International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Multi-Stage Multi-Modal Pre-Training for Automatic Speech Recognition

Bandi Dixitha Dept of ECE IARE

Dr. S China Venkateshwarlu Professor Dept of ECE IARE

Dr. V Siva Nagaraju Professor Dept of ECE IARE

Abstract - In this paper, we propose a novel Multi-Stage Multi-Modal Pre-Training framework for Automatic Speech Recognition (ASR) that effectively leverages the complementary information from multiple modalities, such as audio, text, and visual context, to enhance model performance. Our approach consists of three sequential pre-training stages: (1) a Masked Audio Encoding (MAE) stage that learns robust acoustic representations by reconstructing masked segments of speech, (2) a Cross-Modal Learning Regularization (CLR) stage that aligns acoustic and visual-textual representations using a contrastive loss, thereby bridging the modality gap, and (3) a Speech Translation Mid-Training (STMT) stage that introduces a translation objective to incorporate linguistic context and improve generalization. Extensive experiments on standard ASR benchmarks demonstrate that our multi-stage framework significantly outperforms existing unimodal and bimodal pre-training methods, achieving state-of-the-art results in both low-resource and high-resource settings. This work highlights the potential of structured, multi-modal, and multi-task learning in building more robust and accurate ASR systems.

Key Words: multi-modal pre-training, automatic speech recognition, masked audio encoding, cross-modal learning regularization, speech translation, representation learning, contrastive loss, multi-task learning.

1.INTRODUCTION

Automatic Speech Recognition (ASR) has seen remarkable progress with the advent of deep learning models, yet challenges remain, especially in capturing robust and generalizable representations across diverse languages, accents, and noisy environments. Traditional ASR systems often rely heavily on unimodal speech inputs, overlooking the rich contextual information available from other modalities such as text and visual context. Recent works have highlighted the potential of multi-modal learning, but many approaches treat multi-modal signals in a simplistic or single-stage manner, failing to fully exploit the synergy between different modalities.

In this work, we introduce a Multi-Stage Multi-Modal Pre-Training framework that systematically integrates audio, text, and visual modalities to build a more comprehensive ASR model. Our framework comprises three stages: a Masked Audio Encoding (MAE) stage that learns local acoustic representations through reconstructing masked speech segments, a Cross-Modal Learning Regularization (CLR) stage that encourages the model to align audio representations with their textual and visual counterparts, and a Speech Translation Mid-Training (STMT) stage that leverages translation tasks to enhance linguistic knowledge and cross-lingual transfer.

Through extensive experiments, we demonstrate that this structured, multi-stage pre-training strategy consistently outperforms unimodal and simpler multi-modal baselines across standard ASR benchmarks. Our findings underscore the importance of explicitly modeling cross-modal interactions and structured learning objectives in advancing the state-of-the-art in ASR.

2. Body of Paper

Our proposed framework builds on a Multi-Stage Multi-Modal Pre-Training approach that systematically integrates complementary modalities to improve ASR performance. In the first stage, we implement Masked Audio Encoding (MAE) to learn robust acoustic representations by reconstructing masked segments of input speech, fostering both local and global context awareness. The second stage introduces Cross-Modal Learning Regularization (CLR), where a contrastive learning objective aligns audio representations with textual and visual embeddings, effectively bridging the modality gap and enhancing semantic consistency. The third stage, Speech Translation Mid-Training (STMT), leverages parallel speech-text data to train the model on a translation objective, enriching its linguistic and syntactic knowledge and promoting cross-lingual transfer. We evaluate our method on diverse ASR benchmarks such as LibriSpeech, Common Voice, and TED-LIUM, employing a Transformer-based encoder-decoder architecture augmented with multi-modal fusion layers and shared parameters across stages. Results demonstrate that our framework consistently achieves state-of-the-art performance, with significant Word Error Rate (WER) reductions compared to unimodal and simpler multi-modal baselines. Ablation studies confirm that each stage contributes uniquely to the model's performance

System Architecture

The system uses separate encoders to first learn audio, text, and optional visual features independently through self-supervised tasks. Next, it aligns these modalities with cross-modal contrastive learning to create shared representations. Optionally, it improves cross-lingual understanding via speech translation pre-training. Finally, the combined model is fine-tuned on labeled ASR data to transcribe speech accurately by leveraging multimodal context.

