Computing and Artificial Intelligence 2024, 2(1), 440.
https://doi.org/10.59400/cai.v2i1.440

Article

A comparison of cepstral and spectral features using recurrent neural network for spoken language identification

Irshad Ahmad Thukroo*, Rumaan Bashir, Kaiser Javeed Giri

Department of Computer Science, Islamic University of Science and Technology, Kashmir 192122, India
*** Corresponding author:** Irshad Ahmad Thukroo, thukrooirshad@gmail.com

CITATION

Thukroo IA, Bashir R, Giri KJ. A comparison of cepstral and spectral features using recurrent neural network for spoken language identification. Computing and Artificial Intelligence. 2024; 2(1): 440.
https://doi.org/10.59400/cai.v2i1.440

ARTICLE INFO

Received: 22 December 2023
Accepted: 22 January 2024
Available online: 21 February 2024

Abstract: Spoken language identification is the process of confirming labels regarding the language of an audio slice regardless of various features such as length, ambiance, duration, topic or message, age, gender, region, emotions, etc. Language identification systems are of great significance in the domain of natural language processing, more specifically multi-lingual machine translation, language recognition, and automatic routing of voice calls to particular nodes speaking or knowing a particular language. In his paper, we are comparing results based on various cepstral and spectral feature techniques such as Mel-frequency Cepstral Coefficients (MFCC), Relative spectral-perceptual linear prediction coefficients (RASTA-PLP), and spectral features (roll-off, flatness, centroid, bandwidth, and contrast) in the process of spoken language identification using Recurrent Neural Network-Long Short Term Memory (RNN-LSTM) as a procedure of sequence learning. The system or model has been implemented in six different languages, which contain Ladakhi and the five official languages of Jammu and Kashmir (Union Territory). The dataset used in experimentation consists of TV audio recordings for Kashmiri, Urdu, Dogri, and Ladakhi languages. It also consists of standard corpora IIIT-H and VoxForge containing English and Hindi audio data. Pre-processing of the dataset is done by slicing different types of noise with the use of the Spectral Noise Gate (SNG) and then slicing into audio bursts of 5 seconds duration. The performance is evaluated using standard metrics like F1 score, recall, precision, and accuracy. The experimental results showed that using spectral features, MFCC and RASTA-PLP achieved an average accuracy of 76%, 83%, and 78%, respectively. Therefore, MFCC proved to be the most convenient feature to be exploited in language identification using a recurrent neural network long short-term memory classifier.

Keywords: MFCC; RASTA-PLP; spectral features; RNN-LSTM; SNG

1. Introduction

The technique of recognising the language of a speech segment irrespective of the speaker gender, speaker emotions, speaker tune, and speaker age is known as spoken language identification (LID). Artificial intelligence has achieved new goals in developing intelligent algorithms for processing languages (both text and audio), making humans more interactive with the systems. Many languages have been computationally modelled with a lot of effort [1–3]. In the early days of voice recognition and speaker recognition, speech signals were portrayed as a normal (smooth input without first- and second-order derivatives) input to systems. On a daily basis, LID is utilised in a range of applications, including emergency call routing and multilingual translation systems [4,5], where the response time of the operator is crucial. Despite the use of high-level techniques such as phonotactic and prosody as major sources of information in today's state-of-the-art recognition systems, acoustic

modelling is still employed. The feature extraction procedure is the most effective and critical phase in any pattern recognition activity and depends on the selection and calculation of features. The most often utilised LID characteristics are MFCC features [6–9] which have achieved great accuracy. It also depends on various types of parameters, like demand, computer resources available, and the amount of language data available, all of which impact the reason for picking appropriate language recognition capabilities. Despite significant advances in state-of-the-art voice recognition systems, there is a massive and challenging challenge ahead of us, particularly for languages with limited resources. There has been relatively little study done on the languages spoken in Jammu & Kashmir and Ladakh. To our knowledge, no language identification has been made for the Ladakhi language. In this work, we have worked on six languages, which include five languages of JK and one language of Ladakh. We used two types of features: cepstral and spectral feature vectors. The cepstral features used are MFCC and RASTA-PLP, and the spectral features used are spectral roll-off, spectral flatness, spectral bandwidth, spectral contrast, and spectral centroid individually. By using RNN-LSTM as a backend classifier, it was found that MFCC features perform well in comparison to RASTA-PLP and spectral features on all performance metrics.

The major contribution of the proposed spoken language feature selection and detection model is given below:

- To develop a novel speech corpus for Ladakhi and languages spoken in UT (Jammu & Kashmir).
- To determine various cepstral and spectral features used for spoken language identification.
- To build RNN-LSTM models using three feature vectors.
- To validate the efficiency of the proposed language model on trained datasets using different performance metrics.

The remaining sections of this paper are given here. In Section II, a state-of-the-art literature survey is presented. The III section provides a detailed description of the datasets used in the experimentation. In Section IV, a brief introduction to pre-processing (noise reduction) and feature vectors used in the proposed model is presented. Experiments and results are put forward in Section V. Section VI presents a conclusion and recommendations for future work.

