LaRA: Large Rank Adaptation for Speech and Text Cross-Modal Learning in Large Language Models

Zuhair hasan shaik Pradyoth Hegde Prashant Bannulmath Deepak K T
Indian Institute of Information Technology Dharwad, India
{zuhashaik12, pradyothhegde}@gmail.com {prashantb, deepak}@iiitdwd.ac.in

Abstract

Integrating speech and text capabilities into large language models (LLMs) is a challenging task and we present Large Rank Adaptation (LaRA) for effective cross-modal integration of speech and text in the LLM framework. Unlike conventional LoRA, our method requires significantly larger ranks comparable to the pretrained weights to accommodate the complexities of speech-text cross-modality learning. The approach utilizes HuBERT to convert speech into discrete tokens and fine-tunes the pretrained LLM to adapt to cross-modal inputs and outputs. The work employs a Hi-Fi GAN vocoder to synthesize speech waveforms from the generated speech units. The initial studies use the Librispeech corpus to teach the model the relationships between speech and text, and Daily Talk, which involves dialog conversations, to adapt for interaction. The proposed work demonstrates adaptation for spoken and text conversations. However, the proposed framework can be easily extended to other cross-modal applications.

1 Introduction

In humans, speech is the most preferred mode of communication. It encompasses rich information such as tone, emotion, gender, and speaker information. On the other hand, text is a less complex mode, making it better suited for storage and computational tasks. With the recent advancement of larger deep learning models, the text domain has seen revolutionary changes (Brown et al., 2020; Singhal et al., 2023; OpenAI et al., 2024; Anil and et al., 2024). Applications such as summarizing, conversation, and translation have shown tremendous ability to reason with the input data.

Speech, unlike text, is a high bit rate and complex signal. Understanding its semantics is crucial for language comprehension. Training a large language model (LLM) with speech from scratch is computationally intensive and requires an enormous amount of training data, which is not commonly available (Kalyan, 2023). Considering the capabilities of LLMs, the scope of expanding them to speech domain which already possesses the knowledge in the text domain, can be explored (Hu et al., 2024). This approach to enhancing the capabilities of LLMs is a significant area of research for further exploration.

In this paper, we propose Large Rank Adaptation (LaRA)[1], an approach that utilizes the existing knowledge of pretrained large language models for cross-modal applications. The main contributions of the paper are as follows:

- We present an approach, which involves cross-modal fine-tuning and alternative-cross-modal data modeling techniques during the training phase. This method enables seamless support for both speech and text inputs, as well as their corresponding outputs in either speech or text formats.
- Unlike recent proposed works that require extensive retraining and large computing resources, our approach achieves cross-modal adaptation with relatively modest computational requirements.
- Through extensive experimentation in spoken dialog and speech-text translation tasks, we show that conventional low-rank adaptation methods (LoRA) are insufficient for cross-modal adaptation tasks. Instead, a large rank adaptation (LaRA) method, which is comparable to the pretrained model weights, is necessary to effectively integrate speech and text modalities.

The remaining paper is divided in the following manner. Section 2 extends the introduction and

[1] The code for the proposed work can be found here: https://github.com/Zuhashaik/LaRA

Findings of the Association for Computational Linguistics: EACL 2024, pages 8201–8211
November 12-16, 2024 ©2024 Association for Computational Linguistics

gives an insight into the current literature. Section 3 details the architecture of the model. Section 4 discusses the data corpus and formatting, fine-tuning, and experiments carried out. Section 5 provides a comprehensive analysis of the work. Finally, Section 6 concludes the paper.

2 Related work

Recent advancements in unified models for speech and text tasks have demonstrated the potential for a single architecture to handle diverse input and output modalities effectively. Since we worked on fine-tuning the decoder-only model, we looked into similar work done in the literature. One significant development is the decoder-only model introduced by LauraGPT (Chen et al., 2023), which can perform tasks such as speech recognition, speech-to-text translation, machine translation, text-to-speech synthesis, and speech emotion recognition. This model processes continuous audio inputs and produces discrete audio outputs, exemplifying the capability of a unified model to seamlessly integrate multiple functions.

