ISSN 2278-3091

Volume 11, No.6, November - December 2022

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse071162022.pdf

https://doi.org/10.30534/ijatcse/2022/071162022

Spoken Language Identification using CNN with Log Mel Spectrogram Features in Indian Context

Sreedhar Potla[1], Dr. B. Vishnu Vardhan[2]

[1]Assocaite Professor in Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad
Telangana, India,
sreedharpotla@yahoo.com

[2] Professor in Computer Science and Engineering, JNTUHCEM, Manthani, Telangana, India,
mailvishnu@yahoo.com

Received Date: October 20, 2022 Accepted Date: November 25, 2022 Published Date: December 06, 2022

ABSTRACT

This study demonstrates a novel application of Log Mel Spectrogram coefficients to image classification via Convolutional Neural Networks (CNN). The acoustic features obtained as log mel spectrogram images are used in this article. Log mel spectrogram pictures, a novel technique, ensure that the system is noise-resistant and free of channel mismatch. The majority of Indian languages from our own dataset were used.With the use of auditory features integrated in CNN, we hope to quickly and accurately detect a language. InceptionV3 and Resnet50 models are also used in this study for performance analysis. When compared to the existing system, these approaches achieved significant improvements in language identification accuracy.

Key words Log Mel Spectrogram(LMS), Convolutional Neural Networks (CNN), IncepetionV3, Resnet50, Language identification (LID).

1. INTRODUCTION

The process of detecting language from an audio recording made by an unidentified speaker, regardless of gender, age, or speaking style, is referred to as spoken language identification (SLI). In this paper, we primarily focus on the phonotactic and acoustic representations, which are widely used in LID. Phonotactic representations focus on capturing the statistics of phonemic constraints and patterns for each language. However, the effectiveness of such representations is heavily dependent on the phone recognizer's performance. Acoustic representations primarily capture each language's spectral feature distribution, which is more efficient and does not require prior linguistic knowledge. A front-end feature extractor that forms the frame level representation based on spectral features and a back-end model that constructs the acoustic representation for spoken LID are two important factors for effective acoustic representation. They have been preferred language identification methods due to previous research demonstrating increased efficiency and efficacy using features such as Linear Percepstral Coding Coefficients, Mel Frequency Cepstral Coefficients, Shifted Detlta Coefficients with techniques as Hidden Markov Model, or Gaussian Mixture Model. Alternative methods, such as SDC with GMM or GMM-Universal Background Model, have demonstrated higher accuracy. We employ improved training techniques such as CNN, with acceleration coefficients provided by Log Mel Spectrogram images.

Since, each language has unique qualities that set it apart from others. There are two types of features for audio samples. High-level characteristics are based on morphology and syntax, but contain language discriminating information. Low-level features are commonly based on acoustics, phonotactic and prosodic features. These features are easy to extract and require low computational power. Acoustics, phonotactic and prosodic are the main pillars for extracting low-level features. Acoustic features are any phonemes marked not by a single unique acoustic characteristic, but rather by a set of characteristics, some of which may occur in other phonemes. These characteristics are a consequence of the way the component sounds are produced by the vocal apparatus and are landmarks of word pronunciation. The inclusion of prosodic information ensures that crucial phonotactic components like sound patterns, accents, stresses, tone, rhythm, and pitch are taken into account while recognizing a language using LMS coefficients. Using these attributes guarantees great accuracy for LID since the coefficients are generally unique to its language. We build a language identification model using Sreedhar[1] proposed dataset. Gundeep Singh[2] suggested a method to extract low level features of speech samples using Log Mel Spectrogram coefficients. These settings enabled us to develop a language identification model that outperforms earlier models, while using less processing resources.

The main benefit of using CNN over the models that were previously provided is that it uses a more effective model that uses less computer resources while delivering findings that are more accurate. The advantages of Deep Neural Networks, such as dynamic input support and the usage of un-segmented inputs, become increasingly beneficial in language recognition models as the LMS coefficients ensure that the given features have channel matching and resilience to noise. This study emphasizes the value of LMS coefficients in Deep Neural Networks create a language identification model that may be applied in practical situations.

