Academic Journal of Nawroz University (AJNU), Vol.1, No.1, 2025

https://doi.org/10.25007/ajnu.v1n1a61

AI-Powered Insights and Processing for Kurdish Dialects: Connecting Sorani and Bahdini through Advanced Linguistic Technologies

Zainab Salih Ageed [1], Sarkar Hasan Ahmed [2]

1 Technical College of Informatics, Akre University for Applied Science, Duhok, KRG – Iraq

2 Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, KRG – Iraq

ABSTRACT: Metaheuristic The application of AI-driven techniques to Kurdish dialects, particularly Sorani and Bahdini, presents unique challenges in natural language processing (NLP) due to their phonological and syntactical complexities. This paper critically examines existing research focused on the AI-based analysis and processing of these dialects, comparing various methodologies and algorithms employed. From traditional statistical models to modern deep learning approaches, such as transformer architectures, we explore how AI can effectively differentiate and process these dialects. Our comparative study reveals that transformer-based models offer significant improvements over older techniques, leading to enhanced dialect recognition accuracy. This research also identifies key areas for future exploration, aiming to further refine AI models for improved applications in machine translation, speech recognition, and sentiment analysis across Sorani and Bahdini dialects. By bridging the gap between these dialects through AI, this work paves the way for more inclusive and accurate linguistic technologies for Kurdish-speaking communities.

Keywords: AI-driven Techniques, Kurdish Dialects, NLP, Transformer Models, Deep Learning, Dialect Recognition, Speech Recognition.

1. Introduction

Speech recognition technology, also known as automatic speech recognition (ASR), allows computers to understand and process human speech. This technology has become a key part of many everyday applications, such as virtual assistants like Siri and Alexa, voice-activated customer service, and tools that convert spoken words into written text [1].

The journey of speech recognition began in the mid-20th century with basic systems that could only recognize a few words. However, recent advancements in machine learning, particularly deep learning, have greatly improved the accuracy and capabilities of speech recognition systems. These improvements have been made possible by the use of powerful neural networks, large amounts of data, and faster computers [2].Despite these improvements, speech recognition still faces challenges. It can struggle with background noise, different accents, and understanding the context of conversations. Additionally, issues related to data privacy and fairness in recognition systems are important to address to ensure the technology is used responsibly. Overall, speech recognition technology has the potential to significantly improve how we interact with machines and access information. Continued research and development will help overcome current challenges and make this technology even more useful in our daily lives [3].

Automatic Speech Recognition (ASR) has historically driven the development of many machine learning (ML) techniques, including hidden Markov models, discriminative learning, structured sequence learning, Bayesian learning, and adaptive learning. ASR often serves as a large-scale, real-world application to rigorously test the effectiveness of these ML techniques and to inspire new problems due to the sequential and dynamic nature of speech. Despite its commercial availability in some applications, ASR remains largely an unsolved problem. For most applications, ASR performance still falls short of human-level accuracy [4].

2. Literature review

Optimization Speech recognition, deeply intertwined with deep learning, represents a remarkable fusion of technology and human communication. At its core, deep learning leverages complex neural networks to process vast amounts of data, mimicking the intricate workings of the human brain. In the context of speech recognition, deep learning algorithms analyze patterns in audio signals, allowing machines to

understand and transcribe spoken language with unprecedented accuracy [5]. Figure 1 show how to build automatic speech recognition.

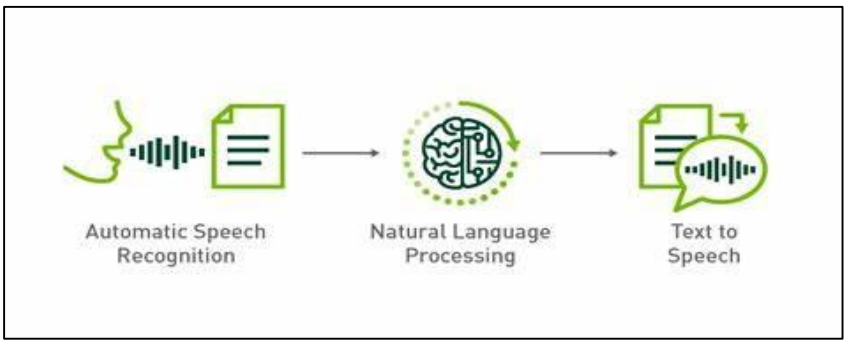

Figure 1: How to build automatic spech recognition [6]

2.1 Deep Learning Algorithms

The use of deep learning, particularly neural networks like Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), has greatly improved the accuracy of speech recognition systems. These models can learn and model complex patterns in speech data, leading to better performance [7]. Deep learning algorithms have predominantly been utilized to advance computer capabilities, particularly in understanding human actions, including speech recognition. Speech, as the primary mode of human communication, has garnered significant attention over the past five decades since the advent of artificial intelligence. Consequently, one of the initial applications of deep learning was in the realm of speech. To date, a substantial number of research papers have been published on the use of deep learning for speech-related applications, specifically speech recognition [8].

