Article

Spoken Kashmiri Recognition with Dual Feature Extraction and Spectrogram Augmentation Using a CNN-gMLP Hybrid Model

Umer Ayub Hajam , Syed Tanzeel Rabani , Akib Mohi Ud Din Khanday , Mehdi Neshat *

Posted Date: 24 September 2024

doi: 10.20944/preprints202409.1876.v1

Keywords: Automatic Speech recognition; Hybrid convolutional neural networks; Kashmiri Language; Spectrogram; Classification

Article

Spoken Kashmiri Recognition with Dual Feature Extraction and Spectrogram Augmentation Using a CNN-gMLP Hybrid Model

Umer Ayub Hajam [1,2,3†] **,Syed Tanzeel Rabani** [4,†] **, Akib Mohi Ud Din Khanday** [5,†*]
and Mehdi Neshat [6,*]

[1] Emplay Analytics Pvt Ltd. Designation, Srinagar, Jammu & Kashmir, India. (umar.hajam@emplay.net);
[2] Birla Institute of Science and Technology, Department of Computer Science & Information Systems, BITS, Pilani, Rajasthan 333031, India.
[3] Department of Artificial Intelligence, Amity Online University, Noida, Uttar Pradesh, 533103, India
[4] Department of Computer Science, Islamic University of Science and Technology, Kashmir, 192122, India. (Tanzeel.Rabani@ieee.org & stanzeelr2013@gmail.com)
[5] Department of Computer Science, Samarkand International University of Technology, Samarkand, 140100, Uzbekistan; (akib.khanday@siut.uz)
[6] Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW, 2007, Australia ; (mehdi.neshat@uts.edu.au)
[*] Correspondence: mehdi.neshat@uts.edu.au
[†] These authors contributed equally to this work.

Abstract: Automatic speech recognition of native languages plays a crucial role in fostering inclusivity and preserving linguistic diversity. The Kashmiri language, an underrepresented Indo-Aryan dialect primarily spoken in the Kashmir Valley, poses substantial challenges for existing AI models due to its phonetic diversity and scant linguistic resources. This study addresses these hurdles by developing a robust spoken Kashmiri recognition system that employs dual feature extraction with spectrogram augmentation and a hybrid Convolutional Neural Network (CNN) and Gated Multi-Layer Perceptron (gMLP) model. Key to this endeavour is the creation of a high-fidelity dataset that captures the phonetic variations across Kashmiri dialects, focusing on twelve specific words. Through the integration of dual feature extraction, spectrogram augmentation, and the innovative hybrid modelling approach, our system attains an impressive 96% accuracy on the test dataset for classifying these twelve spoken words. This research not only enhances the generalization and resilience of spoken Kashmiri recognition systems but also represents a critical step towards advancing technology and safeguarding the Kashmiri language within this underrepresented linguistic domain.

Keywords: Automatic Speech recognition; Hybrid convolutional neural networks; Kashmiri Language; Spectrogram; Classification

1. Introduction

The rapid advancement of speech recognition technology has revolutionized human-computer interaction, enabling machines to understand and process spoken language with increasing accuracy. Despite significant strides in this domain, challenges persist, particularly for underrepresented languages with limited linguistic resources. Automatic Speech Recognition (ASR) systems have become increasingly sophisticated, moving from simple systems that recognized a limited set of sounds to advanced models capable of interpreting natural language in real time. However, much of the progress in ASR has been concentrated on widely spoken languages, leaving many regional and minority languages, like Kashmiri, on the periphery of these technological developments [1,2].Kashmiri, an Indo-Aryan language spoken primarily in the Kashmir Valley, presents unique challenges for speech recognition due to its complex phonetic structure and significant dialectal variations [3**?** ,4]. The language is characterized by rich consonantal distinctions and a variety of dialects, each with its phonetic nuances. These characteristics, coupled with the scarcity of high-quality linguistic data, have hindered the development of effective ASR systems for Kashmiri. Traditional approaches to speech recognition, which rely heavily on extensive training datasets and standardized pronunciation patterns, often fail to capture the linguistic diversity inherent in Kashmiri. This gap underscores the need for specialized models and methodologies that accurately reflect the phonetic richness of underrepresented languages [5]. Recent advancements in ASR technology, particularly in multilingual contexts, have demonstrated the potential of scalable, self-supervised learning models to address the challenges of resource-scarce languages. For example, Google's Universal Speech Model (USM) has shown promising results in scaling ASR to over 100 languages by leveraging large-scale pre-training and multilingual datasets [1]. These models emphasize the importance of incorporating diverse linguistic inputs and advanced feature extraction techniques to improve ASR performance across different languages and dialects. Furthermore, the use of hybrid neural architectures, such as Convolutional Neural Networks (CNNs) combined with Gated Multi-Layer Perceptrons (gMLPs), has proven effective in handling the intricate patterns present in speech data, offering a path forward for languages like Kashmiri that require both local and global feature representations [2,6].

This research aims to bridge the gap in ASR technology for the Kashmiri language by developing a robust spoken recognition system that integrates dual feature extraction techniques—Mel-Frequency Cepstral Coefficients (MFCCs) and partial Mel-Spectrograms—with a hybrid CNN-gMLP model. By focusing on Kashmiri's phonetic intricacies and dialectal variations, this study seeks to enhance the accuracy and generalizability of ASR systems for underrepresented languages, contributing to both linguistic preservation and technological inclusively. Some of the Novel Contributions of this work are:

- Creation of a comprehensive and linguistically diverse dataset that accurately represents the phonetic and dialectal variations of the Kashmiri language. It will serve as a foundational resource for training and evaluating the recognition system.
- Developing a feature extraction technique that captures both spectral and temporal characteristics of Kashmiri speech for enhancing the system's ability to recognize and distinguish the unique phonetic patterns of the language.
- Development of a robust spoken Kashmiri recognition system that effectively addresses the phonetic diversity and dialectical variations within the Kashmiri language.

The article is divided into six sections; in section 2, related work is discussed in detail. Section 3 discusses the proposed methodology for spoken Kashmiri recognition, section 4 discusses the results that are generated from the proposed approach, and section 5 concludes the work by giving some future research directions.

2. Literature Survey

The development of Automatic Speech Recognition (ASR) systems for underrepresented languages like Kashmiri has been a significant area of research, driven by the unique challenges these languages present. Early foundational work by Besacier et al. [5] highlighted the substantial obstacles faced by resource-scarce languages in developing effective ASR systems, particularly due to the lack of high-quality datasets and the need for models capable of adapting to linguistic diversity. Kashmiri, with its complex phonetic structure and rich consonantal distinctions, exemplifies these challenges [3]. The phonetic diversity within the Kashmiri language necessitates the development of datasets encompassing a wide range of dialects and pronunciations, as emphasized by studies such as those by O'Neill and Carson-Berndsen [7,8], which underscore the difficulties ASR systems encounter when processing underrepresented dialects, leading to higher word error rates. Hannun et al. [9] further reinforced that the performance of machine learning models is inherently tied to the quality and diversity of the training data, making comprehensive data collection an essential component in addressing the challenges of ASR for Kashmiri. As research progressed, the focus shifted towards more sophisticated modelling techniques. Carvalho and Gomes [10] demonstrated the effectiveness of combining Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-Spectrograms in audio classification relevant to speech recognition. However, it is important to note that while these techniques are beneficial, they are not novel.

What differentiates this study is the introduction of a partial Mel-Spectrogram, focusing on specific frequency bands crucial for capturing the tonal variations characteristic of Kashmiri phonetics. This targeted extraction approach aims to enhance the robustness of feature representation by mitigating dimensionality and computational load without sacrificing the richness of the audio data. Furthermore, hybrid modeling techniques have shown considerable promise in advancing ASR systems. Dosovitskiy et al. [11] and Liu et al. [6] highlighted the effectiveness of combining Convolutional Neural Networks (CNNs) with Gated Multi-Layer Perceptrons (gMLPs) in handling complex data patterns. These models leverage CNNs' strengths in local feature extraction and gMLPs in capturing global dependencies, making them well-suited for the intricate phonetic structure of languages like Kashmiri. Nevertheless, applying such hybrid models to the Kashmiri language, with its specific phonetic and dialectal challenges, has not been extensively explored, presenting a gap that this research seeks to address. In recent years, addressing data scarcity has become a central focus in developing ASR systems for underrepresented languages. Zhao et al. [12] discussed the use of NN-HMM acoustic models and N-gram Language Models to construct basic ASR systems for low-resource languages. Subsequently, Kipyatkova and Kagirov [13] provided a comprehensive review of methods to solve training data issues, advocating for data augmentation, transfer learning, and crowdsourcing as viable solutions. Zhao and Zhang [14] explored the use of self-supervised models such as wav2vec2.0, HuBERT, and WavLM for ASR in low-resource languages, marking another critical advancement. The introduction of the MAC framework by Min et al. [15] showed significant improvements in character error rates (CER) for languages like Cantonese, Taiwanese, and Japanese, offering insights applicable to Kashmiri ASR. More recent studies by Bartelds et al. [16] demonstrated the effectiveness of data augmentation techniques like self-training and text-to-speech (TTS) in improving ASR systems for underrepresented languages. Yeo et al. [17] proposed a Visual Speech Recognition method using automatic labels from Whisper, significantly increasing training data for low-resource languages. The most recent work by Pratama and Amrullah [18] focused on fine-tuning the Whisper model, achieving substantial Word Error Rate (WER) reductions for low-resource languages with minimal computational cost. Bekarystankyzy et al. [19] extended these findings to agglutinative languages, showcasing the potential of transfer learning in mitigating data scarcity. Despite these advancements, significant challenges remain, particularly in the scarcity of annotated datasets and the pronounced dialectal variations within the Kashmiri language. Wali [4] highlights that the use of multiple scripts—such as Sharada, Devanagari, Perso-Arabic, and Roman—further complicates the development of comprehensive ASR systems. Additionally, Sullivan and Harding [20] emphasize the domain shift in dialect identification, which further complicates the development of robust ASR systems. This literature review illustrates the ongoing challenges and innovations in dataset development, feature extraction, and hybrid modeling to advance ASR systems for underrepresented languages like Kashmiri. While dual feature extraction and spectrogram augmentation are not novel in themselves, their application in a targeted manner within the framework of a hybrid CNN-gMLP model represents an important step toward addressing the unique phonetic challenges of the Kashmiri language. This approach not only aims to improve the accuracy and robustness of spoken Kashmiri recognition but also contributes to the broader goal of technological inclusion and the preservation of linguistic diversity.

