International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 11 Issue 08 August 2024

Study of various Approaches used for Speech Recognition

Baby Priya[1], Asst. Prof. Ayush Kumar[2]
Research Scholar[1], Assistant Professor[2]
Department Of Computer Science & Engineering, Radha Raman Engineering College, Bhopal (M.P.)[1, 2]
bpriya16cs65@gmail.com[1]

Abstract: *This survey paper examines recent advancements in speech recognition technologies and their integration with Natural Language Processing (NLP). The study begins by discussing the evolution of voice recognition systems, highlighting the transition from classical methods to deep learning models. We detail how Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and transformer architectures have set new benchmarks in speech-to-text conversion. The second section addresses the challenges classical NLP models face in interpreting spoken language, emphasizing the need for innovative approaches to tackle dialectal variations, colloquial expressions, and context-dependent nuances. The paper also explores the potential of contextual information, multitask learning, and transfer learning to enhance voice recognition systems. By investigating the synergies between voice recognition and NLP, we underscore the importance of multidisciplinary research in overcoming language understanding and processing barriers. The analysis includes applications such as virtual assistants, transcription services, language translation, and accessibility technologies. The conclusions outline future research directions and emphasize the necessity of collaboration to fully harness the potential of voice recognition while pioneering new advancements in NLP.*

Keywords: *Speech Recognition, Automatic Speech Recognition (ASR), Parsing, NLP*.

1. INTRODUCTION

The study of human-computer interaction has witnessed the emergence of two significant fields within artificial intelligence and natural language processing in recent years: speech recognition and advancements in natural language processing. These breakthroughs have revolutionized how computers interpret and respond to human speech, transcending the limitations of traditional NLP methods.

Speech Recognition: An Overview

Speech recognition, a subfield of Natural Language Processing (NLP), involves converting spoken words into text, enabling computers to comprehend and interpret human speech inputs. This technology has a wide range of applications, including virtual assistants, transcription services, hands-free device administration, and accessibility solutions. Figure 1 illustrates the various components involved in voice recognition and their interplay, facilitating the accurate interpretation of spoken words.

Recent advancements in speech recognition have been propelled by the development of deep learning models. Classical speech recognition systems relied on techniques such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). However, these methods had limitations in handling the variability and complexity of human speech. The advent of deep learning introduced more sophisticated approaches, notably Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and transformer architectures.

RNNs, with their ability to process sequential data, have shown significant promise in speech recognition. Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), variants of RNNs, address the vanishing gradient problem, enabling the capture of long-term dependencies in speech sequences. CNNs, traditionally used in image processing, have also been adapted for speech recognition. By applying convolutional layers to spectrograms (visual representations of audio signals), CNNs can effectively extract hierarchical features from speech data.

Transformers, a more recent development, have further advanced speech recognition. The transformer model, introduced by Vaswani et al., relies on self-attention mechanisms to capture global dependencies in data. This architecture has proven particularly effective in handling the parallelization of training, leading to improved performance and efficiency in speech-to-text conversion.

Challenges in NLP for Spoken Language

While speech recognition has made remarkable strides, classical NLP models often struggle with interpreting spoken language. Challenges include dialectal variations, colloquial expressions, and context-dependent nuances. Spoken language is inherently different from written text, often lacking the structure and formality of written language. This necessitates the development of new methods to bridge the gap between spoken language and NLP.

Enhancing Speech Recognition with NLP Techniques

To address these challenges, researchers are exploring the integration of contextual information, multitask learning, and transfer learning into voice recognition systems. Contextual information, such as speaker identity and conversational context, can significantly enhance the accuracy of speech recognition. Multitask learning, where models are trained on multiple related tasks, helps in capturing shared representations and improving generalization.

Transfer learning, a technique where models pre-trained on large datasets are fine-tuned for specific tasks, has also shown promise. Pre-trained language models, such as BERT and GPT, can be adapted for speech recognition tasks, leveraging their extensive knowledge of language semantics and structure.

The Importance of Multidisciplinary Research

The synergies between speech recognition and NLP underscore the importance of multidisciplinary research. Combining expertise from various fields, including linguistics, computer science, and cognitive psychology, can lead to innovative solutions that break language understanding and processing barriers. By fostering collaboration, researchers can develop more robust and versatile systems that cater to the diverse needs of users. Looking at Figure 1 [3] exposes the various moving pieces that are involved in voice recognition and how they all work together to make it possible for computers to interpret spoken words.

