Multi-Accent Recognition Using Deep Learning Techniques

*A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in The Department of Computer Science &Engineering**

**23AVI3101A-DEEP LEARNING**

Submitted by

**2310030024 : Swetha Oruganti**

Under the guidance of

**Dr. Sumit Hazra**



Department of Computer Science and Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

FEB - 2025.

**Abstract**

The rapid proliferation of voice-activated technologies such as virtual assistants, smart devices, and conversational AI systems has amplified the demand for highly robust Automatic Speech Recognition (ASR) systems. One of the most pressing challenges in this domain is the accurate recognition of speech across diverse global accents. Traditional ASR systems, often trained on limited and accent-biased datasets, tend to exhibit significant performance degradation when exposed to regional or non-standard accents. This limitation not only reduces usability but also raises concerns about inclusivity and equitable access to speech-driven technologies.

Recent advances in deep learning have opened new possibilities for addressing accent variability. End-to-End (E2E) architectures, which directly map acoustic signals to text without relying on handcrafted features, have demonstrated promising results in handling accent shifts. Additionally, multi-task learning frameworks allow ASR models to jointly learn accent classification and speech recognition, thereby improving robustness. Accent embeddings further enhance adaptability by encoding speaker-specific accent features that guide the recognition process.

Moreover, adaptive layer-wise fine-tuning has been explored to adjust pretrained ASR models for accent-specific data without requiring full retraining.Beyond these techniques, researchers are increasingly focusing on large-scale multilingual and multi-accent datasets to improve generalization. Data augmentation strategies, such as accent conversion and synthetic speech generation, are also being used to expose models to wider linguistic variations. Transfer learning has emerged as another powerful tool, enabling ASR systems trained on high-resource languages to adapt more effectively to low-resource accents. Experimental evidence shows measurable reductions in Word Error Rate (WER) and Character Error Rate (CER) across multi-accent benchmarks when these methods are applied, highlighting their practical value.

Looking forward, the development of real-time, accent-agnostic ASR systems remains a crucial research direction. Such systems must not only maintain high recognition accuracy but also operate efficiently under computational constraints, particularly for deployment on edge devices. Achieving this will enable seamless global communication, fostering accessibility, inclusivity, and ensuring that voice-driven technologies can truly serve diverse populations worldwide.

**Introduction**

Automatic Speech Recognition (ASR) has transitioned from a niche technology to a cornerstone of human-computer interaction, powering virtual assistants, transcription services, and hands-free control. However, a critical and persistent challenge remains: performance inconsistency across different accents. An accent encompasses variations in pronunciation, intonation, and rhythm influenced by a speaker's native language, region, or social background. Mainstream ASR systems, often trained predominantly on "standard" accent data (e.g., General American or Received Pronunciation English), suffer from significantly higher Word Error Rates (WER) for speakers with other accents. This bias not only degrades user experience but also exacerbates technological inequity, excluding vast populations from its benefits.

Deep Learning has revolutionized ASR by moving away from traditional pipeline-based models (which handle feature extraction, acoustic modeling, and language modeling separately) to integrated End-to-End (E2E) models. These models, such as Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Transformers, learn to map raw audio waveforms or features directly to text sequences. This paradigm is exceptionally well-suited for multi-accent recognition due to its ability to learn complex, hierarchical representations of speech. Deep learning facilitates the development of models that can discover and leverage latent patterns common to all speech, while also identifying and adapting to accent-specific peculiarities.

This report aims to provide a comprehensive overview of how deep learning techniques are being employed to build robust multi-accent ASR systems. We will review key architectural innovations and training strategies, present recent experimental evidence of their success, propose a detailed methodology for such a project, and conclude with future research trajectories. The ultimate goal is to underscore the potential of these approaches in creating inclusive and universally accessible speech technologies.

**Literature Review/** **Application Survey**

2.1. End-to-End (E2E) Architectures  
E2E models like Listen, Attend and Spell (LAS), Connectionist Temporal Classification (CTC)-based models, and RNN-Transducers have simplified ASR pipelines. For multi-accent tasks, their strength lies in their ability to directly model the relationship between acoustic signals and phonetic units across different accent realizations, without relying on hand-crafted pronunciation dictionaries that may be accent-biased.

