WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI Detected Automatic Speech Recognition

Nikita Chaudhari[1], Prof. Ankush Dhama[2]

[1] Prof. Ramkrishna More College (Autonomous) Pradhikaran Akurdi, Pune, India
E-Mail: purvajachaudhari363@gmail.com
[2] Prof. Ramkrishna More College (Autonomous) Pradhikaran Akurdi, Pune, India
E-Mail: ankushdhamal01@gmail.com

ABSTRACT

Automatic Speech Recognition (ASR) technology has become an essential tool in modern communication, converting spoken language into text with applications in virtual assistants, transcription services, and accessibility tools. With the advancement of deep learning, ASR systems have significantly improved in accuracy and efficiency. However, challenges remain in noisy environments, accented speech recognition, and real-time adaptability. This research explores methods to enhance ASR accuracy and robustness by integrating speech production knowledge, utilizing neural text-to-encoder models, and developing adaptive deep learning architectures such as LSTM beamforming networks. The study investigates convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models while addressing data augmentation strategies and real-world applications in mobile devices and assistive technologies. By improving ASR capabilities, this research contributes to enhancing human-computer interaction across various domains. The dissertation is structured into five chapters: background, literature review, methodology, results, and conclusions, providing a comprehensive exploration of ASR advancements through deep learning.

Introduction

Automatic Speech Recognition (ASR) technology has transformed human-computer interaction by enabling machines to convert spoken language into text with high efficiency. From virtual assistants and transcription services to accessibility solutions for individuals with disabilities, ASR has become a cornerstone of modern communication. Over the years, advancements in deep learning have significantly improved ASR performance, making speech recognition more accurate, faster, and adaptable to various languages and contexts. However, despite these advancements, several challenges persist, particularly in handling noisy environments, diverse accents, and low-resource languages.

Traditional ASR systems relied on statistical methods such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), which, while effective, struggled with complex variations in speech patterns. The emergence of deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has dramatically enhanced ASR capabilities by improving feature extraction and contextual understanding. Moreover, hybrid models that combine CNNs, Long Short-Term Memory (LSTM) networks, and attention mechanisms have further boosted ASR accuracy and robustness.

This research aims to enhance ASR performance by integrating speech production knowledge, leveraging neural text-to-encoder models, and developing adaptive deep learning architectures such as LSTM-based beamforming networks. The study specifically explores several key research questions, including:

1. How can noise-robust ASR models be developed to mitigate the impact of background noise?
2. What techniques can improve ASR accuracy for accented speech and dialectal variations?
3. How can semi-supervised learning and self-supervised techniques reduce dependence on labelled data?
4. What are the benefits of multimodal approaches that integrate audio-visual data for ASR enhancement?

To address these research questions, the study investigates various deep learning architectures, including CNNs, RNNs, and hybrid models. Additionally, it explores data augmentation strategies, including text-to-speech (TTS) synthetic data generation, to enhance ASR performance in low-resource scenarios. Real-world applications of ASR, such as mobile voice assistants, healthcare accessibility tools, and assistive technologies, are also examined. By improving ASR capabilities, this research contributes to the broader goal of enhancing human-computer interaction across multiple domains, from communication and accessibility to education and business. The dissertation is structured into five comprehensive chapters, covering background information, literature review, research methodology, experimental results, and conclusions. Through a deep exploration of ASR advancements powered by deep learning, this research provides new insights into overcoming current limitations and improving ASR technology for diverse real-world applications.

Literature Review

Automatic Speech Recognition (ASR) has undergone a remarkable transformation over the past few decades, evolving from rule-based and statistical models to sophisticated deep learning-driven architectures. The integration of deep learning techniques has led to significant improvements in speech-to-text accuracy, robustness against noise, and adaptability to various speakers and accents. However, despite these advancements, ASR systems still encounter challenges in real-world conditions, including performance degradation in noisy environments, difficulty in recognizing diverse accents, and underutilization of speech production knowledge in model design.

Traditional ASR systems primarily relied on statistical methods such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). These models, although effective in structured and controlled environments, struggled with variability in speech patterns, environmental noise, and speaker diversity. With the advent of deep learning, ASR systems transitioned towards more powerful neural architectures, incorporating techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid models. Theoretical frameworks for ASR integrate machine learning, signal processing, and linguistics, enabling more advanced modeling of acoustic and linguistic features.

Neural network-based approaches have significantly improved ASR performance. Techniques such as Deep Belief Networks (DBNs) combined with HMMs, sequence-to-sequence modeling, and Connectionist Temporal Classification (CTC) have played crucial roles in advancing ASR capabilities. Research has demonstrated notable improvements through deep contextualized acoustic representations, multi-microphone beamforming, and various data augmentation strategies. Additionally, innovations such as LSTM-based adaptive beamforming networks and semi-supervised learning methods have further strengthened ASR robustness, allowing models to generalize better across different acoustic conditions.