Deep learning models can also function as a greedy layer wise unsupervised pre-training method. This approach involves learning hierarchies from features extracted layer by layer. Each layer undergoes training with an unsupervised learning algorithm, using the features from the preceding layer as input for the next. This process enables feature learning to transform previously learned

features at each new layer. With each iteration, a new layer of weights is added to the deep neural network. The layers with these learned weights can eventually be used to initialize a deep supervised predictor [9] .

Deep architectures have proven more efficient in representing non-linear functions compared to shallower ones. Research indicates that fewer parameters are needed to represent a non-linear function in a deep architecture, whereas a shallow architecture requires a larger number of parameters for the same function. This demonstrates the statistical efficiency of deeper architectures [10].

2.2 End-to-End Models

Traditional speech recognition systems used to rely on separate components for acoustic modeling, language

modeling, and phonetic transcription. Modern systems often use end-to-end models that streamline these processes into a single neural network, improving efficiency and accuracy [11]. End-to-end models typically utilize deep learning architectures such as Convolutional Neural Networks (CNNs) for image-related tasks, Recurrent Neural Networks (RNNs) or Transformers for sequence data, and their combinations for more complex tasks [12]. Traditional systems involve separate stages for feature extraction (e.g., MFCCs), acoustic modeling, language modeling, and decoding. End-to-end speech recognition models, like Deep Speech or Transformer-based models, take raw audio as input and directly output the transcribed text [13] . figure 2 show the wave of ASR system:

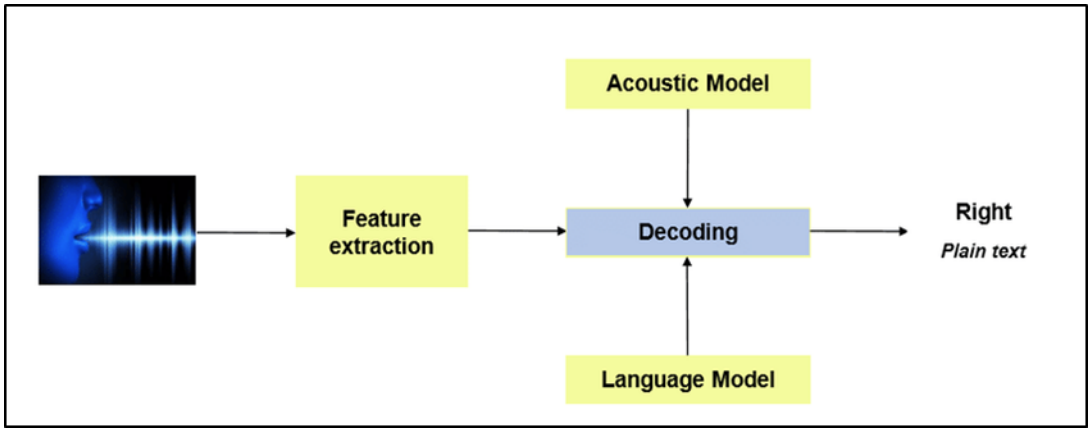

Figure 2: Generic model of an automatic speech recognition (ASR) system [14]

2.3 Large Datasets

The availability of large, high-quality datasets has enabled the training of more robust and accurate models. Crowdsourcing and data collection from various sources have contributed to the development of comprehensive speech databases.

2.4 Real-Time Processing

Advances in computational power and optimization techniques have made it possible to perform speech recognition in real-time. This is crucial for applications like virtual assistants (e.g., Siri, Alexa, Google Assistant) and live transcription services. Real-time processing in deep learning for speech recognition involves the immediate analysis and transcription of spoken language into text as the speech is being delivered. This capability is crucial for applications like virtual assistants, live transcription services, and interactive voice response systems. Here's a detailed exploration of real-time processing in deep learning-based speech recognition [13]:

2.4.1 Low Latency

The system must process audio inputs and generate text outputs with minimal delay to ensure a seamless experience for the user.

