The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

DIALECT RECOGNITION SYSTEM FOR BAGRI RAJASTHANI LANGUAGE USING OPTIMIZED FEATURED SWARM CONVOLUTIONAL NEURAL NETWORK (OFSCNN) MODEL

POONAM KUKANA [1]*, POOJA SHARMA [2], PUNEET SAPRA [3] and NEERU BHARDWAJ [4]

[1, 2, 3] Department of Computer Science and Engineering, University School of Engineering & Technology, Rayat-Bahra University, Mohali, Punjab, India.
[4] Department of Computer Science and Engineering, Chandigarh University, Mohali, Punjab, India.
Email: [1] poonamkukana@gmail.com (*Corresponding Author), [2] pooja.rani@rayatbahrauniversity.edu.in, [3] puneetsapra91@gmail.com, [4] neeru.e13122@cumail.in

Abstract

The dialects of a language hold a significant place in speech processing (SP) applications. The objective of dialect identification is to categorize speech sample data into a specific dialect of a speaker's spoken language. A dialect recognition system must effectively distinguish between different dialects of a standard language, as they tend to possess many similarities. The dialect of a language is not a distinct characteristic, as it can be impacted by the utterer's birthplace. Researchers in the domain of automatic speech recognition (ASR) face difficulties in identifying the speech patterns unique to each dialect or language. The proposed work recognizes the dialects of the Bagri Rajasthani language from undefined expressions of speech. Rajasthani Language is one of the eldest and most famous languages in the Indo-Aryan languages. It comprises the different dialects and for recognizing the dialects, it used dissimilar phases of acoustic and spectral characteristics of the speech signal (SS). The spectral and acoustic features of SSs are measured to design the system. As there is no specific speech dataset for Bagri dialects, the database is built, to verify the Bagri dialects of the Rajasthani language. To improve the accuracy rate, and error rate in recognizing the Bagri dialects, the acoustic and spectral characteristics of speech expressions are joined. To verify several Bagri dialects of the Rajasthani language, different simulations for classification and investigation are carried out i.e., OFSCNN model, GA-NN, etc. The outcomes are important and the accuracy of 96.95 % for the OFSCNN model, 80.63 % for GA-NN, and 93.45% for the Multiclass SVM method is an achievement.

Keywords: Speech Dialect Recognition, Dialects, OFSCNN (Optimized Featured Swarm Convolutional Neural Network), MFCC (Mel Frequency Co-efficient Cepstral), PSO (Particle Swarm Optimization), CNN (Convolutional Neural Network.

1. INTRODUCTION

Speech-processing techniques have become widely utilized in various applications in today's world. Dialect identification refers to the task of determining the region or accent a language belongs to [1]. In this research article, recognition of dialects of the Rajasthani language is discovered. Rajasthani language is one of the eldest languages and goes to the western Indo-Arya language family [2]. The Rajasthani language and its dialects have a distinct geographical distribution. There are two main groups of Rajasthani: the Western Rajasthani group and the Eastern Rajasthani group. The Western Rajasthani group encompasses Bagri and its sub-dialects such as Mewari, etc. The Eastern Rajasthani group, also known as Dhundari, includes

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

Jaipuri, Malvi, etc. In speech-associated submissions, spectral (SF) and prosodic features (PFs) play a critical role in classifying the spoken linguistic and dialect. Spectral features are frequently used in dialect identification as well as other SP techniques [3]. However, it can be challenging to distinguish between similar words from different regions using only spectral features. Prosodic features come into play in these scenarios and provide important cues for identifying and discriminating correspondences in speech. From a human perception perspective (HPP), PFs also offer strong cues for easily identifying dialects. There are different stages of features, as shown in Fig. 1, that shows a significant role in nearly all speech-associated use [4].

The lower most stage characteristics or features are spectral features such as the MFCC method and its differences. These features are simple to extract and implement, but they can be affected by variations in speech utterances**.** In contrast, the topmost features such as lexical features are based on linguistic and spoken words. Prosodic features fall into the category of second-stage features that are extremely useful in providing cues for identifying emotions or dialects in speech-based applications. According to [5], dialect refers to a variation of a standard linguistic that is specific to a certain geographical area. Each dialect has its unique characteristics, which are shaped by its place of origin. However, these variations among dialects can negatively impact the presentation of ASR systems. The term "Rajasthani" (written in Devanagari script as राजस्थानी) denotes a collection of indo-aryan languages and Bagri dialects that are mainly spoken in the Indian state of Rajasthan and nearby regions of Haryana, etc. Additionally, there are utterers in the Punjab and Sindh areas of Pakistan. These Rajasthani diversities are carefully related to, and incompletely understandable with, the Gujarati and Sindhi languages. In Rajasthan, 65.04% of the population speaks राजस्थानी. The level of mutual intelligibility between राजस्थानी and Gujarati varies from 60 percent to 85 percent depending on the particular dialect being considered.

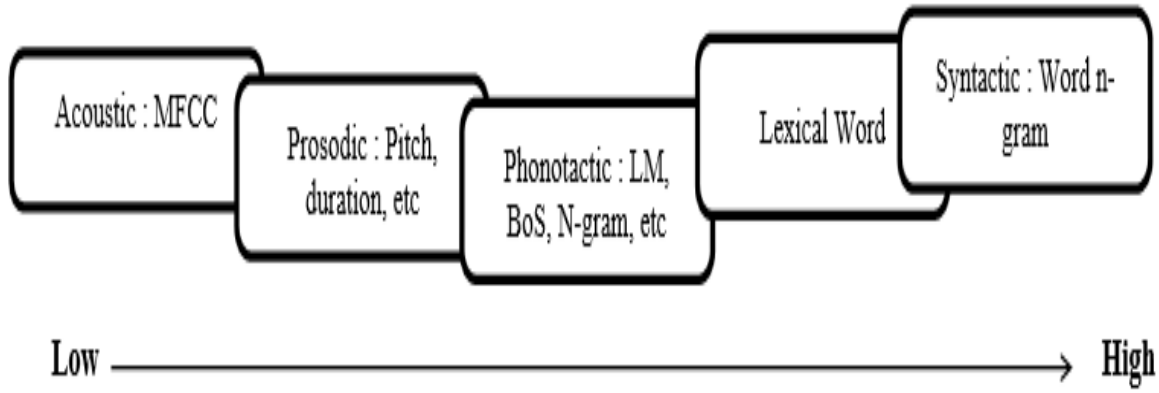

Figure 1: Different Feature Levels

The importance of a database in identifying dialects of language [6] cannot be overstated. It greatly reduces the effort required for language identification. This study involves the creation of a standard database that holds speech samples of dialects spoken in the Rajasthani

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

(राजस्थानी) Language. In Bagri dialects, the acoustic features such as energy and pitch are very low. Several benefits of the DR system that is it is used for forensic operations, news broadcasts, etc.

In this research paper, different acoustic feature techniques [7] are developed to verify the Bagri dialects of the Rajasthani Language from speech samples and calculate the features with mathematical models such as optimized featured swarm convolutional neural network (OFSCNN) model for classification purpose. The proposed work has applied the MFCC algorithm to extract the feature sets. MFCCs represent a set of features that are commonly used in speech dialect recognition. These features capture important information about the spectral shape of the speech signal, which can be used to distinguish between different dialects or accents. After that, implemented the OFSCNN algorithm selects the reliable features and classifies the dialects. The OFSCNN algorithm is a combination of two methods such as PSO and CNN. In the context of speech dialect recognition, the PSO method can be used to optimize the weights of a classifier such as a neural network and CNN to improve the accuracy of dialect recognition. The PSO method can be an effective approach for speech dialect recognition because it can quickly search a large search space of weights to find the optimal solution. By optimizing the weights of a classifier using the OFSCNN method, it is possible to improve the accuracy of dialect recognition and achieve better performance.