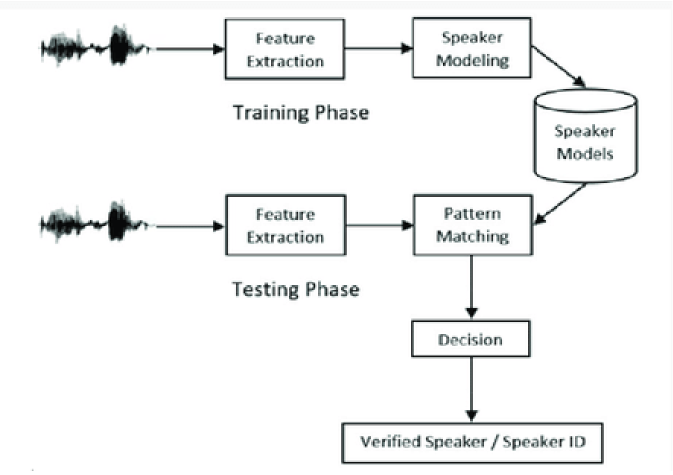

Figure 1: Process of Speech Recognition and Breaking NLP

Alternately, the cutting-edge work that is being done in Breaking NLP [4] is pushing the frontiers of what is assumed to be achievable in terms of natural language comprehension. The investigation of complex models and approaches that go beyond what simple natural language processing systems can do is required. By demonstrating how other methods, such as complex neural architectures and semantic representations, were examined in addition to the conventional standard language comprehension tools, Figure 2 [4] exemplifies the revolutionary nature of Breaking Natural Language Processing (NLP).

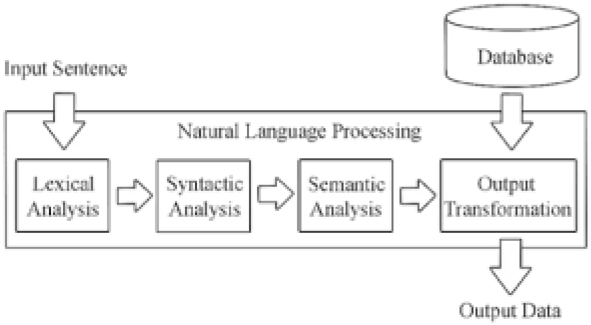

Figure 2: Transforming the landscape of NLP

The purpose of this introduction part is to provide the framework for the rest of the essay, which will investigate Speech Recognition and Breaking NLP in comprehensive depth, covering its applications, challenges, and revolutionary possibilities in the production of future human-computer interaction. Through the exploration of these technical fields, we get an understanding of the intricate mechanisms that make it possible for us to connect with computers, as well as the potential that are made available by Breaking NLP, a revolutionary development in the field of natural language processing.

2. LITERATURE REVIEW

Xue et al. (2019) propose a multi-task learning framework aimed at addressing the issue of catastrophic forgetting in automatic speech recognition (ASR) systems [5]. Catastrophic forgetting occurs when a model trained on new data tends to forget previously learned information. Their framework incorporates multiple tasks during the training process, which helps in retaining knowledge and enhancing the overall performance of ASR systems. This approach is particularly beneficial in dynamic environments where the ASR system needs to adapt continuously without losing prior knowledge.

Madhavaraj and Ramakrishnan (2019) explore data-pooling and multi-task learning to improve speech recognition performance in multiple low-resource languages [6]. By pooling data from various low-resource languages and employing multi-task learning, their method significantly enhances the ASR system's ability to generalize across different languages. This study highlights the potential of collaborative data utilization and multi-task learning to overcome the limitations faced by low-resource languages in speech recognition.

Tomashenko, Caubriere, and Estève (2019) investigate adaptation and transfer learning techniques for end-to-end spoken language understanding from speech [7]. Their study examines how transfer learning can be effectively applied to adapt ASR systems to new domains and languages. By leveraging pre-trained models and fine-tuning them for specific tasks, they demonstrate substantial improvements in spoken language understanding, which is critical for developing robust and versatile ASR systems.

Mendes et al. (2019) focus on the recognition of Latin American Spanish using multi-task learning [8]. Their approach involves training ASR systems on multiple related tasks simultaneously, which helps in capturing shared features and improving recognition accuracy for Latin American Spanish. This study underscores the effectiveness of multi-task learning in enhancing ASR performance for specific language variants.