2. Literature survey

To remedy the prior limitations, similar advancements are being made to propose better LID systems. In LID, the most important activity is the generation of feature maps from a raw speech signal. The feature extraction techniques that have achieved significant accuracy are Linear Prediction Coefficients (LPC) [10,11], cepstral coefficients generated from LPC (LPCC) [12,13], PLP [14,15], RASTA [16], RASTA-PLP [17–19], independent component analysis [20], MFCC [21], and kernel-based approaches [22]. MFCC has made progress in different discourse applications, particularly in LID [23–26], and it achieves high exactness. Notwithstanding, as far as the all-out inclusion aspect (the quantity of MFCC highlights partitioned by the quantity of casings), MFCC actually has significant constraints, bringing about high

time utilization and asset exhaustion. Many examinations [27–31] have taken a gander at the MFCC approach to further develop distinguishing proof exactness and lessen the absolute element aspect to conquer the intricacy, time responsibility, and restricted asset issues.

Mukherjee et al. [32] created a novel version of MFCC named MFCC-2 (second-level MFCC) to govern the MFCC's wide and uneven dimensionality, which has been used to categorize English, Bangla, and Hindi languages. The 399-value feature set of the original MFCC was used to create the MFFC-2 features. The 19 MFCC coefficients, each of which defines a Mel-scale frequency spectrum, are referred to as the number of bands. Between the global extremes of each band, 18 equally spaced classes were designated for each clip, and the number of occurrences of energy values in each such class was recorded. As a consequence, the size of the function stayed constant at $19 \times 18 = 342$ values. In addition, for each clip from the MFCC feature set that was originally produced, the number of energies, mean, and standard deviation for each of the 19 bands were measured. As a consequence, $342 + 3 \times 19 = 399$ values were produced as a feature set. Principal Component Analysis (PCA) was used to evaluate the system's performance with lower-dimensional features. Additional feature sets were assessed, and features with 0% separation functionality were removed.

Many research scientists have helped to hybridize feature extraction methodologies, especially in speech processing applications including native speaker identification, background music identification, and voice recognition. In the case of spoken language identification, however, this domain has yet to be examined. Boussaid and Hassian [33] used PLP, RASTA-PLP, and MFCC to build a speaker recognition system based on 11 Arabic words (along with their first order derivatives). PCA was employed to reduce dimensionality. Hybrid features were employed on the front end, while the FFBPNN classifier was used on the back end. Two gradient learning methods, scaled conjugate (SC) and Levenberg-Marquardt (LM), were used to improve the performance of Feed forward Back Propagation Neural Network (FFBPNN). The TrainSC learning algorithm beats the TrainLM learning technique, according to the findings of the experiments.

Samarpan et al. [34] fostered a hybrid feature selection method based on Harmony Search (HS) and Naked Mole-Rat (NMR) algorithms for SLID. MFCC and RASTA-PLP attributes were removed from a bunch of single-speaker speech datasets for ten dialects (VoxForge, IIT-Madras). The versatile HS procedure has been converged with another nature-roused approach NMR calculation to construct another cross-breed Feature Selection (FS) calculation that picks the best subset of elements while likewise lessening model intricacy to assist it with preparing quicker. To get to the exhibition of this half-component determination method utilizing regular measurements, five classifiers (Random Forest, Multi-layer Perceptron, k-Nearest Neighbour, Nave Bayes, and Support Vector Machine) have been picked in the backend. In contrast with different classifiers, Random Forest acquires the most noteworthy exactness of 99.89% on CSS10, 98.22% on VoxForge, and 99.75% on the IIT-Madras discourse corpus.

Bashir and Quadri [35] proposed a bilingual script identification system for Kashmiri and English machine-generated languages using two features: horizontal

profile coefficients (peaks) and valleys. Experiments were done using machine-generated scripts, and it was found that for the Kashmiri language, lower peaks were achieved as compared to the English language. After working on 500 text lines, it was found that 481 text lines were identified correctly, thus achieving an accuracy of 96.2%.

Thukroo and Bashir [36] proposed a mel-spectrogram-based approach using convolutional neural networks for the officially spoken languages of Jammu and Kashmir and Ladakhi. The dataset contains six languages, i.e., Kashmiri, Ladakhi, Urdu, English, Hindi, and Dogri. Initially, speech segments were converted into mel-spectrograms by using the inverse Fourier transformation to log the Fourier transformation of a time-domain signal, and at the backend, CNN serves as a classifier. Experiments were conducted on recorded speech, IIIT-H, and VoxForge. It was found that while training the model at 100 epochs, an average accuracy of 100% was achieved. One of the main drawbacks of this model is that speech segments were converted into image domain; therefore, the focus has been shifted to image domain rather than linguistic characters such as syntax and semantics of the language. Second, testing was done by using speaker-dependent samples instead of speaker-independent samples. Third, the effect of noise has not been tested in a real domain.

From the above literature survey, it was found that very little work has been done related to low-resource languages like Kashmiri and Ladakhi, both in terms of written and spoken the main challenge of any low-resource language is the creation of a speech and written corpus, as there is no freely available corpus of the above-mentioned languages. Second, whatever work has been done related to the Kashmiri language is either done in the image domain or by using cepstral features, i.e., MFCC; the effect of spectral and temporal features has not been done yet. Third, to check which linguistic features perform better in terms of spoken language has not been performed yet. In this paper, we will focus on which spectral and cepstral features are used in language identification and perform a comparative study to check which feature vector performs better in terms of spoken language identification.