2.2. Adaptive and Accent-Specific Layers  
A common and effective strategy involves designing networks with shared lower layers and adaptive higher layers. The shared layers learn universal speech features (e.g., phonemes, basic acoustics), while the accent-specific layers (often a dedicated layer or a small subnet) are fine-tuned to handle the pronunciation variations of a particular accent. This allows for efficient adaptation to new accents with limited data.

2.3. Advanced Techniques: Multi-Task Learning and Joint Modeling  
Multi-task learning (MTL) forces the model to learn representations that are useful for multiple related tasks simultaneously. In multi-accent ASR, a primary task is speech recognition (transcription), and an auxiliary task can be accent identification. By jointly learning to recognize both what is said and how it is accented, the model develops more robust and generalized features, improving performance on both tasks.

2.4. Representing Accent: i-vectors and Accent Embeddings  
To explicitly provide the model with accent information, researchers use compact numeric representations. i-vectors (originally from speaker verification) are low-dimensional vectors that represent the accent characteristics of an utterance. Accent embeddings are similar learned vectors that are extracted by a neural network. These vectors can be concatenated with the acoustic features at the input of the model, conditioning the entire recognition process on the estimated accent, leading to a significant boost in generalization.

**Techniques to Improve Accent Robustness & Recent Advances**

**3.1. Multi-Domain Training**

One of the most straightforward yet highly effective methods to improve accent robustness is multi-domain training, where models are trained on large and diverse datasets encompassing a wide variety of accents, dialects, and speaking styles. The inclusion of such heterogeneous data prevents the model from overfitting to a single accent domain and equips it with the ability to generalize across multiple variations of the same language. For instance, exposure to American, British, Indian, and African English ensures that the system does not develop a bias toward any single variety. Studies have demonstrated that this approach can reduce Word Error Rate (WER) by up to **25%** on accented speech. Furthermore, large-scale multilingual corpora and crowd-sourced speech collections have made this approach increasingly feasible, enabling ASR systems to move closer to global inclusivity.

**3.2. Layer-wise Adaptation**

Instead of retraining the entire network for every new accent, layer-wise adaptation focuses on fine-tuning only a small subset of layers—often referred to as accent-adaptive layers. This technique is both **computationally efficient** and **data-efficient**, as it minimizes the amount of new accent data required. By freezing the general layers and updating only accent-specific ones, models can quickly adapt to unfamiliar pronunciations and prosodic patterns. Empirical studies have shown that this technique yields around **15% reductions in WER**, making it an attractive option for real-world deployment, particularly in low-resource environments where full retraining is impractical.

**3.3. Leveraging Native Language Data**

For non-native speakers, accents are heavily influenced by their first language (L1). For example, a Spanish speaker learning English may substitute certain sounds, such as pronouncing “v” as “b.” By leveraging L1 data, ASR systems can be trained to anticipate such transfer errors. Multi-task learning frameworks often integrate L1 recognition tasks, where the model simultaneously learns to predict the native language characteristics alongside transcription. This additional supervision provides strong cues for handling accent-specific deviations, significantly boosting recognition performance. Such approaches are particularly useful in global communication platforms where non-native speech dominates.

**Recent Advances and Experimental Results**

**1. Qifusion-Net (2024)**

A breakthrough in accent-robust ASR was introduced with **Qifusion-Net (2024)**, an architecture that employs advanced fusion mechanisms to integrate information from multiple acoustic and linguistic streams. Unlike traditional E2E models that rely on a single feature representation, Qifusion-Net dynamically balances contributions from different modalities, such as phonetic cues, intonation, and contextual embeddings. This architecture achieved an impressive **17–22% relative reduction in Character Error Rate (CER)** on widely used multi-accent benchmarks, thereby setting a new state-of-the-art standard for accent adaptation.

**2. Multi-Accent DNN Models**

Deep Neural Networks (DNNs) enhanced with accent-specific layers have shown significant promise for handling large intra-language variations. For example, in both Mandarin and English, regional dialects introduce considerable phonetic variability. By incorporating accent-dependent parameters into DNN models, researchers have reported tangible accuracy improvements without compromising model generalization. This approach effectively allows the model to share knowledge across accents while still preserving flexibility for accent-specific nuances.