2. LITERATURE SURVEY

The process of detecting language in an audio recording regardless of gender, speaking style, or age is referred to as language identification[2]. The set of features used by SLI is broken up into different levels that match the way humans produce speech. These levels include the acoustic-phonetic information, fantastic information, prosodic information, lexical information, and syntactic information. The system that makes human speech makes possible sound units called phones. The spectral information of the vocal tract is sent by these units. Formant frequencies, articulatory patterns, and spectral components are all parts of a phone's sound. The acoustic-phonetic features of a language make it different from other languages because the number of phones in each language is different. The phonotactic information lists the rules that determine how different phonemes can be put together. Phonetic techniques have been used in conventional language identification systems since different languages have different features like tone, rhythm, accents, and sound patterns.

In designing an accurate SLI for a subcontinent like India, there are several limits related to the choice of features, the availability of speech corpus, and the models that are to be used. Mary [4] proposed a new approach for prosodic feature extraction and representation from a syllable-based segmented speech signal in the Indian context. It is demonstrated that a language recognition task performed on NIST LRE 2003 using prosodic features has the potential to discriminate languages. Jothilakshmi [5] created a hierarchical language identification model based on MFCCs and SDC characteristic vectors with GMM. Her research categories 10 Indian languages hierarchically. Her findings suggest that SLI constructed with SDC and GMM is more efficient than MFCC and GMM. Using the IITKGP-MLILSC corpus, the language identification approach described by Ramu Reddy [6] is assessed on both spectral and prosodic properties that are recorded at all conceivable levels. In Sreedhar[1] speech data from 23 major Indian languages is used for the SLI task through a GMM-UBM. The SLI performance was evaluated using both the GMM method and the GMM-UBM method. In addition, when compared to SLI using the GMM method, this method improved average SLI performance by 5.07%, while consuming only 4.67% of actual training speech data.

Identifying distinguishing features of a language model developed using acoustic characteristics that revealed the most important aspect of a language. To enhance the performance of SLI task, Sreedhar [7] the MFCC variations are used to extract acoustic features, and SDC features are then selected and trained RNN classifiers. Consequently, the SLID system's performance has significantly increased. When a LID switches from MFCC to SDC, its performance gets much better. Sarmah [8] made a model that took parts from both SDC and MFCC and used GMM as a background. This model worked better than either MFCC or SDC alone.

Despite recent developments in LID approaches, performance is far from sufficient [3]. Language features may be latent in the voice signal and mostly depending on its statistical qualities. Existing representations are weak for short duration utterances due to speaker, channel, speech content, and background noise fluctuations. For this, more discriminative and descriptive traits are preferred.

Due to their robust modelling capabilities and access to enormous datasets, deep learning approaches have recently made considerable performance advances in a number of applications, including large-scale speech recognition and picture categorization [9]. Jiang[3] provides a list of SLID deep bottleneck characteristics. Deep Bottleneck Features (DBF) for spoken LID, inspired by Deep Neural Networks (DNN) success in speech recognition. DBFs were uniquely produced by a structured Deep Neural Network (DNN) with an internal bottleneck layer. DNN training forces the activation signals in the bottleneck layer to generate a low-dimensional, compact representation of the original inputs due to the fact that the number of hidden nodes in the bottleneck layer is substantially fewer than in other layers. They demonstrate that DBFs can form a low-dimensional compact representation of the original inputs that is both descriptive and discriminative. This encourages us to concentrate on acoustic properties that are simple to calculate.

According to [10], log Mel spectrogram is the best feature for SLID over MFCC, which gives an idea of how to minimize a step in the flow to achieve accuracy. MFCC is a highly compressed form, typically employing only 20 or 13 coefficients as opposed to 32-64 bands in Mel spectrogram. The MFCC is somewhat less correlated, which might be advantageous for linear models such as Gaussian Mixture Models. In many cases, mel-spectrogram can perform better with more data and robust classifiers, such as Convolutional Neural Networks. From that conclusion, we can eliminate a step in the flow to generate a Mel spectrogram, which can be used to assess SLID performance.