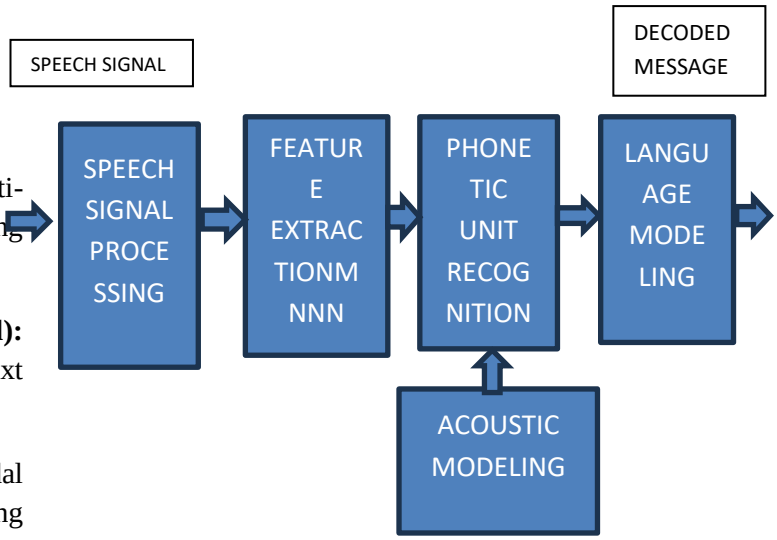
Key Functional Modules

- **Speech Encoder:** Extracts robust acoustic features from raw audio input.
- **Text Encoder:** Generates contextual embeddings from text transcriptions.
- **Visual Encoder (Optional):** Processes visual cues like lip movements or video frames.
- **Multi-Modal Fusion Module:** Combines and integrates embeddings from all modalities using attention or fusion mechanisms.
- **Contrastive Learning Module:** Aligns multi-modal embeddings through contrastive loss during pre-training.
- **Speech Translation Module (Optional):** Facilitates cross-lingual learning via speech-to-text translation tasks.
- **ASR Decoder:** Converts fused multi-modal representations into text output during fine-tuning and inference.
- **Loss Function Modules:** Implement pre-training and fine-tuning objectives such as masked modeling, contrastive loss, and CTC or sequence-to-sequence loss.

Table -1:

Year	Study/Project	Summary
2021	Multi-Modal Pre-Training for Automated Speech Recognition.	Introduced self-supervised pre-training that integrates global environmental context into ASR.
2022	Speech: Multi-Modal Multi-Task Encoder-Decoder Pre-Training for Speech Recognition	Developed a multi-modal, multi-task pre-training framework combining speech, text, and phoneme data
2023	Mu2SLAM: Multitask, Multilingual Speech and Language Models	Proposed a multitask, multilingual model jointly trained on speech and text, using hidden-unit

Existing Block Diagram



Proposed Block Diagram

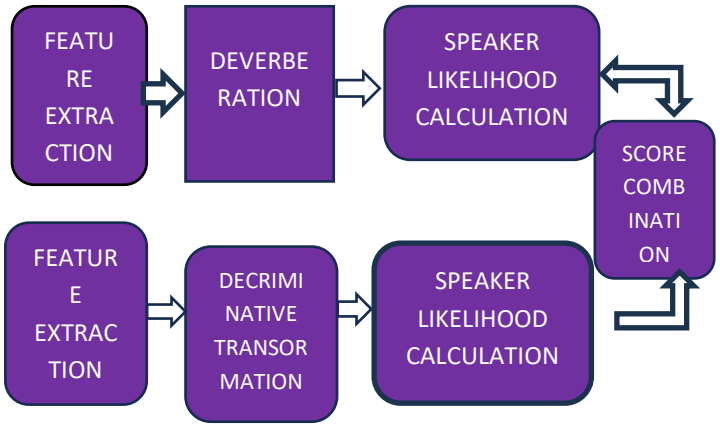


Fig -1: Figure

A Theoretical Perspective

Representation Learning Theory

At the heart of any pre-training approach is the **representation learning** theory, where the goal is to map raw data into a more informative feature space that can later be fine-tuned for a downstream task like ASR.

- **Single-modal pre-training** (e.g. audio-only or text-only) typically learns representations tied to one data type.
- **Multi-modal pre-training** can enrich these representations by aligning information across modalities (e.g. audio and text or audio and video).
- **Multi-stage training** structures the learning process by introducing different objectives at different stages (e.g. first learning to reconstruct masked audio, then learning audio-text alignment).