**3. Joint Speech & Accent Recognition**

A more recent line of research focuses on **joint speech and accent recognition frameworks**, typically built on DNN-HMM hybrid architectures. These systems simultaneously recognize speech content and classify accent identity in **real time**, allowing for dynamic adaptation during interaction. For instance, a conversational AI could instantly detect that a user speaks English with a French accent and subtly adjust its decoding strategy on-the-fly. This dual-task approach not only improves accuracy but also reduces latency, which is critical for real-time applications like customer service bots and virtual assistants.

**4.4. Multi-Domain Fine-Tuning**

Fine-tuning ASR models across multiple accent domains has proven to be an effective strategy for ensuring adaptability in out-of-domain scenarios. For example, models trained on diverse English accents (American, British, Australian, and Indian) and later tested on French-accented English exhibited significant performance gains, demonstrating robustness beyond the training set. This adaptability indicates that deep learning models, when properly fine-tuned, can generalize across unseen accent conditions, a key requirement for building truly universal ASR systems.

**Adopted Methodology**

1. Data Acquisition and Preprocessing:

* Datasets: Utilize multiple diverse datasets (e.g., Common Voice, L2-ARCTIC, ACCENT) containing speech from various accent groups.
* Input: Raw audio waveforms or standard Mel-Frequency Cepstral Coefficients (MFCCs)/Filterbank features will be extracted.
* Preprocessing: Include normalization, silence trimming, and possibly speed perturbation for data augmentation.

2. Model Architecture Design:

A transformer-based E2E architecture will be adopted as the core model. Key modifications will include:

* An accent embedding layer that learns a vector for each accent ID in the training data.
* These embeddings will be concatenated with the acoustic features before being fed into the encoder stack.
* A multi-task learning setup where the model has two output heads: one primary head for grapheme/phoneme prediction (CTC loss) and an auxiliary head for accent classification (Cross-Entropy loss).

3. Training Strategy:

1. Pre-training: Train the entire model on the large, multi-accent dataset.
2. Fine-tuning: For specific low-resource accents, employ layer-wise adaptation by freezing the shared encoder layers and only fine-tuning the accent embedding layer and the final classification layers.

4. Evaluation Metrics:

The model's performance will be rigorously evaluated using:

* Word Error Rate (WER): The primary metric for transcription accuracy.
* Character Error Rate (CER): A finer-grained metric.
* Accent Identification Accuracy: To evaluate the auxiliary task performance.  
  These metrics will be calculated per-accent to identify specific model weaknesses

**Expected Outcomes and Discussion**

The proposed methodology is expected to yield a highly robust multi-accent ASR system capable of handling speech variations from a wide range of speakers across global contexts. By integrating accent embeddings and multi-task learning (MTL), we anticipate a substantial reduction in both Word Error Rate (WER) and Character Error Rate (CER) across all accent groups when compared to baseline models that rely solely on traditional supervised training without accent-specific adaptations.

**Output Text:** The system will deliver highly accurate transcriptions of spoken language, even when the input contains heavy regional or non-standard accents. This will directly improve the usability of voice-driven systems such as virtual assistants, transcription services, and interactive voice response (IVR) applications.

**Accent Robustness:** The model is expected to effectively adapt to a wide variety of pronunciation patterns, reducing errors caused by accent mismatches. For example, vowel shifts in Indian English or consonant substitutions in French-accented English will be better recognized through learned embeddings.

**Generalization:** A key expected benefit of embedding-based learning and MTL is improved generalization. The model will be able to transfer knowledge from high-resource accents to low-resource ones, ensuring robust performance even with limited training data.

**Discussion:** The analysis will explore trade-offs between **model complexity** (e.g., larger architectures with specialized accent modules) and the **performance improvements** they yield. Another crucial aspect will be the **role of data diversity**, as more balanced multi-accent datasets will directly impact fairness and inclusivity. Furthermore, examining how accent embeddings cluster in vector space can provide new insights into linguistic relationships—revealing, for instance, that South Asian and African English accents share certain phonetic characteristics.

