# A Few-Shot Multi-Accented Speech Classification for Indian Languages using Transformers and LLM's Fine-Tuning Approaches

**Jairam R[1,2], Jyothish Lal G[1], Premjith B[1]**
[1]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
[2]RBG AI Research, RBG.AI, SREC Incubation Center, Coimbatore, India.
g_jyothishlal@cb.amrita.edu

## Abstract

Accented speech classification plays a vital role in the advancement of high-quality automatic speech recognition (ASR) technology. For certain applications, like multi-accented speech classification, it is not always viable to obtain data with accent variation, especially for resource-poor languages. This is one of the major reasons that contributes to the underperformance of the speech classification systems. Therefore, in order to handle speech variability in Indian language speaker accents, we propose a few-shot learning paradigm in this study. It learns generic feature embeddings using an encoder from a pre-trained whisper model and a classification head for classification. The model is refined using LLM's fine-tuning techniques, such as LoRA and QLoRA, for the six Indian English accents in the Indic Accent Dataset. The experimental findings show that the accuracy of the model is greatly increased by the few-shot learning paradigm's effectiveness combined with LLM's fine-tuning techniques. In optimal settings, the model's accuracy can reach 94% when the trainable parameters are set to 5%.

*Keywords : Accent-classification, few-shot, LoRA, QLoRA, LLM, Whisper, Dravidian Language*

## 1 Introduction

In this digital era, speech data has become a valuable resource, alongside text data. In the field of speech processing, recent developments in deep learning have made it possible to create end-to-end systems for tasks like speech classification and recognition. Much of the ongoing research in speech processing focuses on constructing end-to-end devices for automatic speech recognition (ASR) capable of handling diverse input data and providing accurate transcriptions. While ASR systems have shown remarkable performance in many cases, they face challenges when it comes to generalizing and adapting to resource-poor or resource-limited languages. Even though there are multilingual ASR and classification systems that have been trained on different Indian languages like Tamil, Kannada, Malayalam, Telugu, and others, their effectiveness is still lacking. This is due to the fact that speech is highly influenced by a variety of factors, such as the speaker's accent, gender, age, and more, and the lack of data covering all these variations (Bachate and Sharma, 2019; Malik et al., 2021).

Apart from gender, the speaker's accent (Huang et al., 2001) is recognized as the second most influential speech variation that affects the performance of speech recognition systems. An accent typically refers to a unique way of speaking or pronouncing a non-native language by a native speaker, influenced by the speaker's demographic background or geographical location. Accent classification, at its core, involves the categorization of regional or demographic accents within spoken language. The speaker's accent classification is considered a preliminary task for enhancing the capabilities of multilingual ASR systems.

The primary objective of this research is to tackle the aforementioned accent variations by introducing a data-driven approach to address the challenge of multi-accented speech classification. This paper proposes a few-shot learning method for multi-accented speech classification tasks as a means to handle the diverse array of accents effectively. The few-shot learning paradigm was chosen for this work due to its ability to learn from small amounts of data, which emphasizes how important it is for successfully tackling accent classification problems. The whisper ASR model serves as the foundational framework for the task of classifying accented speech. In this work, we primarily use the encoder component of the model, to which we attached a classification head, excluding the decoder



*Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 1–9

component.

Furthermore, we used two of the most popular large language model (LLM) adapters, such as Quantized Low-Rank Adaptation (QLoRA) and Low-Rank Adaptation (LoRA), to make the training and fine-tuning processes efficient and memory-friendly. In this work, we utilized the data derived from six Indian languages, sourced from the IndicAccentDB (Darshana et al., 2022) and NISP (Kalluri et al., 2021) datasets. Both of these datasets are significant for their inclusion of multi-accented speech, featuring conversations in English among native speakers.