2.4.2 High Accuracy

Accurate transcription of spoken words, even in noisy environments, with different accents, and varying speaking speeds.

2.4.3 Robustness

The model should handle a variety of audio qualities and background noises without significant degradation in performance.

2.4.4 Scalability

The ability to handle multiple concurrent users or streams without performance bottlenecks.

2.5 Multilingual and Accented Speech

Modern systems have improved in recognizing multiple languages and various accents. This is achieved through transfer learning and data augmentation techniques, which help models generalize better across different speech patterns[15]. Correct pronunciation is often the hardest part for language learners, whether native or non-native. Accented speech adds more variability, which can cause errors in traditional Automatic Speech Recognition (ASR) training methods that rely on phone alignment. Using end-to-end training can help solve this problem [16].

2.6 Challenges in Speech Recognition

Speech recognition technology has made significant strides over recent years, enabling applications such as virtual assistants, automated transcription services, and real-time translation. Despite these advancements, several challenges persist that affect the accuracy and efficiency of speech recognition systems. These challenges include handling diverse accents, mitigating environmental noise, dealing with

homophones, understanding context, achieving real-time processing, managing limited training data, addressing resource constraints, and coping with pronunciation variability. Each of these challenges poses unique difficulties that need to be addressed to improve the reliability and effectiveness of speech recognition technologies [17].

2.6.1 Accurate Recognition in Noisy Environments

While there have been improvements, accurately recognizing speech in noisy or reverberant environments remains a challenge. Background noise and overlapping speech can significantly degrade performance. Accent recognition involves identifying a speaker's regional accent within a specific language based solely on the acoustic signal. This task is considered more challenging than language recognition due to the greater similarity between accents of the same language. Accent recognition has numerous practical applications[18]. It enhances Automatic Speech Recognition (ASR) by accounting for variations in word pronunciation and phonetic differences among speakers with diverse accents. Additionally, accent recognition can determine a speaker's regional origin and ethnicity, enabling the adjustment of speaker recognition features according to regional characteristics [8, 19]. Figure 3 Demonstrate the impact of background noise on audio signals:

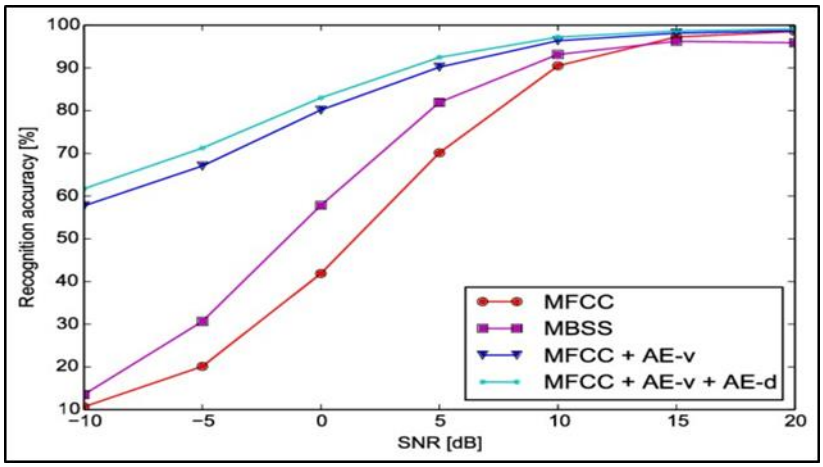

Figure 3: Recognition accuracy in a noisy audio environment [20]

2.6.2 Handling Diverse Accents and Dialects

Despite advancements, speech recognition systems can still struggle with less common accents and dialects. Developing models that can accurately recognize and process diverse speech patterns requires more extensive and varied training data.

2.6.3 Context Understanding

Speech recognition systems often have difficulty understanding context, especially in ambiguous or complex sentences. This can lead to errors in transcription and interpretation, particularly in dynamic conversational scenarios.

2.6.4 Privacy and Security Concerns

The collection and processing of speech data raise significant privacy and security issues. Ensuring that user data is protected and used ethically is a critical challenge for developers and researchers.

2.6.5 Resource Constraints

Deploying advanced speech recognition systems on resource-constrained devices, such as smartphones or IoT devices, is challenging. These systems need to be optimized for low power consumption and limited computational resources while maintaining high accuracy.

