
[SLIDE 1] Title **(0:00–0:30) (include waiting & setting up time)**

“Good afternoon, everyone. Today, I will present my senior project.
Our approach is modular. Flexible. Maintainable.
Let’s begin.”

[SLIDE 2] Purpose **(0:30–0:50)**

“Emotions are complex.
People might hide their feelings or pause before showing emotion.

Our goal is to Analyze emotions using **three channels**.
Facial expressions. Voice tone. Text transcription.

[SLIDE 3] Purpose **(0:50–1:10)**

Instead of building a

tightly-coupled

system,

We propose a **loosely coupled voting mechanism**.

Because real-world problems are not like Kaggle competitions.
It’s not just about accuracy.
We need systems that are flexible and easy to maintain.”

[SLIDE 4] **Traditional Approaches vs. Our Project (1:00 - 1:15)**

Traditional Single-channel systems struggle because They miss the big picture.

Multimodal system combine data too early. Which make it Hard to change. And Heavy to compute.

Our approach keeps channels separate.

We process them **independently** and combine results **after** processing.

This makes our system modular and Flexible.”

Slide 5: Workflow Overview (1:15–1:20)

Slides 6: Input Handling *(1:20 - 1:30)*

“Our system starts with a short video—about 30 seconds.
We split the video into:

Audio stream.

&

Video stream.

Slides 7: Separated Emotional Analysis (model inference) *(1:30 - 1:40)*

Then, we analyze it across three independent channels
using pre-trained models.”

Slides 8: Voice Transcription *(1:40 - 2:10)*

“For transcription, we use two models:

First OpenAI Whisper to Converts audio to text.

then DistilRoBERTa: Analyzes text and detects emotions.

Slides 9: Voice Tone: *(2:10 - 2:30)*

"For voice tone, we use **Wav2Vec 2.0(“o”) Emotion Recognition**.

It detects tone variations, like Pitch, rhythm(ริ ทึ่ม), and intensity.

The model returns emotions every 5 seconds.

Slides 10: Facial Expression: (output) *(2:30 - 2:50)*

The facial expression can switch between two models:

OpenFace and **FaceTorch**.

OpenFace analyzes dynamic movements over time,

like subtle eyebrow raises

or lip changes.

On the other hand, FaceTorch works with individual static frames.

To use FaceTorch,

we extracted one frame per second from the video and ran the analysis frame by frame."

"As you can see here on the screen, this is the output from both models"

Slide 11: Preprocessing & Syncing *(2:50 - 3:00)*

“Now, we prepare the outputs.

Slide 12: Voice Tone & Voice Transcription *(3:00 - 3:10)*

“For **voice tone** and **transcription**,
we map their raw outputs to the same emotion labels

This makes sure both channels speak the same “language” for emotions. (pause).

Slide 13: Facial Expression: *(3:10 - 3:30)*

Now, let’s focus on **facial expression**, where we handle two different models.”

“For **FaceTorch**, which works on static frames,
we align outputs at **1-second intervals**.

But For **OpenFace**, which is continuous—dynamic predictions over time.
We need to group these into **intervals**.

This means both outputs, whether static or dynamic,
are transformed into the same format:
{Time, Emotion, Confidence} (pause).

Slide 14: Syncing *(3:30 - 4:00)*

“The syncing step uses the **transcription timestamps** as the baseline.

We align outputs from all three channels like this picture

[SLIDE 15] Architecture: Voting & Inter-Chunk Processing (4:00 - 4:10)

“Once synced, we use a **voting mechanism** to combine results.

Each channel has a weight based on its reliability:

- **Text**: 50% (because it’s Most reliable for context). [[sacarsm only]]
- **Facial Expression**: 30%. [It can be dlsturbed by light reflection]]
- **Voice Tone**: 20% bc it’s very Sensitive to noise. [[noise and more]]

This balances accuracy and reliability. For final decision.”

[SLIDE 16] Per-Chunk Voting – Weight Adjustment (4:10 - 4:30)

“What happens when channels disagree?

We calculate a **conflict ratio**.

More conflict → Lower weight for that channel.
This reduces the impact of conflicting outputs.

[SLIDE 17] Per-Chunk Voting – Weight Adjustment (4:30 - 4:50)

For example:

If Voice Tone detects Happy, Sad, and Fear:

1 Positive, 2 Negative → Conflict ratio = 0.33.

We adjust the weight to make the prediction stable.”

[SLIDE 18] Inter-Chunk Processing (4:50 - 5:00)

“Emotions don’t jump suddenly in real life.
We smooth transitions across time chunks.

For example:
If emotions jump quickly between Happy → Sad → Happy,
We replace sudden changes with Neutral.

This creates a realistic emotional flow."

Slide 19: Key Challenges and Advantages (5:00 - 5:50)

"Now let's talk about the key advantages of our system.

[SLIDE 20] . Feasibility:

Most state-of-the-art single-channel models are already multi-modal.

Further Combining them into tightly coupled multi-channel systems is too complex.

[SLIDE 21] . Accessibility:

State-of-the-art models like Google API are black boxes.

We only use outputs—no need for model access.

[SLIDE 22] . Flexibility:

We can switch models easily.

For example, OpenFace and FaceTorch worked seamlessly in our system like we show earlier

[SLIDE 23]. Multi-Channel Fusion:

Better than single-channel systems.

And while we may not beat tightly-coupled multimodal systems,

We trade accuracy for flexibility—essential for real-world use.

[SLIDE 24]. Transparency:

Unlike black-box systems, our approach is clear.

We show how each channel contributes.

This builds trust in critical fields like healthcare

Slide 25: Future Work (5:50–6:00)

"To improve the system,
we'll focus on detecting **conflict emotions** like sarcasm or hidden sadness with the condition-based weight-adjusting

Thank you! I’m ready for your questions.”