Our main contributions are as follows:

- The pre-trained ASR model was employed in conjunction with LLM fine-tuning adapters such as Low-Rank Adaptations (LoRA) and Quantized Low-Rank Adaptation (QLoRA) in order to categorize individuals' accents.
- Extensive experiments on the combined IndicAccentDB, NISP, and Gujarati Digits datasets have been conducted to show the efficacy of LLM's fine-tuning techniques. These experiments involve reducing the trainable parameters by setting different low-rank values ranging from 2 to 32.
- The significance of the few-shot learning paradigm was demonstrated by obtaining an average accuracy of 90% under LoRA and 93.3% under QLoRA.
- A multi-class accent classification task using few-shot learning paradigm has been demonstrated using only 2 hours and 30 minutes of training data and observed to have 94% accuracy in the optimal setting where the training parameters were reduced to 5% using LLM's adapters.

The paper contribution is detailed in the sections that follow, which are arranged as follows: Section 2 describes the related works, while Section 3 completely describes the methodology used in this work. Section 4 holds the results and discussion, and we conclude the work in Section 5.

## 2 Related Works

Previous research has extensively investigated how various components of speech change with accents. Notably, spectral features like formant frequencies and temporal features such as intonation and duration's exhibit variation with accent (Arslan and Hansen, 1997; Ferragne and Pellegrino, 2010). To automate accent classification, these features have been combined into different statistical models and machine learning techniques.

Historically, Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) have been commonly used in accent classification (Deshpande et al., 2005; Ghesquiere and Van Compernolle, 2002; Zheng et al., 2005). Some studies looked at how the number of GMM components affected the performance of classification (Chen et al., 2001), while others compared HMMs to Support Vector Machines (SVM) (Tang and Ghorbani, 2003). Further studies explored the impact of GMM component numbers on classification performance (Chen et al., 2001). In (Pedersen and Diederich, 2007), SVM has been used to classify Arabic and Indian accents using the most popular Mel-Frequency Cepstral Coefficients (MFCCs) as the speech feature. Linear models, such as linear discriminant analysis (LDA), have also found applications, as seen in the identification of accents in Australian English (Kumpf and King, 1997).

Conventional statistical models and machine learning models have played a crucial role in accented speech classification. However, deep learning frameworks like deep neural networks (DNNs) and recurrent neural networks (RNNs) have been widely used in the latest speech recognition and synthesis systems (Hinton et al., 2012; Zen and Sak, 2015; Xu et al., 2014; Jiao et al., 2016; Lalitha et al., 2019). Notably, for the accent classification task, Telugu dialect datasets were created and classified using variants of RNN models like LSTM, GRU, and BiLSTM with attention layers (Podila et al., 2022).

However, there have been fewer studies evaluating neural networks for accent identification (Chan et al., 1994) and (Rabiee and Setayeshi, 2010). Nevertheless, in related areas like language identification (LID), neural networks have been thoroughly investigated (Montavon, 2009; Cole et al., 1989; Lopez-Moreno et al., 2014). As a breakthrough, convolutional neural networks (CNNs) and gated recurrent units (GRUs) (Tzudir et al., 2021) have been combined to classify accents with approximately 6 hours of speech data for resource-poor languages. Moreover, transformers have been a crucial breakthrough in both the natural language processing (NLP) and speech processing domains.

Most of the transformer-based studies (Shi et al., 2021; Gao et al., 2021) have been mostly conducted on accented speech recognition and not on the accent classification task. Recently, the few-shot paradigm has gained more popularity among researchers because of its ability to learn from a limited amount of data, which resolves the issues with neural net networks, which require a lot of data to train the model (Shrestha and Mahmood, 2019). To the best of our knowledge, the efficacy of few-shot approaches in accent classification is still unknown, despite the fact that these approaches have been used in audio processing (Keshav et al., 2023; Chou et al., 2018; Arik et al., 2018; Anand et al., 2019).