3. Dataset

Most Asian and African languages are resource-poor, which poses a great challenge for processing such languages in the automation domain. As the paper is related to the processing of resource-scarce languages, one of the main issues was the creation of a sizeable dataset that is not readily available. After the data collection, we trained and tested the model on six phonetically similar languages, which are the official languages of Jammu and Kashmir and the Ladakhi language (which too has official status in the Ladakh region and is also spoken in Tibet and Gilgit-Baltistan). The languages in this dataset belong to four groups, i.e., Tibetan, Indo-Aryan, Germanic, and European. We created our own dataset for four languages spoken in Jammu and Kashmir, i.e., Dogri, Urdu, Kashmiri, and Ladakhi. The dataset is recorded with standard recording devices from different regional radio and TV channels, containing telephonic conversations, formal communication, group discussions, and formal interviews. Recording was done at a sample rate of 16 kHz with a 16-bit resolution. For the other two languages, i.e., Hindi and English, we used VoxForge

and the IIIT-H corpus. A total of 3000 audio files have been used, of which 1800 samples are used for model training and 600 samples are used for testing and validation. All files are in Wav. format with a duration of 5 seconds each. This dataset is intended to support research in the field of speech processing, particularly in the context of low-resource languages.

3.1. Dataset composition

Language Diversity: The dataset is meticulously curated to include a balanced representation of the four languages—Kashmiri, Dogri, Urdu, and Ladakhi. This diversity allows for a comprehensive analysis of the acoustic characteristics unique to each language.

Speaker Demographics: To ensure the dataset's richness and diversity, speakers were selected from various age groups, genders, and socio-economic backgrounds, reflecting the natural variation in spoken language.

Recording Conditions: Recordings were captured in diverse environments, including indoor and outdoor settings, to simulate real-world scenarios and capture the variability in acoustic conditions.

3.2. Data collection process

Native Speakers: Native speakers were recruited to ensure authentic pronunciation and natural speech patterns for each language.

Script Variety: The dataset includes a variety of scripted and spontaneous speech to capture both formal and informal language use.

Recording Equipment: High-quality recording equipment was used to minimize noise interference and maintain the fidelity of the audio recordings.

3.3. Dataset statistics

Size: The dataset includes a substantial number of hours of audio recordings for each language, totaling several hundred hours, providing sufficient data for training and evaluating speech processing models.

Accent and Dialect Variation: The dataset accounts for accent and dialectal variations within each language, enhancing its representativeness.

Speaker Distribution: The dataset ensures a balanced distribution of speakers across different demographic categories, preventing biases in model training.

In conclusion, this audio dataset and its accompanying analysis provide valuable resources for advancing research in speech processing, particularly for languages with limited linguistic resources. The comprehensive nature of the dataset ensures its applicability in various research domains, from automatic speech recognition to linguistic studies of underrepresented languages.

4. Proposed model

Figure 1 depicts the basic architecture of our model, consisting of mainly three components, i.e., pre-processing, feature extraction, and classification. Pre-processing is done to remove high noise and silence using the Spectral Noise Gate (SNG) [37]. The pre-processed signal is then forwarded for feature extraction (using MFCC,

RASTA-PLP, and spectral features), and at the end, RNN-LSTM classifiers are used for classification. The raw input speech signals have some unwanted distortions or noise. Therefore, it cannot be processed directly through the LID model. Hence, the input speech signal should undergo a pre-processing method before the feature extraction process. The pre-processing method analyses the input speech signals to see if some background noise corrupts them. In this paper, the pre-processing of input speech signals is done using the SNG technique.

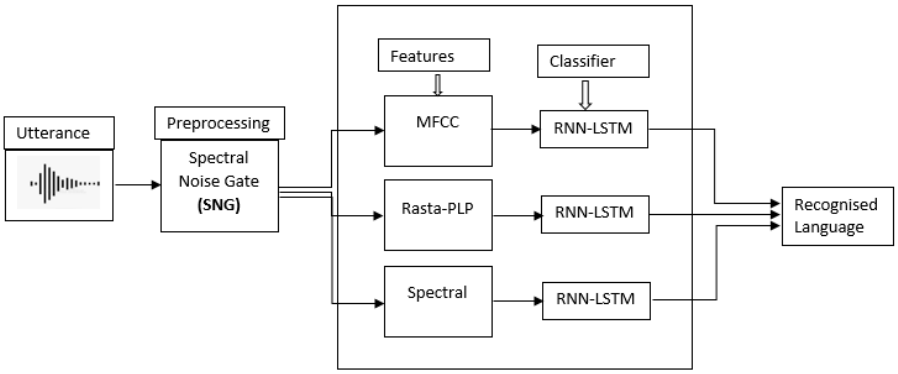

Figure 1. Architecture of model.

4.1. Spectral noise gate (SNG)

It is a common technique used for manipulation and resonant mix, which attenuates a signal corresponding to a certain threshold. This technique is also used in the general noise removal process. The sound spectrum has undergone some attenuation operations. The following steps follow the noise removal processing using SNG:

1) The noise audio clip is processed to calculate Fast Fourier Transform (FFT).
2) Statistics are calculated on the FFT of the noise in terms of frequency.
3) By using the calculated statistics, a threshold value is calculated.
4) FFT is calculated on the input signal.
5) The signal FFT and the threshold are compared to determine a mask value.
6) The determined mask is smoothed through a filter over time and frequency.
7) The smoothed mask is applied to the already calculated FFT of the input signal, and finally, the pre-processed signals are attained by applying inverse FFT.