2.6.6 Bias and Fairness

Ensuring that speech recognition systems are fair and unbiased across different demographics is an ongoing challenge. Biases in training data can lead to systems that perform better for certain groups of people, which can have significant societal implications.

2.7 Literature Review

In 2017 Y. Zhang et al. [13], They proposed method for labeling unsegmented sequences facilitates the training of end-to-end speech recognition systems, as opposed to traditional hybrid models. Despite this, RNNs pose significant challenges due to their computational intensity and complex training requirements. In this paper, they draw inspiration from the advantageous characteristics of both Convolutional Neural Networks (CNNs) and the Connectionist Temporal Classification (CTC) approach. By presenting an advanced end-to-end speech recognition framework for sequence labeling that directly integrates hierarchical CNNs with CTC, thereby obviating the necessity for recurrent connections.

In the same year, T. N. Sainath et al [21], they investigate the direct modeling of the raw time-domain waveform. The study introduces a neural network architecture that conducts multichannel filtering in its initial layer. They demonstrate that this network exhibits robustness to varying target speaker directions of arrival, achieving performance comparable to a model with access to oracle knowledge of the true target speaker direction. Furthermore, they illustrate a method to enhance performance by splitting the first layer to segregate the multichannel spatial filtering operation from a single-channel filterbank that computes frequency decomposition. Additionally, they introduce an adaptive variant that updates spatial filter coefficients at each time frame based on previous inputs.

In 2020 Song and Zhaojuan [22], They focuses on English speech and proposes a novel deep learning-based approach for speech recognition. This approach combines speech features and attributes to enhance accuracy. Initially, a deep neural network supervised learning method is employed to extract high-level speech features. The output of a fixed hidden layer serves as the new speech feature for training a GMM-HMM acoustic model. Additionally, a speech attribute extractor, based on deep neural networks, is trained to identify various speech attributes. These attributes are further classified into phonemes using another deep neural network. Finally, both speech features and speech attribute features are integrated into the same CNN framework using a neural network based on the linear feature fusion algorithm.

P. K. Roy et al. in 2020 [23], Inflammatory exchanges that use misunderstandings to spread divisive beliefs are the hallmark of hate speech on Twitter, which is the subject of this article. Dissemination of hate speech can result in unwanted criminal situations and cause distress to individuals or groups. It targets a variety of protected traits, including gender, religion, race, and disability. This makes it even more important to keep an eye on user posts and take proactive measures to remove hate speech-related information before it gets out of hand. Nevertheless, manually filtering such massive incoming traffic is practically impossible considering Twitter's enormous volume of over six hundred tweets per second and over 500 million tweets each day.

In 2021 A. Alsobhani et al. [24], they embarked on the development of a word-tracking model utilizing speech recognition features and deep convolutional neural networks. The dataset comprises six control words: start, stop, forward, backward, right, and left, spoken by individuals of varying ages. The dataset, contributed by an equal number of men and women, was utilized to train and test our proposed deep neural networks. Data collection took place across diverse locations including streets, parks, laboratories, and markets. The words ranged in duration from 1 to 1.30 seconds across thirty participants. Leveraging Convolutional Neural Networks (CNNs), an advanced deep learning technique, we conducted multi-class classification to identify each word within our pooled dataset.

In 2022 L. Trinh Van et al. [25], They presents their research on speech emotion recognition using deep neural networks, including CNN, CRNN, and GRU. They utilized the Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus, focusing on four emotions: anger, happiness, sadness, and neutrality. The feature parameters for recognition included Mel spectral coefficients, along with other spectrum-related and intensity-related parameters of the speech signal. Data augmentation techniques, such as altering the voice and adding white noise, were employed to enhance the dataset.

In 2022 B. Saritha et al. [26], they examined the impact of impaired speech on speaker identification tasks by using the IIT-G database, they present the first investigation of mismatch effects in training and test situations, which includes differences in speaking styles and sensors. In speaker identification (SI) systems, convolutional neural networks (CNNs) have recently proven to be more effective than conventional techniques. Then, present a unique end-to-end speaker identification system architecture that draws inspiration from a VGG-like network.

