Parameter-efficient Adaptation of Multilingual Multimodal Models for Low-resource ASR

Abhishek Gupta[*] **Amruta Parulekar**[*] **Sameep Chattopadhyay**[*] **Preethi Jyothi**
Indian Institute of Technology Bombay, Mumbai, India
{abhishekumgupta,amrutaparulekar.iitb,sameep.ch.2002}@gmail.com, pjyothi@cse.iitb.ac.in

Abstract

Automatic speech recognition (ASR) for low-resource languages remains a challenge due to the scarcity of labeled training data. Parameter-efficient fine-tuning and text-only adaptation are two popular methods that have been used to address such low-resource settings. In this work, we investigate how these techniques can be effectively combined using a multilingual multimodal model like SeamlessM4T. Multimodal models are able to leverage unlabeled text via text-only adaptation with further parameter-efficient ASR fine-tuning, thus boosting ASR performance. We also show cross-lingual transfer from a high-resource language, achieving up to a relative 17% WER reduction over a baseline in a zero-shot setting without any labeled speech.

1 Introduction

Across the languages of the world, the automation of various speech and text tasks has led to the creation of massive multilingual datasets such as Multilingual LibriSpeech (Pratap et al., 2020), that contain speech, text, and other metadata for a number of different languages. This large-scale collection has catalyzed the emergence of large multilingual automatic speech recognition (ASR) models (Yadav and Sitaram, 2022), which utilize the structural similarities between different languages to learn language-invariant features and boost accuracy. Subsequently, multimodal multilingual models, such as M3P (Ni et al., 2021), that bridge the gap between speech and text using joint representation spaces, have also emerged. These models are trained using large amounts of multilingual speech and text data.

However, less-spoken languages, especially those from developing countries, do not have such large data corpora available (Magueresse et al., 2020), thus hurting model performance for extremely low-resource languages (Chang et al., 2023). Thus, creating targeted models for severely low-resource languages has become crucial. One efficient way to do this is by adapting existing models to the target language using limited amounts of labeled data. Such adaptation has to be done carefully so as to not overfit to the target language characteristics.

Parameter-efficient fine-tuning (PEFT) (Han et al., 2024) techniques have gained wide acceptance where only relevant parts of a model are identified and fine-tuned for a specific downstream task. Text-only adaptation is another sub-area that is gaining popularity for low-resource ASR (Bataev et al., 2023; Vuong et al., 2023). Multimodal models have training pathways for both speech and text data, offering a good framework to combine both approaches. Multilingual models, on the other hand, allow for cross-lingual transfer (Khare et al., 2021), i.e., using a higher resource language to improve performance on a lower resource language.

In this work, we have leveraged the multimodal nature of Meta's SeamlessM4T (Communication et al., 2023) to explore the benefits of speech-based adapter fine-tuning and text-only adaptation. These techniques have been used both in isolation and in combination to identify the best strategy to improve low-resource ASR for a number of Indic languages. We have also exploited the multilingual nature of the model to use higher-resource languages to improve low-resource ASR. Thus, our main contributions include: (a) identifying how to combine speech-based parameter-efficient fine-tuning and text-only adaptation to boost low-resource ASR, (b) identifying a cross-lingual transfer technique that can give more than 17% relative reduction in WER for a low-resource language without using any speech of that language, (c) the use of small amounts of available data to boost the performance of SeamlessM4T (Communication et al., 2023) on six Indic languages, Bengali, Gujarati, Kannada,

[*] These authors contributed equally to this work.

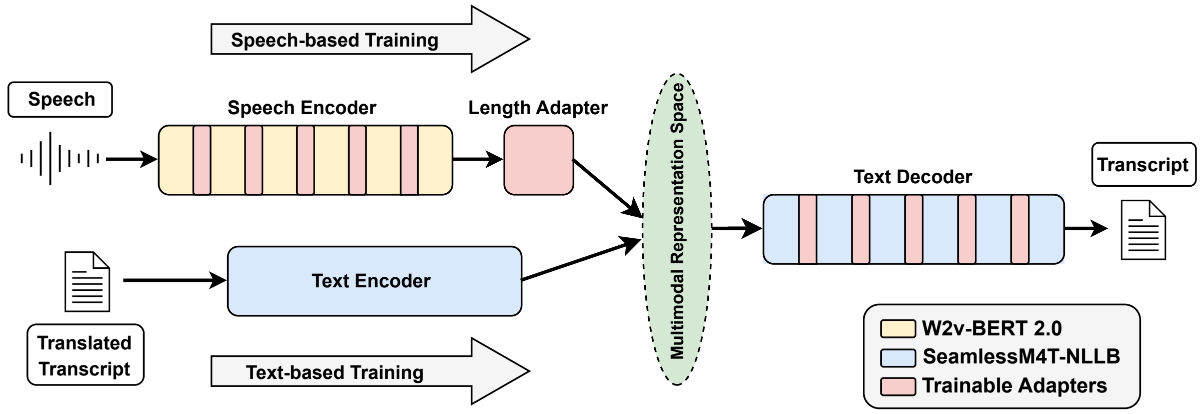

Figure 1: **Parameter-efficient Adaptations for SeamlessM4T:** A multimodal ASR model such as SeamlessM4T can be fine-tuned in a parameter-efficient manner through either speech-based adaptations or text-only adaptation.