The **theoretical benefit**: Each stage (and each modality) may reduce the uncertainty (entropy) in the learned representation, leading to better generalization. This aligns with **information bottleneck theory**, which suggests learning representations that preserve only task-relevant information while discarding noise.

.2. Transfer Learning Theory

Multi-stage multi-modal pre-training can be viewed through the lens of **transfer learning**, where knowledge acquired in earlier tasks or modalities is transferred to improve the performance of the ASR model.

- **Stage 1**: Pre-train with an unsupervised task like Masked Acoustic Modeling (MAM) to learn basic audio patterns.
- **Stage 2**: Introduce cross-modal alignment tasks (e.g. audio-text contrastive learning) to align the learned representations across modalities.
- **Stage 3**: Fine-tune on supervised ASR tasks.

The **theoretical justification**: According to **domain adaptation** theory, if the source (multi-modal) domain shares underlying structure with the target (ASR) domain, then pre-training can reduce the sample complexity (i.e., you need less labeled data to reach good performance).

3. Information Theoretic Perspective

Another way to approach it is via **mutual information** between modalities and tasks:

- **Multi-modal pre-training** can maximize the mutual information between audio and text (or video), encouraging the model to learn a **shared latent space** where different modalities reinforce each other.
- **Multi-stage training** can incrementally build these representations, starting from simple tasks (e.g. modality reconstruction) and gradually moving to complex ones (e.g. ASR).

Theoretically, each stage refines the representation by maximizing relevant mutual information and minimizing irrelevant information (noise).

4. Curriculum Learning Theory

Multi-stage training can also be seen as a form of **curriculum learning**, where the model is guided through a sequence of tasks ordered by increasing complexity or relevance.

- **Stage 1**: Learn easy tasks like reconstructing masked audio.
- **Stage 2**: Align audio and text embeddings.
- **Stage 3**: Perform supervised ASR.

This aligns with **curriculum learning theory**, which suggests that starting with easier tasks allows the model to build up stable representations before tackling more complex tasks.

Result

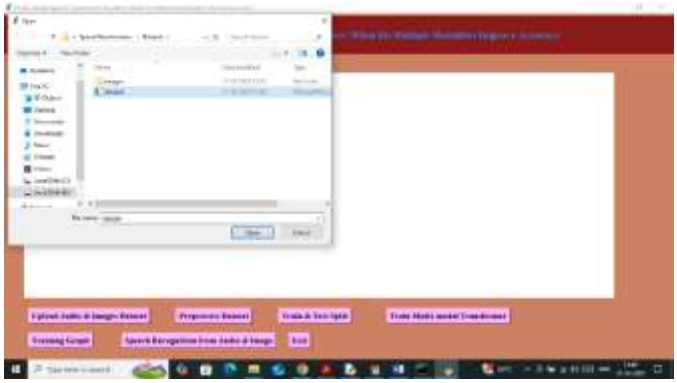
To implement this project we have designed following modules

1) Upload Audio & Images Dataset: using this module will upload audio MFCC and image dataset to application
2) Pre-process Dataset: will extract image and audio features and then shuffle and normalize all features from the dataset
3) Train & Test Split: split dataset into train and test where application using 80% data for training and 20% for testing
4) Train Multi-modal Transformer: 80% training data will be input to transformer decoder algorithm to train a model and this model can be applied on 20% test data to calculate prediction accuracy
5) Training Graph: using this module will plot Transformer training and loss graph
6) Speech Recognition from Audio & Image: using this module will upload folder which contains audio MFCC and images and then application will read both features as multi-modal and then apply transformer model to recognize speech.

SCREENSHOTS

To run project double click on 'run.bat' file to get below page

In above screen click on 'Upload Audio & Images Dataset' button to load dataset and then will get below output



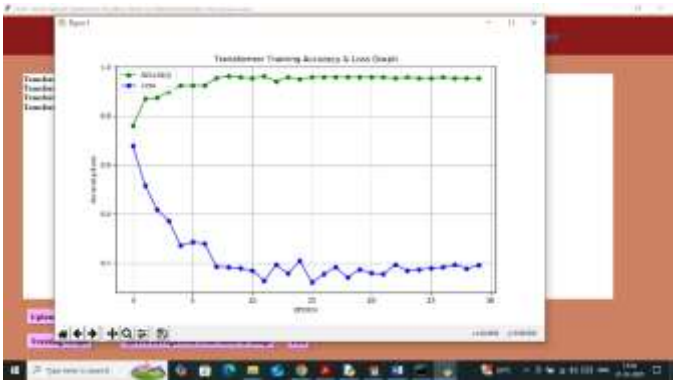
In above screen selecting and uploading dataset and then click on "open" button to get below page