S. Wang [11], proposed a deep neural networks as a signal combining tool for the multi-language identification challenge. They have extracted temporal information through MFCCs and convolutional features, a pipeline comprising bi-LSTM and transfer learning based on Inception-v3 was applied. The deep learning model performs better than their optimized lattice-based model because it is simpler to train and has more flexibility. In Celano [12], used MFCC features and converted into 640x480 sized spectrograms, which is very much suitable for training CNNs. They have also submitted a constrained system based on the ResNet50 and a deep CNN based model. This shows that even with the best system, the task is far from done.The work done by E. Salesky [13], prepared specifications for three distinct systems named Lipsia, Anlirika, and NTR using MFCC features on ResNet50 and RNN. By training on validation data, the best system got a macro-averaged accuracy of 53%.

The work [14], presented a generalization LID techniques to new speakers and to new domains through Log-Mel based pretrained models with triplet entropy loss.

The researchers[10, 14], stated that Log Mel Spectrum features are linked to prosodic characteristics. The resilience and channel mismatch can be tested using Mel feature correlation. Log Mel coefficients provide crucial language-specific information. In B.M. Abdullah[15], the CNN model generates two neural models: a baseline and a robust model for identifying spoken languages used Log Mel features. The research [16], presented SLID task using the CNN-LSTM system, which uses the CTC loss function at the output layer. This work was used in three major Indian languages represented through Log-Mel features and achieved 79.02% accuracy.

The idea for SpecAugment, an augmentation technique that works with the input audio log meal spectrogram rather than the raw audio, was inspired by research on human auditory perception and cognitive neuroscience as well as the recent success of the augmentation methods in the speech and vision domains[17]. This approach operates directly on the log mill spectrogram as if it was a picture, making it straightforward and computationally cheap.

To enhance the outcome of SLID utilizing Mel Spectrum features[2], Log Mel Spectrum features is utilized to improve effectiveness and reduce noise or channel mismatch.

When compared to MFCC, SDC, this work a SLID task that uses Mel Spectrum or LMS characteristics that are more reflective of phonetic information. CNN variants are utilized instead of RNN and GMM because it can handle un-segmented inputs while using less computing resources, making it suited for SLID[7].

The rest of the paper is organized as follows. The proposed method is described in Section 3. Section 4 illustrates how to implement LID utilising Log Mel Spectrum capabilities and Results. Section 5 ends with conclusions.

3. METHODOLOGY

In order to implement the SLID task, we are utilising the CNN model, ResNet50 model, and InceptionV3 model. We have used spoken language speech data used in [1]. The speech features were represented as Mel Frequency Cepstral Coefficients, Mel Spectrograms and Log Mel Sprctrograms.

The structure of the baseline system is shown in Figure 1, that consists of four phases, as listed in [4], beginning with 1. Spoken Data collection, 2. Data Preprocessing. 3. Feature extraction and representation and 4. Language classification. The flow of the process is shown in Figure.1

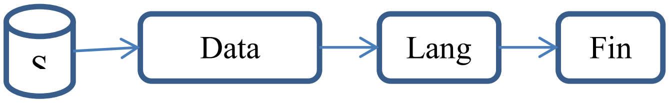


Figure 1: Baseline system of the Spoken Language Identification using CNN/Resnet50/InceptionV3

3.1 Speech Data set

The dataset that was used in this investigation is an Indian language dataset that was referred in Sreedhar [1]. The data set has comprised with 23 distinct Indian languages with durations ranging from 3 to 15 seconds. The summary of the approached dataset is provided in Table 1. We employed only 15 languages from the dataset utilised by Sreedhar [1] because of our limited computational resources.