3. Methodology

Kahmiri language has gained less interest from researchers due to its geographical diversity and delicate issues. Around 8 million people speak Kashmiri, and still, there is no Kashmiri recognition system. To accomplish this challenge, we propose a structured and comprehensive methodology that is illustrated in Figure 1. The key stages of this methodology are Data Collection, Data Analysis and Processing, Feature Extraction, Model Development and Performance Evaluation. These are discussed in detail in following subsections. This methodology ensures that the Automatic Speech Recognition (ASR) system developed is both accurate and reliable, capable of effectively handling the diverse phonetic landscape of the Kashmiri language.

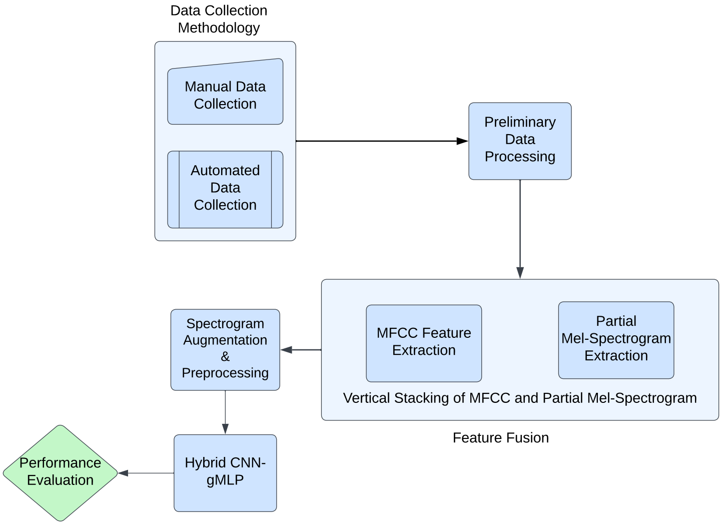

Data Collection Methodology
Manual Data Collection
Automated Data Collection
Preliminary Data Processing
MFCC Feature Extraction
Partial Mel-Spectrogram Extraction
Vertical Stacking of MFCC and Partial Mel-Spectrogram
Feature Fusion
Spectrogram Augmentation & Preprocessing
Hybrid CNN-gMLP
Performance Evaluation

Figure 1. Process Flow for Spoken Kashmiri Recognition System Development Using Hybrid CNN-gMLP Model

3.1. Data Collection

The data collection phase aimed to construct a comprehensive and linguistically diverse dataset for the Kashmiri speech recognition system. The primary objective was to ensure the dataset captured the phonetic and dialectal diversity inherent in the Kashmiri language, focusing on twelve specific words essential for fundamental communication. This section details the rigorous methodology employed, the challenges encountered, and the strategies implemented to ensure an unbiased and representative dataset.

3.1.1. Geographical and Demographic Diversity

To accurately reflect the phonetic diversity of the Kashmiri language, data was gathered from three major regions of Kashmir: North Kashmir (Baramullah, Kupwara), Central Kashmir (Srinagar, Magam), and South Kashmir (Pulwama, Shopian). Each region was chosen for its unique linguistic characteristics:

- **North Kashmir (Baramullah, Kupwara):** Known for its conservative dialects, this region retains older phonetic features of Kashmiri, thus providing a rich source of traditional linguistic elements [21].
- **Central Kashmir (Srinagar, Magam):** The variety of Kashmiri spoken in Srinagar is often regarded as the standard dialect, forming the linguistic baseline against which other dialects are compared [22].
- **South Kashmir (Pulwama, Shopian):** Characterized by softer pronunciations and unique intonations, the dialects here offer a distinct contrast to those of the northern and central regions, contributing to the overall diversity of the dataset [23].

By including these diverse regions, the dataset was designed to encompass the full spectrum of Kashmiri's phonetic and dialectical variations, ensuring comprehensive representation [24,25].

3.1.2. Participant Selection and Data Collection Process

Participants were meticulously selected to represent a balanced mix of age, gender, and regional background, ensuring the dataset captured a wide range of vocal characteristics. Native Kashmiri speakers were recruited through local networks, with selection criteria emphasizing fluency in the regional dialects. This careful selection ensured linguistic authenticity, crucial for developing an effective speech recognition system.

The recordings were conducted using high-quality audio equipment in controlled environments to ensure consistency and clarity across the dataset. Each word was recorded multiple times to capture various pronunciations, intonations, and speech patterns. The Open Speech Corpus tool, an open-source framework for managing and processing speech datasets, was adapted to suit our requirements. This tool played a key role in standardizing the recording process, providing functionalities such as automatic speaker verification, noise reduction, and format conversion, ensuring that all samples met stringent quality standards. Its flexibility allowed for customization in handling diverse dialects and recording conditions, making it an essential component of our data collection methodology [26,27].

3.1.3. Dataset Overview

The collected data consists of twelve specific words frequently used in everyday Kashmiri communication. These words were meticulously selected not only for their phonetic diversity but also for their significance in basic interactions within the Kashmiri-speaking community. Their common usage across different dialects and regions drove their choice, making them essential for developing a comprehensive and effective speech recognition system that can generalize well across various linguistic contexts.

Table 1 provides a detailed list of these words, their pronunciations and the number of voice samples recorded for each. This table highlights the diversity captured in the dataset, both in terms of the phonetic characteristics of the words and the demographic representation of the speakers. The dataset ensures that a wide range of pronunciations, intonations, and speech patterns are represented by including multiple samples for each word. This is critical for training a robust and accurate speech recognition model.

Table 1. List of Kashmiri Words with Pronunciations and Voice Samples

Kashmiri (Koshur / کاشُر	**Pronunciation**	**English**	**Voice Samples**
آ	āh	Yes	70
اَڈسا.	aḍsā	OK	70
بَند	band	Closed	70
بوٚہ	bē	Me	70
خَبَر	khabar	News	70
کیاہ	k'ah	What	70
نَہ	na	No	70
نٔو	nov (m.)	New	70
ٹھیک	'theek	Well/Okay	70
وارَیے	va:ray	Well	70
وچھ	vuch	See	70
یلہٕ	ye:le	Open	70

The dataset comprises 840 samples, which provide a solid foundation for the subsequent stages of data processing and model training. Including multiple samples per word is particularly important as it captures variations in pronunciation that may arise from differences in speaker background, including regional dialects, age, and gender. This diversity is crucial for training a speech recognition system that is accurate and robust enough to handle the inherent variability in human speech.

Furthermore, these specific words were selected by their phonetic characteristics, which cover a broad spectrum of Kashmiri phonology. The dataset includes words representing different types of sounds, such as voiced and unvoiced consonants, fricatives, and vowels, ensuring that the model is exposed to the full range of phonetic diversity in the language. For example, the word "k'ah" (What) contains the glottal stop, a feature that is characteristic of certain Kashmiri dialects, while "ye:le" (Open) involves a long vowel sound, which is another important feature in the language's phonetic structure.

The careful design of this dataset is pivotal to the success of the speech recognition system. By providing a comprehensive and representative sample of Kashmiri speech, the dataset ensures that the model can learn to recognize and interpret the subtle nuances of the language. This is particularly important for underrepresented languages like Kashmiri, where high-quality linguistic resources are limited. The creation of this dataset not only contributes to the development of an effective speech recognition system but also represents a significant step towards the preservation and technological advancement of the Kashmiri language.

3.2. *Preliminary Data Analysis and Processing*

A comprehensive analysis of the collected dataset was conducted to ensure its representativeness and to capture the phonetic diversity inherent in the Kashmiri language. The analysis involved examining key audio signal characteristics, which play a crucial role in understanding the acoustic properties of the dataset. These characteristics are essential for evaluating the dataset's suitability for the development of a robust speech recognition system.

3.2.1. Zero-Crossing Rate (ZCR):

The Zero-Crossing Rate (ZCR) provides valuable insights into the smoothness and texture of the speech signal by measuring the frequency at which the signal changes sign. In our dataset's context, ZCR helps identify the presence of different phonetic elements, particularly unvoiced fricatives, common in various Kashmiri dialects. By analyzing the ZCR values across the dataset, we can assess the distribution of these phonetic elements, thereby ensuring that the dataset captures the full range of phonetic diversity within the Kashmiri language [28].