Section 2 discusses the survey of the related works, Section 3 defines the database design, proposed methodology, parameters, and proposed methods. Section 4 describes the proposed result and comparative analysis with other methods. Section 5 concludes the research work with further enhancement.

2. RELATED WORK

Data analysis, particularly speech recognition, and classification, has been widely researched using traditional machine learning (ML), deep learning (DL), and other techniques [8][9][10][11][12]. Different classifiers, including convolutional neural networks (CNNs) and long-short-term memory (LSTM), have been applied for dialect recognition and identification [13][14]. ML techniques, such as support vector machines (SVMs), multi-layer perceptrons (MLPs), and k-nearest neighbor (KNN), have also been used in works like [15] to classify spoken dialects. Despite this, these techniques have shown efficient outcomes in some signal-processing applications but have produced unsatisfactory results in dialect classification and identification. One of the most researched topics in speech analysis (SA) is dialect recognition (DR). The classification and identification of dialects have been commonly approached using ML methods. Dialect recognition works have utilized CNN, a type of deep neural network (DNN), as suggested in the literature. [16] Proposed a model based on three unique one-dimensional (1D) CNN architectures for Kurdish dialect recognition. The database included three dialects, as the Kurdish language consists of Northern, Hawrami, and Central Kurdish dialects. The 1D CNN method achieved an average accuracy rate, and the 302-based dialect recognition system was also tested on the widely used intonation variation in the English (IViE) database and achieved better accuracy and reduced complexity. [17] Proposed the mel-

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

weighted single-frequency (SF) filtering method for dialect recognition. The SF filtering spectrum improved the definition of speech features, such as burst time and glottal closure, compared to other spectrums. On the UT-podcast database, the proposed representations improved the unweighted average recall by 9.47 percent and 4.69 percent, respectively, on the development and test databases. [18] Implemented a two-layer feed-forward NN technique to recognize Hindi dialects. The method retrieves prosodic and spectral features from speech to distinguish between the dialects. According to the results, the method operates most effectively when all of the features are integrated to form an input feature vector during network training. With these input feature vectors, the dialect recognition (DR) method achieved a recognition score of 79 percent. Currently, the CNN model has increased fame, surpassing cross-breed deep neural networks (DNNs) and hidden Markov Models (HMM) based on an acoustic method. The CNN model can generally handle speech signals, making it a suitable high-quality for an ASR system. [19] Studied the effects of several activations and soft computing methods in the Hindi ASR system. The results showed that the ELU activation function (AF) with Rmsprop optimization (RO) methods provided a better word error rate (WER) of 14.56 percent. [20] Introduced the multi-layer FFNN method to calculate the efficacy of acoustic features for dialect recognition**.** The simulation results showed that merging prosodic and spectral feature sets yielded an 82 percent recognition rate for the dialect**.**

Humans depend on machines from daybreak to end, and machines require input signals to function. Several systems that use speech input in a native tongue have been created. One of them is Punjabi. There are several speeches and dialect recognition systems out there. Still, they all share some basic challenges, the main one being the inability to recognize words in different dialects of Punjabi. Malwa, Doaba, and Majha are the three main dialects of Punjabi spoken in Eastern Punjab. This research aims to explain a novel method that [21] have created that works on Punjabi dialects and to compare its accuracy rate to other methods. [22] Described the design of an isolated word ASR for the Punjabi linguistic. The HTK toolkit depends on the HMM model. It defined the main role of the HTK toolkit, which was utilized in several phases of system implementation. The proposed model has achieved an overall system performance of 95.6 percent and 94.0 percent. [23] Offered methods for creating an effective language model from freely accessible online sources to enhance the ASR accuracy rate. Compared to Google's speech-to-text rate of 55 percent, the proposed model WER of 24.7 percent was the best. They have also shown how our trained ASR model can produce speech corpus semi-supervised, demonstrating a profitable method of developing huge vocabulary corpora for minimum resource languages. Any dialect’s dialects are important in speech processing (SP) applications. [24] Discussed the work that verified the dialects of the Telugu language from undefined utterances of speech. Telugu is one of the oldest and most widely spoken languages in the Dravidian family. It consists of the Telangana, Coastal Andhra, and Rayalaseema dialects. They used several levels of spectral and prosodic aspects of the SS to distinguish the dialects. The spectral features, such as MFCC, etc., and prosodic features, such as pitch and volume of SSs, were measured to design the system. A database was generated to classify the dialects of Telugu because there wasn’t a standard database for linguistics. Two methods for classification and simulation were carried out to verify the distinct dialects of the

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

Telugu language: the HMM and Gaussian mixture matrix (GMM) methods. The outcomes were important, with an accuracy rate of 88.4 percent for the GMM and 86.9 percent for the HMM methods for the MFCC, and the prosodic feature was an accomplishment.

3. PROPOSED BAGRI DIALECTS IN RAJASTHANI SPEECHES OPTIMIZED DEEP LEARNING SYSTEM

All categories of speech dialect recognition systems have different phases: (i) Training, and (ii) Testing. Through the training phase, the system is designed and established. In the test phase, the system is identified. The following sections 3.1 to 3.4 discuss the Rajasthani database generation (3.1), performance metrics (3.2), proposed methodology and methods (3.3), and techniques used in the proposed system (3.4) in the study.

3.1 Rajasthani (राजस्थानी) Speech Database Generation

Rajasthani is the primary language of the state of Rajasthan and is vocal by over 71 million people in the region as well as in some other parts of India and around the world. According to the 2011 census, there are approximately 22 distinct linguistic variants of Rajasthani spoken. Rajasthani is an Indo-Aryan dialect and accent primarily spoken in the states of Rajasthan, Madhya Pradesh, Gujarat, and Haryana in India, and in the regions of Sindh and Punjab in Pakistan. It is carefully connected to and understandable with Sindhi and Gujarati, their parallel dialects. The narrative coherence between Gujarati and Rajasthani ranges from 60 percent to 85 percent, based on the geographical area of their respective dialects. It is also the cultural lateralization of Bagri, which is being developed as the primary language of Rajasthan. In this study, a text-free Rajasthani dialect voice dataset was collected from diverse locations across Rajasthan, particularly in rural and interior areas. According to research, linguistic cues intrinsic to speaking activities differ from those in reading dialogue. Different speaking rates, filled intervals, loudness, tone, apprehensions, echoes, and half-said phrases, among other things, are all visible prosodic indicators in expressive language. These characteristics are considered to transmit dialectal variations more effectively. In this study, a dataset of ten users and fifty speeches in each user's Rajasthani Bagri language [25] was gathered from various locations. Some of the speeches were also collected in real-time using a voice recorder. The details of the dataset used are given in Tables 1 and 2. The input samples were then pre-processed using data speech pre-processing methods to eliminate noise, and the resulting database was saved giving to the dialects. Table 2 represents the Rajasthani Bagri speech dataset with different audio waves, dialects, person (Male, female). Fig. 2 depicts the general organization of SR or identification.