Yu et al. (2019) present a study on cross-language end-to-end speech recognition using transfer learning for the low-resource Tujia language [9]. They demonstrate that by transferring knowledge from high-resource languages to low-resource languages, significant improvements in recognition accuracy can be achieved. This approach is particularly valuable for preserving and promoting the use of low-resource languages in modern technology.

Dong et al. (2019) combine deep transfer learning with multi-task bi-directional LSTM RNN for named entity recognition in Chinese electronic medical records [10]. Their method leverages transfer learning to enhance the model's ability to identify and classify medical entities accurately. This application of transfer learning in the medical domain showcases its versatility and effectiveness in specialized tasks requiring high precision.

Wang et al. (2020) introduce a cross-task transfer learning approach designed to adapt deep speech enhancement models to handle unseen background noise [11]. Their method utilizes paired senone classifiers to enhance the robustness of automatic speech recognition (ASR) systems in noisy environments. By transferring knowledge from models trained on various acoustic conditions to new, unseen noise scenarios, their approach improves the system's ability to accurately recognize speech amidst background interference. This study underscores the effectiveness of transfer learning in increasing the adaptability of ASR systems to a wide range of acoustic challenges. The research demonstrates how leveraging transfer learning can significantly boost the performance of ASR systems in dynamic and noisy conditions, highlighting its potential for advancing speech recognition technology in diverse environments.

Joshi et al. (2020) explore various transfer learning approaches for streaming end-to-end speech recognition systems [12]. They investigate how pre-trained models can be adapted for real-time speech recognition tasks, emphasizing the importance of efficient and scalable transfer learning techniques in deploying ASR systems for live applications.

Wang, Pino, and Gu (2020) focus on improving cross-lingual transfer learning for end-to-end speech recognition by integrating speech translation [13]. Their approach leverages the synergies between speech recognition and translation tasks to enhance cross-lingual performance. This study provides a comprehensive framework for developing multilingual ASR systems capable of handling diverse linguistic inputs.

Shivakumar and Georgiou (2020) revisit the transfer learning approach from adult to children for speech recognition, providing updated evaluations, analyses, and recommendations [14]. Their findings emphasize the critical role of transfer learning in adapting ASR systems to better recognize and interpret children's speech, which differs significantly from adult speech in various aspects.

Kano, Sakti, and Nakamura (2020) propose a multi-task learning framework for end-to-end speech translation, specifically targeting distant language pairs [15]. By integrating speech recognition and translation tasks, their approach improves the accuracy and fluency of translated speech outputs, highlighting the potential of multi-task learning in bridging linguistic gaps.

Woldemariam (2020) explores the application of transfer learning to enhance speech recognition for less-resourced Semitic languages, with a particular focus on Amharic [16]. The study illustrates how transfer learning can substantially boost recognition accuracy in languages that suffer from limited training data. By leveraging pre-trained models and adapting them to Amharic, Woldemariam's approach addresses the challenges of data scarcity and improves the effectiveness of automatic speech recognition (ASR) systems in these under-resourced languages. The research highlights the potential of transfer learning to bridge gaps in training data and offers a viable pathway for developing robust ASR systems for languages with limited resources. This study contributes to advancing speech recognition technology in less-resourced linguistic contexts, demonstrating the significant benefits of transfer learning in overcoming data limitations.

Abad et al. (2020) explore cross-lingual transfer learning for zero-resource domain adaptation in ASR [17]. Their approach aims to adapt ASR systems to new domains without requiring additional labeled data from the target domain. This study underscores the potential of transfer learning to facilitate domain adaptation and improve ASR performance in diverse application areas.

Wang, Sainath, and Weiss (2020) investigate multitask training with text data for end-to-end speech recognition [18]. By incorporating text-based tasks into the training process, their approach enhances the model's ability to learn language patterns and improve speech recognition accuracy. This study highlights the benefits of leveraging text data in training robust ASR systems.

Li et al. (2020) propose improvements to transformer-based speech recognition through unsupervised pre-training and multi-task semantic knowledge learning [19]. Their approach leverages unsupervised learning techniques to pre-train models on large datasets, followed by fine-tuning with multi-task learning, resulting in enhanced recognition performance.