Therefore, the SNG technique produced signal that is used for further feature extraction.

4.2. Feature extraction techniques

The pre-processed speech signals S_{kl}^{pr} are further given as input to the feature extraction process. The feature extraction is a process of deriving appropriate information from speech utterances. This paper uses two types of feature extraction methods such as cepstral features and spectral features. The cepstral features include MFCC, RASTA-PLP, and spectral features such as spectral roll-off, spectral flatness, spectral centroid, spectral bandwidth, and spectral contrast.

4.2.1. Cepstral features

This technique is used to separate the speech signals into their source and its system components. The features are less correlated in the cepstral domain, and it is invariant to amplitude and transition changes.

MFCC [38]**:** It is a predominant feature extraction method for signifying the information of speech utterances. The continuous speech utterances are recognized by matching the input signal $S_{kl}^{pr}(c)$ with a group of words or sentences. The first step is called parameterization. The input signal is transformed to the parameters to reduce the amount of redundant data. The familiar parameters in the recognition system are the MFCC. The MFCC is calculated by applying inverse Discrete Fourier transformation to the log of Fourier transformation of a time-domain signal as described below.

1) Separation of widow: We know that audio signal changes continuously with time, as it very much difficult to deal with continues signals. We do segmentation to make continuous speech signals static for a particular time in order to calculate various cepstral and spectral feature various. A window size of 2048 samples with hope length of 512 has been used.
2) Periodogram Generation: In order to determine various frequencies present in speech segment Fast Fourier Transformation (FFT) is applied to each windowed frame. This is done as analogue with human ear which contains cochlear membrane that vibrates at different places with respect to incoming speech signal frequency. The output after applying FFT on a windowed frame is called periodogram.
3) Filterbanks application: In order to know how much energy is present in a specific frequency range, filterbanks were applied to bundle periodogram bins. This is done by applying mel-filterbanks and the formula for converting hertz to mel scale is given as

$$m = 2595\log10\ (1 + f/700)$$

4) Log application: As human ear is more concerned about the lower frequencies and linear changes, because we don't hear loud volumes linearly. To make it compatible for further we apply logarithms for two reasons, one for getting smaller values and other for channel normalization.
5) Applying Discrete Cosine Transformation: Applying discrete cosine transformation (DCT) on the log filterbank energies is the final step in computing MFCC features. Because our filterbanks are all overlap, and the energies of the filterbanks are highly connected. The DCT de-correlates the energies, allowing for the application of diagonal covariance metrics to represent the features.
The MFCC is calculated in Equation (1).

$$C\big(x(t)\big) = F^{-1}\log\left(F[x(t)]\right) \quad (1)$$

where x(t) is the normal signal in the time domain, F is the discrete Fourier transformation, F^{-1} is the inverse discrete Fourier transformation (Discrete Cosine Transformation) and C represented the cepstrum. The number of MFCC coefficients used here are 128.

RASTA-PLP [36]**:** The short-term speech signals are represented by using Perceptual Linear Prediction (PLP) feature extraction method. RASTA-PLP is the

enhanced form of PLP method, which overcomes the limitations of PLP technique. This improved technique suppresses the adverse frequencies and increases the robustness of PLP in terms of noise. In RASTA-PLP method, critical band spectral resolution is applied for audible spectrum analysis, and a band-pass filter is employed for smoothening spectral variations, which are performed in each-frequency sub-band. This process leads to deriving a new spectral estimation, which is less prone to such variations. Then, the non-linear transformation process is done on the filtered speech signal spectral representation. The RASTA-PLP method tries to include the noise cancellation feature of the human auditory system and it is considered as the main advantage of this feature extraction method for the SLID system. The transformation function calculation of RASTA-PLP is formulated in Equation (2) using IIR filter as:

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - 0.98z^{-1})} \tag{2}$$

The fastest spectral change of the log spectrum is determined by the low cut-off frequency. The fastest spectral change is preserved in the output parameters, which is denoted by the high cut-off frequency. The higher values of the band-pass filter attenuate the convolution noise. The previous outputs are stored in the memory of RASTA filter. The current analysis results depend on these stored output values.

Thus, the RASTA-PLP extracted features are represented as FR_{dr}^{RASTA}, where the total number of RASTA-PLP extracted features are attained as 27.

4.2.2. Spectral features

The spectral features are a kind of frequency-based feature. They are commonly used to classify speech and audio signals. It is obtained by transforming the time-based signal into the frequency domain using FFT.

Spectral roll-off [39]: It is defined as the signal's kl^{th} percentile of spectral distribution. The percentage varies from 80 to 90 percent. It is the frequency obtained below which the magnitude distribution's percentile kl^{th} is concentrated. Spectral roll-off calculation for the given input signal S_{kl}^{pr} is given in Equation (3).

$$SRF = \sum_{qr=0}^{mn_r} \left|S_{kl}^{pr}(qr)\right| = \frac{CR}{100} \sum_{qr=0}^{MN-1} \left|S_{kl}^{pr}(qr)\right| \tag{3}$$

Here CR is the range, qr ranges from two band edges (0, MN−1). Hence, the spectral roll-off extracted features are represented as FR_{dr}^{SPR}.