The necessity of fine-tuning the entire parameters of large models is often computationally expensive and resource-intensive. To address this, a more efficient method involves training low-rank decomposed matrices and attaching them to the dense layers, thereby reducing fine-tuning costs significantly (Hu et al., 2021). This approach has paved the way for adapting large language models (LLMs) for various downstream tasks. For instance, the AudioPaLM model described in (Rubenstein et al., 2023a) combines the linguistic strengths of the text-based PaLM-2 with the paralinguistic capabilities of AudioLM, using separate embedding matrices for audio and text. This dual approach enhances the model's ability to process complex audio and text tasks. Similarly, SpeechGPT, as detailed in (Zhang et al., 2023), employs a three-step training process to fine-tune the Llama-13B pretrained model, incorporating a low-rank adapter to improve performance across different modalities. Other notable works including the Listen, Think, and Understand (LTU) model (Gong et al., 2023b), which extends the Llama LLM for general audio reasoning, and the LTU Audio-Speech (LTU AS) model (Gong et al., 2023a), which combines Whisper and Llama 2 to understand both environmental sounds and speech. These advancements underscore the ongoing efforts to develop comprehensive models capable of addressing a wide array of speech and text-related tasks, pushing the boundaries of what unified multitask learning can achieve.

The work by (Chen et al., 2024) focuses on speech-to-text applications, capturing nuances in audio for tasks like speech translation by adapting a Low-Rank Adaptation (LoRA). Another similar approach is discussed in (Le et al., 2024), emphasizing the importance of detailed audio representation in enhancing speech recognition accuracy.

From the literature we observe that on text-based LLMs interact with the speech modality (Nguyen et al., 2024; Hassid et al., 2024; Rubenstein et al., 2023b; Fathullah et al., 2023; Chou et al., 2023; Maiti et al., 2024). Some studies have also explored the multi-modality in LLM using low-rank adapters. Furthermore, models designed to combine speech and text have typically been pretrained with both types of data. So, our work aims to develop a model that can be easily adapted for speech by leveraging the knowledge from the pretrained text large language model using adapters.

3 Method

As shown in Figure 1, our methodology begins by converting speech input into discrete tokens using a speech tokenizer based on the HuBERT architecture (Hsu et al., 2021). In this work, we use Llama-2 7B[2] as our base LLM (Touvron et al., 2023). The speech tokens are added to the LLM's vocabulary, and the embedding layer is resized accordingly. The embedding for speech tokens are initially set using a Gaussian random distribution and then replaced with HuBERT's quantized hidden representations projected to match the LLM's dimensions using a projection layer. Cross-modal learning involves creating input sequences with speech tokens followed by text tokens and vice versa. The LLM is trained to predict the next token, whether speech or text, based on the previous tokens, allowing it to associate speech units with corresponding text.

For efficient training, Large Rank Adaptation (LaRA) technique is proposed, usage of high-rank adapters to enable the LLM to adapt to the new speech modality while preserving its text generation capabilities. During inference, the input sequence with speech and text tokens is processed,

[2] Model can be found here: https://huggingface.co/meta-llama/Llama-2-7b
This model is open-sourced under the license *llama2*.