Maithili, Malayalam and Odia.

2 Related Work

One of the key challenges in current ASR research is enabling systems to handle multilingual inputs (Yadav and Sitaram, 2022; Kannan et al., 2019) while minimizing resource requirements in terms of training, inference, and storage costs. Currently, the most popular paradigm using multilingual models are to initially pre-train the models in a self-supervised manner on a large multilingual dataset (Babu et al., 2021) before being fine-tuned on a set of target languages (Toshniwal et al., 2018; Bai et al., 2022). A general way of performing such model fine-tuning is by updating all the weights or some specific model components while training. These kinds of methods are parameter inefficient and often cause catastrophic forgetting (Kessler et al., 2021), for all non-target languages. Also, training and storage costs for such methods increase linearly with both the model size and the number of languages.

To mitigate these limitations, recent literature on NLP has introduced several parameter-efficient fine-tuning methods (Xu et al., 2023; Tomanek et al., 2021; Hu et al., 2021), often involving trainable modules called adapters (Houlsby et al., 2019), whose weights are updated while freezing the original backbone. Significant efforts are being made to develop better adapter architectures and efficient training methods (Yu et al., 2023) to utilize contrastive learning (Zhang and Ré, 2024) and meta-learning (Hou et al., 2021). These modules can also be used to adapt multilingual ASR models for a low-resource setting, with Simadapter (Hou et al., 2022) being one of the first models to utilize adapters to leverage cross-lingual features.

In the context of speech recognition, a low-resource setting could refer to any scenario with insufficient training data. This includes challenges such as recognizing atypical speech (Tomanek et al., 2021) or processing less commonly spoken languages. A recent work (Mainzinger and Levow, 2024) demonstrated the benefits of using adapters for very low-resource languages with less than five hours of training data. For the low-resource situation, task- or language-specific adapter modules showcase superior performance (Hu et al., 2024) compared to fine-tuning the model components, but even such approaches are constrained by inherent limitations of the base model.

Over the past few years, considerable effort has gone into developing multilingual ASR foundational models with more generalizable features. These models offer a stronger starting point for low-resource adaptations and enable the use of cross-lingual transfer learning. The exponential growth in computing power has led to the creation of increasingly large language models, which are now used for a wide range of tasks, including as backbones for multimodal ASR models (Rubenstein et al., 2023; Zhang et al., 2023; Chang et al., 2023). For such models, the foundational backbone is expanded using audio tokens generated using techniques like wav2vec (Schneider et al., 2019) and Hubert (Hsu et al., 2021) in order to learn a joint representation in a multimodal space; the token vocabulary is expanded to encompass both text and audio. Note that models with joint multimodal representations are not only useful for ASR but can also be integrated with a vocoder for TTS or conversational chatbots (Zhang et al., 2023).

Multimodal models can be trained with joint text-audio tasks through self-supervision with masked language modeling and denoising objectives; further fine-tuning is often done with ASR and speech-to-text or speech-to-speech translation tasks. One of the most recent examples of such a multilingual multimodal model has been SeamlessM4T (Communication et al., 2023) by Meta AI, which is built upon the NLLB (Team et al., 2022a) backbone and can process speech and text inputs from nearly 100 languages. An implicit advantage of using such multimodal models for low-resource ASR is the ability to benefit from text-only learning for shared parameters. In most cases, there is significantly more text data available than speech data. Thus, the capability to leverage text-only adaptation for ASR models can be highly advantageous in these scenarios.

While there is a lot of prior work in the domain of text-only adaptation for ASR (Vuong et al., 2023; Bataev et al., 2023; Chen et al., 2023; Mittal et al., 2023), and there has been some work on a comparative analysis of various fine-tuning strategies for low-resource ASR (Liu et al., 2024), to the best of our knowledge, our work is the first to explore them for multilingual multimodal models.

3 Methodology

In this work, we leverage a combination of parameter-efficient adaptation, unlabeled textual data, and minimal amounts of transcribed speech to improve ASR performance in low-resource languages using multilingual multimodal models. Figure 1 demonstrates the overall workflow of our proposed pipeline.