Table 1: Rajasthani Speech Sample Dataset Explanation (Poonam 2023)

Language	Speech Sample	Dataset Size	No. of users	No. of Speech Samples
Rajasthani	Rajasthani	19.2 MB	10	50 for each person.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

Table 2: Rajasthani Speech Dataset with Dialects Detail (Poonam 2023)

Rajasthani	Audio wav number	Dialects	person
रोटी जीमौ।	1	रोटी जीमौ।	F
पाणी पीओ।	2	पाणी पीओ।	F
इणनै सुणौ।	3	इणनै सुणौ।	F
ना जाओ।	4	ना जाओ।	F
अठीनै देकौ।	5	अठीनै देकौ।	F
रोटी वणाओ।	6	रोटी वणाओ।	F
बारै जावौ।	7	बारै जावौ।	F
अठै सुवौ।	8	अठै सुवौ।	F
म्हनैं दिखावौ।	9	म्हनैं दिखावौ।	F
मा आओ	10	मा आओ	F
अठी आवौ।	11	अठी आवौ।	F
म्हनै सुणौ।	12	म्हनै सुणौ।	F
बोळा रौ।	13	बोळा रौ।	F
धक्कौ देवौ।	14	धक्कौ देवौ।	F
पाणी उबाळौ।	15	पाणी उबाळौ।	F
फुटरौ काम!	16	फुटरौ काम!	F
चूप रौ!	17	चूप रौ!	F
गोळी खावौ।	18	गोळी खावौ।	F
बोत बड़ीया!	19	बोत बड़ीया!	F
म्हनैं सुणौ।	20	म्हनैं सुणौ।	F
इनै केवौ।	21	इनै केवौ।	F
वौ म्हारौ है।	22	वौ म्हारौ है।	F
वौ कठै है?	23	वौ कठै है?	F
वौ अठै है।	24	वौ अठै है।	F
इने अठै लावौ।	25	इने अठै लावौ।	F
इने वठै मेळौ।	26	इने वठै मेळौ।	F
आपरौ मुं खोलौ।	27	आपरौ मुं खोलौ।	F
इणने आछौ राखौ।	28	इणने आछौ राखौ।	F
होलै वात करौ।	29	होलै वात करौ।	F
म्हनै ठा कोनी।	30	म्हनै ठा कोनी।	F
खानौ बणगौ हैं।	31	खानौ बणगौ हैं।	F
आपरा हाथ धोवौ।	32	आपरा हाथ धोवौ।	F
थै थक गिया।	33	थै थक गिया।	F
वौ कुण है?	34	वौ कुण है?	F
रोटी कौनी जीमौ।	35	रोटी कौनी जीमौ।	F
वौ वठै कौनी।	36	वौ वठै कौनी।	F
टाबर नै पकड़ौ।	37	टाबर नै पकड़ौ।	F
थै कठै जावौ?	38	थै कठै जावौ?	F
थै कदै आया?	39	थै कदै आया?	F
कुण रोवै है?	40	कुण रोवै है?	F
म्हनै ऩमझ कोनीं आवै।	41	म्हनै ऩमझ कोनीं आवै।	F
वै म्हारां रिश्तेदार है।	42	वै म्हारां रिश्तेदार है।	F
थोरै कितरा टाबर है?	43	थोरै कितरा टाबर है?	F
आज घणी गरमी है।	44	आज घणी गरमी है।	F

ISSN 1533-9211

DOI: 10.5281/zenodo.10005831

आज सड़क सूखोड़ी है।	45	आज सड़क सूखोड़ी है।	F
वौ छोरो रोवै है।	46	वौ छोरो रोवै है।	F
आज ठड़ सी है।	47	आज ठड़ सी है।	F
म्हारै कनै थोड़ाक है।	48	म्हारै कनै थोड़ाक है।	F

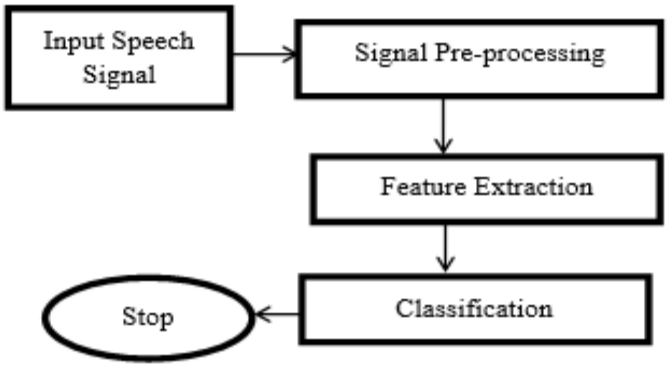

Figure 2: General Structure of the Speech Recognition System

3.2 Performance Metrics

Several performance parameters, such as mean square error (MSE), false acceptance rate (FAR), false rejection rate (FRR), and accuracy rate, have been used to evaluate the research system.

The *accuracy* rate, defined as the percentage of accurate classifications, is one of the most widely used performance parameters. This can be seen in Table 3. [26]

Table 3: Accuracy Classification

		Predict Class	
Actual Class	+	True_positive (T_p)	False_negative (F_n)
	-	False_positive (F_p)	True_negative (T_n)

$$Acc = \frac{T_p + T\_n}{n} \quad \ldots\ldots\ldots\ldots \text{(i)}$$

Where, n signifies the total no. of positive (p), and negative examples (n), and T_p and T_n signify the no. of true positives, and negatives, respectively.

MSE is a performance metric utilized to calculate the precision of a model concerning a test dataset. It is calculated as the average of the squared estimate error across all instances in the test dataset. Typically, the estimate error is determined by associating the true positive value with the predicted value for each instance.

$$MSE = \frac{\sum_{j=1}^{m}\left(Y_j - \lambda(Y_j)\right)^2}{m} \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{(ii)}$$

Here, eq (ii) Yj is the true target value (TV) of the test dataset, $\lambda(Y_j)$ is the predicted target value for test dataset instance Yj, and m is the no. of test illustrations.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

FRR is the amount of the performance of a biometric system (BS), such as fingerprint, facial, iris, or signal recognition. FRR represents the percentage of invalid attempts that are incorrectly rejected as valid by the system. It is a measure of the system's security, as a higher FRR indicates that the system is most likely to allow unauthorized access. The FRR formula is given by:

$$FRR = \frac{no.of\ false\ rejections}{no.of\ false\ rejections + no.of\ true\ acceptances} * 100 \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{(iii)}$$

FAR is a measure of the performance of a biometric system (BS) like a fingerprint, facial, iris, signal recognition, etc. FAR represents the percentage of invalid attempts that are incorrectly accepted as valid by the system. It is a measure of the system's security, as a low FAR indicates that the system is less likely to allow unauthorized access. The FAR formula is given by:

$$FAR = \frac{no.of\ false\ acceptances}{no.of\ false\ acceptances + no.of\ true\ rejections} * 100 \ldots\ldots\ldots\ldots\ldots\ldots \text{(iv)}$$

3.3 Proposed Methodology

Rajasthani Dialect recognition using MFCC, PSO, and CNN involves extracting MFCC features from speech signals, optimizing these features using PSO, and then using a CNN for classification. The implementation is done using MATLAB. The following steps are followed or the present work and are shown in fig 3.