Another significant advancement in ASR is the incorporation of multimodal approaches that integrate both auditory and visual data. These methods enhance ASR performance, especially in challenging environments with high background noise. Audio-visual speech recognition leverages facial expressions and lip movements to complement speech signal analysis, leading to more accurate transcription results.

Despite these technological advancements, research gaps persist in several areas. ASR still struggles with recognizing accented speech and dialectal variations, which limits its effectiveness in global applications. Far-field speech recognition—where the speaker is distant from the microphone—remains a challenge due to reverberation and background noise interference. Additionally, real-time adaptability to changing environments is an ongoing issue, as ASR models often require retraining to perform well in different conditions.

Privacy concerns also remain a significant challenge in ASR technology. While anonymization techniques have been explored, many methods still retain identifiable speaker information, raising ethical and security concerns. Future research should focus on integrating speech production knowledge into ASR models to improve their linguistic and phonetic understanding. Additionally, advancements in semi-supervised learning and self-supervised learning can help reduce reliance on large labeled datasets, making ASR more accessible for low-resource languages. Enhancing multimodal ASR systems by refining audio-visual fusion techniques will further contribute to overcoming current limitations.

Methodology

This study employs a mixed-methods approach, combining quantitative evaluation of ASR performance with qualitative analysis of user experiences. Experimental design assesses the impact of deep learning architectures on ASR accuracy, while comparative analysis identifies the most effective techniques for various conditions.

Data Collection & Sampling

ASR models are tested using established speech corpora (e.g., WSJ, Globalphone) and real-world speech recordings across diverse environments. Crowdsourced speech data ensures accent diversity, while synthetic augmentation (TTS) improves performance in low-resource scenarios. Stratified sampling ensures representation across demographics, with a target of 500 speech samples. Specialized sampling is used for dysarthric and child speech.

Tools & Techniques

The study utilizes deep learning frameworks (TensorFlow, PyTorch) and ASR toolkits (Espresso, FAIRSEQ). Feature extraction is performed using Librosa, and models include CNNs for feature learning, LSTMs for sequential processing, and attention mechanisms for enhanced focus. Noise robustness techniques like MVDR beamforming and curriculum learning improve system performance.

Data Analysis

Performance is primarily evaluated using Word Error Rate (WER), with additional analysis of error patterns, signal-to-noise ratios (SNRs), and multimodal ASR effectiveness. Statistical tests determine significant performance differences, while visualization techniques illustrate ASR trends.

Limitations

Key challenges include accent variability, real-world noise complexity, and limited labelled data for low-resource languages. Computational constraints affect large-scale model training, and standard metrics may not fully capture user experience. Rapid advancements in ASR technology necessitate continuous reassessment.

Results

The results highlight the transformative impact of deep learning on ASR efficiency and performance across diverse conditions. Techniques such as DeCoAR, ACCAN, adaptive beamforming, multimodal fusion, and speech chain models contribute significantly to improving ASR accuracy, robustness, and adaptability. These advancements have broad implications for real-world ASR applications, including mobile speech recognition, assistive technologies, and multilingual speech processing.

Discussion

The findings of this study highlight the significant impact of deep learning on Automatic Speech Recognition (ASR) performance, particularly in accuracy, robustness, and adaptability. Key advancements include Deep Contextualized Acoustic Representations (DeCoAR), which reduces Word Error Rate (WER) by 42% (WSJ eval92) and 19% (LibriSpeech test-clean), demonstrating the effectiveness of deep contextual embeddings.

Adaptive beamforming networks improve ASR robustness by 7.97% in noisy conditions, while curriculum learning (ACCAN) reduces WER by 31.4%, emphasizing the value of progressive training strategies. Multi-task learning for accent adaptation enhances recognition by 15% for seen accents and 10% for unseen accents, and audio-visual fusion boosts accuracy by 7% to 30%, particularly in noisy environments.

A two-stage deep learning approach outperforms single-stage methods in noise reduction, and unsupervised representation learning reduces dependency on labeled data, improving ASR for low-resource languages. Optimizing for WER instead of cross-entropy yields an 8.2% performance boost, while speech chain models integrating ASR and TTS enhance bidirectional learning. Personalized ASR models improve recognition by 71% for non-standard speech with only five minutes of training.

Further improvements include dereverberation techniques, reducing WER by 40%, and accent-aware models, enhancing recognition for diverse speech patterns. Comparative analysis confirms the superiority of two-stage noise reduction, ACCAN (31.4% WER reduction), DeCoAR, and multimodal fusion (up to 30% accuracy improvement).