Ultimately, the findings are expected to confirm that accent-aware deep learning models not only reduce transcription errors but also move ASR closer to being truly **accent-agnostic and globally inclusive**.

**Summary and Future Directions**

The evolution of deep learning has fundamentally transformed the field of Automatic Speech Recognition (ASR), particularly in the domain of multi-accent speech recognition. Traditional ASR systems, often dependent on hand-crafted features and acoustic models, struggled to generalize effectively when confronted with accent variability. In contrast, deep learning—especially End-to-End (E2E) architectures such as Connectionist Temporal Classification (CTC), sequence-to-sequence models, and Transformer-based frameworks—has provided a powerful mechanism to learn directly from raw audio and textual data.Empirical findings consistently report double-digit improvements in Word Error Rate (WER) and Character Error Rate (CER) compared to baseline ASR systems that lack accent awareness. This underscores the fact that deep learning has not only advanced the theoretical frontier but has also made tangible improvements in practical performance.

1. Real-time Streaming and Adaptation:

Future systems must not only be accurate but also low-latency, enabling real-time transcription in live settings such as meetings, lectures, and customer service calls. This requires lightweight models capable of rapid accent adaptation on-the-fly, possibly through incremental fine-tuning or streaming-compatible architectures like Transducers.

1. Unsupervised and Self-Supervised Adaptation:

Collecting labeled data for every accent is impractical. Hence, a promising direction is unsupervised adaptation, where models adjust to novel accents without explicit transcriptions. Self-supervised learning approaches, such as contrastive learning on large-scale unlabeled speech corpora, will play a pivotal role in building accent-agnostic representations.

1. Broader Multilingual and Dialectal Coverage:

Current systems are disproportionately focused on English and a few high-resource languages. A critical future milestone is extending robust ASR to a much wider spectrum of global languages, dialects, and sociolects. This includes indigenous and low-resource languages, ensuring equity and inclusivity across communities.

1. Explainability and Accent-aware Interpretability:

Another important direction is making ASR systems more transparent by analyzing how embeddings and latent spaces represent accents. This could provide linguistic insights while also improving trust in deployment.

Ultimately, the long-term vision is the creation of ASR systems that generalize seamlessly across accents, languages, and domains—truly accent-agnostic, multilingual, and universally accessible speech technologies that democratize voice-based human–computer interaction.

**Conclusion and References**

The variability of human accents poses one of the most persistent challenges in speech recognition. This report has explored how deep learning innovations are driving meaningful progress in overcoming this barrier. Through architectures such as Transformers and RNN-based E2E models, combined with multi-task learning, accent embeddings, adversarial training, and fine-tuning strategies, ASR systems have become significantly more inclusive and accurate.

The impact is not merely technical; it is deeply social and global. By reducing bias against non-standard or regional accents, these systems foster equity in access to speech-enabled technologies, from educational platforms to healthcare assistants. Nevertheless, challenges remain—particularly in low-resource settings, where labeled data is scarce, and in real-time applications, where computational efficiency is critical.

Looking ahead, the convergence of self-supervised learning, streaming architectures, and multilingual expansion is likely to define the next frontier. The path forward requires sustained research and collaboration across linguistics, computer science, and data collection efforts. By continuing to refine these models, we move closer to the vision of seamless, accent-agnostic communication for all global citizens.

9. References

[1] Li, X., et al. (2024). Qifusion-Net: A Novel Fusion Architecture for Multi-Accent Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[2] Deng, J., et al. (2023). Multi-Accent Adaptation in End-to-End Speech Recognition Using Accent Embeddings and Adversarial Training. Proc. Interspeech.

[3] Hansen, J. H. L., & Liu, G. (2022). Unsupervised Accent Adaptation for Speech Recognition Using Self-Training. ICASSP.

[4] Panayotov, V., et al. (2015). Librispeech: An ASR corpus based on public domain audio books. ICASSP.

[5] Zhao, G., et al. (2018). L2-ARCTIC: A non-native English speech corpus. Proc. Interspeech.

[6] Graves, A., et al. (2013). Speech recognition with deep recurrent neural networks. ICASSP.

[7] Vaswani, A., et al. (2017). Attention is All You Need. NeurIPS.