Based on the existing literature, it is evident that while extensive research has been conducted in the field of speech processing techniques, there remains a significant gap in the development of a system that effectively addresses accent variation and performs classification. In our research, we aim to bridge this gap by introducing a novel approach that utilizes a few-shot learning paradigm for accent classification. To the best of our knowledge, this is the first work that poses multi-accented speech classification as a few-shot learning problem to address the diversity in speech variations caused by speakers' accents. This approach is designed to identify the accents of native speakers from spoken non-native English speech datasets.

## 3 Materials and Methodology

### 3.1 Datasets

The IndicAccentDB was first presented in the work MARS (Darshana et al., 2022), where a hybrid CNN was used in multi-accented English speech recognition. IndicAccentDB is comprised of audio recordings containing six English accents spoken by non-native speakers, each originating from six different Indian languages such as Tamil, Telugu, Malayalam, Hindi, Gujarati, and Hindi. Within the dataset, 19 speakers were asked to recite sentences from the Harvard sentences dataset, which is renowned for its phonetically balanced and gender-balanced content (Huang et al., 2001). There are 72 sets in the Harvard Sentences dataset, and each set has 10 moderately long sentences. Together, these 19 speakers contribute 8,180 speech utterances to the IndicAccentDB. The average length of the audio files is about 5 seconds each. The detailed data statistics for the IndicAccentDB are presented in Table 1.

| IndicAccentDB | |
|---|---|
| Accents | No. of Recordings |
| Tamil | 1,640 |
| Malayalam | 1,563 |
| Telugu | 1,614 |
| Gujarati | 298 |
| Hindi | 827 |
| Kannada | 1,486 |

Table 1: Data Statistics for IndicAccentDB corpus

Apart from the IndicAccentDB, we also incorporated publicly available datasets such as NISP (Kalluri et al., 2021) and Gujarati Digits (Dalsaniya et al., 2020). The NISP corpus encompasses speech recordings in five native Indian languages: Tamil, Kannada, Malayalam, Hindi, Telugu, and Indian-accented English. This corpus comprises recordings from 345 speakers, including 126 females and 219 males. It contains a total of 28,268 speech utterances, with 14,691 in English and 13,577 in native languages. In this work, we focused on a subset of the Indian-accented English utterances within this dataset.

Furthermore, the Gujarati digits (Dalsaniya et al., 2020) corpus is specifically designed to support speech recognition systems and features distinct recordings of Gujarati digits. These recordings were collected from various regions of Gujarat, including the Saurashtra, North Zone, South Zone, Central Zone, and Kutch Region, encompassing diverse environmental conditions and background noises. This dataset contains a total of 1,940 speech utterances from 20 different speakers. Table 2 provides a more detailed overview of the data statistics for both the NISP and Gujarati Digits datasets used in this work.

| Corpus | Accents | No. of Recordings |
|---|---|---|
| NISP | Tamil | 280 |
| | Malayalam | 187 |
| | Telugu | 167 |
| | Kannada | 233 |
| | Hindi | 276 |
| Gujarati Digits | Gujarati | 250 |

Table 2: Data Statistics for NISP and Gujarati Digits corpus

In this study, we utilized a subset of the Gujarati Digits dataset in conjunction with the Indi-

cAccentDB and a subset of the NISP datasets. For multi-class classification, a total of six labels were employed for the six Indian languages. As part of the dataset's pre-processing, the audio files were re-sampled to 16 kilohertz. Then, a 400-point Fourier transform was used to make an 80-channel log Mel spectrogram for a 25-millisecond period with a 10-millisecond step. The resulting spectrogram was then used as input for the model's training.

### 3.2 Proposed Methodology

#### 3.2.1 Model Architecture

Whisper (Radford et al., 2023), an advanced automatic speech recognition (ASR) model, currently stands as the state-of-the-art (SOTA) in speech recognition. Trained on an extensive dataset comprising 680,000 hours of multilingual and multitask-labeled data sourced from the web, it adopts a transformer architecture. The model involves both encoder and decoder components to process audio files and generate corresponding textual outputs.