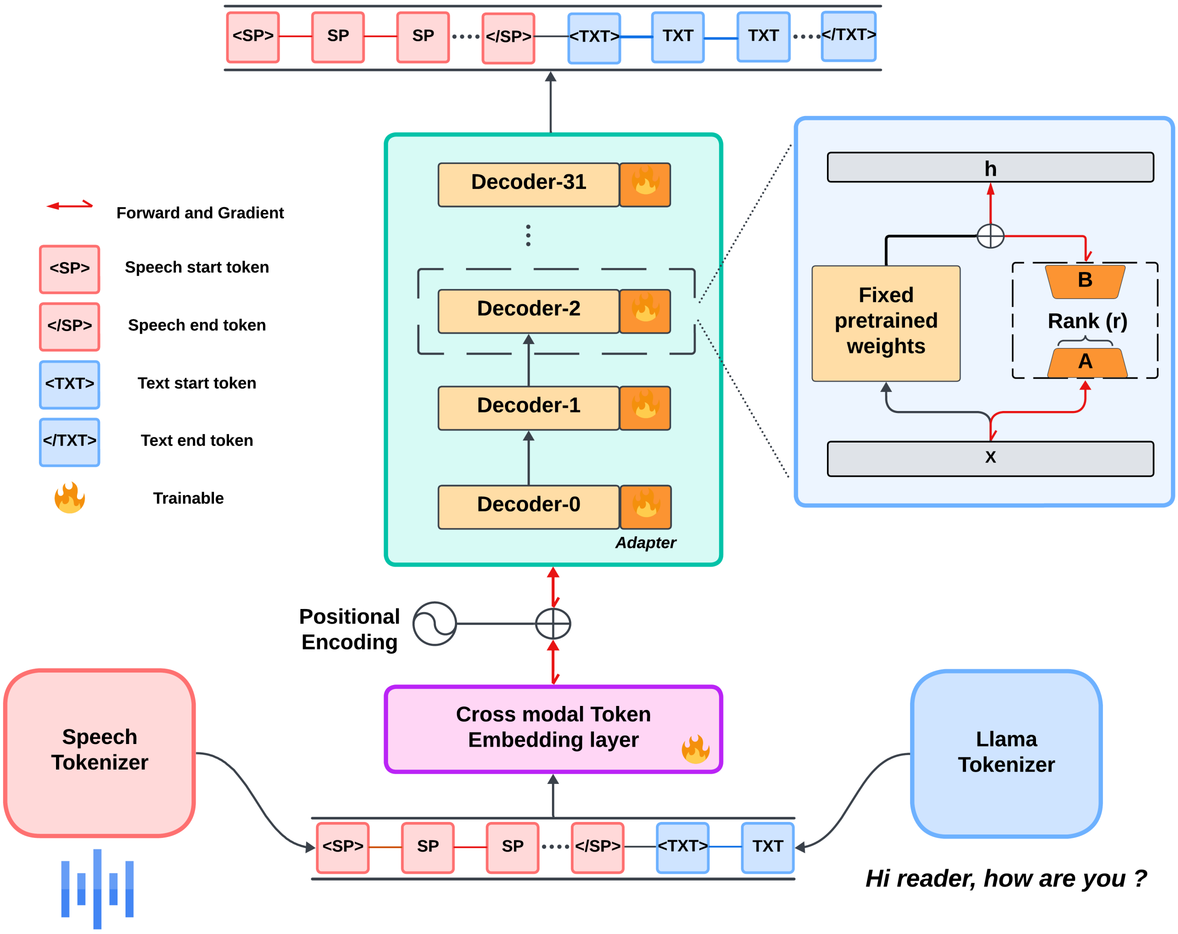


Figure 1: Overall training architecture of the proposed model. The adapters in each decoder block having W_q, W_k, W_v, W_o weight matrices are trained with larger ranks while keeping the pretrained weights fixed.

and the generated speech tokens are converted into speech waveform using a unit HiFi-GAN vocoder (Polyak et al., 2021a), which synthesizes the final speech output.

3.1 Speech Tokenizer

Speech Tokenizer module uses HuBERT architecture which employs a self-supervised learning and vector quantization approach to discretize speech (Hsu et al., 2021). Initially, it utilizes k-means clustering on the model's intermediate representations, or Mel-frequency cepstral coefficients for the initial iteration, to create discrete labels for masked audio segments.

By pre-training on an unlabeled speech corpus in the target language, a HuBERT model can then transform the target speech into hidden representations, computed at each 20-ms frame. Subsequently, a k-means algorithm is employed on these learned representations from the unlabeled speech to get discrete speech representation. The k cluster centroids are then utilized to encode speech utterance into sequences of cluster indices, computed every 20-ms interval (Lakhotia et al., 2021; Polyak et al., 2021b). In the end, adjacent repeated cluster indices are merged to get a discretized speech utterance S represented as, $S = [s_1, s_2, \ldots, s_f], \quad s_i \in \{0, 1, \ldots, k-1\} \quad \forall 1 \leq i \leq f$

where f is the number of frames. In the proposed work we have used three HuBERT variants, they are: mHuBERT-km1000, HuBERT-Base-KM50, and HuBERT-Base-KM100, which compress speech into discrete tokens having k = 1000, 100, and 50 clusters, respectively. When using mHuBERT-km1000, the speech input is quantized into 1000 cluster indices (0 to 999). To add these speech units to the LLM's vocabulary without conflicting with existing numeric tokens in the embedding matrix, we enclose each index in angle brackets, ex.: $< 588 >, < 949 >$. Similarly, for

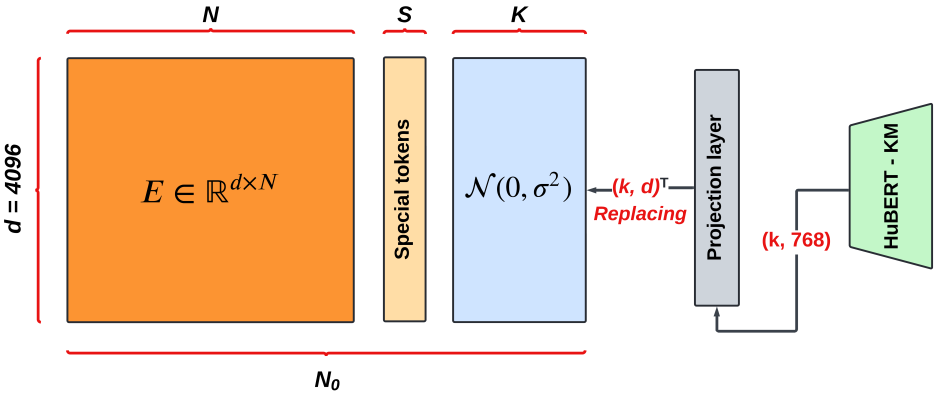


Figure 2: The speech embedding matrix ($d \times k$) is replaced using HuBERT's hidden representations, which are upsampled to the LLM's hidden size via a projection layer to align dimensions.