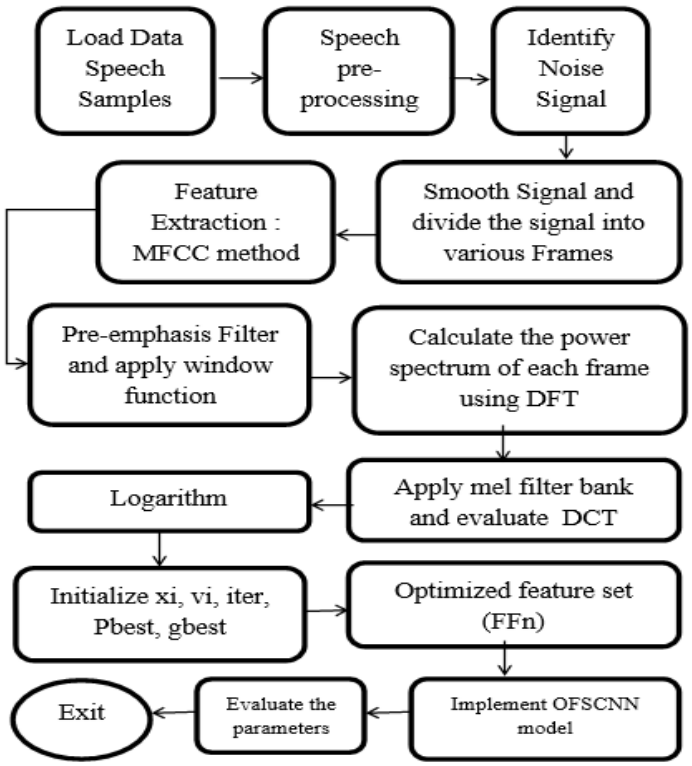

Figure 3: Flow Chart of OFSCNN Model

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

3.3.1 Preprocessing: The speech signals are preprocessed to remove noise and silence, and to segment the signal into frames. Each frame is typically 20-30 ms long with a 50% overlap.

3.3.2 Feature Extraction: MFCCs are extracted from each frame using a filter bank of Mel-scaled triangular filters. The resulting MFCC feature vectors capture the spectral characteristics of the SS.

3.3.3 Feature Optimization: PSO is used to optimize the MFCC feature vectors by adjusting the filter bank parameters to maximize the separability between different speech classes. The PSO algorithm iteratively updates the filter bank parameters until a set of optimized features is obtained.

3.3.4 Classification: The optimized MFCC features are then used as input to a CNN for classification. The CNN consists of multiple convolutional layers, followed by pooling layers, and finally a FCL that outputs the classification result. The CNN is trained on a labeled dataset of speech signals with their corresponding labels.

3.3.5 Evaluation: The metrics of the system are calculated utilizing standard metrics like MSE, FAR, FRR, and accuracy rate, on a separate test dataset.

3.4 Proposed Methods

3.4.1 Feature Extraction using MFCC method

MFCC is a widely used and efficient spectral feature (SF) in speech processing (SP) uses [27]. The set of features, defined as the MFCCs of a SS, describes the entire shape of the spectral cover. In the MFCC, the frequency bands are spaced equally on the Mel scale, which more closely approaches the response of the human auditory system (HAS) than the frequency bands used in a linear demonstration in the normal spectrum shown in eq (v). This is the main difference between the MFCC and normal cepstrum. The MFCC extracts the features that are widely used in speech recognition (SR) and speaking. The steps of the MFCC method are defined in Fig 4, and the technique is evaluated by the formulas below: $mel(f) = 1$

$$125\ In\ (1 + \frac{f}{700}) \ldots\ldots\ldots (v)$$

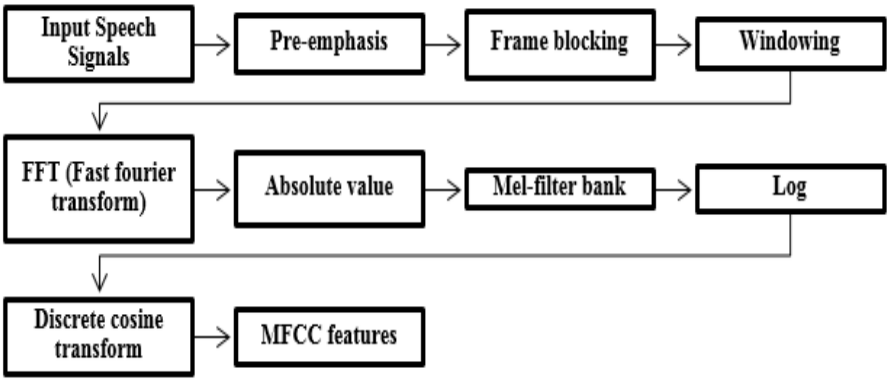

Figure 4: Steps to Extract the MFCC [28] from the Speech Signal

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

The process for extracting features from a speech signal (SS) is described as follows:

1. Segment the SS into short frames of equal size.
2. Apply the FFT to transform the signal into the frequency domain, multiplying each frame by a hamming window to preserve signal continuity.
3. Apply the Mel filter bank (MFB) to generate a power spectrum.
4. Take the LOG of the filter bank energies.
5. Perform discrete cosine transformation (DCT) on the log filter bank energies.
6. Retain only the DCT coefficients between 2-13 and discard the rest.

3.4.2 Feature Selection Using the PSO method

An evolutionary computation method [29] known as bird flocking was introduced in [30]. This method mimics the knowledge-sharing interactions between birds as they search for food. The system consists of a swarm of randomly distributed particles. To find the optimum solution, particles move around the search space, changing their locations based on the practices of both themselves and nearby particles. The recent location and velocity of individual particle j are defined by the vectors yj = (yj1, yj2,…,Yjd) and Wj = (wj1, wj2,…, wjd), respectively, where d represents the dimensionality of the search space.

Each particle updates its position and velocity throughout its movement depending on its own experience and that of its surrounding particles. The best location that a particle has experienced is referred to as its personal best (Pbest), while the population's best location is mentioned to as the global best (Gbest). The particle updates its location and velocity using equations (iv) and (vii) respectively, to find the best solution.

$$y\_(j,d)^(t+1) = y\_(j,d)^t + w\_(j,d)^(t+1) \quad \text{(vi)}$$

$$w\_(j,d)^(t+1) = w * w\_(j,d)^t + c1 * r1 * (Qj,d - y\_(j,d)^t) + c2 * r2 * (Gj,d - y\_(j,d)^t) \quad \text{(vii)}$$

Here, T represents the Tth epoch, d ϵ d signifies the dth dimension in the search space, W represents the inertia weight in the search space and determines the convergence rate (CR) of PSO, C1, and C2 are acceleration constants, and R1 and R2 are random numbers between 0 and 1. Qj,d and Gj,dis the jthdepiction of Pbest and Gbest in the dth dimension, respectively."

3.4.3 CNN Recognition model

It is also used in speech dialect recognition (SDR) tasks, where it acts as a feature extraction (FE) for SS. In SDR, the input to the CNN [31] is typically a spectrogram or an MFCC representation of the SS. In this application, CNN is used to learn and remove the important features from the spectrogram or MFCC representation that are reliable for SR. Fig 4 defines the CNN layers and how CNN works [32].