3.1 Multimodal base model: SeamlessM4T

We use SeamlessM4T (Communication et al., 2023) as our base model for all our experiments. SeamlessM4T, i.e., Massively Multilingual & Multimodal Machine Translation, is a versatile end-to-end model that provides support for multiple tasks, including speech-to-speech translation, speech-to-text translation, text-to-speech translation, text-to-text translation, and automatic speech recognition for up to 100 languages. The model has been trained using over a million hours of unlabeled speech in a self-supervised manner, along with more than 400K hours of human and machine-labeled audio. It supports 96 different languages for input speech and text, as well as output text, and can generate speech in 35 languages.

The SeamlessM4T model architecture is inspired by UnitY (Inaguma et al., 2023), a two-pass modeling framework that, unlike cascaded models, can be jointly optimized. The text encoder and decoder models of SeamlessM4T are initialized by the NLLB model (Team et al., 2022b), a text-to-text translation model. To process speech inputs, the model employs the Wav2Vec-BERT 2.0 speech encoder, which is an enhancement over the original model proposed by Chung et al. (2021) with additional codebooks. The model also includes a modality adapter (Zhao et al., 2022), referred to as the **length adapter**, to align the speech modality with text, projecting it to a unified representation space. Lastly, the model uses a text-to-unit (T2U) component for speech generation that produces discrete speech units from the text output. These units are then transformed into audio waveforms using a multilingual HiFi-GAN unit vocoder (Kong et al., 2020). There are multiple variants of the SeamlessM4T model; we have used SeamlessM4T-medium with a total of 1.2 Billion parameters.

Although the entire model comprises multiple components, our analysis focuses primarily on applying SeamlessM4T for multilingual ASR. The ASR pipeline of SeamlessM4T consists of the speech encoder (311M parameters), the length adapter (46M parameters), and the text decoder (201M parameters). Next, we will elaborate on parameter-efficient fine-tuning of SeamlessM4T (Section 3.2) and how we can use text-only adaptation within such a multimodal model (Section 3.3).

3.2 Parameter-efficient Fine-tuning

The ASR components of SeamlessM4T amount to more than 500M parameters. Full fine-tuning of these components using limited amounts of labeled data for low-resource languages may result in overfitting and degradation of ASR performance. To alleviate these challenges, parameter-efficient fine-tuning paradigms like the *adapter framework* (Houlsby et al., 2019) are very popular, especially for natural language processing tasks. Adapters have also found success in low-resource ASR tasks such as accent adaptation (Tomanek et al., 2021) and cross-lingual adaptation (Hou et al., 2022). Next, we will elaborate on the structure of an existing *length adapter* within SeamlessM4T and the new adapters we introduce in the encoder and decoder layers.

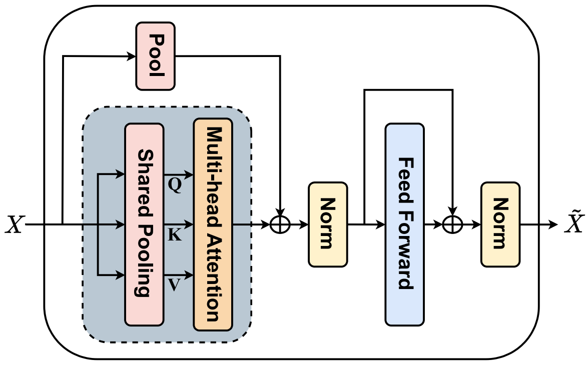

Figure 2: **SeamlessM4T Length Adapter:** Projects speech embedding X to a lower-dimensional representation $\tilde{X}$ in the multimodal space.

3.2.1 The Length Adapter

The length adapter in SeamlessM4T aims to bridge the gap between speech and text representations. It is inspired by the M-adapter architecture (Zhang et al., 2023) and uses a Transformer-based module to adapt speech representations to text. By compressing the speech sequence, the length adapter generates features tailored for multilingual speech-to-text tasks by modeling both global and local dependencies within the speech.

The main part of the original M-adapter architecture, illustrated in Figure 2, is the Multi-head Pooled Self-Attention (MPSA) mechanism. In the original MPSA, convolutional layers pool the input X and are further projected to the inputs of the multi-head attention module using linear transformation matrices. An additional pooling is applied in parallel to X and then added to the output of the attention module before being processed through a feedforward network. These processes together generate a lower dimensional representation of X, denoted by $\tilde{X}$ as the current layer output, addressing any length mismatches between embeddings from different modalities. Unlike the original M-adapter architecture with independent pooling modules for the multi-head attention inputs, the length adapter utilizes a shared pooling module, generating a single $\hat{X}$ for each X to improve efficiency. More formally, given an input sequence $X \in \mathbb{R}^{L \times D}$, where L is the sequence length and D is the embedding dimension, the MPSA mechanism starts by applying shared pooling to the input X to obtain $\hat{X} \in \mathbb{R}^{L' \times D}$. This pooling operation is performed using a 1D convolutional layer with kernel size k, stride s, and padding p. Subsequently, $\hat{X}$ is linearly projected into the query, key, and value matrices, denoted as Q, K, and V, respectively.