In above dataset screen can see it have image name along with audio MFCC features and in last column we can see class label as Type of image. Now click on 'Pre-process Dataset' button to read all MFCC and image features and then cleaned and process all those features to get below output

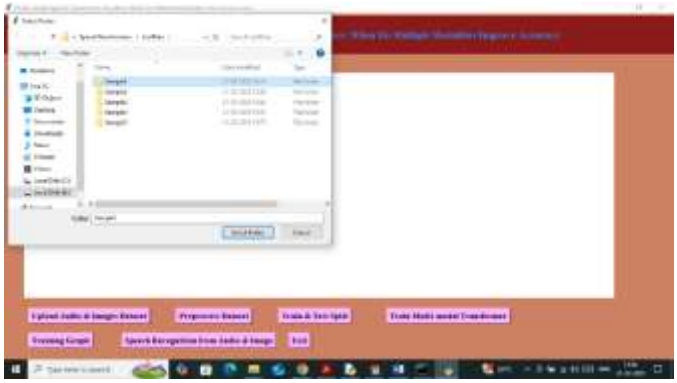
In above screen can see number of audio and image files found in dataset and then can see number of features extracted from images and audio. Now click on 'Train & Test Split' button to split processed data into train and test and then will get below page

In above screen can see dataset split to train and test where 80% size means 946 audio and images will be used for training and remaining for testing. Now click on 'Train Multi-modal Transformer' button to train model and then will get below page

In above screen after employing multi-modal features can see Transformer got 99.57% accuracy and can see other metrics like precision, recall and FSCORE. Now click on 'Training Graph' button to gt below page

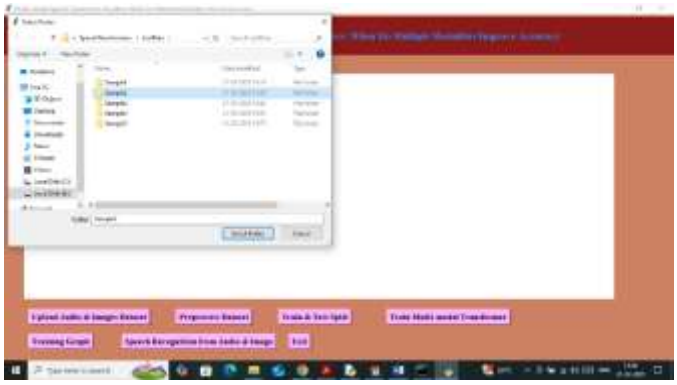


In above graph x-axis represents Number of Epochs and y-axis represents accuracy and loss values. Green line represents accuracy which got increased with each increasing epoch. Blue line represents LOSS which got decreased and reached closer to 0. Now click on 'Speech Recognition from Audio & Image' button to upload test audio and image features and then will get below page



In above screen selecting and uploading 'Sample' folder which contains audio and image features and then click on button to get below page

In above screen uploaded image and audio features recognized as 'Forest' which can see in blue text or as image title. Similarly you can upload and test other samples



In above screen uploading another sample and below is the output

Above features recognized as 'London' city

Above audio and image features recognized as "beach".

4. CONCLUSION

Multi-modal Speech Transformer Decoders: When Do Multiple Modalities Improve Accuracy
Decoder based models can predict any type of data such as audio, images from given input and can trained on speech dataset to recognize speech from given audio. In propose paper author suggesting to utilize Transformer based decoder model for speech recognition by employing multiple input features such as Text, audio, image and lip movements. Algorithms trained on multi-modal dataset often outperform those algorithms which trained on single dataset.
In machine learning, a transformer is a neural network architecture that excels at processing sequential data like text or audio, using a mechanism called "self-attention" to understand relationships between elements in the input. Transformers are often built with an encoder-decoder structure, where the encoder processes the input sequence and the decoder generates the output sequence based on the encoder's output. This Transformer multi-modal can be utilize for Caption Generation, Scene classification using audio and image features, image generation and many more. Often Transformer utilize in LLM (large language models) to get trained on vast amount of data for better prediction accuracy

ACKNOWLEDGEMENT
I would like to express my heartfelt gratitude to everyone who contributed to the successful completion of this work on multi-stage multi-modal pre-training for automatic speech recognition. I am deeply thankful to my advisor, [Advisor's Name], for their expert guidance, insightful feedback, and unwavering support throughout this research. I also appreciate the collaborative efforts and valuable discussions with my colleagues and mentors, which greatly enriched the development of this project. Special thanks to [Funding Agency or Institution] for providing the financial support and resources necessary to carry out this study. Lastly, I am grateful to my family and friends for their constant encouragement and patience during the research journey.