In 2023 B. Saritha et al. [27], They present a unique framework for SER that incorporates data augmentation methods before extracting relevant feature sets from individual utterances and choosing the best suitable features for discrimination. Next, using multilingual databases, the chosen feature vector is fed into the Normalized 1D CNN for emotion recognition. This work evaluates an XGB classifier's

performance on data from a corpus trained on a different corpus in order to determine how effective it is at multilingual emotion recognition. The testing results show that our suggested SER architecture worked better than previous SER methods.

In 2024 N. Barsainyan and D. K. Singh [27], presents a machine learning-based automatic speech recognition framework with four main components: speech acquisition, preprocessing, feature extraction, and classification. They presents a machine learning-based automatic speech recognition framework with four main components: speech acquisition, preprocessing, feature extraction, and classification.

In the same year V. A. Kherdekar and S. A. Naik [28], presents a feature fusion approach for speech recognition in mathematical statements. By merging time, frequency, and cepstral domain characteristics, feature-level fusion improves speech recognition precision. The 600.wav files in the dataset, which includes recordings of ten people speaking six different kinds of mathematical phrases, were utilized to train the model.

3. Discussion

The following a simple discussion about the result of each referenced mentioned in table 1:

[13] By evaluating the approach on the TIMIT phoneme recognition task, they demonstrate that the proposed model is both computationally efficient and competitive with existing baseline systems. Furthermore, they argue that CNNs, when provided with appropriate context information, are capable of effectively modeling temporal correlations. Through comprehensive assessment on the TIMIT phoneme recognition task, our analysis illustrates that the proposed model not only showcases computational efficiency but also contends strongly with the established baseline systems. Furthermore, we posit that CNNs exhibit the prowess to adeptly capture temporal correlations when furnished with appropriate context information.

[21] They illustrate that these methodologies can be executed with greater efficiency in the frequency domain. Their analysis observe that such multichannel neural networks yield a relative word error rate enhancement of over 5% when contrasted with a conventional beamforming-based multichannel ASR system, and over 10% when compared to a single-channel waveform model.

[22] The experimental findings demonstrate the efficacy of the proposed English speech recognition algorithm, which leverages a deep neural network integrating multiple features. By amalgamating both speech features and speaker attributes directly at the input layer of the deep neural network, the algorithm effectively merges the two methods. This integration substantially enhances the performance of the English speech recognition system.

[23] The suggested DCNN model exhibited superior performance compared to its predecessors, utilizing tweet text alongside GloVe embedding vectors to encode tweet semantics through convolutional operations. In the most favorable scenario, the model attained precision, recall, and F1-score metrics of 0.97, 0.88, and 0.92, respectively.

[24] The devised deep neural network attained an impressive word classification accuracy of 97.06% when confronted with an entirely unfamiliar speech sample. CNN served as the tool for both training and testing the dataset. Our research distinguishes itself from numerous other studies, which often rely on readily available and relatively uniform datasets consisting of isolated words. In contrast, our dataset encompasses speech samples gathered from diverse and noisy environments, presenting a blend of isolated and continuous speech forms.

[25] The results indicate that the GRU model achieved the highest average recognition accuracy of 97.47%, surpassing the performance of existing studies on speech emotion recognition using the IEMOCAP corpus. In addition to selecting suitable models and feature parameters, data augmentation has proven effective, particularly when the available data is limited. Although data augmentation increases memory usage and training time, it significantly enhances the recognition system's performance.

[26] The suggested design outperforms statistical techniques, improving speaker identification accuracy by 8%. The results show that this method works better than the most advanced speaker recognition methods, however there is a discernible drop in performance when compared to the matched case.

[27] retrieved the local representation has been automatically from provided voice samples using deep learning methods. The methodologies provided are unable to identify the most valuable traits from difficult voice inputs. To overcome this limitation. This study evaluated the effectiveness of an XGB classifier for multi-lingual emotion recognition by testing its performance on data from a corpus trained on a different corpus. The testing outcomes showed that our proposed SER architecture performed better than existing SER approaches.

[27] Speech recognition has the potential to significantly improve the lives of those who struggle with speech. They present an automatic speech recognition framework driven by machine learning, with speech acquisition, preprocessing, feature extraction, and classification serving as its main building blocks.

[28] Training the suggested feature extraction model involved using both a sequential and KNN classifier. According to the findings, the suggested feature extraction approach outperformed merely utilizing the time, frequency, and cepstral domains for the sequential classifier and KNN in terms of accuracy.