$$\begin{aligned}
\hat{X} &= \text{SharedPooling}\,(X) \\
Q &= \hat{X}W^Q, && \text{where } Q \in \mathbb{R}^{L' \times D}, \\
K &= \hat{X}W^K, && \text{where } K \in \mathbb{R}^{L' \times D}, \\
V &= \hat{X}W^V, && \text{where } V \in \mathbb{R}^{L' \times D}.
\end{aligned}$$

where the new sequence length L' is given by:

$$L' = \left\lfloor \frac{L + 2p - k}{s} \right\rfloor + 1.$$

We hypothesize that the length adapter module could potentially learn prosodic characteristics of languages, such as phoneme durations, by mapping speech embeddings — which include both segmental and suprasegmental information — to text embeddings that contain only content information. Learning certain prosodic characteristics like durations can be particularly beneficial for extremely low-resource languages that lack sufficient data for learning fine-grained contextual and syntactical information.

3.2.2 Encoder and Decoder Adapters

In addition to the pre-existing length adapter (Figure 2) in the SeamlessM4T architecture, we inserted additional trainable adapter layers within the encoder and decoder modules to adapt this multilingual model for low-resource languages. The adapter modules, following the architecture proposed in (Houlsby et al., 2019), initially project the original D_1-dimensional features into an intermediate space of dimension D_2. A non-linearity, specifically GeLU (Hendrycks and Gimpel, 2023) in our implementation, is then applied, after which the features are projected back to the original D_1 dimensions. To adjust the number of parameters for these adapters, we can change the intermediate dimension D_2. By decreasing the value of D_2, the number of trainable parameters in the adapters is reduced accordingly.

In our current experimental setup, we have inserted adapters after every Conformer layer in the encoders and after every Transformer layer in the text decoder. By setting the intermediate dimension D_2 to one-fourth of D_1 for all adapters, we introduce 6 million new trainable parameters each in the encoder and decoder modules.

Formally, the operations inside the i^{th} speech encoder layer can be summarized as:

$$\begin{aligned}
\mathbf{H} &= \text{MultiHeadAttn}(\mathbf{h}^{i-1}, \mathbf{h}^{i-1}, \mathbf{h}^{i-1}) \\
\mathbf{C} &= \text{Convolution}(\mathbf{H}) \\
\hat{\mathbf{h}}^{\mathbf{i}} &= \text{FFN}(\mathbf{C}) \\
\mathbf{h}^{i} &= \text{Adapter}(\hat{\mathbf{h}}^{i})
\end{aligned}$$

Similarly, the operations inside the i^{th} decoder layer can be summarized as:

$$\begin{aligned}
\mathbf{D} &= \text{MultiHeadAttn}(\mathbf{d}^{i-1}, \mathbf{d}^{i-1}, \mathbf{d}^{i-1}) \\
\hat{\mathbf{D}} &= \text{MultiHeadAttn}(\mathbf{d}^{i-1}, \mathbf{h}^{\ell}, \mathbf{h}^{\ell}) \\
\hat{\mathbf{d}}^{\mathbf{i}} &= \text{FFN}(\hat{\mathbf{D}}) \\
\mathbf{d}^{i} &= \text{Adapter}(\hat{\mathbf{d}}^{i})
\end{aligned}$$

where ℓ is the last encoder layer, and MultiHeadAttn(Q, K, V) is the standard multi-head attention implementation (Vaswani, 2017) with Q, K, and V denoting queries, keys, and values, respectively.

During our experiments, we fine-tuned the encoder adapters and length adapters on labeled ASR data, while the decoder was fine-tuned using ASR and machine translation (MT) data, thereby leveraging the text-to-text pipeline of SeamlessM4T.

3.3 Text-only Adaptation

The text decoder in the SeamlessM4T model is shared between the ASR pipeline and the text-to-text translation pipeline, allowing it to be trained for both tasks. This shared component in multimodal models possesses the ability to transfer knowledge from one task to another, thereby simultaneously enhancing the performance of multiple tasks. We hypothesize that we can improve the ASR performance for a target language by fine-tuning the text decoder adapters via text-to-text translation into that language. This allows us to perform a purely text-only fine-tuning of ASR models and is especially beneficial for languages where speech data is scarce. With the latest advancements in NLP, the quality of machine-translation models has greatly improved, allowing these models to be utilized to augment the existing parallel text using machine-translated text for these languages.

In our text-only fine-tuning experiments, we fine-tuned the decoder adapters on an English-to-target language translation task to help them learn the relevant syntactical features for the target language.