Spectral flatness [39]: The spectral flatness finds the differentiation between noise and harmonic-like sounds. The spectral flatness is nearly zero for harmonic sounds and around one for noise-like sounds. In power spectrum, it measures the uniformity in the frequency distribution. It is computed as the ratio of the geometric mean to the arithmetic mean.

$$SFT = \frac{\prod_{qr=0}^{MC-1} \left|S_{kl}^{pr}(qr)\right|^{\frac{1}{MC}}}{\frac{1}{MC} \sum_{qr=0}^{MC-1} \left|S_{kl}^{pr}(qr)\right|} \tag{4}$$

Therefore, the spectral flatness extracted features are represented as FR_{dr}^{flat}.

Spectral centroid [39]: It is a spectral characterization metric used in Digital Signal Processing (DSP). It specifies where the spectrum's center of mass is located. It's calculated as the weighted mean of the signal's accessible frequencies, and it's defined by a Fourier transform using their magnitudes as weights, as in Equation (5).

$$SCD = \frac{\sum_{qr=0}^{MC-1} fe(qr)mg(qr)}{\sum_{qr=0}^{MC-1} mg(qr)} \quad (5)$$

Here, the weighted frequency value or the magnitude is denoted by $mg(qr)$, and its central frequency is denoted by $fe(qr)$ ranging from 0 to MC-1(band edges). Hence, the spectral centroid extracted features are represented as FR_{dr}^{cen}.

Spectral bandwidth [39]**:** In speech signal, the spectral bandwidth is defined as the bandwidth of signal at one-half the peak maximum used to determine the narrowness of a wave spectrum is used in music classification and environmental sound recognition.

Hence, the spectral centroid extracted features are represented as FR_{dr}^{cen}.

Spectral Contrast [39]: It is defined as the difference between the peak values and valley values of the spectrum, used in music and music mood classification. Hence, the spectral contrast extracted features are represented as FR_{dr}^{cont}.

5. Experiments and results

One of the most prominent architectures of RNN (recurrent neural network) is LSTM, which processes the entire sequence, such as audio and video. Instead of processing a single data point, such as an image, to cover long-term dependencies, As RNN is facing vanishing gradient problems when training, LSTM overcame this problem to provide better results for problems that require longer sequences of data. LSTM is insensitive to the gap length of input in comparison to RNN, Markov models, and other sequence learning methods. We use the RNN-LSTM architecture, which contains five layers, i.e., input and output layers, with three hidden layers.

The number of neurons in the input layer is 128 in order to match the shape of the input feature vector, i.e., (128,409). The input is followed by three hidden layers with nodes 128, 64, and 32, respectively, activated by the non-linear function ReLu (Rectified Linear Activation Function). A dropout of 20%, 40%, and 40% neurons has been carried out to prevent overfitting. The Softmax function is used as an output layer to determine the number of classes, i.e., six. As we have calculated 128 MFCC features for each speech segment, to make it compatible with the input layer, we use 128 nodes. Then we gradually decrease the nodes to 64 and 32 to make it suitable for the output layer. The model is optimised by RMSprop (Root Mean Squared Propagation) using sparse categorical cross-entropy as a loss function, given by Equations (6) and (7). One of the reasons for using RMSprop as an optimizer is that it converges much faster to the local minima. In addition, it can overcome getting stuck in saddle points of any of the dimensions, which may be quite common in multi-dimensional functions, and works well with sparse problems by providing easy tuning parameters in comparison to stochastic gradient descent (SGD).

$$S(x)_i = \frac{\exp(x)_i}{\sum_{j=0}^{5} \exp(x_j)} \quad (6)$$

From the above equation, S represents Softmax function, the input and output is represented by x_i and x_j respectively. The exp() represents exponentiation for both input and output vectors. Value of J ranges from 0 to 5 to determine number of categories in which an audio sample is to be classified.

$$CCE = -\sum_{j=0}^{5} t_{j \log(S(x)_j)} \quad (7)$$

In the above equation $S(x)_j$ defines Softmax probability for the *j*-th class and t_j represents the truth label.

The distribution of data for our model is 60%, and 20% of the data is allocated for training and testing, containing 1800 and 600 files, respectively, and the remaining 20% of the data is used for validating the model's performance, containing 600 audio files. We train our model by choosing the optimal batch size and number of epochs as 64 and 100, respectively. The architecture of our model is shown in **Table 1**, which has MFCC trainable parameters of 272806. We did experiments on six languages, i.e., Hindi, Dogri, Kashmiri, Urdu, English, and Ladakhi. The data to be trained is first converted into a sequence of acoustic feature vectors like MFCC, RASTA-PLP, and spectral features (spectral roll-off, spectral flatness, spectral bandwidth, spectral centroid, and spectral contrast). The acoustic features serve as input to the RNN-LSTM classifier, which has RMSprop as an optimizer. The loss and accuracy of both training and testing data are represented by **Figures 2** and **3**, respectively.

Table 1. Model architecture of Recurrent Neural Network (Long Short Term Memory).