Zhuang et al. (2020) deliver an extensive survey on transfer learning, exploring its fundamental principles, methodologies, and diverse applications across various domains, including speech recognition and natural language processing (NLP) [20]. This comprehensive review serves as a crucial resource for gaining a thorough understanding of transfer learning's scope and depth. The survey examines the core concepts and techniques of transfer learning, detailing how they are applied to different fields and highlighting both the potential benefits and challenges associated with this approach. By providing a detailed analysis of transfer learning's role in advancing technologies such as speech recognition and NLP, Zhuang et al.'s work offers valuable insights into the current state and future directions of this transformative method. Their survey is instrumental for researchers and practitioners seeking to grasp the full impact and versatility of transfer learning.

Tang et al. (2021) propose a general multi-task learning framework to leverage text data for speech-to-text tasks [21]. Their approach integrates multiple related tasks, enhancing the model's ability to generalize across different speech and text inputs. This study emphasizes the importance of multi-task learning in developing versatile and robust ASR systems.

Zhao et al. (2021) introduce self-attention transfer networks for speech emotion recognition [22]. By combining self-attention mechanisms with transfer learning, their approach improves the model's ability to recognize emotional cues in speech, highlighting the potential of advanced neural architectures in emotion recognition tasks.

Rongali et al. (2021) investigate the use of transfer learning for end-to-end spoken language understanding, emphasizing how pre-trained models can be adapted for complex language understanding tasks [23]. Their study highlights the adaptability of transfer learning techniques in improving the performance of end-to-end automatic speech recognition (ASR) systems. By leveraging models that have been pre-trained on extensive data, the research demonstrates how these models can be fine-tuned to better handle the nuances of spoken language, leading to more accurate and effective language understanding. This approach showcases the flexibility of transfer learning in addressing diverse challenges within spoken language processing and underscores its potential to significantly enhance the capabilities of ASR systems. Rongali et al.'s work offers valuable insights into how transfer learning can be utilized to refine and optimize spoken language understanding technologies.

Lahiri et al. (2021) investigate multilingual speech recognition using knowledge transfer across learning processes [24]. By transferring knowledge between languages, their approach improves recognition accuracy for multiple languages, highlighting the potential of knowledge transfer in multilingual ASR systems.

Padi et al. (2021) propose an enhanced method for speech emotion recognition by combining transfer learning with spectrogram augmentation [25]. Their approach leverages transfer learning to boost the model's capability in identifying emotions in speech, especially in low-resource settings where data is limited. By using pre-trained models on large datasets, the system can transfer knowledge to the emotion recognition task, improving its performance. Additionally, spectrogram augmentation techniques are applied to diversify the training data, further enhancing the model's robustness. This dual strategy significantly improves emotion recognition accuracy, making it more effective in real-world applications where emotional nuance is critical.

Rolland et al. (2022) investigate the application of multilingual transfer learning to enhance automatic speech recognition (ASR) systems for children [26]. Their study highlights the unique challenges of recognizing children's speech, such as higher pitch, variability in pronunciation, and limited training data. By leveraging transfer learning, they develop ASR models that can effectively adapt to children's speech patterns across various languages. This approach involves utilizing pre-trained models on adult speech and transferring the learned knowledge to improve the recognition accuracy for children's speech. Their findings demonstrate that multilingual transfer learning significantly enhances the performance of ASR systems for children, making them more reliable and effective in diverse linguistic contexts.

Latif et al. (2022) introduce a multitask learning approach that utilizes augmented auxiliary data to improve speech emotion recognition [27]. Their method integrates auxiliary data sources alongside multiple related tasks to enhance the model’s capability in accurately detecting emotions in speech. By combining these elements, the approach leverages diverse data and task interactions, which collectively contribute to a more robust and precise emotion recognition system. This multitask learning framework allows the model to generalize better across different emotional contexts and improves its performance even in scenarios with limited labeled data. The study underscores the effectiveness of using augmented auxiliary data and multitask learning techniques to advance the accuracy of emotion recognition in speech applications.

Sullivan et al. (2022) examine the use of transfer learning and language model decoding to enhance automatic speech recognition (ASR) for non-native English speakers [28]. Their study focuses on adapting ASR systems to better handle the unique challenges associated with non-native speech, such as diverse accents and pronunciation variations. By applying transfer learning, they leverage pre-trained models to improve recognition accuracy for non-native speech patterns. Additionally, their approach incorporates advanced language model decoding techniques to refine the system's ability to understand and transcribe non-native speech more effectively. This dual strategy addresses common issues in ASR for non-native speakers, such as mispronunciations and accent-related distortions, resulting in a more reliable and accurate speech recognition system for diverse linguistic backgrounds.