ISSN 1533-9211

DOI: 10.5281/zenodo.10005831

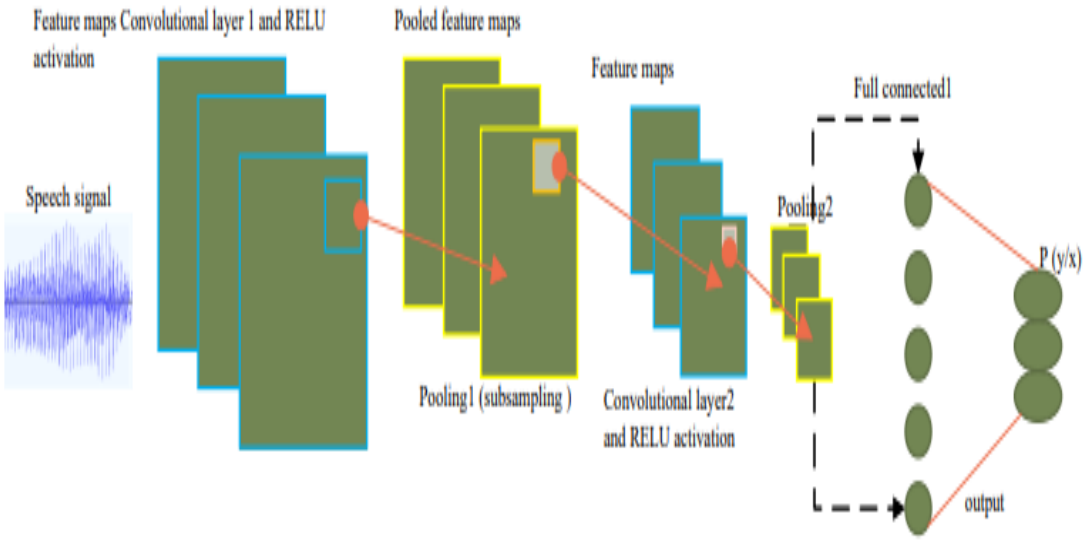

Figure 5: The Architecture of CNN layers

The convolutional layers (CL) of the network apply filters to the input and process feature maps that consider reliable features. The pooling layers (PL) then reduce the spatial dimensions (SD) of the feature maps and produce a compact representation of the features.

Finally, the fully connected layers (FCL) of the network use the compact representation to produce the final output, such as a transcription or a prediction of the spoken words. The network is qualified using great amounts of labeled speech samples, where the correct transcriptions or word predictions serve as the ground truth.

In general, the use of CNNs in speech recognition has been exposed to be efficient in enhancing the accuracy rate of speech recognition systems, especially when combined with other methods such as recurrent neural networks (RNNs) and attention mechanisms.

Here is a general outline of the steps involved in using a CNN [33] for speech dialect recognition:

1. Data Gathering: The chief phase is to gather a huge database of speech signals from different dialects. The speech signals should be labeled with the corresponding dialect.
2. Pre-processing: The speech signals are then pre-processed to remove any noise or artifacts. This may involve techniques such as filtering, windowing, and normalization.
3. Feature extraction: The pre-processed speech signals are then transformed into a suitable representation for input to the CNN. This may involve converting the speech signals into a spectrogram or a Mel-Frequency Cepstral Coefficient (MFCC) representation.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

Network Architecture Design: The next step is to design the CNN architecture. This involves deciding on the number of CLs, PLs, FCLs, AFs, and normalization layers to use.

4. Training: The CNN is then trained using the extracted features and the corresponding dialect labels as inputs. The network is trained using an optimization procedure, like stochastic gradient descent (SGD), to diminish the difference between the expected and ground truth outputs.
5. Testing: After the training process is complete, the CNN is established on a separate test dataset to evaluate its performance. The test dataset should also be pre-processed and transformed into the same representation used for training.
6. Deployment: Finally, the trained CNN can be deployed as a speech dialect recognition system. This may involve integrating the CNN with other components such as a front-end speech processing module or a back-end database.

This is a general outline of the steps involved in using a CNN for speech dialect recognition. The specific details of each step, such as the exact architecture of the CNN or the specific pre-processing techniques used, may vary depending on the precise necessities of the task and the obtainable information.

3.4.4 OFSCNN Method

The OFS-CNN is a DL model that combines the strengths of CNNs and PSO to enhance the accuracy of speech DR. The OFSCNN algorithm involves the following steps:

1. Data Preprocessing: The speech signals are preprocessed to extract features such as MFCCs.
2. Feature with Optimization: PSO is used to optimize the set of features used by CNN to improve the accuracy of the classification. The PSO algorithm searches for the optimal set of features that maximize the accuracy of the classification.
3. Training: The optimized features are used to train a CNN. It consists of several layers of convolutional, and PLs that learn to extract features from speech signals. The weights of the CNN are trained using a back propagation algorithm to minimize the classification error.
4. Testing: The trained CNN is used to classify speech signals in the testing dataset. The CNN takes the preprocessed speech signals as input and outputs the predicted dialect.

The OFSCNN algorithm has several advantages over traditional approaches to speech dialect recognition. The PSO algorithm optimizes the set of features used by the CNN, which improves the accuracy of the classification. The CNN construction is considered to learn complex features from the speech signals, which allows the model to capture subtle differences between dialects. Additionally, the OFSCNN algorithm is highly automated and can learn from large datasets, which reduces the need for manual tuning and improves the scalability of the approach.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

In summary, the OFSCNN algorithm is an effective method for speech dialect recognition (SDR) that combines the strengths of CNNs and PSO to improve the accuracy of the classification.

Pseudo Code: OFSCNN algorithm for Bagri Dialect Speech Recognition System

Input: Speech Data Samples DS1

Output: DF as dialects, Accuracy, FAR, FRR, MSE.

1. Load Input Data DS1, OFSCNN_training_model;
2. Preprocess the Input DS1;
3. Identify Ns (noise signal);

SNR_val =SNR(Ns,Signal_noise);

4. SSignal (Smooth Signal) = smooth (Ns);
5. Divide the signal into various Frames F
6. For each Frame F(i)
7. Module 1 Feature Extraction Using MFCC
   - Apply a pre-emphasis filter to the speech DS1 to boost high-frequency components and reduce the effect of noise.
   - Frame the Speech DS1 into overlapping frames of length N and with a hop-size of M.

Pre-emphasis Filter :

*Y(n) = x(n) – alpha*x(n-1);*

- Apply window Function to each frame to reduce spectral leakage.

*W(n)=0.54-0.46*cos(2*pi*n/);*

- Evaluate the power spectrum of each frame using the DFT (discrete fourier transformation).
- Apply mel filter bank to the power spectrum to obtain the mel spectrogram.

$$Hm(k)=\begin{cases} 0 & k < f(m-1) \\ k - \frac{f(m-1)}{f(m)-f(m-1)), } & f(m-1) \le k < f(m) \\ \frac{f(m+1)-k}{f(m+1)} - f(m); & f(m) \le k < f(m+1) \\ 0 & k \ge f(m+1) \end{cases}$$

Where f(m) is the center frequency of the mthmel filter.

- Evaluate the DCT (discrete cosine transformation) of the log mel spectrogram.

DCT:

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

x(k) = sum_n = 0^n-1, x(n)*cos(pi*k*(n+0.5/N);

for k= 0,1,…..N-1;

- Logarithm:

Log_Spec = log(mel_spec);

8. Module 2 Optimized Swarm Convolution Neural Network (OSCNN)

Initialize xi, vi, iterations, Pbest, Gbest \\ vi (velocity), xi(position), iterations (feature extracted data).