3.2.2. Root Mean Square (RMS) Energy:

RMS Energy serves as a quantitative measure of the speech signal's power, offering insights into the intensity and prominence of different speech segments. In our analysis, RMS Energy is crucial for distinguishing between voiced and unvoiced speech segments, particularly in a language like Kashmiri that exhibits significant tonal variations. By evaluating the RMS Energy across different samples, we can ensure that the dataset reflects the dynamic range of vocal expressions, which is vital for the accuracy of the speech recognition system [29].

3.2.3. Spectral Entropy:

Spectral Entropy measures the complexity and unpredictability of the speech signal's frequency spectrum, providing a deeper understanding of the signal's structure. This metric is particularly useful in differentiating between speech and non-speech segments within the dataset, especially in varying acoustic environments. High Spectral Entropy values indicate more complex and less predictable signals, which may correspond to speech segments in noisy conditions. Analyzing Spectral Entropy allows us to assess the dataset's robustness in handling different acoustic conditions, thereby improving the system's performance in real-world scenarios [30].

Table 2 summarizes key metrics derived from the dataset, providing a quantitative overview of its characteristics. The average duration of the audio samples ranges from 0.574 to 1.536 seconds, ensuring the inclusion of both brief and extended pronunciations. The range of ZCR values reflects the presence of both voiced and unvoiced segments, while the RMS Energy values indicate varying levels of vocal intensity across the samples. These metrics underscore the dataset's ability to capture the phonetic variability necessary for developing a robust and accurate Kashmiri speech recognition system.

Table 2. Dataset Metrics for Kashmiri Speech Recognition

Metric	Min Value	Max Value
Average Duration (s)	0.574	1.536
Zero-Crossing Rate	0.0515	0.3160
RMS Energy	0.0156	0.1411

The data analysis conducted here lays the groundwork for further processing stages, such as model training and evaluation. By thoroughly understanding the acoustic properties of the dataset, we can ensure that the subsequent stages of development, including feature extraction and classification, are based on a robust and representative dataset. This careful analysis is fundamental to improving the accuracy and generalization of the Kashmiri speech recognition system [26].

The waveform depicted in Figure 2 represents the acoustic signal of the Kashmiri word "Band" (Closed). This waveform highlights key temporal features of the speech signal, including variations in amplitude that correspond to different phonetic elements. The sharp increases in amplitude indicate voiced segments, where the vocal cords are actively vibrating, producing higher energy output [31].

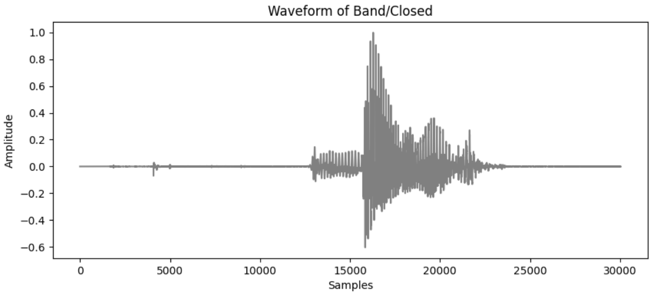


Figure 2. Waveform of the word "Band/Closed"

Understanding waveforms is critical in the context of data analysis, as they provide a visual representation of the signal's temporal structure. This analysis helps identify patterns that are essential for the accurate recognition of speech sounds. The amplitude variations, for instance, correlate with the RMS Energy values, which further supports the differentiation between voiced and unvoiced speech sounds within the dataset [32]. Additionally, analyzing the waveform helps us understand the periodicity of the signal, which is reflected in the ZCR values and is key to distinguishing between different types of speech sounds, such as fricatives and stops [33].

This detailed analysis of the dataset's acoustic properties ensures that the Kashmiri speech recognition system is built on a solid foundation capable of accurately capturing the linguistic diversity and complexity of the language.

3.3. Feature Extraction

The effectiveness of our spoken Kashmiri recognition system is grounded in the advanced data processing techniques employed, particularly the dual feature extraction and spectrogram augmentation. High-quality and diverse datasets are crucial for developing robust models that generalize to new data. Our approach to optimizing the dataset includes carefully crafted steps such as dual feature extraction, spectrogram augmentation, and standard scaling to enhance model performance.

3.3.1. Existing Methods:

The use of Mel-Frequency Cepstral Coefficients (MFCCs) is well-established in the field of speech recognition due to their effectiveness in capturing the power spectrum of audio signals [10]. However, despite their popularity, MFCCs exhibit certain limitations, particularly in noisy environments. Their focus on short-term power spectra can result in a loss of temporal dynamics and transient characteristics, which are crucial for accurate speech recognition in languages with complex phonetic structures, such as Kashmiri [34,35]. Moreover, MFCCs' susceptibility to background noise often degrades the performance of speech recognition models, necessitating the exploration of alternative or supplementary feature extraction methods.

Recent advancements have addressed these limitations by combining MFCCs with other feature extraction techniques. For example, the integration of Principal Component Analysis (PCA) with MFCCs has enhanced accuracy and reduced data dimensionality in Indonesian speech recognition systems [34]. Similarly, the use of Mel-Spectrograms, which provide a detailed representation of both spectral and temporal variations, has gained traction in conjunction with deep learning models like Convolutional Neural Networks (CNNs) [36]. The combination of Constant-Q Transform (CQT) with Mel-Spectrograms has also been proposed to capture more intricate spectral features, demonstrating potential improvements over traditional MFCC methods [37].

3.3.2. Challenges and Limitations:

Despite these innovations, several challenges persist. The hybrid approaches that combine MFCCs with deep learning models, such as CNNs, often require significant computational resources, limiting their applicability in real-time or resource-constrained environments [38]. Additionally, methods that rely solely on Mel-Spectrograms or similar features may offer rich spectral information but often at the cost of computational efficiency and temporal precision. This trade-off highlights the need for a balanced approach that leverages the strengths of both MFCCs and spectrogram-based features while mitigating their respective weaknesses.

3.3.3. Rationale for Our Approach:

Given the challenges identified in existing methodologies, we propose a dual feature extraction approach that combines MFCCs with partial Mel-Spectrograms. Our approach is specifically designed to balance the spectral richness provided by Mel-Spectrograms with the temporal precision of MFCCs, while also addressing the computational efficiency concerns. Unlike traditional Mel-Spectrogram techniques that utilize the full frequency range, our method focuses on specific Mel bands most relevant to the phonetic characteristics of the Kashmiri language. This targeted extraction reduces dimensionality and computational load, capturing the most critical spectral information without compromising efficiency.

Our experiments tested multiple Mel frequency bands for the partial Mel-Spectrogram, including several combinations from low to high Mel bins. These experiments involved testing ranges like 10-50 Mel, 30-60 Mel, and 50-120 Mel bins. Among these, the range of 20-60 Mel bins consistently provided the highest accuracy. This particular range proved optimal for capturing Kashmiri's critical phonetic elements, leading to an accuracy rate of 86% in our Random Forest classifier. This performance indicates the effectiveness of our dual feature extraction approach, balancing the spectral richness of the Mel-Spectrogram with the temporal precision of MFCCs.

$$\text{Partial Mel-Spectrogram}(t, f) = 20\log_{10}\left(|\text{STFT}(x(t)) \cdot M(f)|^2\right)[f_{\text{start}} : f_{\text{end}}] \tag{1}$$

where $\text{STFT}(x(t))$ is the Short-Time Fourier Transform of the signal $x(t)$ and $M(f)$ represents the Mel filter bank. This partial representation is critical for understanding speech's tonal and dynamic elements, particularly in capturing the subtle phonetic nuances of Kashmiri [39]. The choice of the 20-60 Mel bin range was informed by extensive empirical testing, demonstrating that this range effectively captures the tonal variations distinctive to the Kashmiri language. While MFCCs offer a strong foundation for capturing the cepstral characteristics of speech, their susceptibility to noise and loss of temporal detail are well-documented concerns [9]. By combining MFCCs with partial Mel-Spectrograms, we aim to balance these limitations with the strengths of Mel-Spectrograms, resulting in a more robust feature set. The final feature vector is constructed by stacking the partial Mel-Spectrogram and MFCC features vertically, as shown below:

$$\mathbf{F}_{\text{combined}} = \begin{bmatrix} \mathbf{F}_{\text{Partial Mel-Spectrogram}} \\ \mathbf{F}_{\text{MFCC}} \end{bmatrix} \tag{2}$$

This vertical stacking effectively captures both spectral and temporal variations by treating them as sequential layers, significantly improving the robustness and accuracy of the speech recognition system. By integrating the temporal precision of MFCCs with the spectral richness provided by the partial Mel-Spectrogram, our approach ensures that the most critical acoustic features are preserved and emphasized. This layered representation allows the model to leverage complementary strengths from both feature types: MFCCs contribute to capturing the fine-grained temporal dynamics, which are essential for distinguishing between phonemes that may be temporally close but spectrally different, while the partial Mel-Spectrogram provides a detailed view of the harmonic structure and formants, which are crucial for recognizing vowel quality and tonal variations.

