Multi-modal Speech Transformer Decoders: When Do Multiple Modalities Improve Accuracy?

Yiwen Guan (yguan2@wpi.edu), Viet Anh Trinh, Vivek Voleti, and Jacob Whitehill

Department of Computer Science, Worcester Polytechnic Institute, MA 01609, USA

ABSTRACT

We investigate how different input modalities (audio, image, lip movements, etc.) impact speech recognition performance in decoder-only discrete-token language models:

- Does adding more context and more modalities increase accuracy?
- Does the accuracy benefit depends on the audio noise?
- Do *synchronized* vs. *unsynchronized* modalities behave differently?

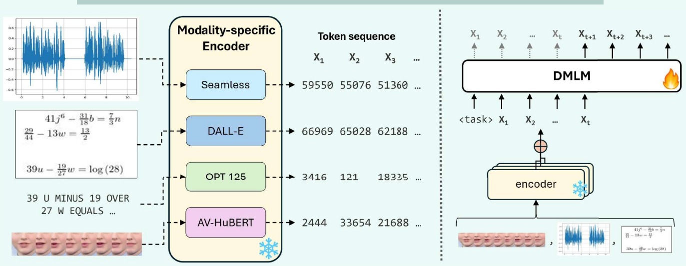
Key results:

✓ **Integrating more modalities can increase accuracy** but the benefit depends on the amount of audio noise.

✓ **Image context provides its greatest benefit at *moderate* audio noise levels**; moreover, it exhibits a different trend compared to inherently synchronized modalities like lip movements.

✓ **Filtering the most relevant visual information improves accuracy** on both synthetic (3-Equations) and real-world datasets (SlideAVSR).

ARCHITECTURE: Discrete Multi-modal Language Model (DMLM)



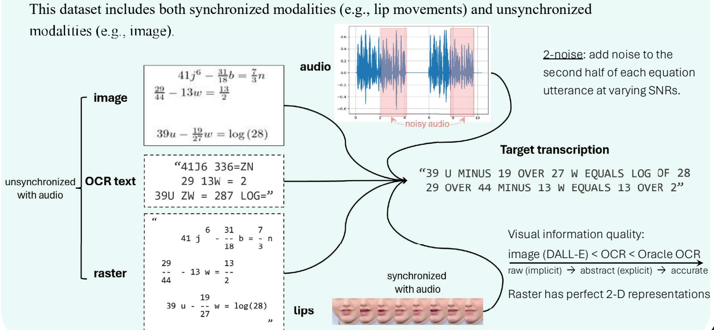
Pipeline:

1. Tokenize input data with frozen modality specific encoders into discrete token sequence.
2. Concatenate the sequence to a task description (e.g., "[ASR]")
3. Pass the entire sequence to the DMLM.

3-EQUATIONS DATASET

We create a dataset (3-Equations) on which we can control its characteristics precisely, and simulate anticipated situations, such as with different modality SNRs. (N=10,000)

This dataset includes both synchronized modalities (e.g., lip movements) and unsynchronized modalities (e.g., image).



EXPERIMENTS

Evaluation metrics:

- **Speech recognition:** Word Error Rate (WER)
- **Impact of additional modalities:** We introduce a new metric for evaluating the impact of additional modalities in ASR: Relative WER Benefit (↑) = $(WER_A - WER_{X+A})/WER$

  A: audio, X: additional modalities added to audio-only model

This metric evaluates how much the WER is reduced relatively when additional modalities are incorporated.

Methodology:

We compute WER for all combinations of modalities (I+A→T, O+A→T, etc.) for each noise level in SNR=(+∞,20,10,5,0,-5,-10,-20,-∞)dB, and evaluate the relative WER benefit (%) for each combination.

