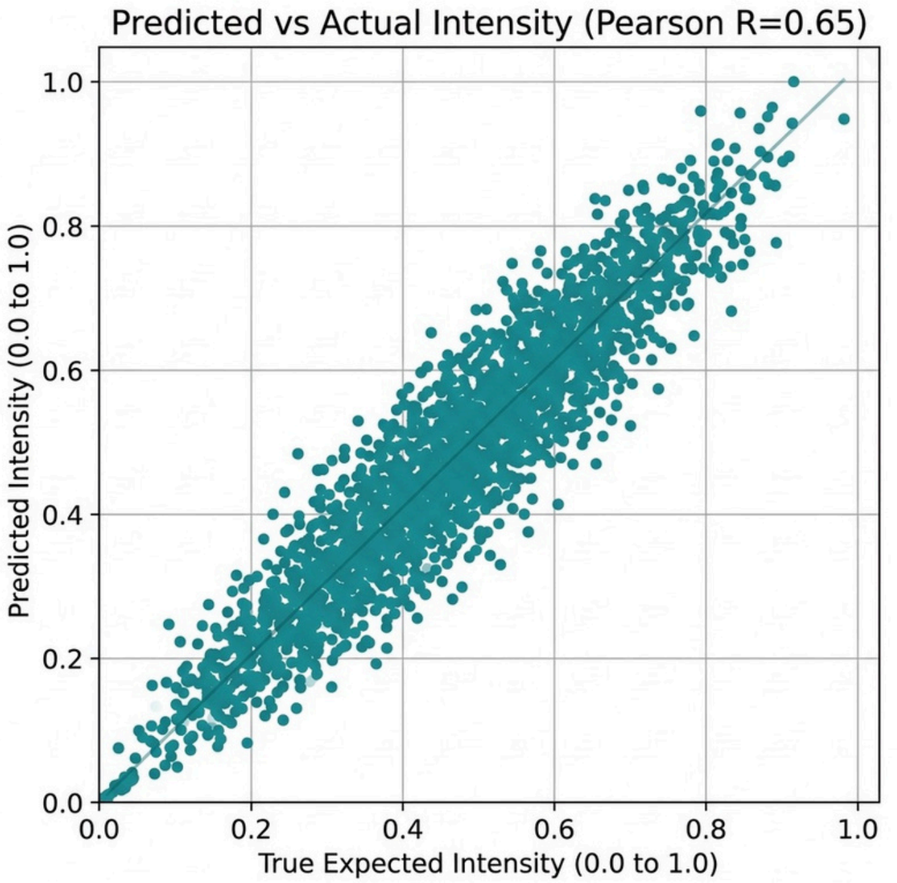
Amplifier Override Mechanism

- To minimize Mean Squared Error (MSE), neural networks naturally become cautious, constantly pulling extreme predictions (like a 0.95 screaming intensity) closer to the dataset's overall average.
- We built an override switch. If the raw data shows extreme physical action (like a massive spike in audio volume or a wide-open jaw), the system bypasses the neural network’s conservatism and forces an extreme emotional score.