6.Comparison Among Reviewed Works

Ten publications covering various algorithms are included in the following table, as shown below in Table 1:

Table 1: Comparison between reviewed works

	Ref.	Year	Method	Results
1.	[13]	2017	Hybrid speech recognition systems that integrate CNNs with Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs)	Drawing inspiration from the respective advantages of both CNNs and the CTC approach, we introduce a novel end-to-end speech framework tailored for sequence labeling. This framework seamlessly integrates hierarchical CNNs with CTC, thereby circumventing the necessity for recurrent connections.
2.	[21]	2017	Conduct multichannel enhancement simultaneously with acoustic modeling within a deep NN framework.	The analysis reveals that these multichannel neural networks yield a notable relative improvement in word error rate, exceeding 5% compared to a conventional beamforming-based multichannel ASR system, and surpassing 10% when contrasted with a single-channel waveform model.
3.	[22]	2020	train the GMM–HMM By using a neural network based on the linear feature fusion algorithm	The experimental findings indicate that the suggested English speech recognition algorithm, employing a deep neural network with various features, seamlessly integrates two approaches by merging speech features and speaker attributes at the input layer. This integration leads to a substantial improvement in the performance of the English speech recognition system.
4.	[23]	2020	an automated system is developed using the Deep Convolutional Neural Network (DCNN).	The suggested DCNN model outperformed the previous models by using the tweet text with the GloVe embedding vector to capture the semantics of the tweets using a convolution operation. For the best case, the model achieved precision, recall, and F1-score values of 0.97, 0.88, and 0.92, respectively.
5.	[24]	2021	applying speech recognition features with deep convolutional neuro-learning and training CNN	employed CNN for both training and testing dataset. The research stands out from numerous other studies, which frequently rely on pre-existing datasets of isolated words that are relatively uniform in nature. In contrast, the dataset comprises speech samples collected from various noisy environments and conditions, encompassing two types of speech: isolated words and continuous speech.
6.	[25]	2022	deep neural-network models CNN, CRNN, and GRU were used for emotion recognition	The GRU model, with 153 parameters, achieved the highest recognition accuracy of 97.47%, outperforming the CNN and CRNN models. For both the CNN and GRU models, increasing the number of parameters from 128 (S1 set) to 153 (S2 set) resulted in higher average recognition accuracy. However, for the CRNN model, the average recognition accuracy decreased as the number of parameters increased from 128 to 153. The GRU model's ability to retain information from the past while avoiding the vanishing gradient problem proved advantageous in this scenario.
7.	[26]	2022	proposes a novel architecture based on a VGG-(CNNs) have surpassed traditional techniques in (SI) systems	They introduce a novel architecture inspired by a VGG-like network for an end-to-end speaker identification system. This architecture surpasses statistical methods, achieving an 8% improvement in speaker identification accuracy. The results demonstrate that the proposed approach is more accurate than state-of-the-art speaker identification techniques, albeit with noticeable performance degradation compared to the matched scenario.
8.	[27]	2023	an XGB classifier for multi-lingual emotion recognition	In this work, the performance of an XGB classifier for multilingual emotion recognition was assessed using data from a separate corpus, which served as the training set. According to the testing results, our suggested SER architecture outperformed other SER approaches.
9.	[27]	2024	machine learning-based framework for automatic speech recognition	Presented a framework grounded in machine learning for the automatic recognition of speech. At its core, the framework comprises four main components: speech acquisition, preprocessing, feature extraction, and classification
10.	[28]	2024	using a trained sequential classifier and KNN classifier	presented a technique for speech detection in mathematical expressions based on feature fusion. According to the findings, the suggested feature extraction approach outperformed merely utilizing the time, frequency, and cepstral domains for the sequential classifier and KNN in terms of accuracy.

4. **Conclusion**:

Convolutional Neural Networks (CNNs) are effective models for reducing spectral variations and modeling spectral correlations in acoustic features for automatic speech recognition (ASR). The CNN is employed for both training and testing the data. Numerous researchers depend on CNN to differentiate words in speech recognition, as evidenced in references [24], [25] and [26, 27] while [21] their analysis reveals that these multichannel neural networks yield a notable relative improvement in word error rate, exceeding 5% compared to a conventional beamforming-based multichannel ASR system, and surpassing 10% when contrasted with a single-channel waveform model. [23] aims to enhance hate speech detection on Twitter by leveraging tweet texts as input for prediction, thus streamlining the manual feature extraction process. This approach achieved commendable prediction accuracy on an imbalanced dataset and surpassed the performance of existing models.

