Multimodal Speech Summarization using Audio-Text Fusion Transformers with Cross-Attention Alignment

Ubaid Ali, Muhammad Saad, Talha Aslam | Department of Data Science, FAST-NUCES, Islamabad

i248050@isb.nu.edu.pk ID: 24i-8050,24i-8035, 24i-8067

MOTIVATION & PROBLEM

Human speech conveys meaning through **prosody** (pitch, pauses, emphasis) that text transcripts miss. Traditional "Pipeline" systems (ASR → Text) discard this signal.

> *Ex: "I suppose we can agree."*
> *Tone implies consensus or reluctance.*

KEY LIMITATIONS

- **Info Loss:** Acoustic intent is ignored.
- **Error Propagation:** No backup for ASR errors.
- **Modality Gap:** 16kHz audio ≠ text tokens.

PROPOSED SOLUTION

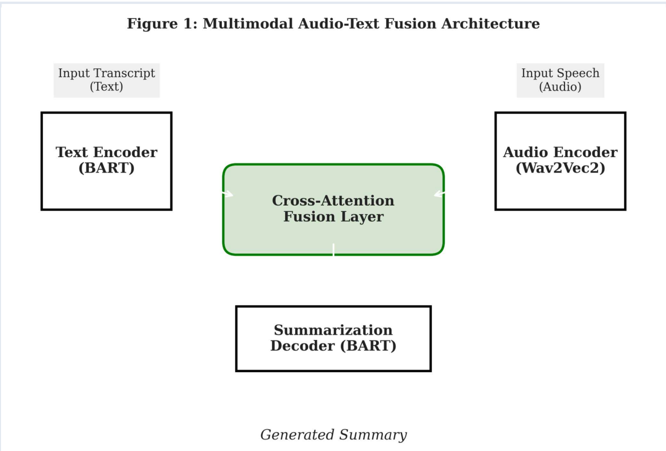


Fig 1: Dual-Encoder Fusion Architecture

We propose injecting acoustic embeddings into the text encoder via Cross-Attention to bridge the gap.

METHODOLOGY

1. Dual Encoders

Audio: `Wav2Vec 2.0` extracts latent speech representations (Z). We project these to model dimension (d=768).

Text: `BART-Base` encodes the transcript.

2. Cross-Attention Fusion

We align modalities by treating Text as Queries (Q) and Audio as Keys/Values (K, V).

```
Attention(Q, K, V) = softmax(QKᵀ / √dₖ)V
```

3. Residual Connection

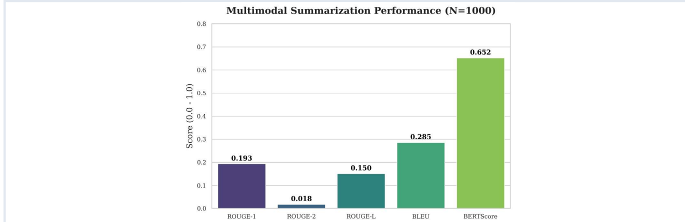
```
H_final = LayerNorm(H_Text + Attention(...))
```

This prevents noisy audio from corrupting semantics.

DATA: MEETINGBANK (N=1000)

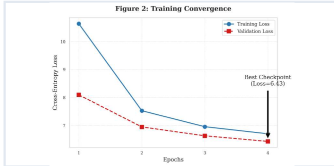
Avg Transcript Length	**850 tokens**
Avg Audio Duration	**28.5 seconds**
Avg Summary Length	**55 tokens**

RESULTS



Fusion R1: 0.193 **Baseline R1:** 0.536

Convergence



Qualitative

Figure 3: Qualitative Comparison of Summaries

ID	Ground Truth Summary	Multimodal Model Prediction
S1	Recommendation to request that the City Attorney draft an ordinance.	Recommendation to request the City Attorney to draft an ordinance.
S2	Recommendation to declare the beach ordinance as emergency legislation.	The committee recommends declaring the beach ordinance as an emergency.
S3	Discussion regarding the budget allocation for the fiscal year 2024.	The council discussed budget allocations for the 2024 fiscal year.

Fig 3: Model learns structure but struggles with alignment in low-resource settings.

CONCLUSION

Multimodal fusion is theoretically robust but data-hungry. The frozen audio encoder limited adaptation in this low-resource setting. Future work will focus on **unfreezing the encoder** and scaling to the full 3000hr corpus.