4 Experimental Setup

4.1 Dataset

The **IndicVoices** dataset (Javed et al., 2024) was utilized for all our experiments. This dataset is a multilingual, multi-speaker collection of natural and spontaneous speech in 22 Indian languages. It comprises 9% read speech, 74% extempore speech, and 17% conversational speech. Among these languages, Maithili is classified as a zero-shot language for SeamlessM4T, while Bengali is the sole high-resource Indic language. The remaining languages are categorized as low-resource languages for the model (Communication et al., 2023). One of the main reasons for using this dataset is that it is among the most comprehensive open-source, multilingual speech datasets for Indic languages covering many low-resource languages and one of the few published after the release of SeamlessM4T, ensuring there is no data leakage between the evaluation sets and the SeamlessM4T training data.

4.1.1 Transcribed Speech Data

The speech data and the corresponding transcripts from the IndicVoices dataset were used for the ASR fine-tuning experiments. The dataset, primarily consisting of extempore speech recorded under natural conditions, is characterized by a significant amount of noise and includes occasional disfluencies. For each language, 5 hours of speech were selected for the training set, sourced from an average of 336 speakers, to simulate an extremely low-resource setting. On average, each of the test and validation sets had 1 hour of speech by 68 and 206 speakers respectively. The out-of-vocabulary (OOV) rate of the test set was calculated to determine the amount of test-train domain overlap in the data. The OOV rates for Gujarati, Bengali, Kannada, Maithili, Malayalam, and Odia test sets were 39%, 35%, 58%, 41%, 53%, and 37%, respectively, averaging to an OOV of 43.87% on the test sets, further demonstrating the challenging nature of the task.

4.1.2 Text-only Data

The **IndicTrans2** (Gala et al., 2023) model was used to translate all the transcriptions present in the IndicVoices dataset to obtain parallel English-X text. Another set of parallel text data was created by using only the transcriptions of the 5-hour speech data in the training set for every language. For Bengali, Gujarati, Kannada, Maithili, Malayalam, and

Components	Learnable	Maitihli		Malayalam		Kannada		Gujarati		Odia		Bengali	
fine-tuned	Parameters	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER
None	-	82.20	43.39	56.15	20.65	69.29	29.11	41.03	24.50	42.81	17.38	37.70	18.44
Length adapter	46M	54.97	26.10	52.82	18.14	55.48	20.38	33.91	16.40	35.48	13.75	35.90	17.08
Text Decoder	201M	54.56	26.21	54.04	19.28	54.3	20.57	33.62	17.12	35.14	13.48	36.14	17.95
Speech Encoder	311M	43.87	17.79	46.99	13.45	47.91	14.93	27.79	11.58	29.82	9.24	29.07	12.09

Table 1: **Fine-tuning a Multimodal Model:** Comparison of WER (%) and CER (%) after ASR fine-tuning of SeamlessM4T with 5 hours of labeled speech, without adaptations; the first row presents the pre-fine-tuning results.

Text-only	Learnable	Maithili		Malayalam		Kannada		Gujarati		Odia		Bengali	
Adaptation	Parameters	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER
None	-	82.20	43.39	56.15	20.65	69.29	29.11	41.03	24.50	42.81	17.38	37.70	18.44
5hr Transcript	6M	71.32	37.92	**53.96**	**18.94**	70.52	32.54	35.67	19.19	38.77	**14.84**	**35.28**	**16.77**
Full Transcript	6M	**68.24**	**36.84**	55.30	20.43	**68.13**	**26.91**	**35.45**	**18.66**	**38.39**	16.22	35.44	17.73

Table 2: **Text-only Adaptation:** Comparison of WER (%) and CER (%) after text-only adaptation on SeamlessM4T with Eng-X parallel text using the full dataset and a 5-hour subset; the first row presents the pre-adaptation results.

Odia, the number of tokens in the 5-hour text sets were 40k, 43k, 30k, 42k, 34k, and 34k, respectively, while those in the large text set were 785k, 118k, 297k, 834k, 398k and 503k respectively. Thus, on average, each of the larger text data sets contained 489000 tokens for every language, while each of the smaller sets contained only 37261 tokens.

4.1.3 Implementation Details

The SeamlessM4T model comprises a speech encoder with 12 Conformer blocks and a text decoder with 12 Transformer blocks, with a model dimension $D_1 = 1024$. Two D_2 configurations were tested: $D_2 = 256$ (about 500K parameters per adapter layer, totaling 6M parameters) and $D_2 = 2048$ (matching adapter parameters with the length adapter, totaling 50M parameters). Text-only adaptation needed roughly 200 epochs of fine-tuning, while ASR fine-tuning required up to 40 epochs. All experiments were performed with a learning rate of 5×10^{-6} and a batch size of 16.