Generate random particle(p)

For each particle (i)

Evaluate FFN (fi)

Update pbest, gbest

End for

While iteration

For each particle I

Update vi, xi

If (xi>limit then xi= limit)

Evaluate FFn fi

//Collect optimized feature vector

Update Pbest, Gbest as TFV

End if

End for

End while.

End for.

9. Load optimized feature set TFV and simulate
10. Generate classified labels
11. Simulate labels from F as DF //Identified Dialect signal frame
12. Show dialects speech frame DF
13. Compute the performance of all the classified frames vs input //, Accuracy, FAR, FRR, MSE

MSE = $\frac{1}{N}\sum_{j=1}^{n}(yj - y'j)^2$ …… (i)

Accuracy = Tp+Tn/Tp+Tn+Fp+Fn *100 % ……..(ii)

FAR= Fp/Fp+tn

FRR=Fn/tp+fn

14. Stop.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

4. EXPERIMENT SETUP AND RESULT ANALYSIS

Speech data for Bagri dialects of the Rajasthani Language. Rajasthani language is recorded using a mobile application and signal data pre-processing is applied to eliminate the attack and filter the unwanted noises. Gathered 10 users and 50 speeches per person speech samples for training and testing purposes. OFSCNN (optimized feature swarm CNN) model is utilized as a speech dialect recognition model in this research work and it is implemented by the GUIDE (graphical user interface environment). To allocate the fitness to the candidate DL architecture. MFCC method trains the CNN with a minimum batch size of 128. The softmax cross entropy loss is used as the LF (loss function). Simulations are achieved to extract feature sets and design training and testing steps using the MATLAB tool. It works under the WINDOW operating system (OS). DL toolbox is used for the design of the research method with MATLAB. When huge architecture created by the PSO method does not execute due to memory limitation, then 0 fitness is assigned to the candidate solution (CS).

Above discussed these metrics are calculated and reported to calculate the performance of the dialect recognition system of the Rajasthani language using MFCC, PSO, and CNN. The dialect recognition system is a feature selection-based system. Several feature vectors from the sample set are identified and saved for comparing and matching at the last stage. Input Speech feature sets are matched with train sets. Several features are used for speech dialect recognition. OFSCNN method is used in this system for better performance. This method sub-divided the feature sets into groups and recognizes the match based on sub-divided feature sets.

Fig 6 defines the uploaded speech sample. After that, it identifies the noisy speech sample defined in fig 7. Fig 8 shows the filtered speech sample. It means to smooth the speech sample using a mel filter. Fig 9 shows the features extracted speech sample. This research work used the MFCC technique to extract speech properties.