HuBERT-Base-KM100 the tokens are from 0 to 99, and HuBERT-Base-KM50 for we get tokens from 0 to 49.

3.2 Cross modal Token Embedding layer

The embedding layer E in a large language model (LLM) acts as a lookup table $E \in R^{d \times N}$, where d is the dimension of the embedding and N is the vocabulary size. After tokenizing the speech input, the speech tokens are added to the LLM's vocabulary. The total number of tokens becomes N_o, which can be expressed as $N_o = N + K + S$. Here K represents the cluster indices, and the additional four special tokens (S=4) account for the start `<sp>` and end `</sp>` speech tokens, as well as the start `<txt>` and end `</txt>` text tokens.

Now we resize the embedding matrix with the vocabulary size N_o which can be represented as $E \in R^{d \times N_o}$, where entries from N^{th} index to $N_o - 1$ are initialized using a Gaussian distribution. This resizing ensures the matrix can handle the expanded vocabulary, and the Gaussian distribution to ensure they start with moderate values, promoting consistency during training and preventing issues from extreme values.

Transfer Learning: We then replace the $d \times k$ embedding matrix of speech tokens `<0>` to `<k-1>` with HuBERT's quantized hidden representations to leverage the knowledge from HuBERT. However, since HuBERT's representation dimension is 768 and the LLM's dimension is 4096, we use a projection layer to upsample from 768 to 4096 dimensions.

This projection layer helps align the dimensions, ensuring the embeddings fit seamlessly into the LLM's architecture, is illustrated in Figure 2. The embedding layer remains trainable to maximize the cross-modal understanding.

3.3 Cross-modal Learning

In our approach, we represent speech as a sequence of tokens, where each token represents a cluster index obtained through HuBERT's vector quantization process. Let's denote the speech tokens as <s_1>, <s_2>, .. <s_f> where f is the number of frames in the speech and <s_i> represents the i^{th}th speech token. Similarly, we have text tokens denoted as <t_1>, <t_2>, .. <t_l>, where l is the length of the text sequence and <t_j> represents the j^{th} text token.

To train the LLM and make it understand speech units, we create a cross-modal input consisting of speech tokens followed by text tokens as shown below

$$< s_p >< s_1 >< s_2 > ... < /sp >< txt ><$$
$$t_1 >< t_2 > ... < /txt >< sp > ...$$

The goal of the autoregressive model is to maximize the probability of the next token x_t (which can be either a speech token s_i or a text token t_j) given $t - 1$ tokens. This can be expressed as:

$$P(x_t \mid x_1, x_2, \ldots, x_{t-1}) \tag{1}$$

This formulation ensures that the prediction of the next token (whether it's a speech token or a text token depends on all previously generated tokens up to that point). This process helps the LLM learn to associate speech units with corresponding text, improving its ability to generate accurate text output from speech input and vice versa.

```
<sp><454><951><386> ... <544></sp> <txt> UNDERSTAND SAID TOM ... MEET</txt>
<sp><454><373><374> ... <39></sp> <txt> THERE WAS A MAN ..WAS THIS</txt><sp> ... so on
```

(a) Data modeling of the Librispeech-360 dataset for training.

```
<A><sp><497><63><991><635>... <497></sp></A> <B><txt>Yes, are you Mr. James?</txt></B>
<A><sp><497><63><991><873> ... <497></sp></A> <B><txt>That's right.</txt></B> <A><sp> ... so on
```

(b) Data modeling of the DailyTalk dataset for training. The dataset contains conversations between two people and they are represented by <A> and <B>.

Figure 3: Alternative-Cross-Modal data modeling for different datasets.

3.4 Large Rank Adaptation (LaRA)

We propose Large Rank Adaptation (LaRA) for cross-modal learning, allowing us to use pre-trained language models without needing to retrain them from scratch on a large dataset with extensive computing power. LaRA builds on the idea of Low-Rank Adaptation (LoRA) but modifies it slightly in concept and inference. For a pre-trained weight matrix $W_0 \in R^{d\times k}$, we update it using a low-rank decomposition $W_0 + \Delta W = W_0 + \beta\alpha$, where $\beta \in R^{d\times r}$ and $\alpha \in R^{r\times k}$, with the rank r being less than $\min(d, k)$ but not significantly smaller as in LoRA.

Unlike LoRA, which uses a rank $r \ll \min(d, k)$, we believe this is not effective for cross-modal adaptation. Instead, we propose using a rank r that is less than but close to $\min(d, k)$ that can be represented as $r < \min(d, k)$. This makes the adaptor weights comparable to the base model weights, enabling better new modality learning. We have applied LaRA for four weight matrices in the self-attention module which are W_q, W_k, W_v, W_o.