Table 1: Indian languages speech data set details

Language	#Male Speakers	#Female Speakers	Total Duration (in mins)
Tamil	6	10	161.11
Telugu	10	5	136.53
Urdu	15	2	100.05
Gujarati	12	4	126.64
Hindi	10	9	166.24
Indian English	4	10	105.09
Kannada	5	10	145.40
Konkani	8	6	140.17
Ladkh	4	8	92.56
Leptcha	7	5	59.24
Malayalam	11	5	152.02
Manipuri	9	6	172.41
Odiya	6	5	93.25
Panjabi	12	4	169.10
Sanskrit	9	4	59.28

We have randomly selected 7800 audio recordings consisting of 3sec, 5sec, and 8sec duration samples from different age group and gender.

3.2 Data Preprocessing and feature representation

It is the second phase in our process, and its major goal is to generate a frequency spectrum employing various steps with no information losing. The Figure 2, shows detailed activities starting from data preprocessing, buffering, windowing, Fast Fourier Transformation, filtering, Mel scaling [18] and further followed by log( ) Mel Spectrogram phases as the final entity.

In this phase, the audio samples undergo pre-emphasis function, which balances the frequency spectrum, because the size of high-frequency and low-frequency waves may cause difficulties during Fourier operations shown as in [19]. In order for the audio sample to be utilized in Fourier processing without altering frequency contours, it must be divided into minute time intervals. Windowing occurs post-frame to overcome the irrelevant assumptions provided by the short-time Fourier transformation. The Fast Fourier Transform of each frame is computed to convert the time domain to the frequency domain. Now, a Mel scale is produced by splitting the full frequency spectrum into equally spaced frequencies. Finally, the Mel Spectrogram is constructed for each window by decomposing the signal's magnitude into components corresponding to Mel scale frequencies. The log( ) is applied on Mel Spectrogram to generate Log Mel Spectrogram as a final outcome.

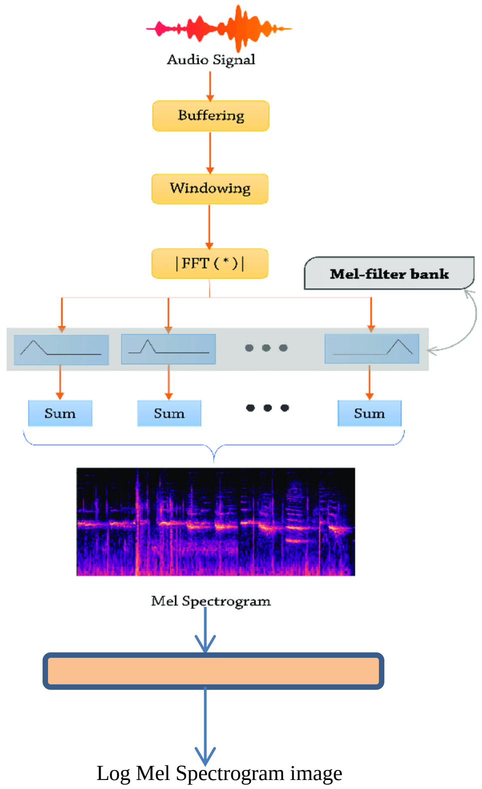


Figure 2: Data Preprocessing and Log Mel Spectrogram generation

3.3 Language Classification model

We have performed SLID task using three different models as basic CNN model, ResNet50 model and InceptionV3 model.

3.3.1 CNN classification model

As the CNN model is a kind of Deep Learning system which can take an input picture, attach importance to the image in the form of customizable weights and biases, and differentiate between various characteristics or objects included within the Log Mel Spectrogram images. ConvNets need a far less amount of pre-processing than do other classification methods due to the nature of their architecture. The development of a ConvNet, which is comparable to the function as a strategy of Neurons in the human brain, was inspired by the layout of the Visual Cortex. This makes it viable for effective language classification. The detailed description of the CNN model is demonstrated in Figure 3. Preparation of the CNN model has CNN has four convolution layers, with the first, third, and fourth layers using maximum pooling. In this case, we have 3*3 kernel size for convolution layers and 2*2 kernel size for pooling layers.