5 Experiments and Results

5.1 System A: Pure ASR Fine-tuning

We use the name *System A* to refer to the standard speech-to-text fine-tuning of SeamlessM4T using labeled speech and the ASR objective. The results of this experimental setup are summarized in Table 1. From the results, it is evident that fine-tuning the length adapter requires fewer parameters while providing similar benefits to text decoder fine-tuning across both metrics. Additionally, the ASR fine-tuning of the speech encoder proves to be significantly beneficial, although it involves training a substantially larger number of parameters.

In order to reduce the computational and storage requirements, the fine-tuning was substituted with language-specific adaptations, wherein adapters were introduced in the encoder and decoder, and these were fine-tuned in various combinations using transcribed speech data while freezing the base model. Table 3 depicts the results for the adaptations on System A. The results demonstrate that larger encoder adapters with 50M parameters are the most beneficial in enhancing the ASR performance, achieving WER and CER close to full fine-tuning of the model and the adapters while reducing trainable parameters by 90%. Additionally, Table 3 indicates that for the same number of trainable parameters, speech-based training of encoder adapters performs much better than that of decoder adapters. The performance of the length adapter fine-tuning surpasses that of the decoder adapters but falls short compared to the encoder adapters.

5.2 System T-A: Using Text-only Adaptation

The parallel English-target language text data generated by translating the transcripts of IndicVoices data was used to fine-tune the decoder adapters on an English-to-target language MT objective. Table 2 shows the ASR word error rates (WERs) with the complete transcription data and a smaller 5-hour text data subset (described in Section 4.1) to check the comparative benefits of text-only adaptation, without any ASR fine-tuning. For most languages, using the larger text corpus led to better performance. However, the smaller parallel dataset, with significantly fewer tokens, demonstrated comparable performance to that of the complete corpus. This suggests that text-only adaptation can be effective for multilingual multimodal models, even with very limited amounts of data.

Language	Component Fine-tuned	None		Length adapter		Encoder adapter		Decoder adapter		Len+Enc adapter		Encoder adapter (L)		All Components	
	Learnable Parameters	-		46 M		6 M		6 M		52 M		50 M		571 M	
	System	A	T-A	A	T-A	A	T-A	A	T-A	A	T-A	A	T-A	A	T-A
Maithili	WER	82.20	68.24	54.97	54.74	52.95	48.14	63.52	58.39	47.92	45.98	46.08	**44.60**	42.58	46.54
	CER	43.39	36.84	26.10	27.10	22.86	21.58	31.60	29.70	20.56	20.47	19.20	**19.52**	17.14	20.78
Malayalam	WER	56.15	55.3	52.82	52.51	49.71	50.14	56.03	53.71	48.22	48.19	47.81	**47.75**	47.38	45.9
	CER	20.65	20.43	18.14	18.87	15.34	16.35	20.21	20.00	14.76	15.46	14.12	**14.92**	13.86	13.38
Kannada	WER	69.29	68.13	55.48	53.83	52.54	53.29	62.88	58.71	49.36	48.24	49.14	**47.75**	45.48	43.5
	CER	29.11	26.91	20.38	20.94	16.95	18.84	23.76	23.44	15.63	16.51	15.26	**14.92**	14.06	14.18
Gujarati	WER	41.03	35.45	33.91	34.41	29.20	**27.72**	38.88	35.53	28.03	27.73	28.09	27.90	25.56	26.31
	CER	24.50	18.66	16.40	17.41	11.96	**12.05**	19.28	17.80	12.63	12.35	12.00	12.50	11.28	11.67
Odia	WER	42.81	38.39	35.48	34.99	32.03	32.97	38.55	36.24	30.09	31.18	30.04	**28.92**	30.54	30.17
	CER	17.38	16.22	13.75	14.62	10.57	11.25	14.50	14.57	10.11	11.32	10.01	**9.92**	10.37	10.30
Bengali	WER	37.70	35.44	35.90	35.09	29.65	28.77	38.10	35.60	29.96	**28.50**	29.30	31.92	28.12	27.62
	CER	18.44	17.73	17.08	17.22	12.76	12.58	18.59	17.72	13.06	**12.38**	12.52	14.63	12.12	11.91

Table 3: **Parameter-efficient Adaptation Results:** Comparison of WER (%) and CER (%) between different parameter-efficient adaptation methods for SeamlessM4T. System A refers to pure ASR fine-tuning, while system T-A refers to text-only adaptation followed by ASR fine-tuning. The best results for System A are underlined while the best results for System T-A are in **bold** for every language. The overall best results have been highlighted.