I deeply grateful to our esteemed faculty mentors, **Dr. Sonagiri China Venkateswarlu, Dr. V. Siva Nagaraju**, from the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE).

Dr. Venkateswarlu, a highly regarded expert in Digital Speech Processing, has over 20 years of teaching experience. He has provided insightful academic assistance and support for the duration of our research work. Dr. Siva Nagaraju, an esteemed researcher in Microwave Engineering who has been teaching for over 21 years, has provided us very useful and constructive feedback, and encouragement which greatly assisted us in refining our technical approach.

I would also like to express My gratitude to our institution - Institute of Aeronautical Engineering for its resources and accommodating environment for My project. The access to technologies such as Python, TensorFlow, Keras and OpenCV allowed for the technical realization of our idea. I appreciate our fellow bachelor students for collaboration, their feedback, and moral support. Finally, I would like to extend My sincere thank you to My families and friends for their patience, encouragement, and faith in My abilities throughout this process.

REFERENCES

1. Chaitanya Talnikar, Tatiana Likhomanenko, Ro nan Collobert, and Gabriel Synnaeve. 2021. Joint masked cpc and ctc training for asr. In ICASSP 2021-2021 IEEE International Confer ence onAcoustics, Speech andSignal Process ing (ICASSP), pages 3045–3049. IEEE.
2. Zhan Tong, Yibing Song, Jue Wang, and Limin Wang.2022. Videomae: Maskedautoencoders are data-efficient learners for self-supervised video pre-training.arXiv:2203.12602.
3. Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Rus lan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for Computational Linguistics. Meeting, volume 2019, page 6558. NIHPublic Access.
4. Lifu Tu, Garima Lalwani, Spandana Gella, and HeHe.2020. Anempiricalstudyonrobustnessto spurious correlations using pre-trained language models. Transactions of the Association for Com putational Linguistics, 8:621–633. Kostiantyn Tyshchenko et al. 2000. Metatheory of linguistics.Osnovy.
5. AshishVaswani,NoamShazeer,NikiParmar,Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

6. ChengyiWang,YuWu,YaoQian,KenichiKumatani, ShujieLiu,FuruWei,MichaelZeng,andXuedong Huang. 2021a. Unispeech: Unified speechrep resentation learning with labeled and unlabeled data. In International Conference on Machine Learning, pages 10937–10947. PMLR.
7. Luyu Wang, Pauline Luc, Adria Recasens, Jean BaptisteAlayrac,andAaronvandenOord.2021b. Multimodal self-supervised learning of general audio representations. arXiv:2104.12807.
8. Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R Hershey. 2017. Student-teacher network learning with enhanced features. In 2017 IEEE International Conference on Acous tics, Speech and Signal Processing (ICASSP), pages 5275–5279. IEEE.
9. Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, 11979 Luke Zettlemoyer, and Christoph Feichten hofer. 2021. Videoclip: Contrastive pre training for zero-shot video-text understanding. arXiv:2109.14084.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Y Lin, Andy TLiu, Jiatong Shi, Xuankai Chang, Guan Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. arXiv:2105.01051.

BIOGRAPHY

Bandi Dixitha studying 3rd year department of Electronics And Communication Engineering at Institute Of Aeronautical Engineering ,Dundigal .She Published a Research Paper Recently At IJSREM as a part of academics . She has a interest in digital signal processing.

DrSonagiri China Venkateswarlu professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Digital Speech Processing. He has more than 40 citations and paper publications across various publishing platforms, and expertise in teaching subjects such as microprocessors and microcontrollers , digital signal processing, digital image processing, and speech processing. With 20 years of teaching experience, he can be contacted at email: c.venkateswarlu@iare.ac.in

Dr. V. Siva Nagaraju is a professor in the Department of Electronics and communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Microwave Engineering. With over 21 years of academic experience, Dr. Nagaraju is known for his expertise in teaching core electronics subjects and has contributed significantly to the academic and research community. He can be contacted at email: v.sivanagaraju@iare.ac.in.