Model: "sequential_8"		
Layer (type)	**Output shape**	**Param #**
lstm_8 (LSTM)	(None, 128)	245760
dropout_24 (Dropout)	(None, 128)	0
dense_32 (Dense)	(None, 128)	16512
dense_33 (Dense)	(None, 64)	8256
dropout_25 (Dropout)	(None, 64)	0
dense_34 (Dense)	(None, 32)	2080
dropout_26 (Dropout)	(None, 32)	0
dense_35 (Dense)	(None, 6)	198
Total params: 272,806		
Trainable params: 272,806		
Non-trainable params: 0		

Table 2 shows the model's performance on validation data (20% of the data corpus) using standard performance metrics, i.e., recall, precision, F1 score, macro accuracy, weight accuracy, and accuracy, using three different feature vectors, i.e., MFCC, RASTA-PLP, and spectral. It was found that Kashmir shows the highest precision of 98% and Hindi shows the lowest precision of 61% by using MFCC features. By using RASTA-PLP features, English shows the highest precision of 96%, while Urdu shows the lowest precision of 47%. By using spectral features, English shows the highest precision of 98%, while Urdu shows the lowest precision of 50%. Overall, Kashmiri shows the highest average precision of 95%, while Urdu shows the lowest average precision of 64%. By analysing the various performance metrics in **Table 2**, it is found that MFCC performed well in comparison to RASTA-PLP and spectral features.

Table 2. Performance of different cepstral and spectral features using standard performance metrics.

	MFCC			**RASTA_PLP**			**SPECTRAL**		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Recall	F1-Score
Dogri	00.99	00.93	00.96	00.98	00.88	00.92	00.61	00.85	00.71
English	00.60	00.62	00.61	00.46	00.96	00.62	00.86	00.98	00.91
Hindi	00.61	00.61	00.61	00.86	00.65	00.74	00.96	00.88	00.92
Kashmiri	00.98	00.98	00.98	00.82	00.90	00.86	00.85	00.97	00.90
Ladakhi	00.90	00.90	00.90	00.78	00.82	00.65	00.60	00.75	00.67
Urdu	00.92	00.95	00.93	00.67	00.47	00.63	00.82	00.50	00.62
Macro average	00.83	00.83	00.83	00.76	00.78	00.73	00.78	00.83	00.79
Weighted average	00.83	00.83	00.83	00.76	00.78	00.73	00.78	00.83	00.79
Accuracy	**00.83**			**00.76**			**00.78**		

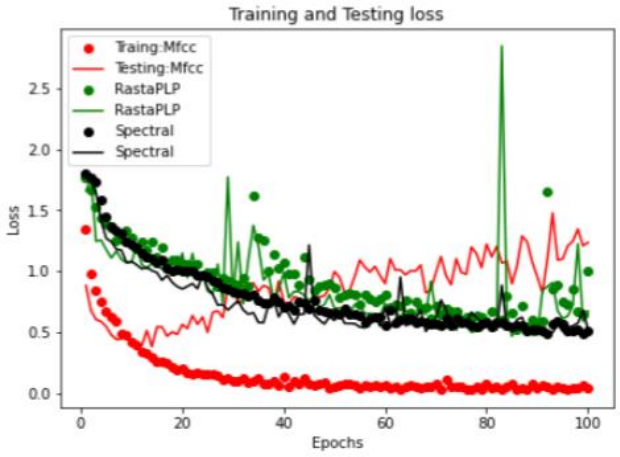

Figure 2. Cross-Category loss values of three feature vectors using training and testing data.

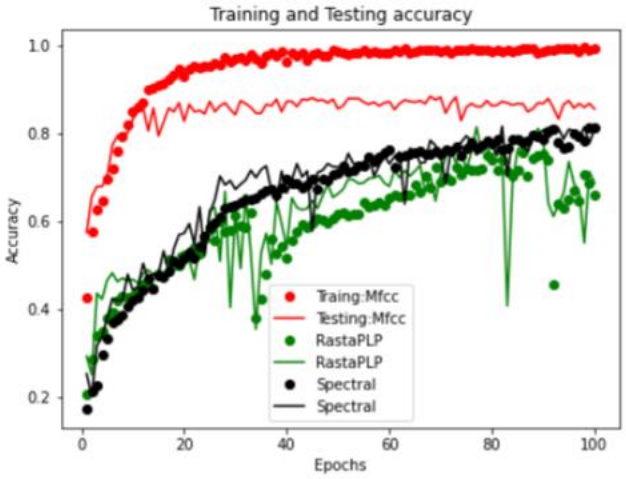

Figure 3. Accuracy of three feature vectors using training and testing data.