In this study, we leveraged the pre-trained whisper-large-v2 ASR model for a classification task. This variant of the whisper is a multi-lingual model with 1550 million parameters. The encoder within the whisper model undergoes a specific architectural sequence: initial processing through a short stem consisting of two convolution layers, utilizing a filter width of 3, and activation by the GELU activation function. The second convolution layer introduces a stride of two. Following this, sinusoidal position embeddings are incorporated into the stem's output.

Subsequently, the encoder applies transformer blocks. Notably, the transformer employs pre-activation residual blocks. A concluding layer normalization step is then applied to the output of the encoder. The decoder component was omitted, as the absence of a transcription task negated its necessity. Instead, we utilized a classification head on top of the encoder to facilitate classification tasks. Then this model is optimized using the LLM fine-tuning adapters and trained as seen in Figure 1.

#### 3.2.2 LLM's Fine-Tuning Techniques

Recently, large language models (LLMs) have gained increased attention within the research community, particularly in the field of natural language processing (NLP) applications. This surge in popularity has led to a gradual expansion of their utility into other domains, including computer vision and speech. This paradigm involves large-scale pre-training on diverse web data, followed by fine-tuning for specific downstream tasks. However, fine-tuning LLMs for such tasks often necessitates substantial computing resources, rendering them inaccessible to many.

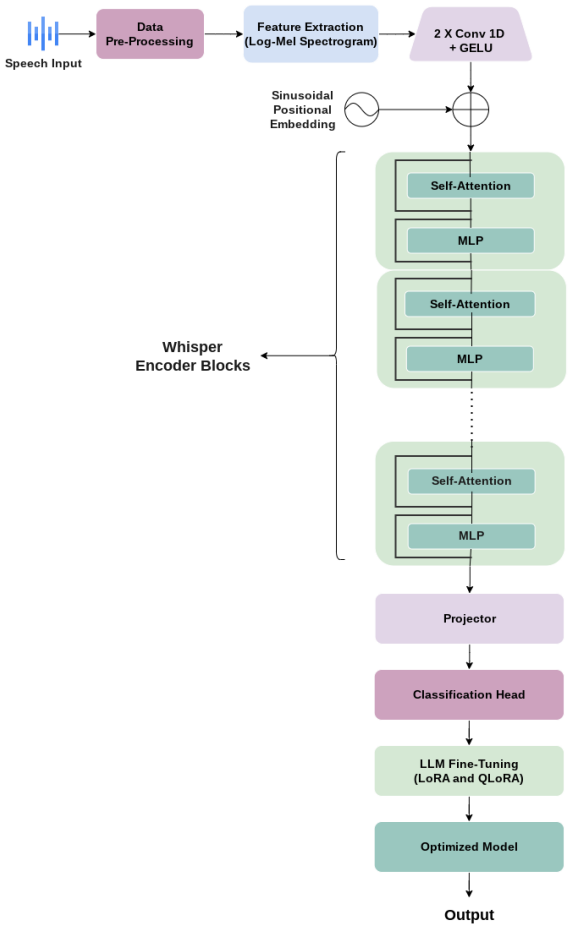


Figure 1: The proposed model includes data processing, feature extraction, encoder block of Whisper, and a classification head added on top of it, which is followed by LLM's fine-tuning techniques.

Parameter Efficient Fine-Tuning (PEFT) (Liu et al., 2022) addresses this issue by loading and fine-tuning the model in a memory-efficient manner while ensuring the model's performance. Despite the fact that these methods were initially applied to language models, they could be modified to improve the usability and accessibility of sophisticated models like Whisper in a variety of domains, including speech processing.

The PEFT methodology proves invaluable in fine-tuning LLMs. It achieves this by selectively

fine-tuning only a small subset of parameters in a pre-trained model, significantly mitigating computational and storage expenses. In our work, we leverage two popular PEFT methods: LoRA (Low-Rank Adaptation) (Hu et al., 2021) and its evolution, QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2021), for the fine-tuning of the Whisper language model.