Figure 3 presents the distinct visual patterns captured by the partial Mel-Spectrogram and MFCC features. The partial Mel-Spectrogram highlights detailed spectral properties over a specific Mel bin range, crucial for distinguishing nuanced phonetic elements. This focused frequency analysis is particularly effective in capturing the complex tonal variations inherent in the Kashmiri language. To ensure that our approach was effective and appropriate, we conducted a comparative analysis of different feature extraction techniques, including full-range Mel-Spectrograms and the combination of MFCCs and full Mel-Spectrograms. The results of this analysis indicate that while these alternative methods provided valuable insights, the partial Mel-Spectrogram combined with MFCCs offered superior performance in terms of both accuracy and computational efficiency, particularly for the phonetic challenges presented by the Kashmiri language. The effectiveness of such hybrid approaches in various domains has been well-documented, further validating our methodology [40].

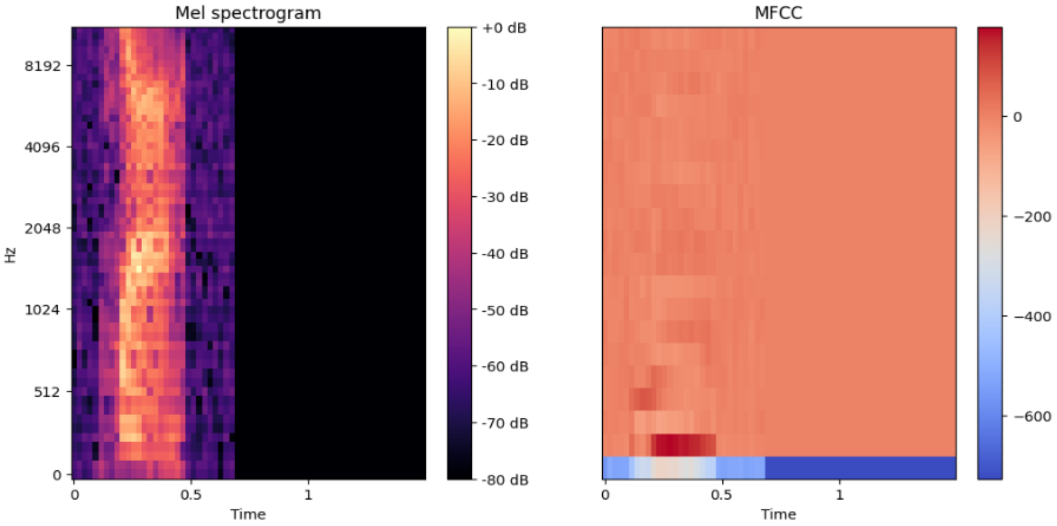


Figure 3. Partial Mel-Spectrogram (left) and MFCC (right) visualizations.

Table 3. Performance comparison of different feature sets and models using various classifiers.

Feature Set	Accuracy	Precision	Recall
MFCC + Partial Mel-Spectrogram (Random Forest)	0.86	0.87	0.87
MFCC + Partial Mel-Spectrogram (Ridge Classifier)	0.80	0.83	0.82
MFCC + Partial Mel-Spectrogram (K-Nearest Neighbors)	0.61	0.61	0.62
MFCC + Partial Mel-Spectrogram (Decision Tree)	0.77	0.77	0.78
MFCC only (Random Forest)	0.79	0.80	0.80
Mel-Spectrogram only (Random Forest)	0.75	0.77	0.77

The complementary nature of MFCC and partial Mel-Spectrogram features is underscored by their respective entropy values, with the partial Mel-Spectrogram capturing more detailed and diverse spectral information [41]. This dual feature extraction approach not only improves the accuracy of the recognition system but also enhances its robustness, enabling the model to generalize effectively across different speakers and conditions. We combined MFCC and partial Mel-Spectrogram features with spectrogram augmentation techniques, such as Gaussian noise and time-stretching, to enhance model robustness. These augmentations introduced variability, allowing the model to generalize better across different acoustic environments and addressing dataset limitations [42**?** ,43]. After augmentation, standard scaling was applied to normalize feature distributions, ensuring comparability across feature types and improving the stability and accuracy of the model during training [42,43]. By leveraging the strengths of both MFCC and partial Mel-Spectrogram features, our approach offers a significant advancement in the accuracy and generalization of the Kashmiri speech recognition model, providing a solid foundation for further research and development in this field.

3.4. Deep Learning Models

3.4.1. Convolutional Neural networks (CNN)

A meticulously organized array of acoustic feature values represented mathematically as $\mathbf{Z} \in \mathbb{R}^{c \times b \times f}$, encompasses an intricate configuration characterized by a specific number of channels denoted as c, an expansive frequency bandwidth articulated as b, and a determined time length expressed as f. Within this framework, the convolutional layer embarks on a transformative journey by convolving the aforementioned $\mathbf{Z}$ with an array of k distinct filters, denoted collectively as $\{\mathbf{W}_i\}_k$, where each individual filter $\mathbf{W}_i$ is formulated as a $3D$ tensor residing in the realm of $\mathbb{R}^{c \times m \times n}$, showcasing a width along the frequency dimension that is precisely m and a length along the frame dimension that is accurately n. The outcome of this convolutional endeavour results in k preactivation feature maps that merge into a singular 3D tensor, denoted as $\mathbf{C} \in \mathbb{R}^{k \times b_C \times f_C}$, with each feature map $\mathbf{C}_i$ being meticulously computed through a specific mathematical operation, where the symbol $*$ serves as the conduit for the convolution operation, and b_i signifies a bias parameter that nuances the output.

$$\mathbf{C}_i = \mathbf{W}_i * \mathbf{Z} + b_i, \quad i = 1, \cdots, k \tag{3}$$

It is essential to highlight three pivotal points that warrant attention: first, the sequence length f_C of the tensor C post-convolution is meticulously ensured to be congruent with the sequence length f of the input matrix Z by implementing a strategic zero-padding technique along the frame axis before the commencement of each convolution operation; second, in our model, the convolution stride is deliberately selected to be 1, ensuring a consistent and fluid convolutional process across all operations; and third, we consciously opt against the application of limited weight sharing, which typically partitions frequency bands into discrete groups of restricted bandwidths, instead favouring a holistic convolutional approach that traverses **Z** along both the frequency axis and the time axis, thereby culminating in a straightforward 2D convolution that is widely recognized and utilized in the domain of computer vision.

$$\tilde{\mathbf{C}}_i = \max(0, \mathbf{C}_i) \tag{4}$$

The pre-activation feature maps, denoted as **C**, undergo a transformative process where they are subjected to a variety of nonlinear activation functions that significantly alter their values and representation. In the subsequent sections, we will introduce a trio of distinct activation functions, elucidating their unique functionalities within the context of a convolutional layer, and it is imperative to highlight that all the operations we will describe below are executed on an element-wise basis for clarity and precision. The Rectifier Linear Unit, commonly referred to as ReLU, is a piecewise linear activation function characterized by its behaviour of yielding a value of zero whenever the input falls below zero while simultaneously returning the input itself when it is non-negative. Formally, if we consider a single feature map represented as $\mathbf{C}_i$, the definition of a ReLU function can be articulated as follows:

$$\left[\hat{\mathbf{C}}_i\right]_{r,\ell} = \max_{j=1}^{p}\left\{\left[\tilde{\mathbf{C}}_i\right]_{r\times s+j,t}\right\} \tag{5}$$

Following the intricate process of applying these element-wise nonlinearities, the resulting features are then directed into a max-pooling layer, which is responsible for extracting the maximum value from a set of p adjacent units, thereby summarizing the most salient information. We focus our pooling operations exclusively along the frequency axis, as this approach has been shown to effectively mitigate spectral variations within the same speaker's output and in comparisons between the outputs of different speakers, as supported by previous research findings.

Notably, pooling along the temporal axis has been demonstrated to yield less significant benefits, indicating a preference for frequency-based pooling methods. In particular, let us denote the i-th feature map prior to pooling as $\mathbf{C}_i$ and its post-pooling counterpart as $\hat{\mathbf{C}}_i$, then the value at position (r, t) in the pooled feature map, denoted as $\left[\hat{\mathbf{C}}_i\right]_{r,t}$, is computed based on the following criteria: where the variable s refers to the step size utilized in the pooling operation, and p represents the size of the pooling window, ensuring that all the values $\left[\tilde{\mathbf{C}}_i\right]_{r\times s+j,t}$ involved in the max calculation share the same temporal index t. Consequently, it is essential to note that the feature maps that emerge from the pooling process maintain identical sequence lengths when compared to their precursory versions before the pooling operation was applied, thereby preserving the overall structure of the data while enhancing its representative quality.

3.4.2. Gated Multi-Layer Perceptrons (gMLPs)

Liu et al. have unveiled a groundbreaking network architecture known as gMLP [44], which is ingeniously built upon Multi-Layer Perceptrons (MLPs) principles while ingeniously integrating sophisticated gating mechanisms that enhance its functionality and adaptability. Their extensive research and experimentation compellingly illustrated that the performance of gMLP can stand shoulder to shoulder with that of the widely recognized Transformers when applied to vital language and vision tasks that hold significant importance in the field of artificial intelligence. Interestingly, the results of their empirical analyses have brought to light the fascinating notion that self-attention, a hallmark feature of Vision Transformers, may not actually be a critical requirement for achieving high performance, as gMLP manages to reach comparable levels of accuracy without it. In a direct comparison with the BERT model, the gMLP architecture not only matches the performance of Transformers in terms of pretraining perplexity but also demonstrates superior capabilities in certain downstream assignments related to NLP [45], showcasing its versatility and effectiveness. Overall, the empirical findings derived from their comprehensive studies strongly suggest that gMLP possesses a remarkable scalability that is comparable to that of Transformers, especially when considering the utilization of expanded datasets and enhanced computational resources that are available in modern research environments.