6. Conclusion and future work

Language identification of an audio signal is a basic task in multi-lingual speech processing systems, as it acts as a front end for various NLP tasks dealing with speech. The paper presents the comparison of various cepstral and spectral features in the process of spoken language identification using RNN-LSTM. The model has been

trained and tested on the official languages of Jammu and Kashmir (Union Territory) and Ladakhi. Various features of audio signals, such as cepstral and spectral, are used to model the language identification system. The dataset that is used in experimentation consists of some readily available portions for Hindi and English and self-created portions for the other four languages. The model has been trained on 100 epochs with RMSProp as an optimizer. We evaluate relative performance in terms of common performance criteria like precision, recall, f1-score, and accuracy. On average, MFCC, RASTA-PLP, and spectral characteristics are recognized at 83%, 76%, and 78%, respectively. The results indicate that utilising a recurrent neural network (LSTM) classifier, MFCC is the best feature for language identification. The model thus created is first of its nature related to the given set of languages, as it contains a good number of languages for which least work has been done and the Ladakhi language for which it is the first of its kind to map it with automatic processing. The main challenge was the dataset, which we handled gracefully. One of the issues with our dataset is that it lacks generality, i.e., sampling bias, limited context, temporal dynamics, feature representation, and domain shift, which lowers the accuracy of the model as the dataset is homogeneous, which needs to be considered for future work. Moreover, the model may be updated by using CNN, GRU, and attention mechanisms along with expanded datasets to make the system more relevant regarding the extraction of feature sets and subsequent linguistic categorization.

Author contributions: Conceptualization, IAT and RB; methodology, IAT; software, IAT; validation, IAT and RB; formal analysis, IAT; investigation, IAT; resources, IAT; data curation, IAT; writing—original draft preparation, IAT, RB and KJG; writing—review and editing, IAT, RB and KJG; visualization, IAT; supervision, RB and KJG. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. China Bhanja C, Laskar MA, Laskar RH. Modelling multi-level prosody and spectral features using deep neural network for an automatic tonal and non-tonal pre-classification-based Indian language identification system. Language Resources and Evaluation. 2021, 55(3): 689-730. doi: 10.1007/s10579-020-09527-z
2. Lee HS, Tsao Y, Jeng SK, et al. Subspace-Based Representation and Learning for Phonotactic Spoken Language Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020, 28: 3065-3079. doi: 10.1109/taslp.2020.3037457
3. Chandak C, Raeesy Z, Rastrow A, et al. Streaming language identification using combination of acoustic representations and ASR hypotheses. arXiv. 2020. doi.org/10.48550/arXiv.2006.00703
4. Gemmeke JF, Van Hamme H, Cranen B, et al. Compressive Sensing for Missing Data Imputation in Noise Robust Speech Recognition. IEEE Journal of Selected Topics in Signal Processing. 2010, 4(2): 272-287. doi: 10.1109/jstsp.2009.2039171
5. Wang P, Tan K, Wang DL. Bridging the Gap Between Monaural Speech Enhancement and Recognition With Distortion-Independent Acoustic Modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020, 28: 39-48. doi: 10.1109/taslp.2019.2946789
6. Albadr MAA, Tiun S, Ayob M, et al. Mel-Frequency Cepstral Coefficient Features Based on Standard Deviation and Principal Component Analysis for Language Identification Systems. Cognitive Computation. 2021, 13(5): 1136-1153. doi: 10.1007/s12559-021-09914-w
7. Biswas M, Rahaman S, Kundu S, et al. Spoken Language Identification of Indian Languages Using MFCC Features. Machine Learning for Intelligent Multimedia Analytics. Published online 2021: 249-272. doi: 10.1007/978-981-15-9492-