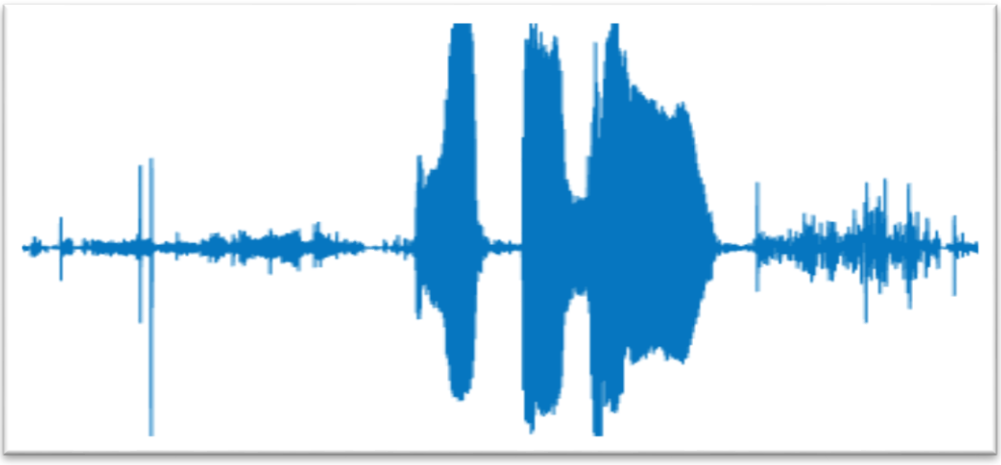

Figure 6: Original Sample

ISSN 1533-9211

DOI: 10.5281/zenodo.10005831

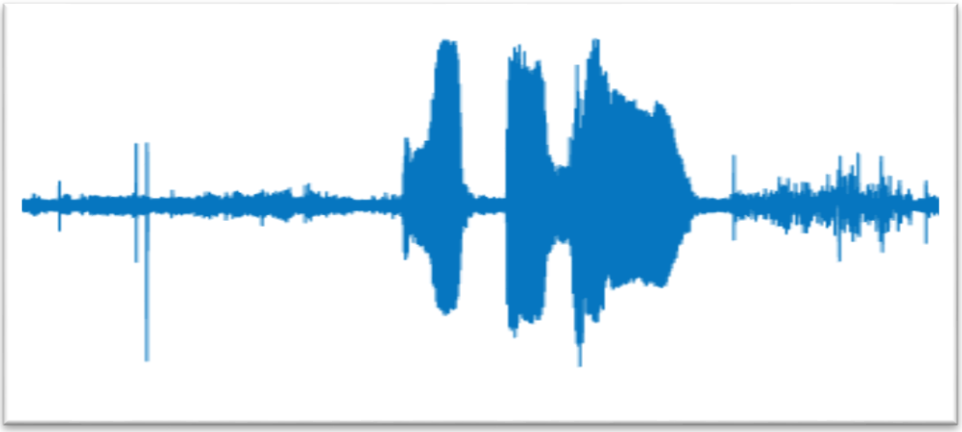

Figure 7: Noisy Speech Sample

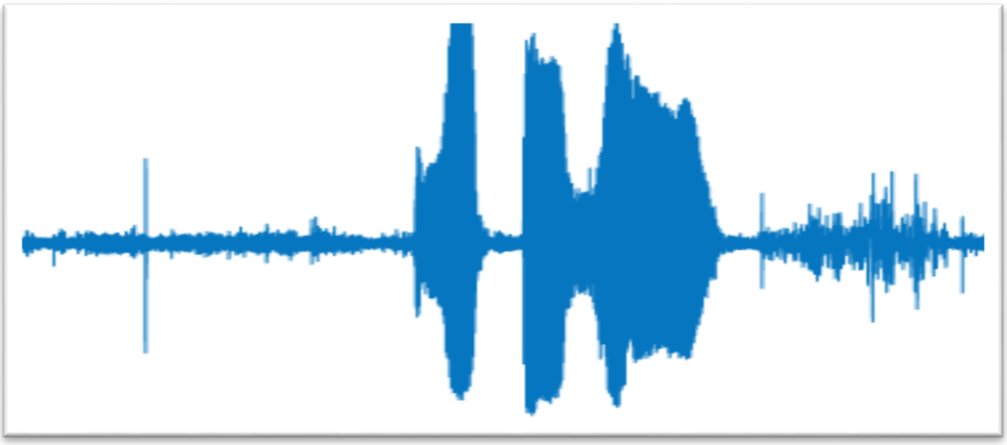

Figure 8: Filtered Speech Sample

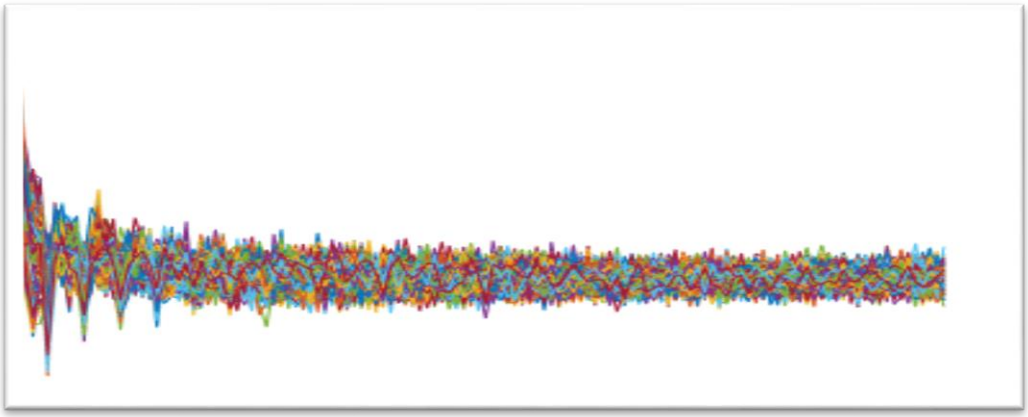

Figure 9: Feature Extracted Sample

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

From table 4, it is considered that. 96.95 percent of accuracy, 0.015 of MSE, 0.0012 FAR, and 0.97 FRR of Bagri dialect speech recognition are attained with MFCC features shown in table 4. From these outcomes, it can be considered that features alone are comparatively successful in the classification of Bagri dialects. Where the MFCC method has extracted reliable features. From all simulations considered it is measured that the utilization of consequent features has established an enhanced dialect presentation over rare feature sets. The proposed OFSCNN model has achieved the highest accuracy rate of 96.96 percent as compared with existing models such as the Multi SVM [34] accuracy rate of 93.45%, and the GA-NN [35] model accuracy rate of 80.63 % shown in table 5 and fig 10. The OFSCNN model has attained a higher accuracy rate and reduced the error probabilities. The proposed model has calculated error rates such as MSE, FAR, and FRR. The proposed model has achieved less error rate than MSE, FAR, and FRR.

Table 4: Parameters Attained by the Proposed Model (OFSCNN)

Parameters / Models	Accuracy	MSE	FAR	FRR
OFSCNN	96.95	0.015	0.0012	0.97

In this article, three different dialect recognition models are implemented with the OFSCNN, Multi-SVM, and GA-NN classifier models in the Bagri Dialect speech recognition. The highest of 96.95% accuracy is attained with the different feature vectors and minimum error probabilities. The reason for the improved accuracy rate may be due to the use of PSO optimizer and lower performance using the Multi-SVM [36] and GA-NN model shown in fig 10.

Table 5: Comparative Analysis with Different Proposed and Existing Models

Parameters | Models	OFSCNN	Multi-SVM	GA-NN
Accuracy (%)	96.95	93.45	80.63

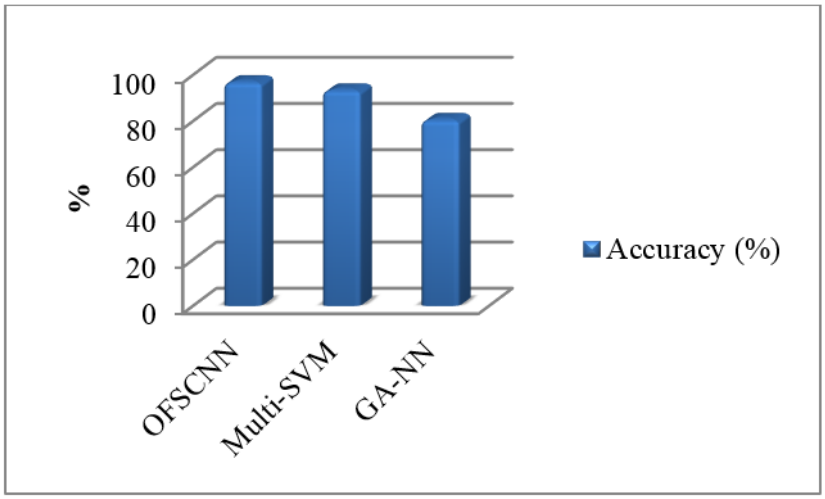

Figure 10: Comparison of the Speech recognition: Accuracy (%)

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

5. CONCLUSION AND FUTURE SCOPE

In the conclusion, the use of the OFSCNN model for speech dialect recognition has proven to be highly effective. The model uses the advanced DL method to accurately classify the Bagri dialects, resulting in highly accurate recognition results. The Rajasthani dialect recognition system developed using MFCC, PSO, and CNN has shown promising results in accurately classifying the various dialects of the Rajasthani language. The system involves the acquisition of speech signals, preprocessing, feature extraction using MFCC, optimization using PSO, and classification using a CNN. The system was evaluated using an accuracy rate of 96.95%, MSE of 0.015, FAR of 0.0012, and FRR of 0.97 metrics. The results demonstrate that the system achieved high accuracy in dialect recognition with low values for MSE, FAR, and FRR. The OFSCNN model has achieved high accuracy as compared to other methods such as GA-NN, and the Multi-SVM method.

The proposed system can be further extended for other Indian languages to accurately identify various dialects of the language. The system has potential applications in speech recognition, natural language processing, and automated translation systems. With further research and development, it may become an indispensable tool for businesses, governments, and individuals seeking to better understand and communicate with speakers of different languages and dialects.

DECLARATION

Conflicts of interest/ Competing interests: No conflicts of interest that could affect the objectivity or impartiality of this research.

Data Availability Statement: Data is available at https://github.com/poonamkukana/Speech-Dataset-of-Rajasthani-language.git

Funding sources: This research did not receive any specific grant from any funding agencies.

Author Contributions: Er. Poonam Kukana carried out the experiment and wrote the manuscript with support of Dr. Pooja Sharma, Dr. Puneet Sapra and Dr. Neeru Bhardwaj. All authors discussed the results and contributed to the final manuscript.

References

1) Ibrahim, N. J., Idris, M. Y. I., Yakub, M., Yusoff, Z. M., Rahman, N. N. A., &Dien, M. I. (2019). Robust feature extraction based on spectral and prosodic features for classical Arabic accents recognition. *Malaysian Journal of Computer Science*, 46-72.

2) Stroński, K., Tokaj, J., &Verbeke, S. (2019, September). A diachronic account of converbal constructions in Old Rajasthani.In *Historical Linguistics 2015. Selected papers from the 22nd International Conference on Historical Linguistics, Naples, 27–31 July 2015* (pp. 424-441). Amsterdam: John Benjamins.

3) Tong, R., Ma, B., Zhu, D., Li, H., &Chng, E. S. (2006, May). Integrating acoustic, prosodic and phonotactic features for spoken language identification. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Vol. 1, pp. I-I).IEEE.

4) Chittaragi, N. B., &Koolagudi, S. G. (2017, August). Acoustic features based word level dialect classification using SVM and ensemble methods. In *2017 Tenth International Conference on Contemporary Computing (IC3)* (pp. 1-6). IEEE.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

5) Hirschberg, J. B., Biadsy, F., & Collins, M. (2010).Dialect recognition using a phone-GMM-supervector-based SVM kernel.