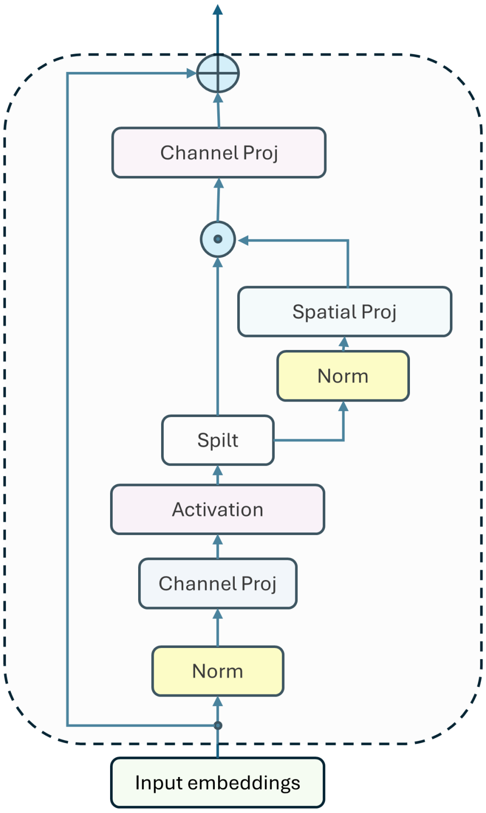


Figure 4. A comprehensive depiction of the gMLP architecture incorporating the Spatial Gating Unit (SGU). The architecture is comprised of a series of L blocks characterized by uniform structure and dimensions. All projection operations within the model are linear in nature, and " $\odot$ " denotes element-wise multiplication (linear gating).

The architecture of gMLP [44], is elegantly constructed from a series of L blocks, all of which share identical dimensions and configurations, resulting in a harmonious and cohesive design. In this context, denote $X \in \mathbb{R}^{n\times d}$ as the representations of tokens, where the sequence length is denoted by n, and the dimensionality is represented by d, encapsulating the essence of each token's information and attributes. Each block is meticulously defined in the manner outlined below. Here, σ represents an activation function (GeLU) known for its effectiveness in neural networks. The matrices U and V serve the purpose of establishing linear projections along the channel dimension, akin to the mechanisms utilized in the FFNs found within the architecture of Transformers.

$$\zeta = \sigma(XU), \quad \tilde{\zeta} = s(\zeta), \quad Y = \tilde{\zeta}V \tag{6}$$

An essential component of the previously described formulation is the layer, denoted as $s(\cdot)$, which plays a crucial role in capturing the intricate spatial interactions among the tokens in the scenario where s functions as an identity mapping, the transformation described above simplifies to that of a conventional Feed-Forward Network, wherein individual tokens undergo processing in isolation, devoid of any inter-token communication that could enhance their contextual understanding.

Consequently, one of the primary objectives is to ingeniously devise an effective s that is adept at capturing the complex and multifaceted spatial interactions between different tokens.

$$f_{W,\beta}(\zeta) = W\zeta + \beta \tag{7}$$

The overall architecture of each block draws inspiration from the concept of inverted bottlenecks, whereby $s(\cdot)$ is defined explicitly as a spatial depthwise convolution, which is instrumental in promoting efficient information flow. It is worth noting that, in contrast to the Transformer architectures, our model elegantly circumvents the necessity for position embeddings, as the requisite positional information is inherently captured within the operations of $s(\cdot)$.

$$s(\zeta) = \zeta \odot f_{W,\beta}(\zeta) \tag{8}$$

In the context of the discussion, where the symbol $\odot$ signifies the operation of element-wise multiplication, we have come to recognize that for the purpose of maintaining stability throughout the training process, it is of utmost importance to set the initial values of the weights, denoted by W, to be extremely close to zero while assigning the biases, represented by β, to the value of one; this specific configuration leads to the outcome that $f_{W,\beta}(\zeta)$ is approximately equal to the vector of ones, denoted as $\mathbf{1}$, which in turn implies that at the very inception of training, the function $s(\zeta)$ closely resembles the input ζ. By implementing this careful initialization strategy, we effectively guarantee that each gMLP block operates in a manner akin to a standard feedforward neural network, or FFN, during the initial phases of training; at this juncture, every individual token is treated in isolation, processed independently without any immediate interaction with others, and only as the learning progresses does the model begin to gradually introduce spatial information that interconnects and influences the tokens over the period of training.

3.4.3. Hybrid CNN-gMLP Model

The hybrid CNN-gMLP model presented in this study builds upon a rich history of advancements in both Convolutional Neural Networks (CNNs) and Gated Multi-Layer Perceptrons (gMLPs), particularly in the domain of speech recognition. Over the past decade, CNNs have established themselves as a powerful tool for local feature extraction, especially in tasks requiring structured data analysis, such as images and spectrograms. Their ability to capture spatial hierarchies in data has made them a natural choice for tasks like speech recognition, where both temporal and spectral features are critical [46,47]. CNNs have been effectively used in various audio processing tasks, demonstrating robust performance in extracting meaningful features from raw audio signals and improving the accuracy of speech recognition systems [48,49]. The introduction of gMLPs has added a new dimension to modelling sequential data. Unlike traditional MLPs, which primarily focus on individual inputs independently, gMLPs are designed to capture global dependencies across input sequences. This capability makes them particularly well-suited for tasks such as language modelling and, more recently, speech recognition [50,51]. The gMLP model has demonstrated competitive results in tasks that require the integration of context across long sequences, outperforming traditional models in scenarios that demand a deep understanding of the global structure of the data [52,53]. The decision to combine CNNs and gMLPs in this study was driven by the need to leverage the strengths of both models while mitigating their limitations. CNNs excel at capturing localized patterns within the data, such as the fine-grained temporal and spectral features of speech signals. Still, they often struggle with capturing long-range dependencies [47]. Conversely, gMLPs are adept at modeling these long-range dependencies but may not capture local features with the same level of granularity as CNNs [50]. By integrating these two approaches, the hybrid model aims to provide a more comprehensive understanding of the speech signal, capturing both local and global features essential for accurate speech recognition. A key innovation of our approach is the use of a dual feature extraction technique, combining Mel-Frequency Cepstral Coefficients (MFCCs) with partial Mel-Spectrograms. MFCCs are widely recognized for their ability to capture the power spectrum of audio signals, offering a detailed representation of the temporal aspects of speech. However, their performance can be limited in noisy environments, and they may miss important spectral details. To address this, we complement MFCCs with partial Mel-Spectrograms, which provide a more comprehensive view of the spectral content by focusing on specific Mel frequency bands relevant to the phonetic characteristics of the Kashmiri language. This dual feature extraction method enhances the robustness of the hybrid CNN-gMLP model by capturing both the spectral richness and the temporal precision required for accurate recognition. The dual feature extraction technique employed in this study serves as a robust input for the CNN, which processes these features to extract detailed local patterns. The output of the CNN is then passed through the gMLP layers, which are responsible for capturing the broader contextual information across the entire input sequence. This combination is particularly effective in addressing the phonetic challenges of Kashmiri, as it allows the model to balance the need for detailed spectral analysis with the requirement for global context understanding [6,11].

In designing this hybrid architecture, we implemented three convolutional blocks with increasing filter sizes to capture progressively more abstract features from the input data. The first block, with 32 filters, captures basic spectral patterns, while the subsequent blocks, each with 64 filters, extract more complex features. Pooling layers follow these convolutional layers to reduce the dimensionality of the feature maps, thereby enhancing computational efficiency and focusing the model's attention on the most salient features. After the CNN layers, the extracted features are flattened and passed through a dense layer, which serves as a bridge to the gMLP layers. The gMLP component, consisting of four layers, employs layer normalization and gating mechanisms to effectively model the dependencies across the input sequence, ensuring that the model can generalize well across different speakers and dialects.

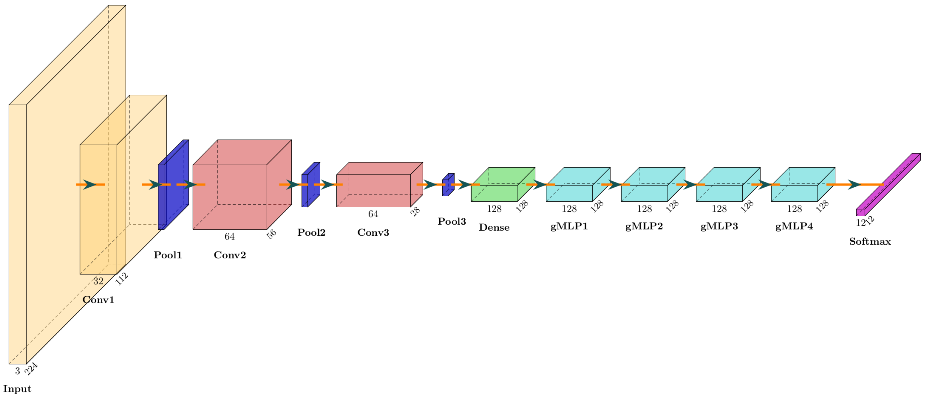


Figure 5. Architecture of the Hybrid CNN-gMLP Model for Spoken Kashmiri Recognition.