3.5 Inference

Unit-vocoder: We use the unit HiFi-GAN (Polyak et al., 2021a) to decode the speech signal from the discrete speech token generated by LLM. During the inference phase, the model utilizes the cross-modal learning acquired during training to efficiently process new input data. The input sequence consists of both speech tokens, delineated by <sp>, and text tokens, delineated by <txt>. These tokens are initially parsed, and based on their type, the model generates either speech or text tokens accordingly. Speech tokens are then converted into speech units and fed into a unit HiFiGAN-based vocoder. The vocoder synthesizes the generated speech, providing the final output for the given input sequence. This comprehensive process seamlessly integrates the model's understanding of both speech and text modalities to produce coherent and natural-sounding speech output.

4 Experiments

This section describes the experimental setup and datasets used for evaluating the proposed approach. The experiments were conducted using the LibriSpeech-360 and DailyTalk datasets, which contain speech recordings and corresponding transcriptions. It also provides details on the dataset statistics, experimental configurations, and hyper-parameter settings used for training and evaluation.

4.1 Dataset

Librispeech: We utilized the LibriSpeech-360[3] variant (Panayotov et al., 2015), sourced from audiobooks, which offers a dataset consisting of approximately 360 hours of audio recordings, each accompanied by its respective transcription and we present a summary of their statistics in Table 1. This vast corpus directly links speech and text, providing a rich training ground for our model to understand the relationship between speech and text tokens. The dataset is structured in an alternative-cross-modal format, alternating between speech sequences and text sentences. Each data point consists of speech tokens enclosed within speech start and end tokens, followed by a text sentence, also enclosed within text start and end tokens. This format ensures clear boundaries between speech and text modalities. The data modeling of the LibriSpeech-360 is illustrated in the Figure 3a.

[3] LibriSpeech ASR corpus: https://www.openslr.org/12

Dataset	Libri-speech	Daily talk
hours	360 hours	20 hours
Total sentences	28,539	2,541
Train Split:		
Speech Tokens	20 M	1.1 M
Text Tokens	4.1 M	0.15 M
Validation Split:		
Speech Tokens	0.3 M	0.12 M
Text Tokens	63.2k	17.5k

Table 1: This table illustrates the statistics of the datasets utilized in our work.

DailyTalk: We also utilized the DailyTalk dataset[4] (Lee et al., 2022), containing approximately 20 hours of speech conversations and their transcripts. We present a summary of their statistics in Table 1. This dataset was employed to teach the alignment between conversational speech and text in a cross-modal context. The conversations are structured to indicate the speaker, denoted as <A> or <B>, with their corresponding speech tokens enclosed within speech start and end tokens, or text sentences enclosed within text start and end tokens. This fixed configuration maintains consistency in the dataset structure, facilitating effective training of the model. The data modeling follows a format similar to LibriSpeech and alternative-cross-modal approach, as illustrated in Figure 3b.

4.2 Experimental setup

For the proposed work we adapted HuBERT in speech tokenizer. The base LLM used was Llama-2 7B, which has 32 decoder layers and a vocabulary size (N) of 32,000 with a hidden dimension (d) of 4096. Extensive experimentation was performed on the following aspects:

- Since speech is a continuous signal, we need to discretize it before feeding it to the model when adapting cross-modality. We employed three models: mHuBERT-km1000, HuBERT-Base-KM50, and HuBERT-Base-KM100, which discretize speech into units having 1000, 100, and 50 clusters, respectively.
- As mentioned in the previous section, we attempted transfer learning by replacing the randomly initialized ((speech token embeddings H(<0>, <1> ..<S_{k-1}>) where H is the hidden state of the word embedding matrix for the respective <S_{k-1}>th token)) with Hubert's hidden states.
- The impact of using increasing ranks on the model's learning capability was investigated, as LaRA (Large Rank Adaptation) was discovered during this process.

The hyperparameters and training configurations employed in our experiments are presented in Table 2. The training and inference were carried out utilizing a computational setup consisting of four V100 Tesla GPUs, each equipped with 32GB of VRAM (Video RAM). Three of the GPUs, collectively providing 96GB of VRAM, were dedicated to the training phase, while the remaining 1 GPU, with its 32GB of VRAM, was allocated for inference and testing purposes.

Hyper parameter	Value
Rank-(r)	$r = 2^{n-1}, n \leq 11$
Scaling_factor (ψ)	1
Dropout	0.2
Learning Rate	$2 \times e{-5}$
Batch size	1
adam_beta1	0.9
adam_beta2	0.999
adam_epsilon	$1 \times e{-8}$
rms_norm_eps	$1 \times e{-5}$

Table 2: Hyperparameters for the Training, where $\psi = \alpha/rank$ and ψ used to give weight to the resultant of LaRA matrices as follows, h = $W_0 + \psi(\beta\alpha)$.