2_12

8. Wicaksana VS, S.Kom AZ. Spoken Language Identification on Local Language using MFCC, Random Forest, KNN, and GMM. International Journal of Advanced Computer Science and Applications. 2021, 12(5). doi: 10.14569/ijacsa.2021.0120548
9. Athiyaa N, Jacob G. Spoken language identification system using MFCC features and gaussian mixture model for Tamil and Telugu Languages. International Research Journal of Engineering and Technology(IRJET). 2019, 6(4): 4243–4248.
10. Das A, Guha S, Singh PK, et al. A Hybrid Meta-Heuristic Feature Selection Method for Identification of Indian Spoken Languages From Audio Signals. IEEE Access. 2020, 8: 181432-181449. doi: 10.1109/access.2020.3028241
11. Das HS, Roy P. Bottleneck Feature-Based Hybrid Deep Autoencoder Approach for Indian Language Identification. Arabian Journal for Science and Engineering. 2020, 45(4): 3425-3436. doi: 10.1007/s13369-020-04430-9
12. Qu D, Wang B, Wei X. Language identification using vector quantization. In: Proceedings of the 6th International Conference on Signal Processing; 26–30 August 2002; Beijing, China. 492–495. doi: 10.1109/ICOSP.2002.1181100
13. Maity S, Kumar Vuppala A, Rao KS, et al. IITKGP-MLILSC speech database for language identification. 2012 National Conference on Communications (NCC). Published online February 2012. doi: 10.1109/ncc.2012.6176831
14. Sarthak, Shukla S, Mittal G. Spoken Language Identification Using ConvNets. Ambient Intelligence. Published online 2019: 252-265. doi: 10.1007/978-3-030-34255-5_17
15. Lopez-moreno I, Gonzalez-dominguez J, Plchot, D. Martinez O, et al. Google Inc ., New York, USA ATVS-Biometric Recognition Group, Universidad Autonoma de Madrid, Spain Brno University of Technology, Czech Republic Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain. 2014. pp. 0–4.
16. Hermansky H, Morgan N. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing. 1994, 2(4): 578-589. doi: 10.1109/89.326616
17. Hermansky H, Morgan N, Bayya A, Kohn P. RASTA-PLP speech analysis technique. In: Proceedings of the ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing; 23–26 March 1992; San Francisco, CA, USA. pp. 121-124. doi: 10.1109/ICASSP.1992.225957
18. Kingsbury BED, Morgan N. Recognizing reverberant speech with RASTA-PLP. In: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing; 21–24 April 1997; Munich, Germany. pp. 1259–1262. doi: 10.1109/ICASSP.1997.596174
19. Razia Sulthana A, Mathur A. A State of Art of Machine Learning Algorithms Applied Over Language Identification and Speech Recognition Models. International Virtual Conference on Industry 40. Published online 2021: 123-132. doi: 10.1007/978-981-16-1244-2_10
20. Ghanghor N, Krishnamurthy P, Thavareesan S, et al. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In: Chakravarthi B, Priyadharshini R, Kumar MA, et al., Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages; 20 April 2021; Kyiv, Ukraine. Association for Computational Linguistics; 2021. pp. 222–229.
21. Anusuya MA, Katti SK. Speech recognition by machine: A review. International Journal of Computer Science and Information Security. 2009. 6(3): 181–205.
22. Schutte KT. Parts-Based Models and Local Features for Automatic Speech Recognition [PhD thesis]. Massachusetts Institute of Technology; 2009.
23. Deshwal D, Sangwan P, Kumar D. Feature Extraction Methods in Language Identification: A Survey. Wireless Personal Communications. 2019, 107(4): 2071-2103. doi: 10.1007/s11277-019-06373-3
24. Han W, Chan CF, Choy CS, Pun KP. An efficient MFCC extraction method in speech recognition. In: Proceedings of the 2006 IEEE International Symposium on Circuits and Systems (ISCAS); 21–24 May 2006; Kos, Greece. pp. 145–148. doi: 10.1109/ISCAS.2006.1692543
25. Dewi Renanti M, Buono A, Ananta Kusuma W. Infant cries identification by using codebook as feature matching, and MFCC as feature extraction. Journal of Theoretical and Applied Information Technology. 2013, 56(3): 437–442.
26. Trang H, Tran Hoang Loc, Huynh Bui Hoang Nam. Proposed combination of PCA and MFCC feature extraction in speech recognition system. 2014 International Conference on Advanced Technologies for Communications (ATC 2014). Published online October 2014. doi: 10.1109/atc.2014.7043477
27. Ahmed AI, Chiverton JP, Ndzi DL, et al. Speaker recognition using PCA-based feature transformation. Speech Communication. 2019, 110: 33-46. doi: 10.1016/j.specom.2019.04.001

28. Krishna SR, Rajeswara R. SVM based emotion recognition using spectral features and PCA. International Journal of Pure and Applied Mathematics. 2017, 114(9): 227–235.
29. Sabab MdN, Chowdhury MAR, Nirjhor SMMI, et al. Bangla Speech Recognition Using 1D-CNN and LSTM with Different Dimension Reduction Techniques. Emerging Technologies in Computing. Published online 2020: 158-169. doi: 10.1007/978-3-030-60036-5_11
30. Saleh MAM, Ibrahim NS, Ramli DA. Data reduction on MFCC features based on kernel PCA for speaker verification system. WALIA Journal. 2014, 30(S2): 56–62.
31. 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). Published online March 2018. doi: 10.1109/icoiact.2018.8350748
32. Mukherjee H, Obaidullah SM, Santosh KC, et al. A lazy learning-based language identification from speech using MFCC-2 features. International Journal of Machine Learning and Cybernetics. 2019, 11(1): 1-14. doi: 10.1007/s13042-019-00928-3
33. Boussaid L, Hassine M. Arabic isolated word recognition system using hybrid feature extraction techniques and neural network. International Journal of Speech Technology. 2017, 21(1): 29-37. doi: 10.1007/s10772-017-9480-7
34. Guha S, Das A, Singh PK, et al. Hybrid Feature Selection Method Based on Harmony Search and Naked Mole-Rat Algorithms for Spoken Language Identification From Audio Signals. IEEE Access. 2020, 8: 182868-182887. doi: 10.1109/access.2020.3028121
35. Bashir R, Quadri S. Identification of Kashmiri script in a bilingual document image. 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013). Published online December 2013. doi: 10.1109/iciip.2013.6707658
36. Thukroo IA, Bashir R. Spoken Language Identification System for Kashmiri and Related Languages Using Mel-Spectrograms and Deep Learning Approach. 2021 7th International Conference on Signal Processing and Communication (ICSC). Published online November 25, 2021. doi: 10.1109/icsc53193.2021.9673212
37. van Keeken A. Understanding Records. A Field Guide to Recording Practice. Second Edition. By Jay Hodgson. New York: Bloomsbury, 2019. 233 pp. ISBN 978-1-5013-4237-0. Popular Music. 2021, 40(1): 172-174. doi: 10.1017/s0261143021000192
38. Deshwal D, Sangwan P, Kumar D. A Language Identification System using Hybrid Features and Back-Propagation Neural Network. Applied Acoustics. 2020, 164: 107289. doi: 10.1016/j.apacoust.2020.107289
39. Sharma G, Umapathy K, Krishnan S. Trends in audio signal feature extraction methods. Applied Acoustics. 2020, 158: 107020. doi: 10.1016/j.apacoust.2019.107020