In the context of training the neural networks, the most fundamental procedure involves the iterative updating of weight matrices through the use of gradient descent. Nevertheless, when it comes to LLMs, the process of updating the weight matrix (*Wo*) in the pre-trained model presents challenges in terms of both compute and memory resource use.

LoRA (Hu et al., 2021) effectively addresses this by introducing two low-rank matrices (*B and A*) as an update matrix that approximates the large weight matrix. During training, *Wo* remains fixed without receiving gradient updates, while the smaller matrices *B* and *A* house the trainable parameters for LLMs fine-tuning. Consequently, when inputs are processed, they undergo multiplication by both *Wo* and the newly introduced update matrices (*B and A*), with the loss computed by aggregating the output vectors of *Wo, B, and A*.

QLoRA (Dettmers et al., 2021) is a further enhanced version of LoRA with improved memory efficiency. In QLoRA, the pre-trained model is loaded onto GPU memory using quantized 4-bit weights, a notable advancement from the 8-bit weights employed in LoRA. Importantly, QLoRA maintains comparable effectiveness to its predecessor, LoRA.

## 4 Results and Discussion

### 4.1 Experiments

The experiments were carried out in a hardware environment equipped with a T4-XLarge, 4 cores, 16 GB of RAM, 1 GPU, and 40 GB of disk space. We conducted experiments utilizing the 'whisper-large-v2' model with two Large Language Model (LLM) settings: one employing the LoRA adapter and the other employing the QLoRA adapter.

Following preprocessing, the audio files underwent segmentation into training, testing, and validation sets. The specific details of this segmentation are outlined in Table 3. The whisper-large-v2 model was loaded with 8-bit precision into memory using the INT8 and bitsandbytes Python libraries during training with LoRA. On the other hand double-quantization was used to load the model with 4-bit precision in QLoRA, and a datatype of NF4 (normalfloat4) was used to reduce perplexity.

| Language | Total Recordings | Train | Test | Validation |
|---|---|---|---|---|
| Tamil | 1,920 | 403 | 1,355 | 162 |
| Malayalam | 1,750 | 201 | 1,415 | 156 |
| Telugu | 1,781 | 437 | 1,188 | 134 |
| Kannada | 1,719 | 229 | 1,309 | 181 |
| Gujarati | 548 | 205 | 233 | 110 |
| Hindi | 1,103 | 289 | 674 | 140 |
| Total | 8,821 | 1,764 | 6,174 | 883 |

Table 3: Train, Test, and Validation split statistics

The whisper model was trained under both LoRA and QLoRA configurations, employing varying rank projections from 2, 4, 8, 16, 24, and 32 with a training epoch of 10. The training process utilized a batch size of 8 and a learning rate of 10-3. During training, around 2 hours and 30 minutes of data were used, while testing was conducted on approximately 8 hours of data.

### 4.2 Discussion

The effectiveness of the whisper model along with the LLM's fine-tuning techniques in multi-accented speech classification has been evaluated in this section through qualitative analysis of corpora (Darshana et al., 2022; Kalluri et al., 2021; Dalsaniya et al., 2020). Precision, Recall, F1-Score, and Accuracy are the primary metrics we use in our evaluation. Precision measures how often the model correctly predicts positive samples among all positive predictions. Recall measures the accuracy of the model's positive predictions among the actual positive samples. Accuracy measures the total number of correct predictions made by the model for the entire corpus, whereas Precision and Recall are combined to score the model's accuracy for each class in the F1-Score.

| Rank | LoRA (Accuracy) | QLoRA (Accuracy) |
|---|---|---|
| 32 | 93% | 96% |
| 24 | 91% | 95% |
| 16 | 90% | 95% |
| 8 | 85% | 94% |
| 4 | 89% | 86% |
| **2** | **92%** | **94%** |