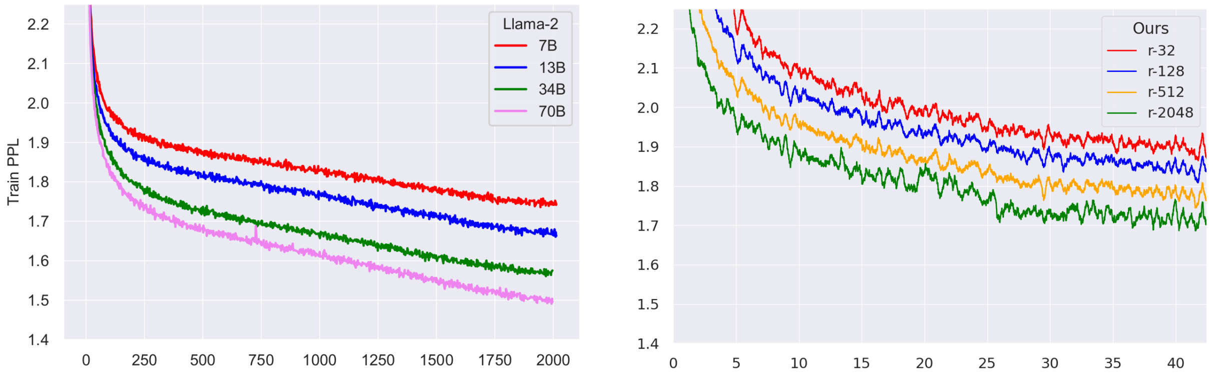
5 Analysis

We employed the cross-entropy loss for training the network, which is defined as:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{2}$$

where y_i represents the ground truth probability distribution over the vocabulary, and $\hat{y}i$ denotes the predicted probability distribution generated by the model. Minimizing this cross-entropy loss is equivalent to maximizing the conditional probability of

[4] DailyTalk: https://github.com/keonlee9420/DailyTalk Both datasets are licensed under CC BY 4.0.



(a) Llama-2: Train-loss / Billion Tokens

(b) Llama-2-LaRA (Ours): Train-loss / Million Tokens

Figure 4: The training of the base model (Llama-2-7B) with increasing parameters, as well as the training of our proposed work with increasing ranks. In sub-figure b, we can't see the saturation of loss while increasing the ranks.

the next token x_t (which can be either a speech token s_i or a text token t_j) given the previous tokens $x_1, x_2, \ldots, x_{t-1}$ as mentioned in equation 1. This objective reflects the model's cross-modal understanding between speech and text, as it learns to associate speech units with corresponding text by optimizing the likelihood of generating the correct token sequence from the cross-modal input consisting of both speech and text tokens.

5.1 Text-Speech Embedding Space

Initially, word embeddings and speech embeddings are unrelated. By utilizing the Alter cross-modal data modeling format for training auto-regressive models on cross-modal understanding, our approach can learn patterns between speech and text sequences. This allows the model to capture linear relationships between text and speech tokens, bringing their embeddings closer together in the shared embedding space, as shown in Figure 5. In figure 4 we illustrate base model (llama-2) training and our model (cross-modal) training with increasing ranks.

5.2 Model performance on larger ranks

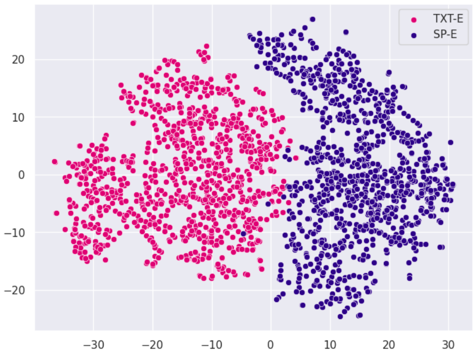
When utilizing base models, adapting them to different tasks is essential due to computational and data limitations that make it impractical to train entire models from scratch.

LoRA (Low-Rank Adaptation) introduces adapters with rank $r \ll \min(d, k)$, allowing efficient training with limited data and computational resources. However, this method is effective only for adapting models within the same modality, as different modalities have embeddings or hidden states that do not lie in the same space, and thus, the model lacks understanding between them.