4. Results and Discussion

The hybrid CNN-gMLP model was trained using a learning rate of 0.000235 with a step decay schedule, optimizing the model parameters over 50 epochs. The Adam optimizer, known for its adaptive learning rates and efficiency in handling sparse gradients, was used to further enhance the robustness of the model. A dropout rate of 0.378036 was implemented to prevent overfitting, ensuring that the model remained generalizable to unseen data [54,55].

Table 4. Hyper-parameters for training Hybrid CNN-gMLP Model

Parameter	**Value**
Learning rate	0.000235
Learning rate schedule	Step decay
Number of gMLP layers	4
Optimizer	Adam
Dropout rate	0.378036
Epochs	50

Experimental evaluations of this hybrid model demonstrated significant improvements in recognizing spoken Kashmiri words, achieving an overall accuracy of 96%. This performance underscores the effectiveness of the hybrid approach, as the model successfully captured both the detailed local features and the global context necessary for accurate speech recognition. The results indicate that the integration of CNNs and gMLPs, combined with dual feature extraction, offers a promising direction for advancing speech recognition systems, particularly in the context of underrepresented languages like Kashmiri, where the need for both detailed and context-aware processing is paramount [47,50]. The performance of the hybrid CNN-gMLP model in recognizing spoken Kashmiri words was assessed using a combination of key metrics: accuracy, precision, recall, and F1-score. These metrics were chosen for their comprehensive ability to evaluate the effectiveness of the model in a speech recognition classification task. While accuracy provides a general measure of the model's overall performance, it can often be misleading, especially in the presence of class imbalance, as it may not reflect the true discriminative power of the model. Precision and recall offer more granular insights by focusing on the model's performance for individual classes. Precision measures the proportion of true positive predictions among all positive predictions, thus indicating the model's capability to minimize false positives [56]. Recall, on the other hand, evaluates the proportion of true positive predictions among all actual positives, reflecting the model's ability to capture relevant instances and minimize false negatives [56]. The F1-score, defined as the harmonic mean of precision and recall, balances these metrics, which is crucial in contexts like speech recognition where both types of errors can significantly impact user experience [57]. Additionally, the confusion matrix serves as an essential tool for visualizing the performance of the classification model, providing detailed insights into the distribution of correct and incorrect predictions across different classes [56].

The hybrid CNN-gMLP model was trained over 50 epochs, demonstrating rapid convergence with training accuracy nearing 100% after approximately 20 epochs and the validation accuracy stabilizing around 96%. This indicates that the model effectively learned from the training data without significant overfitting, as evidenced by the close alignment between the training and validation curves. The graph in Figure 6 illustrates the training and validation accuracy over the course of the epochs.

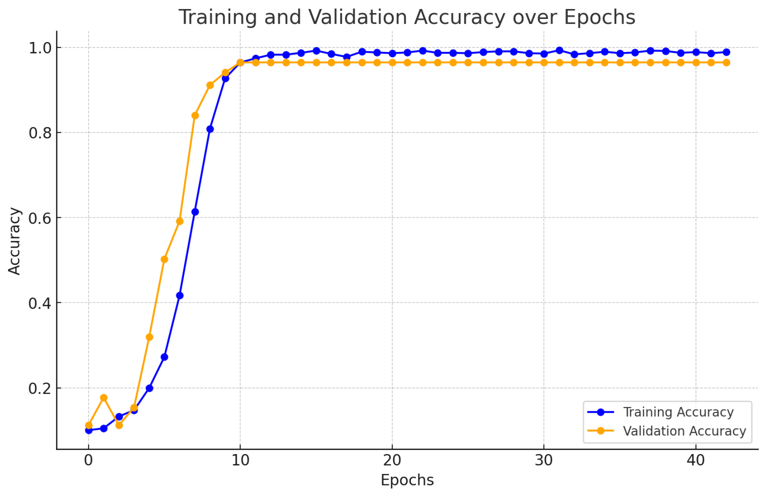
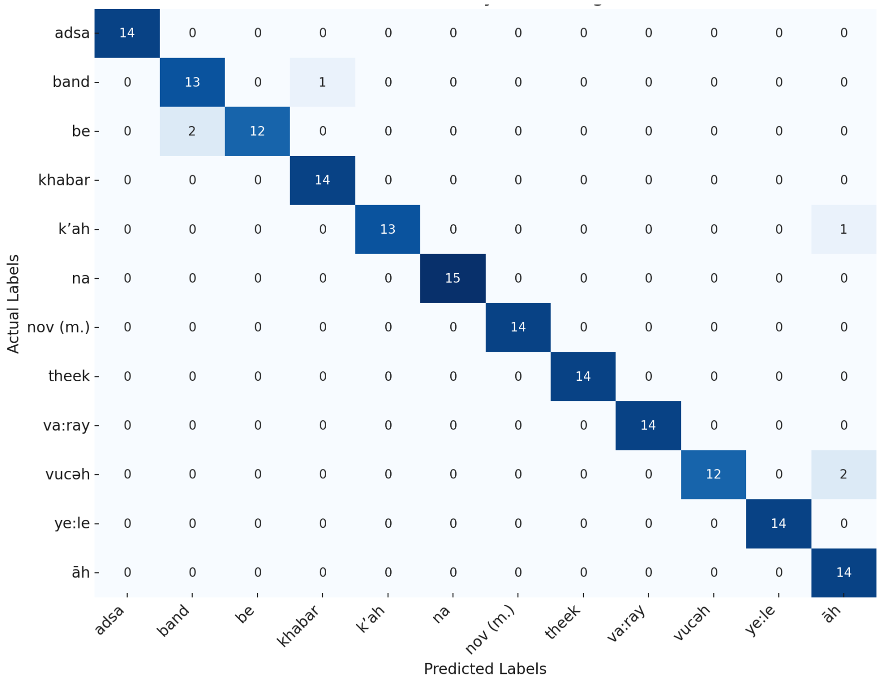


Figure 6. Training and Validation Accuracy over Epochs

The model's effectiveness is further underscored by the confusion matrix shown in Figure 7, which details the model's ability to accurately classify the twelve distinct spoken words in the dataset. The overall accuracy of 96% is complemented by high precision, recall, and F1 scores across most categories. Notably, the model achieved perfect precision and recall for words such as "adsa," "nov (m.)," "khabar," "na," and "ye:le." However, slight misclassifications were observed in words like "k'ah" and "band," suggesting potential areas for further improvement through additional data augmentation or model fine-tuning. The classification report, presented in Table 5 and Table 6 summarizes the model's performance metrics, revealing a macro average precision of 0.97, recall of 0.96, and an F1-score of 0.96. These results affirm the robustness of the hybrid CNN-gMLP architecture in addressing the phonetic diversity of the Kashmiri language, providing a solid foundation for future advancements in spoken Kashmiri recognition.


Actual Labels
Predicted Labels

	adsa	band	be	khabar	k'ah	na	nov (m.)	theek	va:ray	vucəh	ye:le	āh
adsa	14	0	0	0	0	0	0	0	0	0	0	0
band	0	13	0	1	0	0	0	0	0	0	0	0
be	0	2	12	0	0	0	0	0	0	0	0	0
khabar	0	0	0	14	0	0	0	0	0	0	0	0
k'ah	0	0	0	0	13	0	0	0	0	0	0	1
na	0	0	0	0	0	15	0	0	0	0	0	0
nov (m.)	0	0	0	0	0	0	14	0	0	0	0	0
theek	0	0	0	0	0	0	0	14	0	0	0	0
va:ray	0	0	0	0	0	0	0	0	14	0	0	0
vucəh	0	0	0	0	0	0	0	0	0	12	0	2
ye:le	0	0	0	0	0	0	0	0	0	0	14	0
āh	0	0	0	0	0	0	0	0	0	0	0	14

Figure 7. Confusion Matrix for Hybrid CNN-gMLP Model

Table 5. Summary of Classification Metrics for Hybrid CNN-gMLP Model on each word

Metric	Precision	Recall	F1-Score
adsa	1.00	1.00	1.00
āh	0.82	1.00	0.90
nov (m.)	1.00	1.00	1.00
khabar	0.93	1.00	0.97
be	1.00	0.86	0.92
vuch	1.00	0.86	0.92
na	1.00	1.00	1.00
band	0.87	0.93	0.90
va:ray	1.00	1.00	1.00
k'ah	1.00	0.93	0.96
ye:le	1.00	1.00	1.00
theek	1.00	1.00	1.00

Table 6. Classification Report based on Evaluation Metrics

Metric	Precision	Recall	F1-Score
Macro Avg	0.97	0.96	0.96
Weighted Avg	0.97	0.96	0.96
Accuracy		**0.96**	

These findings underscore the effectiveness of the hybrid CNN-gMLP approach in tackling the complexities inherent in the phonetic diversity of spoken Kashmiri, establishing a strong foundation for continued research and development in this domain.

