Question 3: Novel Research Challenge in Speech Understanding

Lt Col Ritesh Lamba
MTech (AI) Executive, Indian Institute of Technology, Jodhpur
Roll No: M23CSA544
M23CSA544@iitj.ac.in

Multimodal Code-Switching Speech Understanding in Low-Resource Environments

1. Problem Statement & Significance

While working on speech processing tasks, one challenge that stood out to me was how poorly models handle code-switched speech — especially in real-world scenarios like Hindi-English conversations. This problem becomes even harder when the environment is noisy and there's little labeled training data available. It made me think: can we do better if we also look at the speaker's visual cues, like lip movements?

Our current models largely treat speech as audio-only. But in many real situations — video calls, interviews, mobile assistants — the speaker's face is visible. We're not leveraging that at all. This leads me to propose a speech understanding system that can handle multilingual, code-switched speech using both audio and visual inputs.

Why it matters: Solving this could make voice assistants more accessible for rural and multilingual users, especially those who naturally switch between languages. It would also open up more inclusive technology for education, public service and entertainment.

2. Proposed Algorithm & Methodology

What I propose: A model that combines what the user says with how their lips move while speaking — even if they're switching languages mid-sentence.

Components:

- **Audio Encoder:** Something like Wav2Vec 2.0 or Whisper to convert raw audio into embeddings.
- **Visual Encoder:** A lightweight video transformer (like TimeSformer or LipFormer) to extract lip movement patterns.
- **Fusion Module:** A transformer that fuses both modalities and predicts text with language tags — e.g., "[EN] I am going to [HI] bazaar."

Training Plan:

- Start with contrastive pretraining using YouTube clips with people speaking and visible faces.
- Fine-tune using smaller, annotated datasets like Hindi-English conversations.
- Use a combined loss that handles both transcription and language tagging.

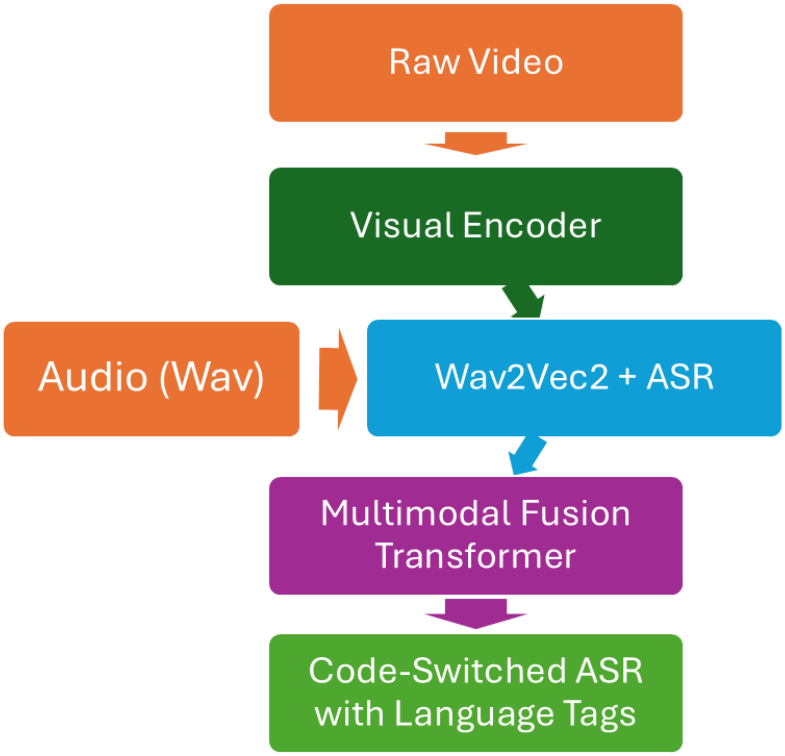
Raw Video
Visual Encoder
Audio (Wav)
Wav2Vec2 + ASR
Multimodal Fusion Transformer
Code-Switched ASR with Language Tags

Figure 1: Proposed Multimodal ASR Architecture

3. Evaluation Strategy

Datasets:

- **MuST-C-VIS:** TED talks with video and multilingual audio.
- **CMU-MOSEI:** Speech and facial emotion data.
- **Indic Code-Switch Corpus:** Annotated Hindi-English speech.
- **YouTube clips:** Using VAD and subtitles to auto-align speech and video.

Metrics:

- WER (Word Error Rate) per language

- CER (Code-switching Error Rate)
- BLEU (if we extend to translation later)
- MOS (Mean Opinion Score) to judge real-world intelligibility

Testing Plan:

- Compare against Whisper and lip-reading-only baselines.
- Evaluate under poor audio conditions and partial face occlusions.

4. Broader Implications

If we solve this, it could really shift how speech understanding systems are built — from being purely audio-based to leveraging vision too. It can improve:

- **Scientific research:** by showing how audio-visual fusion helps code-switching.
- **Commercial tools:** like multilingual smart assistants in AR devices or video conferencing tools.
- **Social impact:** by giving underserved populations a voice interface that actually understands them.

Final Thought: This idea excites me because it blends multiple modalities, addresses a real problem in India and beyond, and could set the stage for the next generation of inclusive speech technology.