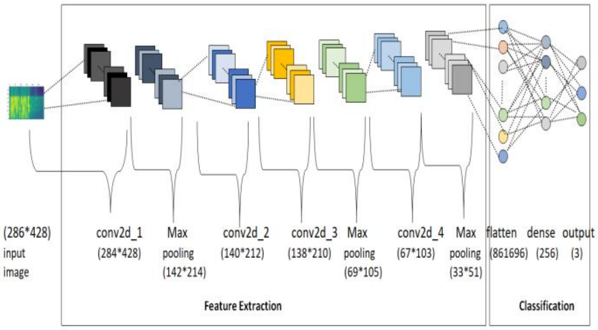
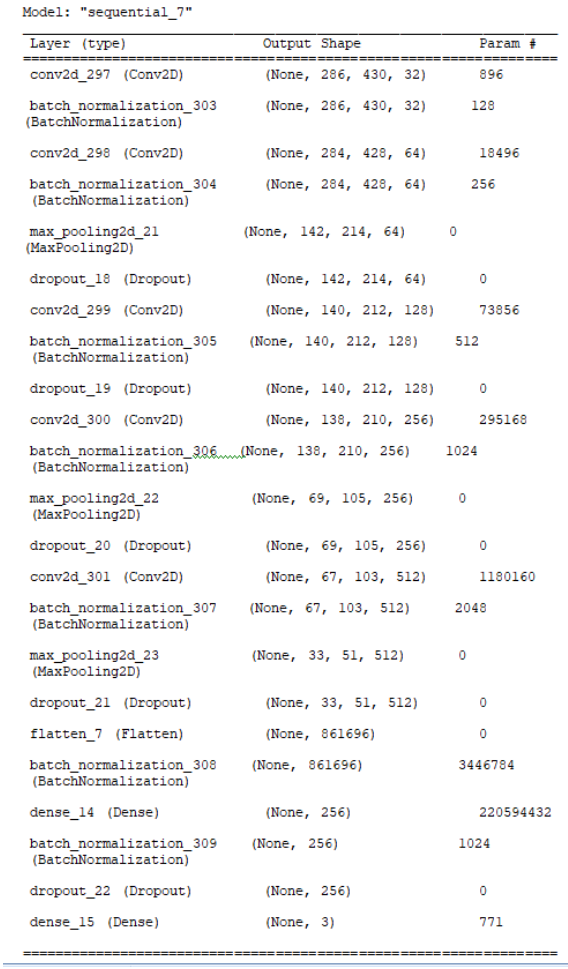


Figure 3: CNN architecture representation

The Table 2. describe the implementation details of our CNN model using a sample LMS image with the dimesions of 286 x 428 x 3.

Table 2: CNN SLID classification model implementation



```
Model: "sequential_7"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_297 (Conv2D)          (None, 286, 430, 32)      896

 batch_normalization_303      (None, 286, 430, 32)     128
(BatchNormalization)

 conv2d_298 (Conv2D)          (None, 284, 428, 64)      18496

 batch_normalization_304      (None, 284, 428, 64)     256
 (BatchNormalization)

 max_pooling2d_21          (None, 142, 214, 64)     0
(MaxPooling2D)

 dropout_18 (Dropout)         (None, 142, 214, 64)      0

 conv2d_299 (Conv2D)          (None, 140, 212, 128)     73856

 batch_normalization_305    (None, 140, 212, 128)    512
 (BatchNormalization)

 dropout_19 (Dropout)         (None, 140, 212, 128)     0

 conv2d_300 (Conv2D)          (None, 138, 210, 256)     295168

 batch_normalization_306   (None, 138, 210, 256)    1024
 (BatchNormalization)

 max_pooling2d_22           (None, 69, 105, 256)     0
 (MaxPooling2D)

 dropout_20 (Dropout)         (None, 69, 105, 256)      0

 conv2d_301 (Conv2D)          (None, 67, 103, 512)      1180160

 batch_normalization_307    (None, 67, 103, 512)     2048
 (BatchNormalization)

 max_pooling2d_23           (None, 33, 51, 512)      0
 (MaxPooling2D)

 dropout_21 (Dropout)         (None, 33, 51, 512)       0

 flatten_7 (Flatten)          (None, 861696)            0

 batch_normalization_308     (None, 861696)           3446784
 (BatchNormalization)

 dense_14 (Dense)             (None, 256)               220594432

 batch_normalization_309     (None, 256)              1024
 (BatchNormalization)

 dropout_22 (Dropout)         (None, 256)               0

 dense_15 (Dense)             (None, 3)                 771

=================================================================
```