4.1. Comparative Analysis and Validation

The model is validated using 5 fold cross validation and the comparative study is being performed with other state of art approaches. Table ??, shows the comparison of our approach with previous stte of art approaches that are being used for automatic speech recognition. In this Hybrid CNN-gMLP model is designed with a combination of CNN and gMLP layers, achieves the highest accuracy of 96% on the Kashmiri Speech Dataset. This model uniquely leverages dual feature extraction methods (MFCC and Mel-Spectrogram) to capture both local and global features that make it robust for handling the phonetic diversity of the Kashmiri language. While as CNN + LSTM Hybrid model achieves a slightly lower accuracy of 91.3% on the TIMIT dataset that emphasizes its strength in handling sequential data through LSTM to capture phonetic variations effectively. The Deep Speech (RNN) and Wav2Vec (CNN) models do not report accuracy but focus on word error rate (WER) as a key performance indicator in this Wav2Vec achieve a superior WER of 8.5% due to its unsupervised pre-training, compared to 10.55% for Deep Speech. Models like Listen, Attend, and Spell and the Hybrid CNN-RNN Emotion Recognition focus on different aspects of speech recognition, such as large vocabulary and emotional speech detection. The latter attains 89.2% accuracy on emotion datasets which demonstrates its effectiveness in contextual speech analysis. Each model is characterized by its distinctive architectural choices and optimizations for specific speech recognition challenges, from handling continuous speech to phonetic and emotional diversity.

Table 7. Comparison of Hybrid CNN-gMLP Model with Other Speech Recognition Models

Model	Architecture	Dataset Used	Accuracy	Precision	Recall	F1-Score	Unique Features
Hybrid CNN-gMLP	CNN + gMLP (4 layers)	Kashmiri Speech Dataset	96%	0.97	0.96	0.96	Dual feature extraction (MFCC + Mel-Spectrogram), local and global feature capture, robust for phonetic diversity of Kashmiri language.
CNN + LSTM Hybrid	CNN + LSTM	TIMIT	91.3%	0.90	0.89	0.89	Local feature extraction via CNN combined with sequential modeling by LSTM, effective in handling phonetic variations in speech data [46].
Deep Speech (RNN)	RNN-based end-to-end	Various speech datasets	N/A	N/A	N/A	WER: 10.55%	End-to-end speech recognition, large-scale training with parallelization for scalability, robust handling of continuous speech [9].
Wav2Vec (CNN)	CNN for pre-training	LibriSpeech	95.5%	N/A	N/A	WER: 8.5%	Unsupervised pre-training with contrastive loss to improve robustness on noisy and limited data [58].
Listen, Attend, and Spell	Attention-based neural net	Large Vocabulary Dataset	N/A	N/A	N/A	WER: 13.1%	Attention mechanisms for large vocabulary speech recognition, efficient handling of conversational speech [59].
Hybrid CNN-RNN Emotion Recognition	CNN + RNN (LSTM/GRU)	Speech Emotion Dataset	89.2%	0.87	0.88	0.87	Captures local features and temporal dependencies, well-suited for emotional and contextual speech analysis [48].

While performing 5-fold cross-validation experiment our model's performance was consistently robust across all folds. The macro average precision, recall, and F1-score remained high, with precision averaging 0.97, recall at 0.96, and an F1-score of 0.96 across the five folds. Also, the weighted averages for these metrics mirrored the macro averages which reflect the model's balanced performance across various classes, with precision at 0.97, recall at 0.96, and an F1-score of 0.96. The model also achieved a stable and impressive average accuracy of 96% which indicates that its predictive capabilities are consistent and generalize well to unseen data. Each fold in the cross-validation process demonstrated negligible variance in these metrics that underscore the model's robustness and reliability across different subsets of the dataset.

4.2. Limitations

While the proposed hybrid CNN-gMLP model demonstrates impressive performance in recognizing spoken Kashmiri, achieving a 96% accuracy, but there are few limitations of this study. These are discussed as follows

- **Dataset Size and Diversity:** The dataset used in this study, although representative of the phonetic and dialectal diversity of the Kashmiri language, is limited to 12 commonly used words. This narrow vocabulary may not reflect the full range of complexities in real-world speech patterns and interactions. Moreover, the dataset contains a total of 840 samples, which may limit the generalization of the model to more complex speech recognition tasks involving extended vocabulary and spontaneous speech.
- **Dialectal Variations:** Although efforts were made to capture dialectal diversity from different regions of Kashmir (North, Central, and South Kashmir), there may still be unaccounted variations in pronunciation, intonation, and accent that could affect the model's performance. The current model may not generalize well to speakers from less-represented dialects or regions.
- **Environmental Conditions:** The dataset was collected in controlled environments to ensure high audio quality. However, in practical applications, speech recognition systems are often used in noisy or variable acoustic environments. The model's robustness in such conditions is yet to be fully evaluated, and additional noise-robust techniques, such as advanced denoising methods, may be necessary to enhance its performance in real-world settings. Feature Extraction Methodology: The dual feature extraction approach, combining MFCCs and partial Mel-Spectrograms, showed significant improvement in capturing the phonetic nuances of Kashmiri speech. However, alternative feature extraction techniques, such as wavelet transforms or more advanced self-supervised learning models like HuBERT, could potentially enhance performance further, especially in noisy or resource-constrained environments.
- **Computational Complexity:** While the hybrid CNN-gMLP model balances local and global feature extraction effectively, the computational cost associated with this approach could limit its deployment on resource-constrained devices, such as mobile phones or embedded systems. Optimizing the model for efficiency, without compromising accuracy, remains a challenge.
- **Bias and Fairness:** Although participant selection aimed to capture a diverse range of age, gender, and regional backgrounds, there is potential for demographic bias in the dataset. The performance of the model across different demographic groups (e.g., gender or age) has not been explicitly evaluated and may reveal disparities that need to be addressed.
- **Application to Continuous Speech:** The model was trained and tested on isolated word recognition, which differs significantly from continuous speech recognition tasks. Future research should explore extending the model's capabilities to handle continuous speech, where word boundaries are less defined and contextual dependencies play a more critical role.

5. Conclusion

This research marks a significant advancement in the field of Automatic Speech Recognition (ASR) for the Kashmiri language, a language historically underrepresented in technological developments. By developing a hybrid Convolutional Neural Network (CNN) and Gated Multi-Layer Perceptron (gMLP) model, this study successfully integrates dual feature extraction methodologies—Mel-Frequency Cepstral Coefficients (MFCCs) and partial Mel-Spectrograms—to capture the critical spectral and temporal characteristics necessary for accurate Kashmiri speech recognition. The model's ability to achieve an impressive 96% accuracy in classifying twelve distinct spoken words underscores the robustness and effectiveness of our approach. Our work aligns with global efforts in language preservation and technological inclusivity. This alignment underscores the broader relevance and potential impact of our research, particularly in preserving linguistic diversity through advanced technological solutions. Future research could explore advanced noise reduction techniques, such as deep learning-based denoising methods, which have proven effective in other low-resource language ASR tasks. Integrating alternative feature extraction methods, such as wavelet transforms or phonetically-informed HuBERT models, could also enhance the robustness of feature representations, particularly in noisy or variable conditions.

Funding: The authors did not receive any external funding for the development of this study.

Data Availability Statement: Data will be available on Request

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Team, G.R. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. *arXiv preprint arXiv:2303.01037* **2023**.
2. Singh, A.; Kadyan, V.; Kumar, M.; Bassan, N. A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications* **2020**, *79*, 3673–3704.
3. Koul, O.N. *Kashmiri: A Descriptive Grammar*; Routledge, 2009.
4. Wali, K.; Koul, O.N. *Kashmiri: A Cognitive-Descriptive Grammar*; Routledge, 1997.
5. Besacier, L.; Barnard, E.; Karpov, A.; Schultz, T. Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech Communication* **2014**, *56*, 85–100. doi:10.1016/j.specom.2013.07.001.
6. Liu, H.; Dai, Z.; So, D.R.; Le, Q.V. Pay Attention to MLPs. *Advances in Neural Information Processing Systems* **2021**, *34*, 9204–9215.
7. O'Neill, J.; Carson-Berndsen, J. Challenges in ASR Systems for Underrepresented Dialects. Proceedings of the Annual Conference on Computational Linguistics, 2023, pp. 233–241.
8. Adams, O.; Neubig, G.; Cohn, T.; Bird, S. Learning a Lexicon and Translation Model from Phoneme Lattices. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 2379–2389. doi:10.18653/v1/P17-1218.
9. Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenger, R. Deep Speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* **2014**. doi:10.48550/arXiv.1412.5567.
10. Aggarwal, A.; Srivastava, A.; Agarwal, A.; Chahal, N.; Singh, D.; Alnuaim, A.; Alhadlaq, A.; Lee, H.N. Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. *Sensors* **2022**, *22*, 2378. doi:10.3390/s22062378.
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR), 2020. doi:10.48550/arXiv.2010.11929.
12. Zhao, J.; Shi, G.X.; Wang, G.B.; Zhang, W. Automatic Speech Recognition for Low-Resource Languages: The Thuee Systems for the IARPA Openasr20 Evaluation. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021, pp. 113–119. doi:10.1109/ASRU51503.2021.9688260.
13. Kipyatkova, I.; Kagirov, I. Analytical Review of Methods for Solving Data Scarcity Issues Regarding Elaboration of Automatic Speech Recognition Systems for Low-Resource Languages. *Information Technologies and Computing Systems* **2022**, *21*, 2–15. doi:10.15622/ia.21.4.2.