Padi et al. (2022) present a multimodal emotion recognition approach that utilizes transfer learning from speaker recognition and BERT-based models [29]. Their method integrates various modalities to enhance the precision of emotion recognition in speech. By leveraging transfer learning from pre-trained speaker recognition models and BERT-based language models, the approach improves the system’s ability to identify and interpret emotional cues with greater accuracy. Combining these modalities allows the model to capture a broader range of emotional signals and contextual nuances, leading to more reliable emotion detection. This study highlights the effectiveness of using transfer learning in multimodal applications, demonstrating its potential to significantly advance emotion recognition technology by incorporating diverse sources of information.

Yadav and Sitaram (2022) present a comprehensive survey on multilingual models for automatic speech recognition (ASR), focusing on different transfer learning and multi-task learning strategies [30]. Their review explores the various techniques and methodologies used to develop ASR systems that can effectively handle multiple languages. They discuss the challenges encountered in multilingual ASR, such as managing linguistic diversity, accent variations, and limited data resources, while also highlighting recent advancements in the field. The survey provides an overview of how transfer learning can enhance cross-lingual performance and how multi-task learning approaches can improve language adaptability and accuracy. By covering both the current state-of-the-art methods and ongoing issues, their survey offers valuable insights into the evolution and future directions of multilingual ASR systems.

Kheddar et al. (2023) explore deep transfer learning techniques to advance automatic speech recognition (ASR), with a focus on enhancing generalization across various languages and domains [31]. Their research highlights how transfer learning can significantly improve the adaptability and robustness of ASR systems. By leveraging deep learning models pre-trained on large, diverse datasets, their approach enables ASR systems to better handle linguistic and contextual variations. The study demonstrates that transfer learning enhances the system's performance by allowing it to generalize more effectively to new languages and domains, thereby improving accuracy and reliability. This work underscores the potential of deep transfer learning to address challenges in ASR and make systems more versatile and resilient in real-world applications.

Steinmetz (2023) investigates the application of transfer learning with L2 speech to enhance automatic speech recognition (ASR) for dysarthric speech [32]. Their study focuses on transferring knowledge from models trained on non-dysarthric speech to those recognizing dysarthric speech, which is affected by motor disorders. This approach aims to improve ASR systems' ability to accurately understand and transcribe speech that has been impaired due to motor difficulties. By leveraging pre-existing models trained on clearer, non-dysarthric speech, the method helps address the specific challenges associated with dysarthric speech, such as altered articulation and reduced clarity. Steinmetz's work demonstrates how transfer learning can be effectively utilized to bridge the gap between different types of speech, ultimately enhancing the recognition accuracy for individuals with speech impairments.

Nga et al. (2023) introduce a cyclic transfer learning approach to enhance Mandarin-English code-switching speech recognition [33]. Their method employs cyclic training between the two languages to boost recognition accuracy in code-switching situations, where speakers alternate between Mandarin and English within the same conversation. By iteratively training the model on both languages, this approach effectively captures the nuances and complexities of bilingual speech patterns. The cyclic process helps the model better understand and process the transitions between languages, improving its ability to handle mixed-language inputs. This study demonstrates the effectiveness of cyclic transfer learning in addressing the unique challenges posed by code-switching, ultimately leading to more accurate and reliable speech recognition in bilingual contexts.

Tun et al. (2023) explore the use of multimodal transfer learning for assessing oral presentations [34]. Their approach integrates both speech and visual data to improve the evaluation process. By combining these modalities, their method enhances the accuracy and comprehensiveness of oral presentation assessments. This multimodal approach leverages transfer learning techniques to incorporate pre-trained models from different domains, allowing for a more nuanced analysis of presentation skills. The study highlights how transfer learning can be effectively applied in educational settings to assess various aspects of oral presentations, such as delivery, clarity, and visual engagement. Tun et al.'s work underscores the potential of multimodal transfer learning to provide richer, more detailed evaluations, showcasing its value in enhancing educational assessment methods.

Zheng and Zhang (2023) introduce an enhanced multi-label transfer learning model designed for intelligent speech systems [35]. Their approach tackles the complexities of multi-label classification in speech applications, where multiple labels or categories need to be assigned to a single speech input. By leveraging transfer learning, their model effectively utilizes pre-trained knowledge to improve performance in these challenging scenarios. This improvement addresses issues such as overlapping labels and context-dependent classifications, showcasing how transfer learning can be adapted to handle intricate tasks in speech recognition. The study highlights the versatility and effectiveness of transfer learning in managing complex multi-label problems, demonstrating its potential to advance the capabilities of intelligent speech systems.