5. References

[1] M. Johnson, S. Lapkin, V. Long, P. Sanchez, H. Suominen, J. Basilakis, et al., "A systematic review of speech recognition technology in health care," BMC medical informatics and decision making, vol. 14, pp. 1-14, 2014.

[2] R. Jampala, D. S. Kola, A. N. Gummadi, M. Bhavanam, and I. R. Pannerselvam, "The Evolution of Voice Assistants: From Text-to-Speech to Conversational AI," in 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), 2024, pp. 1332-1338.

[3] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 9, pp. 1-28, 2018.

[4] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, pp. 1060-1089, 2013.

[5] A. Izbassarova, A. Duisembay, and A. P. James, "Speech recognition application using deep learning neural network," Deep Learning Classifiers with Memristive Networks: Theory and Applications, pp. 69-79, 2020.

[6] X. Lu, S. Li, and M. Fujimoto, "Automatic speech recognition," Speech-to-speech translation, pp. 21-38, 2020.

[7] U. Kamath, J. Liu, and J. Whitaker, Deep learning for NLP and speech recognition vol. 84: Springer, 2019.

[8] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," IEEE access, vol. 7, pp. 19143-19165, 2019.

[9] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in 2013 IEEE international conference on acoustics, speech and signal processing, 2013, pp. 6645-6649.

[10] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in 2013 IEEE international conference on acoustics, speech and signal processing, 2013, pp. 8599-8603.

[11] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," Symmetry, vol. 11, p. 1018, 2019.

[12] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMs," in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2017, pp. 2592-2596.

[13] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," arXiv preprint arXiv:1701.02720, 2017.

[14] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic speech recognition using advanced deep learning approaches: A survey," Information Fusion, p. 102422, 2024.

[15] Y. Peng, J. Zhang, H. Zhang, H. Xu, H. Huang, S. Li, et al., "Multilingual approach to joint speech and accent recognition with DNN-HMM framework," in 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2021, pp. 1043-1048.

[16] T. Viglino, P. Motlicek, and M. Cernak, "End-to-End Accented Speech Recognition," in Interspeech, 2019, pp. 2140-2144.

[17] M. Y.-C. Jiang, M. S.-Y. Jong, N. Wu, B. Shen, C.-S. Chai, W. W.-F. Lau, et al., "Integrating automatic speech recognition technology into vocabulary learning in a flipped English class for Chinese college students," Frontiers in Psychology, vol. 13, p. 902429, 2022.

[18] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, et al., "Automatic speech recognition and speech variability: A review," Speech communication, vol. 49, pp. 763-786, 2007.

[19] M. Djellab, A. Amrouche, A. Bouridane, and N. Mehallegue, "Algerian Modern Colloquial Arabic Speech Corpus (AMCASC): regional accents recognition within complex socio-linguistic environments," Language Resources and Evaluation, vol. 51, pp. 613-641, 2017.

[20] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, et al., "Icassp 2023 deep noise suppression challenge," IEEE Open Journal of Signal Processing, 2024.

[21] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, et al., "Multichannel signal processing with deep neural networks for automatic speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, pp. 965-979, 2017.

[22] Z. Song, "English speech recognition based on deep learning with multiple features," Computing, vol. 102, pp. 663-682, 2020.

[23] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, "A framework for hate speech detection using deep convolutional neural network," IEEE Access, vol. 8, pp. 204951-204962, 2020.

[24] A. Alsobhani, H. M. ALabboodi, and H. Mahdi, "Speech recognition using convolution deep neural networks," in Journal of Physics: Conference Series, 2021, p. 012166.

[25] L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," Sensors, vol. 22, p. 1414, 2022.
[26] B. Saritha, S. K. Sharma, T. T. V. N. Manoj, R. H. Laskar, and M. Choudhury, "DNN Based Speaker Identification System Under Multi-Variability Speech Conditions," in 2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), 2022, pp. 233-236.
[27] N. Barsainyan and D. K. Singh, "Optimized cross-corpus speech emotion recognition framework based on normalized 1D convolutional neural network with data augmentation and feature selection," International Journal of Speech Technology, vol. 26, pp. 947-961, 2023.
[28] V. A. Kherdekar and S. A. Naik, "Feature Fusion Extraction Method for Improvement of Recognition of Continuous Speech: A Feature Fusion Method for Recognition of Continuous Speech," in 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2024, pp. 1-7.