6) Prasad, A., Srinivas, Y., &Brahmaiah, P. (2010). Gender based emotion recognition system for telugu rural dialects using hidden markov models. *arXiv preprint arXiv:1006.4548*

7) Sinha, S. (2015).Analysis and recognition of dialects of Hindi speech. *Int J Sci Res ComputSciEng*, *3*(5).

8) Alsayadi H A, Al-Hagree S, Alqasemi FA, Abdelhamid AA (2022) Dialectal Arabic Speech Recognition using CNN-LSTM Based on End-to-End Deep Learning. In: Proceedings of the International *Conference on Emerging Smart Technologies and Applications. eSmarTA 2022*, pp. 1-8.

9) Singh N, Kumar M, Singh B, Singh J (2022) DeepSpacy-NER: an efficient deep learning model for named entity recognition for Punjabi language. *Evolving Systems*, pp 1-11.

10) Vashisht V, Pandey AK, Yadav SP (2021) Speech recognition using machine learning. *IEIE International Transactions on Smart Processing & Computing*, *10*(3): 233-239.

11) Xu, Y. (2022). English speech recognition and evaluation of pronunciation quality using deep learning. *Mobile Information Systems*, *2022*, 1-12.

12) Al-Jumaili, Z., Bassiouny, T., Alanezi, A., Khan, W., Al-Jumeily, D., &Hussain, A. J. (2022, August).Classification of Spoken English Accents Using Deep Learning and Speech Analysis. In *Intelligent Computing Methodologies: 18th International Conference, ICIC 2022, Xi'an, China, August 7–11, 2022, Proceedings, Part III* (pp. 277-287). Cham: Springer International Publishing.

13) Das, A., Kumar, K., & Wu, J. (2021, June).Multi-Dialect Speech Recognition in English Using Attention on Ensemble of Experts.In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6244-6248).IEEE.

14) Chittaragi, N. B., &Koolagudi, S. G. (2020).Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms. *Language Resources and Evaluation*, *54*(2), 553-585.

15) Aljuhani, R. H., Alshutayri, A., &Alahdal, S. (2021). Arabic Speech Emotion Recognition From Saudi Dialect Corpus. *IEEE Access*, *9*, 127081-127085.

16) Ghafoor, K. J., Rawf, K. M. H., Abdulrahman, A. O., &Taher, S. H. (2021). Kurdish dialect recognition using 1D CNN. *Aro-the Scientific Journal of Koya University*, *9*(2), 10-14.

17) Kethireddy, R., Kadiri, S. R., Alku, P., &Gangashetty, S. V. (2020). Mel-weighted single frequency filtering spectrogram for dialect identification. *IEEE Access*, *8*, 174871-174879.

18) Sinha, S., Jain, A., &Agrawal, S. S. (2014). Speech processing for Hindi dialect recognition.In *Advances in Signal Processing and Intelligent Recognition Systems* (pp. 161-169).Springer International Publishing.

19) Choudhary, T., Bansal, A., &Goyal, V. (2022). Investigation of CNN-based acoustic modeling for continuous Hindi speech recognition. In *IoT and Analytics for Sensor Networks: Proceedings of ICWSNUCA 2021* (pp. 425-431). Springer Singapore.

20) Sinha, S., Jain, A., &Agrawal, S. S. (2015). Acoustic-phonetic feature based dialect identification in Hindi Speech. *International journal on smart sensing and intelligent systems*, *8*(1), 235-254.

21) Singh, R., & Sharma, A. (2018).Identification system for different Punjabi dialects using random forest. *Int. J. Comput. Sci. Eng*, *6*, 254-259.

22) Dua, M., Aggarwal, R. K., Kadyan, V., &Dua, S. (2012).Punjabi automatic speech recognition using HTK. *International Journal of Computer Science Issues (IJCSI)*, *9*(4), 359.

23) Changrampadi, M. H., Shahina, A., Narayanan, M. B., & Khan, A. N. (2022).End-to-End Speech Recognition of Tamil Language. *Intelligent Automation & Soft Computing*, *32*(2).

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10005831

24) Shivaprasad, S., &Sadanandam, M. (2021).Dialect recognition from Telugu speech utterances using spectral and prosodic features. *International Journal of Speech Technology*, 1-10.

25) Poonamkukana (no date) Poonamkukana/speech-dataset-of-rajasthani-language: Contains 500 audio recordings of 50 sentences from 10 speakers of Rajasthani dialect.,GitHub. Available at: https://github.com/poonamkukana/Speech-Dataset-of-Rajasthani-language.git (Accessed: February 13, 2023).

26) Hegde, P., Chittaragi, N. B., Mothukuri, S. K. P., &Koolagudi, S. G. (2020).Kannada dialect classification using cnn. In *Mining Intelligence and Knowledge Exploration: 7th International Conference, MIKE 2019, Goa, India, December 19–22, 2019, Proceedings 7* (pp. 254-259). Springer International Publishing.

27) Ittichaichareon, C., Suksri, S., &Yingthawornsuk, T. (2012, July).Speech recognition using MFCC.In *International conference on computer graphics, simulation and modeling* (Vol. 9).

28) Li, Y., Chang, S., & Wu, Q. (2022).A short utterance speaker recognition method with improved cepstrum–CNN. *SN Applied Sciences*, *4*(12), 330.

29) Albadr, M. A. A., Tiun, S., Ayob, M., & Al-Dhief, F. T. (2022). Particle Swarm Optimization-Based Extreme Learning Machine for COVID-19 Detection. *Cognitive Computation*, 1-16.

30) Kennedy, J., &Eberhart, R. C. (1942).Particle swarm optimization.IEEE Int. Conf. Neutral Networks, Australia.

31) Bantupalli, K., &Xie, Y. (2018, December). American sign language recognition using deep learning and computer vision. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 4896-4899). IEEE.

32) Alsobhani, A., ALabboodi, H. M., & Mahdi, H. (2021, August).Speech Recognition using Convolution Deep Neural Networks. In *Journal of Physics: Conference Series* (Vol. 1973, No. 1, p. 012166). IOP Publishing.

33) Kapoor, S., & Kumar, T. (2022).Fusing traditionally extracted features with deep learned features from the speech spectrogram for anger and stress detection using convolution neural network. *Multimedia Tools and Applications*, *81*(21), 31107-31128.

34) Ali, Ahmed, Dehak, Najim., Cardinal, Patrick., Khurana, Sameer., Yella, Sree. Harsha., Glass, James., Renals, Steve. (2015). Automatic dialect detection in arabic broadcast speech. *arXiv preprint arXiv:1509.06928*.

35) Aggarwal, R. K., & Dave, M. (2011, December). Application of genetically optimized neural networks for hindi speech recognition system. In *2011 World Congress on Information and Communication Technologies* (pp. 512-517). IEEE.

36) Bansal, S. R., Wadhawan, S., &Goel, R. (2022).mRMR-PSO: A hybrid feature selection technique with a multiobjective approach for sign language recognition. *Arabian Journal for Science and Engineering*, *47*(8), 10365-10380.

37) Kukana, P., Sharma, P., Bhardwaj, N. Optimized Featured Swarm Convolutional Neural Network (OFSCNN) Model based Dialect Recognition System for Bagri Rajasthani Language. Preprint at https://www.researchgate.net/publication/369726292, DOI: 10.21203/rs.3.rs-2752584/v1 (March 2023)