Zhou et al. (2024) present a multitask co-training framework designed to enhance speech translation by simultaneously leveraging speech recognition and machine translation tasks [36]. Their innovative approach integrates these tasks to improve both the accuracy and fluency of translated speech outputs. By training models on multiple related tasks, the framework enables better alignment between spoken input and its translated text, resulting in more coherent and contextually accurate translations. This multitask co-training method helps capture the nuances of both speech recognition and translation processes, leading to improvements in overall performance. The study demonstrates how combining these tasks can address the challenges of translating spoken language, highlighting the potential for more effective and natural-sounding speech translation systems.

Ta and Le (2024) examine transfer learning techniques aimed at enhancing speech accent recognition in low-resource environments, with a focus on the Vietnamese language [37]. Their study illustrates how transfer learning can significantly boost recognition accuracy for accented speech when resources are limited. By leveraging pre-trained models and adapting them to the specific nuances of Vietnamese accents, their approach addresses the challenges of limited training data and accent variability. The research demonstrates that transfer learning effectively transfers knowledge from more resource-rich contexts to improve performance in recognizing accented speech. This approach not only improves the accuracy of speech recognition systems in low-resource settings but also highlights the potential of transfer learning to address language-specific challenges in accent recognition.

Kheddar et al. (2024) present a comprehensive survey on cutting-edge deep learning methods for automatic speech recognition (ASR) [38]. Their review explores a range of advanced techniques, focusing on transfer learning and multi-task learning, and highlights the most recent advancements in ASR technology. By examining the latest developments, their survey provides insights into how these techniques are being applied to enhance ASR systems, including improvements in accuracy, adaptability, and efficiency. The study covers various strategies for leveraging pre-trained models and integrating multiple tasks to address complex speech recognition challenges. This survey serves as a valuable resource for understanding current trends and future directions in ASR, showcasing how deep learning innovations are driving progress in the field.

Hassan et al. (2024) introduce a deep bidirectional LSTM model that is enhanced through transfer-learning-based feature extraction for dynamic human activity recognition [39]. Their method integrates transfer learning with deep learning techniques to significantly improve the accuracy of recognizing various human activities. By leveraging pre-trained models for feature extraction, their approach enables the LSTM network to better capture and interpret complex activity patterns. This combination enhances the model's ability to accurately monitor and classify dynamic activities, even in diverse and challenging conditions. Their study demonstrates how incorporating transfer learning can refine deep learning models and boost performance in activity recognition applications, offering a more reliable solution for monitoring human activities in real-world scenarios.

Kumar and Yadav (2024) investigate the use of multiview learning techniques to improve speech recognition for low-resource languages [40]. Their approach harnesses multiple perspectives or "views" of the data to enhance the accuracy of automatic speech recognition (ASR) systems. By integrating diverse types of data, such as acoustic features and linguistic information, their method addresses the specific challenges associated with recognizing speech in languages with limited resources. This multiview learning framework helps overcome the scarcity of training data and the unique linguistic characteristics of low-resource languages, leading to more accurate and robust speech recognition. Their study demonstrates how leveraging multiple data views can significantly improve ASR performance, making it a promising solution for enhancing recognition capabilities in underrepresented languages.

3. CONCLUSION

This survey paper has provided a comprehensive overview of recent advancements in speech recognition technologies and their integration with Natural Language Processing (NLP). We traced the evolution of voice recognition from traditional methods to modern deep learning models, highlighting the significant impact of Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and transformer architectures on improving speech-to-text accuracy. The discussion then shifted to the limitations faced by classical NLP models in interpreting spoken language, particularly with regard to dialectal variations, colloquial expressions, and context-dependent subtleties. We explored how contextual information, multitask learning, and transfer learning can address these challenges and enhance voice recognition systems.

Our analysis demonstrated the crucial role of multidisciplinary research in bridging gaps between voice recognition and NLP, emphasizing the synergies that can drive forward advancements in language understanding and processing. Applications in virtual assistants, transcription services, language translation, and accessibility technologies illustrate the practical benefits of these advancements.