Results:

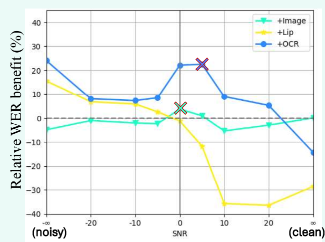


Figure 1: Relative WER benefit (%) of adding one more modality (I+A, L+A, and O+A)

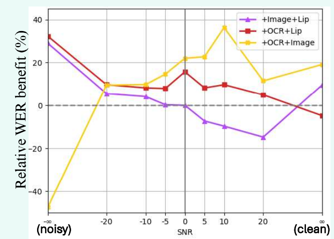


Figure 2: Relative WER benefit (%) of adding two more modalities (I+L+A, O+L+A, and O+I+A)

❑ Experiment I: The benefit of adding additional modalities to audio-only model (Fig. 1 & Fig. 2 & Fig. 4):

- ➢ **Integrating more modalities can improve ASR accuracy;**
- ➢ **OCR modality provide the best complementary benefit;**
- ➢ **In general, 3-modality models works better than 2-modality models.**

❑ Experiment II: Trend of benefit provided by each modality across noise levels (Fig.1 & Fig. 4):

- ➢ **Unsynchronized modalities (image and OCR) provide the greatest benefit at medium SNRs (0dB~10dB);**
- ➢ **Synchronized modality (lip movements) has larger benefit when there's more noise.**

❑ Experiment III: The benefit of adding different (implicit/explicit) visual modalities (Fig. 3):

- ➢ **Better visual representation can lead to better supplementary benefit;**
- ➢ **Oracle OCR, which is the most abstract and accurate visual information, provides the greatest overall benefit.**
- ➢ **The model has difficulties with 2-D representations, even it's as perfect as raster representation.**

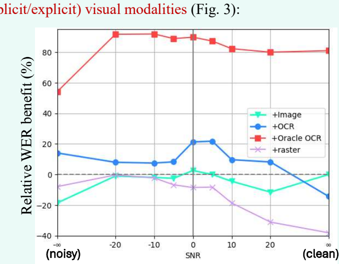
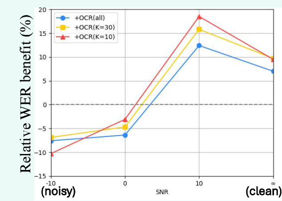


Figure 3: Relative WER benefit (%) of adding different visual modalities (I, O, Oracle OCR, and R)

Notations: audio (A), image (I), OCR (O), lip (L), raster (R)

SlideAVSR[2]: An audio-visual dataset of AI paper explanation videos, which include transcribed speech, synchronized slides, and OCR keywords.



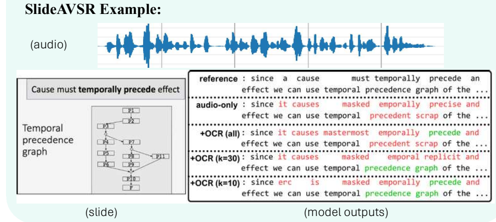
K value: top K words ranked by frequency of word occurrences in English Wikipedia (FQRanker)[2]. Only these K words will be kept as input.

Figure 4: Relative benefit (%) of adding filtered OCR with K values on SlideAVSR

❑ Experiment IV: Impact of irrelevant information (Fig. 4):

- ➢ **Irrelevant information may hurt the performance, filtering relevant or long-tail words can help. Model with OCR (K=10) has the best overall benefit on SlideAVSR.**

SlideAVSR Example:



CONCLUSIONS

✓ Fusing additional modalities enhances speech recognition performance.

✓ In different noise levels, unsynchronized modalities (image and OCR) exhibit a different trend from synchronized modalities (lip movements)

✓ Filtering relevant visual information enhances performance.

✓ More abstract and accurate visual modality improves accuracy more with supplementary visual information.

✓ Our work is the first to show the benefit of combining audio, image, and lip movements in one model for speech recognition.

FUTURE WORK

- Extend to other backbone language models of other architecture.
- Exploration of different visual encoders and visual modalities.
- Try other input (prompt) strategies like interleaving.
- Extend our findings to other real-world datasets.

Acknowledgement

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL #2019805, and also from an NSF CAREER grant #2046505. The opinions expressed are those of the authors and do not represent views of the NSF.

https://arxiv.org/pdf/2409.09221

References

[1] V. A. Trinh, R. Southwell, Y. Guan, X. He, Z. Wang, and J. Whitehill, "Discrete multimodal transformers with a pretrained large language model for mixed-supervision speech processing," arXiv preprint arXiv:2406.06582, 2024.

[2] H. Wang, S. Kurita, S. Shimizu, and D. Kawahara, "Slideavsr: A dataset of paper explanation videos for audio-visual speech recognition," arXiv preprint arXiv:2401.09759, 2024.