Performance Metrics

MAE , Pearson R , Alignment Verdict



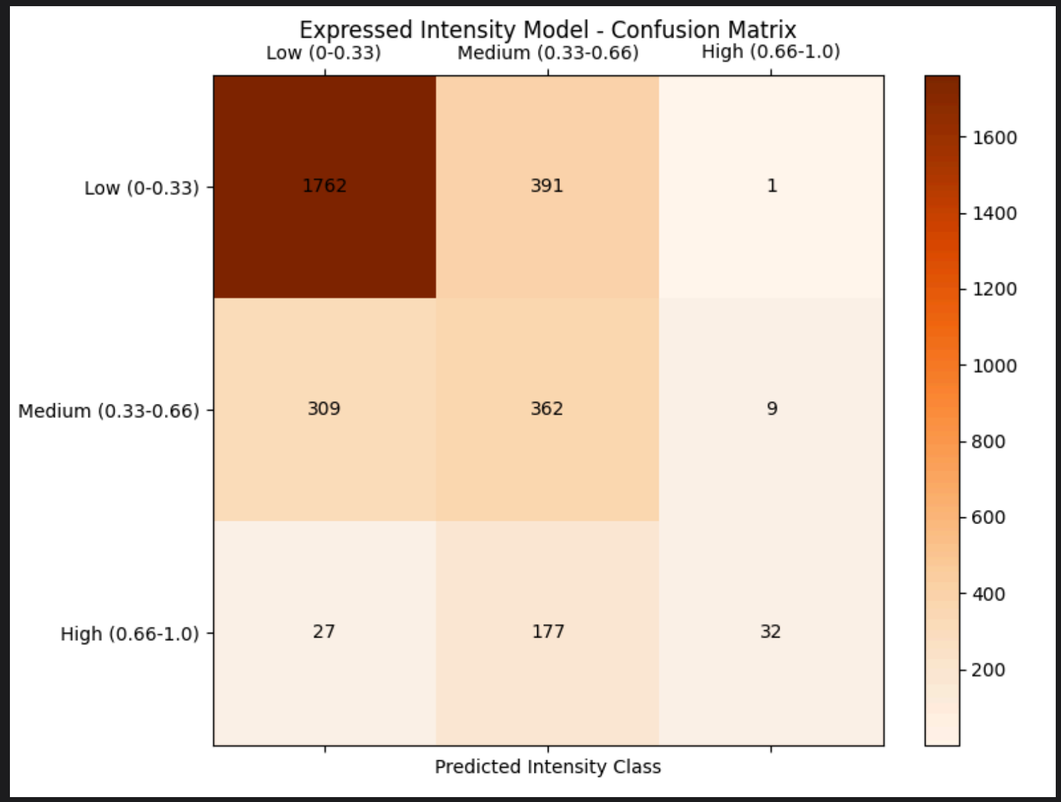


Text Brain (Expected Intensity)	AV Brain (Expressed Intensity)	Tri-Modal Ensemble (Complete Model)
Pearson R **> 0.65**	Pearson R —	Pearson R
MAE: 0.13	**MAE: 0.17**	**MAE: ~0.25**

Classification Report

	precision	recall	f1-score	support
Low (0-0.33)	0.84	0.82	0.83	2154
Medium (0.33-0.66)	0.39	0.53	0.45	680
High (0.66-1.0)	0.76	0.14	0.23	236
accuracy			0.70	3070
macro avg	0.66	0.50	0.50	3070
weighted avg	0.73	0.70	0.70	3070

Confussion Matrix



Alignment Verdict: Z-Score Calibration

```
z_expected = (i_expected – 0.5) / 0.2
z_expressed = (i_expressed – 0.5) / 0.2
alignment = z_expressed – z_expected
```

Networks squeeze predictions toward the mean (0.5). Z-score normalization re-scales both outputs to a comparable spread before computing the gap.

UNDER-EXPRESSIVE	CORRECTLY EXPRESSIVE	OVER-EXPRESSIVE
`alignment < –0.15`	`–0.15 ≤ alignment ≤ 0.15`	`alignment > 0.15`
The actor was too flat / monotone relative to the script's emotional gravity.	Perfect alignment between script expectation and physical execution.	The actor overacted means physical delivery far exceeded script demands.

Live Dashboard: User Interface

Alignment Verdict

UNDER-EXPRESSIVE / CORRECTLY EXPRESSIVE / OVER-EXPRESSIVE with explanation text

Emotion Label

Dominant emotion from j-hartmann/emotion-english-distilroberta-base (e.g. Surprise, Anger)

Intensity Alignment Gap

Bar chart: Expected Intensity (blue) vs Expressed Intensity (red) , clearly shows the numerical gap

Temporal Attention Graph

Line graph of attention weights across 50 frames, shows which frames the model focused on

Limitations

Deployability

Future Scope

Current Limitations

Hardware Variance

Microphone noise affect LibROSA feature extraction. LayerNorm mitigates but doesn't eliminate.

Lighting Sensitivity

MediaPipe FaceLandmarker drops to 0.0 feature values in heavily backlit or dark rooms, nullifying the visual branch.

Limited Expression Bias

CMU-MOSEI is primarily Western/English. Indian communicative nuances (head nods, cadence) may not map correctly — needs future fine-tuning on localized data.

Threshold Subjectivity

Z-score thresholds (±0.15) were manually calibrated. A learned classifier could make these adaptive per context.

Deployability at Plaksha

- Total model weights < 10MB (3 .pth files)
- Gradio UI
- Hosted live on HuggingFace Spaces (free)
- CPU/MPS inference , < 3 seconds per video
- Zero external API calls in production

Future Scope

- Fine-tune on Indian/multilingual dataset
- Add Whisper ASR for real-time transcript
- Learned adaptive threshold classifier
- Expand to multi-speaker dialogue scenes

Thank You

Manan Singla found that he could be a great actor.

You might be the next one!

Perform now at : https://huggingface.co/spaces/abhi-s/Multimodal_Emotion