3.3.2 Resnet50 classification model

Resnet50 is a deep convolutional neural network with 50 layers. Here we use a pre trained Resnet50 model with an ImageNet weight but in classification layer we use dense layer with size 512.

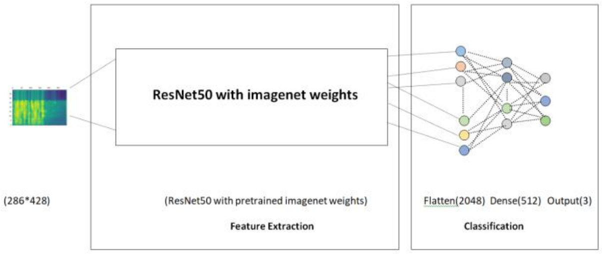
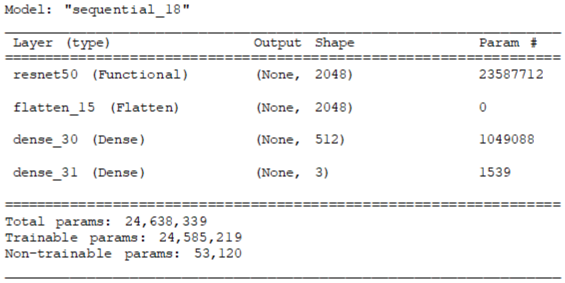

ResNet50 with imagenet weights
(286*428)
(ResNet50 with pretrained imagenet weights)
Feature Extraction
Flatten(2048) Dense(512) Output(3)
Classification

Figure 4: Architecture of Resnet50 classification model

The Figure 4. show structure of the Resnet50 approach and implementation details are given in Table 3.

Table 3: ResNet50 SLID classification model implementation



```
Model: "sequential_18"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 resnet50 (Functional)       (None, 2048)              23587712

 flatten_15 (Flatten)        (None, 2048)              0

 dense_30 (Dense)            (None, 512)               1049088

 dense_31 (Dense)            (None, 3)                 1539

=================================================================
Total params: 24,638,339
Trainable params: 24,585,219
Non-trainable params: 53,120
_________________________________________________________________
```

3.3.2 InceptionV3 classification model

The third model used for SLID task here is InceptionV3 classification model. The InceptionV3 model is createdwith the gal of enabling deeper networks while keeping model parameters under 25 million as compared with 60 million for Alexnet[20]. This model is trained with an ImageNet weight used at the end of dense layers of size as 1024. The Figure 5. present the architecture of the InceptionV3 classification model.

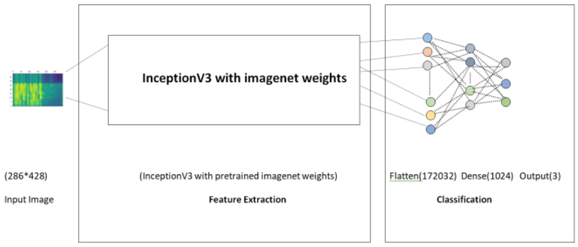
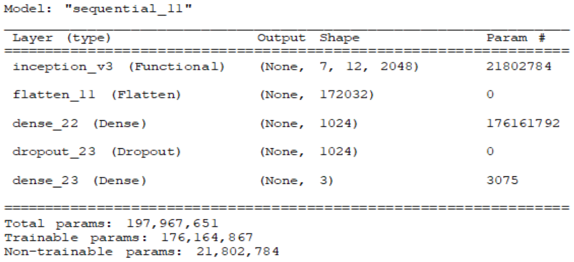