Table 5: Accuracy comparison between LoRA and QLoRA

| Rank and Trainable Parameters | Languages | LoRA | | | QLoRA | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| Rank - 32, Trainable Parameters-80% | Tamil | 0.93 | 0.96 | 0.95 | 0.97 | 0.95 | 0.96 |
| | Malayalam | 0.95 | 0.93 | 0.94 | 0.96 | 0.96 | 0.96 |
| | Telugu | 0.93 | 0.92 | 0.92 | 0.96 | 0.96 | 0.96 |
| | Kannada | 0.95 | 0.98 | 0.97 | 0.96 | 0.98 | 0.97 |
| | Gujarati | 0.89 | 0.77 | 0.82 | 0.95 | 0.93 | 0.94 |
| | Hindi | 0.90 | 0.88 | 0.89 | 0.91 | 0.92 | 0.92 |
| Rank - 24, Trainable Parameters-61% | Tamil | 0.93 | 0.94 | 0.93 | 0.96 | 0.96 | 0.96 |
| | Malayalam | 0.90 | 0.95 | 0.93 | 0.95 | 0.97 | 0.96 |
| | Telugu | 0.90 | 0.89 | 0.90 | 0.96 | 0.93 | 0.95 |
| | Kannada | 0.93 | 0.96 | 0.95 | 0.96 | 0.97 | 0.97 |
| | Gujarati | 0.88 | 0.70 | 0.78 | 0.99 | 0.86 | 0.92 |
| | Hindi | 0.85 | 0.76 | 0.80 | 0.88 | 0.92 | 0.90 |
| Rank-16, Trainable Parameters-40% | Tamil | 0.91 | 0.91 | 0.91 | 0.95 | 0.95 | 0.95 |
| | Malayalam | 0.91 | 0.93 | 0.92 | 0.95 | 0.95 | 0.95 |
| | Telugu | 0.86 | 0.86 | 0.86 | 0.93 | 0.92 | 0.93 |
| | Kannada | 0.95 | 0.96 | 0.96 | 0.97 | 0.97 | 0.97 |
| | Gujarati | 0.84 | 0.58 | 0.69 | 0.95 | 0.93 | 0.94 |
| | Hindi | 0.81 | 0.83 | 0.82 | 0.91 | 0.94 | 0..92 |
| Rank-8, Trainable Parameters-20% | Tamil | 0.82 | 0.90 | 0.86 | 0.94 | 0.98 | 0.96 |
| | Malayalam | 0.88 | 0.96 | 0.92 | 0.96 | 0.94 | 0.95 |
| | Telugu | 0.82 | 0.75 | 0.79 | 0.93 | 0.93 | 0.93 |
| | Kannada | 0.90 | 0.94 | 0.92 | 0.96 | 0.97 | 0.96 |
| | Gujarati | 0.76 | 0.41 | 0.53 | 0.96 | 0.83 | 0.89 |
| | Hindi | 0.79 | 0.67 | 0.73 | 0.91 | 0.88 | 0.90 |
| Rank-4, Trainable Parameters-10% | Tamil | 0.86 | 0.91 | 0.89 | 0.90 | 0.91 | 0.91 |
| | Malayalam | 0.87 | 0.95 | 0.91 | 0.88 | 0.87 | 0.87 |
| | Telugu | 0.93 | 0.84 | 0.88 | 0.84 | 0.85 | 0.85 |
| | Kannada | 0.93 | 0.96 | 0.94 | 0.92 | 0.92 | 0.92 |
| | Gujarati | 0.93 | 0.73 | 0.82 | 0.66 | 0.62 | 0.64 |
| | Hindi | 0.86 | 0.77 | 0.81 | 0.69 | 0.70 | 0.70 |
| Rank-2, Trainable Parameters-5% | Tamil | 0.93 | 0.94 | 0.93 | 0.96 | 0.95 | 0.95 |
| | Malayalam | 0.91 | 0.95 | 0.93 | 0.95 | 0.96 | 0.95 |
| | Telugu | 0.94 | 0.86 | 0.90 | 0.93 | 0.92 | 0.93 |
| | Kannada | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| | Gujarati | 0.91 | 0.75 | 0.82 | 0.93 | 0.84 | 0.88 |
| | Hindi | 0.78 | 0.86 | 0.82 | 0.85 | 0.90 | 0.87 |

Table 4: Multiclass Classification Report for LoRA and QLoRA.