Looking ahead, future research should focus on refining these technologies and exploring new methodologies to further improve voice recognition systems. Collaboration across fields will be essential to fully leverage the potential of these technologies and to continue pushing the boundaries of NLP innovation.

REFERENCES

[1] Shivakumar, Prashanth Gurunath, and Panayiotis Georgiou. "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations." Computer speech & language 63 (2020): 101077.

[2] Kamath, Uday, John Liu, and James Whitaker. Deep learning for NLP and speech recognition. Vol. 84. Cham, Switzerland: Springer, 2019.

[3] Han, Zhijie, Huijuan Zhao, and Ruchuan Wang. "Transfer learning for speech emotion recognition." In 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), pp. 96-99. IEEE, 2019.

[4] Wang, Changhan, Juan Pino, and Jiatao Gu. "Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation." arXiv preprint arXiv:2006.05474 (2020).

[5] Xue, Jiabin, Jiqing Han, Tieran Zheng, Xiang Gao, and Jiaxing Guo. "A multi-task learning framework for overcoming the catastrophic forgetting in automatic speech recognition." arXiv preprint arXiv:1904.08039 (2019).

[6] Madhavaraj, A., and A. G. Ramakrishnan. "Data-pooling and multi-task learning for enhanced performance of speech recognition systems in multiple low resourced languages." In 2019 National Conference on Communications (NCC), pp. 1-5. IEEE, 2019.

[7] Tomashenko, Natalia, Antoine Caubriere, and Yannick Estève. "Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech." In Interspeech 2019, pp. 824-828. ISCA, 2019.

[8] Mendes, Carlos, Alberto Abad, Joao Paulo Neto, and Isabel Trancoso. "Recognition of Latin American Spanish Using Multi-Task Learning." In INTERSPEECH, pp. 2135-2139. 2019.

[9] Yu, Chongchong, Yunbing Chen, Yueqiao Li, Meng Kang, Shixuan Xu, and Xueer Liu. "Cross-language end-to-end speech recognition research based on transfer learning for the low-resource Tujia language." Symmetry 11, no. 2 (2019): 179.

[10] Dong, Xishuang, Shanta Chowdhury, Lijun Qian, Xiangfang Li, Yi Guan, Jinfeng Yang, and Qiubin Yu. "Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN." PloS one 14, no. 5 (2019): e0216046.

[11] Wang, Sicheng, Wei Li, Sabato Marco Siniscalchi, and Chin-Hui Lee. "A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6219-6223. IEEE, 2020.

[12] Joshi, Vikas, Rui Zhao, Rupesh R. Mehta, Kshitiz Kumar, and Jinyu Li. "Transfer learning approaches for streaming end-to-end speech recognition system." arXiv preprint arXiv:2008.05086 (2020).

[13] Wang, Changhan, Juan Pino, and Jiatao Gu. "Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation." arXiv preprint arXiv:2006.05474 (2020).

[14] Shivakumar, Prashanth Gurunath, and Panayiotis Georgiou. "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations." Computer speech & language 63 (2020): 101077.

[15] Kano, Takatomo, Sakriani Sakti, and Satoshi Nakamura. "End-to-end speech translation with transcoding by multi-task learning for distant language pairs."IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 1342-1355.

[16] Woldemariam, Yonas. "Transfer learning for less-resourced semitic languages speech recognition: the case of Amharic." In Proceedings of the 1st joint workshop on spoken language technologies for under-resourced languages (SLTU) and collaboration and computing for under-resourced languages (CCURL), pp. 61-69. 2020.

[17] Abad, Alberto, Peter Bell, Andrea Carmantini, and Steve Renais. "Cross lingual transfer learning for zero-resource domain adaptation." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6909-6913. IEEE, 2020.

[18] Wang, Peidong, Tara N. Sainath, and Ron J. Weiss. "Multitask training with text data for end-to-end speech recognition." arXiv preprint arXiv:2010.14318 (2020).

[19] Li, Song, Lin Li, Qingyang Hong, and Lingling Liu. "Improving Transformer-Based Speech Recognition with Unsupervised Pre-Training and Multi-Task Semantic Knowledge Learning." In Interspeech, pp. 5006-5010. 2020.

[20] Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. "A comprehensive survey on transfer learning." Proceedings of the IEEE 109, no. 1 (2020): 43-76.

[21] Tang, Yun, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. "A general multi-task learning framework to leverage text data for speech to text tasks." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6209-6213. IEEE, 2021.