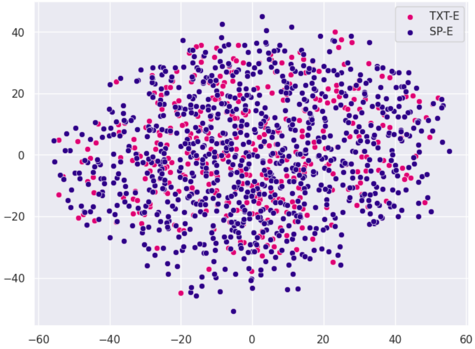
Rank	Text	Speech	Sp-Txt	Trainable %
1	1.23	0.22	1.97	3.76
4	1.18	0.20	1.95	3.80
16	1.16	0.17	1.92	3.97
32	1.16	1.17	1.90	4.20
64	1.16	0.17	1.88	4.66
128	1.16	0.17	1.84	5.56
512	1.16	0.18	1.79	10.60
1024	1.17	0.18	1.77	16.54
2048	-	0.19	1.74	26.34

Table 3: This table represents the validation loss for text, speech, and speech-text cross-modal learning at different ranks, along with the percentage of trainable parameters. Blue indicates the first least validation loss across text and speech. Red highlights where the loss begins to increase, suggesting saturation in the case of Speech and Text, for the speech-text (Sp-Txt) modality, where the loss consistently decreases, indicating no saturation.

As shown in Table 3 and Figure 4b, increasing the ranks of cross-modal adapters leads to better performance. However, most prior work focuses on model architectures (Wu et al., 2023; Shen et al., 2024) rather than experimenting with adapter ranks because of computational reasons. As a result, they overlook the potential benefits of using higher ranks for cross-modal adapters and blindly employ low-rank ($r \ll \min(d, k)$) adapters instead. Our findings suggest that using higher capacity cross-modal adapters with larger ranks can significantly improve cross-modal performance by better capturing inter-modality relationships. Still keeping $r < \min(d, k)$ such that we don't overpass the base model parameters. Lower ranks perform better (Hu et al., 2021) when the modality is the same, and our model should address task-specific problems or



(a) Before training



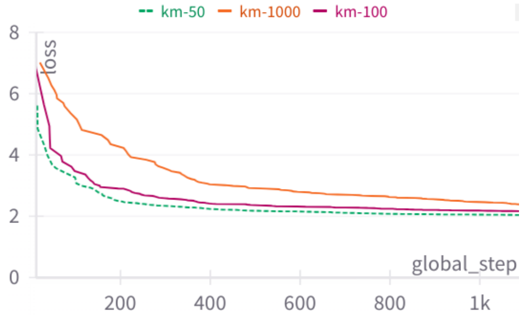
(b) After training

Figure 5: t-SNE visualization of the same text embeddings (TXT-E) and speech embeddings (SP-E) before and after the cross-modal training procedure.

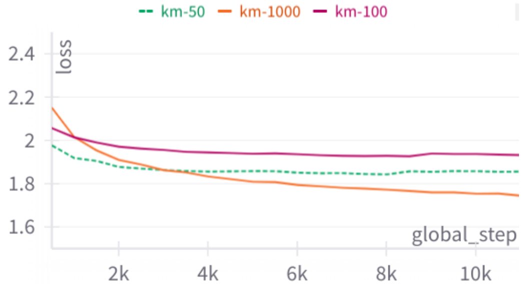
situations that require little tweaking to the model.

5.3 K-means Cluster Indices

We experimented with different k-means cluster sizes: 50 (HuBERT-km50), 100 (HuBERT-km100), and 1000 (HuBERT-km1000). We initially assumed that models with fewer clusters would perform better, as they would face less uncertainty in token prediction and could more effectively understand the relationships between speech tokens. However, the results, illustrated in Figure 6a, showed a different trend. In the early training stages, HuBERT-km1000 exhibited higher training and validation loss compared to HuBERT-km50 and HuBERT-km100, which was expected due to the increased number of tokens to predict. Surprisingly, as training progressed, the loss for HuBERT-km1000 decreased and eventually fell below the losses for HuBERT-km50 and HuBERT-km100. This unexpected result prompted further analysis. The higher cluster indices in self-supervised speech models like HuBERT capture more detailed speaker, gender, and acoustic information (Sicherman and Adi, 2023). The HuBERT models with larger cluster sizes effectively preserve speaker and gender information while still capturing phoneme details. Our findings mention that HuBERT-km1000 benefits from learning richer representations, encompassing more speaker, gender, and acoustic details, which enhances its ability to understand and predict speech tokens. This leads to a reduction in loss over time, despite the increased complexity of predicting more tokens.



(a) Training loss



(b) Validation loss

Figure 6: Training and Validation losses for HuBERT models with different k-means cluster sizes

In summary, while we initially favored models with fewer clusters, our experimental results and insights from related research suggest that higher cluster indices can enhance speech representation, ultimately improving model performance on spoken language modeling tasks.