Table 4 shows the outcomes of multi-class classification using LoRA and QLoRA adapters at different rank values. Notably, the whisper model performs well in both LoRA and QLoRA settings across most rank values. Particularly, it excels when double quantized and operating in 4-bit precision under QLoRA settings.

The rank values in these adapters represent the low-rank matrix dimension learned during fine-tuning, impacting the model's trainable parameters. The rank values parameter in both LoRA and QLoRA is used to reduce the number of trainable parameters. Reducing the trainable parameters minimizes the computational cost and memory usage of the model. Optimal performance occurs at a rank value of 2 for both LoRA and QLoRA, utilizing only 5% of trainable parameters compared to the pre-trained model's total parameters.

Table 5 highlights significantly improved performance for both LoRA and QLoRA at this optimal rank value of 2, achieving an overall accuracy of 92% and 94%, respectively. Throughout the experiments, the model consistently performs better when trained on roughly two and a half hours of speech data, emphasizing the significance of the few-shot learning paradigm.

## 5 Conclusion and Future Works

In this work, we aimed to mitigate the impact of accent variation on speech classification systems. Our approach leveraged a data-driven method employing few-shot learning to perform multi-accented speech classification across six Indian language speakers' English accents. This was achieved by utilizing the IndicAccentDB alongside subsets from the NISP and Gujarati Digits Corpora, utilizing the pre-trained whisper ASR model.

Furthermore, we demonstrated the efficacy of LLM fine-tuning techniques such as LoRA and QLoRA, which make it possible to fine-tune large language models in a manner that is both memory-efficient and computationally efficient. As can be shown in Table 5, our studies produced remarkable overall accuracies of 92% and 94% when the settings were optimized.

Future endeavors will focus on encompassing other speech variations, such as age group and gender, using the few-shot learning paradigm. This approach will be particularly valuable in scenarios where data availability for these diverse attributes is limited, continuing to enhance the robustness of speech classification systems.

## References


Prashant Anand et al. 2019. Few shot speaker recognition using deep neural networks. *arXiv preprint arXiv:1904.08775*.

S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou. 2018. Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*, pages 10019–10029.

L. M. Arslan and J. H. Hansen. 1997. Frequency characteristics of foreign accented speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1123–1126, Munich, Germany. IEEE.

R. P. Bachate and A. Sharma. 2019. Automatic speech recognition systems for regional languages in india. *International Journal of Recent Technology and Engineering*, 8(2):585–592.

M. V. Chan, X. Feng, J. A. Heinen, and R. J. Niederjohn. 1994. Classification of speech accents with neural networks. In *Proceedings of the IEEE World Congress on Computational Intelligence, IEEE International Conference on Neural Networks*, volume 7, pages 4483–4486. IEEE.

T. Chen, C. Huang, E. Chang, and J. Wang. 2001. Automatic accent identification using gaussian mixture models. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 343–346, Madonna di Campiglio, Italy. IEEE.

S.-Y. Chou, K.-H. Cheng, J.-S. R. Jang, and Y.-H. Yang. 2018. Learning to match transient sound events using attentional similarity for few-shot sound recognition. *arXiv preprint arXiv:1812.01269*.

R. A. Cole, J. W. Inouye, Y. K. Muthusamy, and M. Gopalakrishnan. 1989. Language identification with neural networks: A feasibility study. In *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 525–529. IEEE.

Nikunj Dalsaniya et al. 2020. Development of a novel database in gujarati language for spoken digits classification. In *Advances in Signal Processing and Intelligent Recognition Systems: 5th International Symposium, SIRS 2019, Trivandrum, India, December 18–21, 2019, Revised Selected Papers 5*. Springer Singapore.