[22] Zhao, Ziping, Keru Wang, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Shihuang Sun, Haishuai Wang, Jianhua Tao, and Björn W. Schuller. "Self-attention transfer networks for speech emotion recognition." Virtual Reality & Intelligent Hardware 3, no. 1 (2021): 43-54.

[23] Rongali, Subendhu, Beiye Liu, Liwei Cai, Konstantine Arkoudas, Chengwei Su, and Wael Hamza. "Exploring transfer learning for end-to-end spoken language understanding." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 15, pp. 13754-13761. 2021.

[24] Lahiri, Rimita, Kenichi Kumatani, Eric Sun, and Yao Qian. "Multilingual speech recognition using knowledge transfer across learning processes." arXiv preprint arXiv:2110.07909 (2021).

[25] Padi, Sarala, Seyed Omid Sadjadi, Ram D. Sriram, and Dinesh Manocha. "Improved speech emotion recognition using transfer learning and spectrogram augmentation." In Proceedings of the 2021 international conference on multimodal interaction, pp. 645-652. 2021.

[26] Rolland, Thomas, Alberto Abad, Catia Cucchiarini, and Helmer Strik. "Multilingual transfer learning for children automatic speech recognition." (2022).

[27] Latif, Siddique, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller. "Multitask learning from augmented auxiliary data for improving speech emotion recognition." IEEE Transactions on Affective Computing 14, no. 4 (2022): 3164-3176.

[28] Sullivan, Peter, Toshiko Shibano, and Muhammad Abdul-Mageed. "Improving automatic speech recognition for non-native English with transfer learning and language model decoding." In Analysis and Application of Natural Language and Speech Processing, pp. 21-44. Cham: Springer International Publishing, 2022.

[29] Padi, Sarala, Seyed Omid Sadjadi, Dinesh Manocha, and Ram D. Sriram. "Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models." arXiv preprint arXiv:2202.08974 (2022).

[30] Yadav, Hemant, and Sunayana Sitaram. "A survey of multilingual models for automatic speech recognition." arXiv preprint arXiv:2202.12576 (2022).

[31] Kheddar, Hamza, Yassine Himeur, Somaya Al-Maadeed, Abbes Amira, and Faycal Bensaali. "Deep transfer learning for automatic speech recognition: Towards better generalization." Knowledge-Based Systems 277 (2023): 110851.

[32] Steinmetz, Hillel. "Transfer Learning Using L2 Speech to Improve Automatic Speech Recognition of Dysarthric Speech." Master's thesis, University of Washington, 2023.

[33] Nga, Cao Hong, Duc-Quang Vu, Huong Hoang Luong, Chien-Lin Huang, and Jia-Ching Wang. "Cyclic Transfer Learning for Mandarin-English Code-Switching Speech Recognition." IEEE Signal Processing Letters (2023).

[34] Tun, Su Shwe Yi, Shogo Okada, Hung-Hsuan Huang, and Chee Wee Leong. "Multimodal Transfer Learning for Oral Presentation Assessment." IEEE Access (2023).

[35] Zheng, Ruonan, and Rui Zhang. "Classification of intelligent speech system and education method based on improved multi label transfer learning model. "International Journal of System Assurance Engineering and Management (2023): 1-10.

[36] Zhou, Yue, Yuxuan Yuan, and Xiaodong Shi. "A multitask co-training framework for improving speech translation by leveraging speech recognition and machine translation tasks." Neural Computing and Applications 36, no. 15 (2024): 8641-8656.

[37] Ta, Bao Thang, and Nhat Minh Le. "Transfer learning methods for low-resource speech accent recognition: A case study on Vietnamese language." Engineering Applications of Artificial Intelligence 132 (2024): 107895.

[38] Kheddar, Hamza, Mustapha Hemis, and Yassine Himeur. "Automatic speech recognition using advanced deep learning approaches: A survey." Information Fusion (2024):102422.

[39] Hassan, Najmul, Abu Saleh Musa Miah, and Jungpil Shin. "A Deep Bidirectional LSTM Model Enhanced by Transfer-Learning-Based Feature Extraction for Dynamic Human Activity Recognition." Applied Sciences 14, no. 2 (2024): 603.

[40] Kumar, Aditya, and Jainath Yadav. "Multiview Learning-Based Speech Recognition for Low-Resource Languages." Automatic Speech Recognition and Translation for Low Resource Languages (2024): 375-403.