Performance evaluations reinforce the efficiency of modern ASR architectures: DeCoAR improves WER by 42% (WSJ) and 19% (LibriSpeech), beamforming networks enhance noise robustness by 7.97%, and text-to-speech augmentation matches hybrid ASR models in low-resource settings. The Espresso toolkit accelerates decoding by 4–11x while maintaining state-of-the-art accuracy.

These findings underscore the transformative potential of deep learning, demonstrating that integrating advanced neural architectures, adaptive learning, and multimodal approaches makes ASR systems more accurate and robust for real-world applications.

Conclusion

This study highlights how deep learning has revolutionized automatic speech recognition (ASR), improving accuracy, robustness, and adaptability across diverse conditions. Techniques such as deep contextualized acoustic representations (DeCoAR), curriculum learning (ACCAN), and multi-task learning for accent adaptation have significantly reduced word error rates (WER). Audio-visual fusion enhances recognition in noisy environments, while text-to-speech augmentation supports ASR in low-resource scenarios. Additionally, personalized ASR models demonstrate significant improvements for non-standard speech with minimal training data. These findings affirm that deep learning has made ASR more efficient, inclusive, and effective for real-world applications.

Future Scope

Future research should focus on integrating speech production knowledge into ASR models, improving multimodal fusion techniques, and advancing self-supervised learning to reduce dependency on labeled data. Enhancing privacy-preserving ASR techniques and optimizing deep learning architectures for real-time efficiency will be critical for broader adoption. Addressing challenges in child speech recognition, refining adaptive curriculum learning strategies, and developing user-centered evaluation metrics will further advance ASR performance. Finally, rapid personalization methods for non-standard speech will improve accessibility, ensuring ASR systems serve a wider range of users effectively.

REFERENCES

1. D. Amodei et al., “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin,” in Proc. Int. Conf. Machine Learning (ICML), 2016.
    A. Baevski et al., “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.
2. Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A Neural Probabilistic Language Model,” J. Mach. Learn. Res., vol. 3, pp. 1137–1155, 2003.
3. W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, Attend and Spell,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2016.
4. J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-Based Models for Speech Recognition,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2015.
5. R. Collobert et al., “Natural Language Processing (Almost) from Scratch,” J. Mach. Learn. Res., vol. 12, pp. 2493–2537, 2011.

6. L. Dong, S. Xu, and B. Xu, “Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2018.
    A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks,” in Proc. Int. Conf. Machine Learning (ICML), 2006.
7. Graves, A. Mohamed, and G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2013.
8. Hannun et al., “Deep Speech: Scaling up End-to-End Speech Recognition,” arXiv preprint arXiv:1412.5567, 2014.
9. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.
10. G. Hinton et al., “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, 2012.
11. S. Karita et al., “A Comparative Study on Transformer vs RNN in Speech Applications,” in Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU), 2019.
12. Y. Kim, H. Lee, and S. Kim, “Joint CTC-Attention Based End-to-End Speech Recognition Using Multi-Task Learning,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2016.
13. Y. Liu et al., “Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2020.
14. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in Proc. Interspeech, 2019.
15. S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, “Learning Problem-Agnostic Speech Representations from Self-Supervised Learning,” in Proc. Interspeech, 2019.
    A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” arXiv preprint arXiv:2212.04356, 2022.
16. Vaswani et al., “Attention Is All You Need,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.
17. Y. Zhang et al., “Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition,” in Proc. NeurIPS, 2020.
18. P. N. Garner, J. Dines, T. Hain, and A. El Hannani, “Real-World Speech Recognition with Deep Neural Networks,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2014.
19. S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/Attention Architecture for End-to-End Speech Recognition,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1240–1253, 2017.
20. X. Li, X. Wu, Y. Yang, and M. Zhang, “Multi-Modal Speech Recognition with Audio-Visual Fusion and Transformer Models,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2021.
21. G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, and D. Dimitriadis, “English Conversational Telephone Speech Recognition by Humans and Machines,” in Proc. Interspeech, 2017.
22. Geng, H. Lv, J. Zhang, and J. Tao, “Robust End-to-End Speech Recognition with Adaptive Curriculum Learning,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2020.
    A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: Large-Scale Speaker Verification in the Wild,” in Proc. Interspeech, 2017.
23. M. Ravanelli and Y. Bengio, “Speaker Recognition from Raw Waveform with SincNet,” in Proc. IEEE Spoken Lang. Technol. Workshop (SLT), 2018.
24. T. N. Sainath and J. Cui, “Self-Supervised Learning for End-to-End Speech Recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2020.
25. Liu, J. Zhang, and C. Yang, “Improving Speech Recognition with Self-Supervised Pretrained Acoustic and Language Models,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2021.
26. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in Proc. Int. Conf. Learning Representations (ICLR), 2015.