S. Darshana, H. Theivaprakasham, G. Jyothish Lal, B. Premjith, V. Sowmya, and K. Soman. 2022. Mars: A hybrid deep cnn-based multi-accent recognition system for english language. In *Proceedings of the First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR)*, pages 1–6, Hyderabad, India.

S. Deshpande, S. Chikkerur, and V. Govindaraju. 2005. Accent classification in speech. In *Proceedings of the Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pages 139–143, Buffalo, NY, USA. IEEE.

T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.

E. Ferragne and F. Pellegrino. 2010. Formant frequencies of vowels in 13 accents of the british isles. *Journal of the International Phonetic Association*, 40(01):1–34.

Qiang Gao et al. 2021. An end-to-end speech accent recognition method based on hybrid ctc/attention transformer asr. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

P.-J. Ghesquiere and D. Van Compernolle. 2002. Flemish accent identification based on formant and duration features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I–749, Orlando, FL, USA. IEEE.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. Lora: Low-rank adaptation of large language models. Preprint or Unpublished.

C. Huang, E. Chang, and T. Chen. 2001. Accent issues in large vocabulary continuous speech recognition. In *Proceedings of ACL*, Beijing, China. Microsoft Research China. Technical Report MSR-TR-2001-69.

Y. Jiao, M. Tu, V. Berisha, and J. Liss. 2016. Online speaking rate estimation using recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China. IEEE.

Shareef Babu Kalluri et al. 2021. Nisp: A multi-lingual multi-accent dataset for speaker profiling. In *Proceedings of ICASSP*. IEEE.

S. Keshav, G. Jyothish Lal, and B. Premjith. 2023. Multimodal approach for code-mixed speech sentiment classification. In *Proceedings of Seventh ICMEET-2022: Advances in Signal Processing, Embedded Systems and IoT*, pages 553–563, Singapore. Springer Nature Singapore.

K. Kumpf and R. W. King. 1997. Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In *Proceedings of EuroSpeech*, volume 4, pages 2323–2326.

S. Lalitha, Shikha Tripathi, and Deepa Gupta. 2019. Enhanced speech emotion detection using deep neural networks. *International Journal of Speech Technology*, 22:497–510.

H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965.

I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno. 2014. Automatic language identification using deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5337–5341, Florence, Italy. IEEE.

Mishaim Malik et al. 2021. Automatic speech recognition: A survey. *Multimedia Tools and Applications*, 80:9411–9457.

G. Montavon. 2009. Deep learning for spoken language identification. In *Proceedings of the NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, pages 1–4, Whistler, BC, Canada.

C. Pedersen and J. Diederich. 2007. Accent classification using support vector machines. In *Proceedings of the 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*, pages 444–449. IEEE.

Rama Sai Abhishek Podila et al. 2022. Telugu dialect speech dataset creation and recognition using deep learning techniques. In *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE.

A. Rabiee and S. Setayeshi. 2010. Persian accents identification using an adaptive neural network. In *Proceedings of the Second International Workshop on Education Technology and Computer Science*, pages 7–10, Wuhan, China. IEEE.

Alec Radford et al. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR.

Xian Shi et al. 2021. The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Ajay Shrestha and Ausif Mahmood. 2019. Review of deep learning algorithms and architectures. *IEEE access*, 7:53040–53065.

H. Tang and A. A. Ghorbani. 2003. Accent classification using support vector machine and hidden markov model. In *Advances in Artificial Intelligence*, pages 629–631. Springer.

M. Tzudir, S. Baghel, P. Sarmah, and S. M. Prasanna. 2021. Excitation source feature based dialect identification in ao — a low resource language. In *Proceedings of Interspeech 2021*, pages 1524–1528.

Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee. 2014. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 21(1):65–68.

H. Zen and H. Sak. 2015. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474, Brisbane, Australia. IEEE.

Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon. 2005. Accent detection and speech recognition for shanghai-accented mandarin. In *Proceedings of Interspeech*, pages 217–220, Lisbon, Portugal. Citeseer.