5.4 MMLU Evaluation on Text-Only Performance

We conducted an additional experiment using the MMLU dataset. We compared the performance of the Llama-2 7B baseline model with our Llama-2-LaRA models to determine if there were any significant declines in text generation capabilities. As shown in Table4, our results show a slight decrease in performance across most subjects, with the exception of formal logic, where the performance remained consistent with the baseline. This slight decline can be attributed to the shift in text embedding vectors caused by cross-modal learning,

Subject	Base model	LaRA
Astronomy	0.39	0.36
College Biology	0.44	0.38
College CS	0.31	0.27
High School CS	0.44	0.32
College Mathematics	0.31	0.27
Formal Logic	0.24	0.26
Average	**0.35**	**0.31**

Table 4: Comparison of MMLU performance between the Base Model (llama-2-7B) and llama-LaRA-7B. The table shows a slight decrease in text-only performance for most subjects.

as illustrated in the t-SNE plot in Figure 5. Given the multimodal capabilities of the model, which now supports both speech and text, this minor decrease is acceptable and expected.

The results indicate that while there is a slight reduction in performance, the overall text generation capability of the model remains largely unaffected by the integration of speech modality.

6 Conclusion and Future Work

In this paper, we presented a large-rank adapter for Llama-2 7B LLM. Our approach shows the adaptation of the speech modality for the pretrained LLM with a significantly larger rank of 2048. The findings underscore the importance of adapting existing text-based LLMs to incorporate speech modality without the need for training from scratch. Future work will focus on extending this framework to other cross-modal applications, further enhancing the versatility and applicability of large language models in multimodal contexts.

7 Ethical considerations

All datasets utilized in this work (LibriSpeech and DailyTalk) are licensed under the Creative Commons BY 4.0 license, ensuring their ethical and legal use for research purposes. Additionally, the base language model employed, Llama-2 7B, is open-sourced under the Llama2 license, further promoting transparency and responsible development.

However, risks exist with large language models like biases in the training data or misuse of generated content. Careful evaluation and safeguards are needed before real-world deployment. Integrating speech raises privacy concerns since voice data contains personal identifiers. Anonymization techniques and strict data protocols must protect user privacy.

While focusing on technical aspects, it is important to consider the broader ethical implications of increasingly capable AI systems. Continuous efforts towards transparency, accountability, and responsible AI development benefiting society are crucial.

Limitations

We solely focused on cross-modal integration between speech and text modalities. The exploration of vision modality and its integration with speech and text is left for future work. We limited our experiments to the llama-2 7B base model, as our primary objective was to demonstrate the effectiveness of cross-modal integration and the LaRA approach. Choosing a single base model allowed us to conduct consistent and controlled experiments to showcase our findings. Considering multiple base models would have introduced additional computational constraints and complexity, which we aimed to avoid in this initial study.

Acknowledgements

We sincerely thank the funding agencies, including the Technology Innovation Hub on Autonomous Navigation (TiHAN) Foundation at IIT Hyderabad, the Ministry of Electronics and Information Technology (MeitY), and the Science and Engineering Research Board (SERB), Government of India, for their valuable support.

References

Rohan Anil and et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqi Zheng, et al. 2023. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *arXiv preprint arXiv:2310.04673*.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. IEEE.

Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. 2023. Toward joint language modeling for speech units and text. *Preprint*, arXiv:2310.08715.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2023. Prompting large language models with speech recognition abilities. *Preprint*, arXiv:2307.11795.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023a. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James R Glass. 2023b. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2024. Textually pretrained speech language models. *Preprint*, arXiv:2305.13009.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and EnSiong Chng. 2024. Large language models are efficient learners of noise-robust speech recognition. *arXiv preprint arXiv:2401.10446*.

Katikapalli Subramanyam Kalyan. 2023. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Chenyang Le, Yao Qian, Long Zhou, Shujie Liu, Yanmin Qian, Michael Zeng, and Xuedong Huang. 2024. Comsl: A composite speech-language model for end-to-end speech-to-text translation. *Advances in Neural Information Processing Systems*, 36.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2022. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. *Preprint*, arXiv:2207.01063.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. Voxtlm: unified decoder-only models for consolidating speech recognition/synthesis and speech/text continuation tasks. *Preprint*, arXiv:2309.07937.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. Spirit-lm: Interleaved spoken and written language model. *Preprint*, arXiv:2402.05755.

OpenAI, Josh Achiam, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021a. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021b. Speech resynthesis from discrete disentangled self-supervised representations. In *INTERSPEECH 2021-Annual Conference of the International Speech Communication Association*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023a. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah

Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023b. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. 2024. Multimodal instruction tuning with conditional mixture of lora. *Preprint*, arXiv:2402.15896.

Amitay Sicherman and Yossi Adi. 2023. Analysing discrete self supervised speech representation for spoken language modeling. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *Preprint*, arXiv:2309.05519.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.