Moreover, text-only adaptation can be combined with ASR fine-tuning using labeled speech. We refer to the resulting ASR system with text-only adaptation, followed by ASR fine-tuning, as *System T-A*. Table 3 shows our overall results comparing System A and System T-A. We observe that text-only adaptation followed by ASR fine-tuning is more beneficial than pure ASR fine-tuning, as in System A. The trends of System T-A matched those of System A, with the larger encoder adaptation showing the best performance across all languages except Bengali, the only high-resource language in our study. This suggests that for low-resource languages with limited text and speech data, the most effective strategy is to first use text-only decoder adaptation, followed by speech-based encoder adaptation. It must also be noted that the results of using this strategy are comparable to those after full ASR fine-tuning of the entire model, with a $>90\%$ reduction in the number of trainable parameters, from 571M to 50M.

5.3 Cross-lingual Transfer

We hypothesize that the length adapter could capture content-agnostic prosodic characteristics of a language without overfitting on its syntax. Consequently, fine-tuning this adapter using data from a closely related high-resource language might enhance the model's predictions for a low-resource target language. The target languages chosen for this experiment were Maithili and Odia, categorized as zero-shot and low-resource languages for SeamlessM4T, respectively. Bengali, a language belonging to the same Eastern Indo-Aryan language family (Eberhard et al., 2020) as Maithili and Odia, was selected as the high-resource *pivot*. To further justify our choice of the pivot, we examined the genetic distance between the pivot and target languages using lang2vec (Malaviya et al., 2017). Genetic distance (Bjerva et al., 2019) refers to the measure of divergence between languages based on their evolutionary relationship. The results showed that Bengali was quantifiably close to both target languages. The labeled Bengali speech was used to fine-tune the length adapter and encoder adapters individually and in combination. Separately, Kannada speech was used for length adapter fine-tuning to check if any benefits are obtained with an unrelated language. We also combined this with the text-only adaptation of target language text data to check if both approaches complement each other. Table 4 summarizes the performance of the cross-lingual systems with both the target low-resource languages. Length adapter fine-tuning outperforms encoder adaptation for cross-lingual transfer.

Language 1 (Target)	Language 2 (ASR Fine-tuning)	Genetic Distance	Text- only Adaptation	ASR fine-tuned Component	Number of Parameters	WER	CER
Maithili	None	-	No	None	-	82.2	43.39
	Bengali	0.625	No	Length Adapter	46M	79.77	40.04
			No	Encoder Adapter	50M	81.81	41.61
			No	Len. + Enc. Adapter	52M	80.81	40.44
			Yes	Length Adapter	6M+46M	**72.52**	**39.31**
	Kannada	1.000	No	Length Adapter	46M	80.29	38.37
			No	Encoder Adapter	50M	85.25	41.58
Odia	None	-	No	None	-	42.81	17.38
	Bengali	0.375	No	Length adapter	46M	41.05	15.07
			No	Encoder Adapter	50M	43.67	16.03
			No	Len. + Enc. Adapter	52M	42.4	15.27
			Yes	Length Adapter	6M+46M	**35.45**	**13.92**
	Kannada	1.000	No	Length Adapter	46M	41.21	14.08
			No	Encoder Adapter	50M	44.01	14.59

Table 4: **Results for cross-lingual transfer via ASR adaptation:** Comparison of WER(%) and CER(%) on low-resource languages with cross-lingual transfer through ASR adaptation of SeamlessM4T. The genetic distances between the (language 1, language 2) pairs suggest that Bengali is related to both the target languages; Kannada, despite being an Indic language, is genetically unrelated to both Maithili and Odia.

Additionally, we obtained an overall 17% reduction in relative WER for Odia, compared to the base model, by inserting decoder adapters fine-tuned on target language text data into the model whose length adapter was fine-tuned on Bengali ASR data. Thus, for low-resource languages without any speech data, ASR performance may be boosted by length adapter fine-tuning with a closely related pivot language coupled with text adaptation.

6 Discussion

We observe that for decoder adapters, it is more beneficial to use text-only adaptation compared to ASR-based training; the latter's benefit is mainly derived via the encoder layers. This emphasizes the role played by text data in improving the decoder's ability to enhance the internal language model of the ASR system. We also observed that 5-hour text data adaptation, having on average 92% fewer tokens than the full text, performed comparably to full-text data adaptation. This indicates that even limited amounts of text data can significantly boost ASR.

For a given target language with labeled speech, we found that fine-tuning the encoder adapters was the most accurate and parameter-efficient strategy. However, for cross-lingual zero-shot settings with no labeled data in a target language, we found it beneficial to fine-tune the length adapter with data in a related language rather than fine-tuning encoder adapters; the latter led to overfitting to the related language rather than enabling transfer to the target language. Text-based adaptation led to further improvements in the cross-lingual setting, indicating that even without speech data, ASR for low-resource languages can be improved by fine-tuning the length adapter. Lastly, a curious observation was that higher cross-lingual transfer was seen for genetically closer language pairs, with Odia-Bengali outperforming Maithili-Bengali in terms of relative WER reduction.