14. Zhao, J.; Zhang, W. Improving Automatic Speech Recognition Performance for Low-Resource Languages with Self-Supervised Models. *IEEE Journal of Selected Topics in Signal Processing* **2022**, *16*, 1105–1115. doi:10.1109/JSTSP.2022.3184480.
15. Min, Z.; Ge, Q.; Li, Z.; Weinan, E. MAC: A Unified Framework Boosting Low Resource Automatic Speech Recognition. *arXiv preprint arXiv:2302.03498* **2023**.
16. Bartelds, M.; San, N.; McDonnell, B.; Jurafsky, D.; Wieling, M.B. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. *arXiv preprint arXiv:2305.10951* **2023**.
17. Yeo, J.H.; Kim, M.; Watanabe, S.; Ro, Y. Visual Speech Recognition for Languages with Limited Labeled Data Using Automatic Labels from Whisper. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2023, pp. 6650–6654. doi:10.1109/ICASSP48485.2024.10446720.
18. Pratama, R.S.A.; Amrullah, A. Analysis of Whisper Automatic Speech Recognition Performance on Low Resource Language. *Pilar Nusa Mandiri Journal of Computer and Information Science* **2024**, *20*, 1–8. doi:10.33480/pilar.v20i1.4633.
19. Bekarystankyzy, A.; Mamyrbayev, O.; Anarbekova, T. Integrated End-to-End Automatic Speech Recognition for Agglutinative Languages. *Proceedings of the ACM on Human-Computer Interaction* **2024**, *8*, 1–19. doi:10.1145/3663568.
20. Sullivan, T.; Harding, E. Domain Shift in Speech Dialect Identification: Challenges and Solutions. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2023, pp. 510–517. doi:10.1109/ICASSP.2023.10096345.
21. Bhatt, R.M. *The Kashmiri Language*; Vol. 46, *Studies in Natural Language and Linguistic Theory*, Springer, 1999. doi:10.1007/978-94-015-9279-6_2.
22. Koka, N.A. Social Distribution of Linguistic Variants in Kashmiri Speech. *Academy Publication* **2016**.
23. Zampieri, M.; Nakov, P. *Phonetic Variation in Dialects*; Cambridge University Press, 2021. https://doi.org/10.1017/9781108565080.004.
24. Rather, S.N.; Singh, N. Phonetic and Phonological Variations in Kashmiri Language Dialects. *International Journal of Linguistics* **2017**, *9*, 12–20. doi:10.5296/ijl.v9i5.11781.
25. Kachru, B.B. Kashmiri. In *The Indo-Aryan Languages*; Cardona, G.; Jain, D., Eds.; Routledge, 2003; pp. 895–930.
26. Bhat, M.A.; Hassan, S. Prosodic Features of Kashmiri Dialect of Maraaz: A Comparative Study. *IRE Journals* **2020**.
27. Open Speech Corpus Contributors. Open Speech Corpus Tool. https://github.com/open-speech-corpus/open-speech-corpus-tool, 2023. Accessed: 2024-08-12.
28. Rabiner, L.; Juang, B.H. *Fundamentals of Speech Recognition*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, USA, 1993.
29. Journals, I. Zero Crossing Rate and Energy of the Speech Signal of Devanagari Script. *IOSR Journals* **2015**.
30. Shen, J.L.; Hung, J.W.; Lee, L.S. Robust entropy-based endpoint detection for speech recognition in noisy environments. Proc. 5th International Conference on Spoken Language Processing (ICSLP 1998), 1998, p. paper 0232. doi:10.21437/ICSLP.1998-527.
31. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed.; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 2009.
32. Wang, D.; Brown, G.J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*; IEEE Press/Wiley, 2016.
33. Ganie, H.A.; Hasnain, S.K. Computational Linguistics and the Kashmiri Language: Issues and Challenges. *International Journal of Computer Applications* **2015**, *128*, 1–6. doi:10.5120/ijca2015906615.
34. Winursito, A.; Hidayat, R.; Bejo, A. Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition. 2018 International Conference on Information and Communications Technology (ICOIACT), 2018. doi:10.1109/ICOIACT.2018.8350748.
35. Hidayat, R. Frequency Domain Analysis of MFCC Feature Extraction in Children's Speech Recognition System. *Infotel Journal* **2022**, *14*, 28–36. doi:10.20895/infotel.v14i1.740.
36. Setianingrum, A.; Hulliyah, K.; Amrilla, M.F. Speech Recognition of Sundanese Dialect Using Convolutional Neural Network Method with Mel-Spectrogram Feature Extraction. 2023 8th International

Conference on Information Technology, Information Systems and Mechatronics (CITSM), 2023. https://doi.org/10.1109/CITSM60085.2023.10455447.

37. Permana, S.D.H.; Rahman, T.K.A. Improved Feature Extraction for Sound Recognition Using Combined Constant-Q Transform (CQT) and Mel Spectrogram for CNN Input. 2023 International Conference on Mechatronics, Robotics, and Automation (ICMERALDA), 2023. doi:10.1109/ICMERALDA60125.2023.10458162.
38. Li, Q.; Yang, Y.; Lan, T.; Zhu, H.; Wei, Q.; Qiao, F.; Liu, X.; Yang, H. MSP-MFCC: Energy-Efficient MFCC Feature Extraction Method With Mixed-Signal Processing Architecture for Wearable Speech Recognition Applications. *IEEE Access* **2020**. doi:10.1109/ACCESS.2020.2979799.
39. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Walker, J.; Zhu, Z. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. Proceedings of The 33rd International Conference on Machine Learning, 2016, pp. 173–182. doi:10.5555/3045390.3045410.
40. Joshi, D.; Pareek, J.; Ambatkar, P. Comparative Study of MFCC and Mel Spectrogram for Raga Classification Using CNN. Indian Journal of Science and Technology, 2023. doi:10.17485/ijst/v16i11.1809.
41. Hafiz, N.F.; Mashohor, S.; Shazril, M.H.S.E.M.A.; Rasid, M.F.A.; Ali, A. Comparison of Mel Frequency Cepstral Coefficient (MFCC) and Mel Spectrogram Techniques to Classify Industrial Machine Sound. *Proceedings of the 16th International Conference on Software, Knowledge, Information Management, and Applications (SKIMA)* **2023**. doi:10.1109/SKIMA59232.2023.10387339.
42. Abayomi-Alli, O.O.; Damaševičius, R.; Qazi, A.; Adedoyin-Olowe, M.; Misra, S. Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review. *Electronics* **2022**, *11*, 3795. doi:10.3390/electronics11223795.
43. Qazi, A.; Damaševičius, R.; Abayomi-Alli, O.O. Combining Transformer, Convolutional Neural Network, and Long Short-Term Memory Architectures: A Novel Ensemble Learning Technique That Leverages Multi-Acoustic Features for Speech Emotion Recognition in Distance Education Classrooms. *Applied Sciences* **2022**, *12*, 1–21. doi:10.3390/app11221001.
44. Liu, H.; Dai, Z.; So, D.; Le, Q.V. Pay attention to mlps. *Advances in neural information processing systems* **2021**, *34*, 9204–9215.
45. Yu, P.; Artetxe, M.; Ott, M.; Shleifer, S.; Gong, H.; Stoyanov, V.; Li, X. Efficient language modeling with sparse all-mlp. *arXiv preprint arXiv:2203.06850* **2022**.
46. Passricha, V.; Aggarwal, R. A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition. *Journal of Intelligent Systems* **2018**, *27*, 555–563. doi:10.1515/jisys-2018-0372.
47. Wubet, Y.A.; Lian, K.Y. Voice Conversion Based Augmentation and a Hybrid CNN-LSTM Model for Improving Speaker-Independent Keyword Recognition on Limited Datasets. *IEEE Access* **2022**, *10*, 114222–114234. doi:10.1109/ACCESS.2022.3200479.
48. Atila, O.; Şengür, A. Attention Guided 3D CNN-LSTM Model for Accurate Speech-Based Emotion Recognition. *Applied Acoustics* **2021**, *175*, 108260. doi:10.1016/j.apacoust.2021.108260.
49. John Lorenzo Bautista, Y. Lee, H.S. Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation. *Electronics* **2023**, *11*, 3935. doi:10.3390/electronics11233935.
50. Hu, K.; Sainath, T.N.; Pang, R.; Prabhavalkar, R. Deliberation Model Based Two-Pass End-To-End Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2020**, *28*, 123–134. doi:10.1109/TASLP.2020.9053606.
51. Li, G.; Sun, Z.; Hu, W.; Cheng, G.; Qu, Y. Position-aware relational transformer for knowledge graph embedding. *IEEE Transactions on Neural Networks and Learning Systems* **2023**. doi:10.1109/TNNLS.2023.3201305.
52. Yang, C.H.H.; Gu, Y.; Liu, Y.C.; Ghosh, S.; Bulyko, I.; Stolcke, A. Generative Speech Recognition Error Correction With Large Language Models and Task-Activating Prompting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2023**, *31*, 234–245. doi:10.1109/TASLP.2023.10389673.
53. Bai, Y.; Yi, J.; Tao, J.; Tian, Z.; Wen, Z.; Zhang, S. Fast End-to-End Speech Recognition Via Non-Autoregressive Models and Cross-Modal Knowledge Transferring From BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2021**, *29*, 456–467. doi:10.1109/TASLP.2021.3082299.
54. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
55. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **2014**, *15*, 1929–1958.
56. Powers, D.M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* **2020**.

57. Goutte, C.; Gaussier, E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science, vol 3408. Springer, Berlin, Heidelberg, 2005, pp. 345–359. doi:10.1007/978-3-540-31865-1_25.
58. Baevski, A.; Schneider, S.; Auli, M. Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv preprint arXiv:2006.11477* **2020**. doi:10.48550/arXiv.2006.11477.
59. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 4960–4964. doi:10.1109/ICASSP.2016.7472621.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.