7 Conclusion

In this work, we explored the combination of parameter-efficient ASR fine-tuning and text-only adaptation techniques to enhance ASR for low-resource Indic languages using a multi-lingual multi-modal base model (SeamlessM4T). We find that a limited amount of text data was sufficient for adaptation, text-based adaptation was superior to ASR fine-tuning of decoder adapters, and encoder adapters were most effective in limited speech settings. In cross-lingual settings, however, the length adapter (and not the encoder adapter) was most successful, and text adaptation was additionally beneficial. Future work will focus on developing a better understanding of the interplay between different adapters within multimodal models.

8 Acknowledgements

The last author would like to gratefully acknowledge the support of the Amazon IITB AI ML Initiative and the consortium project on “Speech Technologies in Indian Languages" under National Language Translation Mission (NLTM), MeitY, Government of India.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech*.

Junwen Bai, Bo Li, Yu Zhang, Ankur Bapna, Nikhil Siddhartha, Khe Sim, and Tara Sainath. 2022. Joint unsupervised and supervised training for multilingual asr. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6402–6406.

Vladimir Bataev, Roman Korostik, Evgeny Shabalin, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Text-only domain adaptation for end-to-end asr using integrated text-to-mel-spectrogram generator. In *Interspeech*.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent?

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2023. When is multilinguality a curse? language modeling for 250 high- and low-resource languages.

Chang Chen, Xun Gong, and Yanmin Qian. 2023. Efficient text-only domain adaptation for ctc-based asr. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2020. *Ethnologue: Languages of the World*, 23 edition. SIL International, Dallas, Texas.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey.

Dan Hendrycks and Kevin Gimpel. 2023. Gaussian error linear units (gelus).

Wenxin Hou, Yidong Wang, Shengzhou Gao, and Takahiro Shinozaki. 2021. Meta-adapter: Efficient cross-lingual adaptation with meta-learning. In *ICASSP2021*, pages 7028–7032.

Wenxin Hou, Han Zhu, Yidong Wang, Jindong Wang, Tao Qin, Renjun Xu, and Takahiro Shinozaki. 2022. Exploiting adapters for cross-lingual low-resource speech recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:317–329.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

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 Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Qing Hu, Yan Zhang, Xianlei Zhang, Zongyu Han, and Xiuxia Liang. 2024. Language fusion via adapters for low-resource speech recognition. *Speech Communication*, 158:103037.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2023. UnitY: Two-pass direct speech-to-speech translation with discrete units. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15655–15680, Toronto, Canada. Association for Computational Linguistics.

Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vaijayanthi, Krishnan Karunganni, Pratyush Kumar, and Mitesh Khapra. 2024. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10740–10782, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Anjuli Kannan, Arindrima Datta, Tara Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Ankur Bapna, Zhifeng Chen, and Seungji Lee. 2019. Large-scale multilingual speech recognition with a streaming end-to-end model.

Samuel Kessler, Bethan Thomas, and Salah Karout. 2021. Continual-wav2vec2: an application of continual learning for self-supervised automatic speech recognition. *ArXiv*, abs/2107.13530.

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Proc. Interspeech 2021*, pages 1529–1533.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-tuning ASR models for very low-resource languages: A study on mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 170–176, Bangkok, Thailand. Association for Computational Linguistics.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

Ashish Mittal, Sunita Sarawagi, and Preethi Jyothi. 2023. In-situ text-only adaptation of speech models with low-overhead speech imputations. In *The Eleventh International Conference on Learning Representations*.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, interspeech$_2$020.$ISCA$.

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. 2023. Audiopalm: A large language model that can speak and listen.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022a. No language left behind: Scaling human-centered machine translation.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022b. No language left behind: Scaling human-centered machine translation.

Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadsy. 2021. Residual adapters for parameter-efficient ASR adaptation to atypical and accented speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6751–6760, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2018. Multilingual speech recognition with a single end-to-end model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 4904–4908. IEEE Press.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Tyler Vuong, Karel Mundnich, Dhanush Bekal, Veera Elluru, Srikanth Ronanki, and Sravan Bodapati. 2023. AdaBERT-CTC: Leveraging BERT-CTC for text-only domain adaptation in ASR. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 364–371, Singapore. Association for Computational Linguistics.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

Zhongzhi Yu, Yang Zhang, Kaizhi Qian, Cheng Wan, Yonggan Fu, Yongan Zhang, and Yingyan (Celine) Lin. 2023. Master-asr: achieving multilingual scalability and low-resource adaptation in asr with modular learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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.

Michael Zhang and Christopher Ré. 2024. Contrastive adapters for foundation model group robustness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jinming Zhao, Hao Yang, Gholamreza Haffari, and Ehsan Shareghi. 2022. M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation. In *Proc. Interspeech 2022*, pages 111–115.