InceptionV3 with imagenet weights
(286*428)
Input Image
(InceptionV3 with pretrained imagenet weights)
Feature Extraction
Flatten(172032) Dense(1024) Output(3)
Classification

Figure 5: Architecture of InceptionV3 classification model

The implementation details of the InceptionV3 classification model are described in Table 4.

Table 4: Performance of InceptionV3 SLID classification model



```
Model: "sequential_11"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 inception_v3 (Functional)   (None, 7, 12, 2048)       21802784

 flatten_11 (Flatten)        (None, 172032)            0

 dense_22 (Dense)            (None, 1024)              176161792

 dropout_23 (Dropout)        (None, 1024)              0

 dense_23 (Dense)            (None, 3)                 3075

=================================================================
Total params: 197,967,651
Trainable params: 176,164,867
Non-trainable params: 21,802,784
```

4. RESULTS

In this work, we verified the performances of three SLID classification models as the CNN classification model, ResNet50 classification model and InceptionV3 classification model. We have prepared separate image dataset for both training and testing based on the major Indian language dataset used[1] using Mel Spectrum images and Log Mel Spectrum images. We have also verified the performances of our models against Kaggle dataset[21].

The SLID accuracy values are presented in Table 5. Mel Spectrograms of 5 sec and 10 sec speech samples are used to demonstrate the effectiveness of the proposed models. Using the Mel spectrogram and the ResNet50 classification model, we find the greatest SLID performance as 93.4% for 10 second samples from the Indian languages dataset[1]. The best SLID performance found as 89.1% for 10 second samples on the Foreign languages dataset[21] through the ResNet50 classification model. The SLID results of ResNet50 classifier are better when compared to the CNN classifier and the InceptionV3 classifier.

Table 5: Performance of CNN model, ResNet50 model and an InceptionV3 model for SLID using Mel Spectrum features

SLID classifier	SLID Accuracy % for Indian Languages[1]		SLID accuracy % for Foreign Languages[25]	
	Speech Duration		Speech Duration	
	5 sec	10 sec	5 sec	10 sec
CNN	82.3%	87.7%	63.5%	76.7%
Resenet50	91.2%	93.4%	84.6%	89.1%
InceptionV3	83.2%	85.3%	68.6%	70.4%

The methods are based on CNN, ResNet50 and InceptionV3 using Log Mel spectrum features and were experimented on Indian Language dataset[1] and Kaggle foreign language data set[21]. The SLID task results are presented in Table 6. For SLID accuracy of 96.7% and 94.8% were achieved using ResNet50 on 10sec and 5sec speech data respectively. When ReNet50 was fed with Log Mel spectrogram features on Indian Dataset[1], an improvement of 45.1% is observed. An accuracy improvement of 6.1% and 2.7% were noticed when ResNet50 with Log Mel scale features and Mel spectrum features used respectively against SLID task carried out using RNN with SDC_MFCCΔ [7]. All three classifier shows that the SLID performance using Indian language data set[1] are superior when compared to Foreign language dataset[21].

Table 6: Performance of CNN model, ResNet50 model and InceptionV3 model for SLID using Log Mel Spectrum features

SLID classifier	SLID Accuracy % for Indian Languages[1]		SLID accuracy % for Foreign Languages[25]	
	Speech Duration		Speech Duration	
	5 sec	10 sec	5 sec	10 sec
CNN	80.6%	85.4%	79.3%	82.7%
Resenet50	94.8%	96.7%	87.3%	89.1%
InceptionV3	84.2%	86.3%	80.2%	81.9%

5. CONCLUSIONS

In this work, both Mel Spectrograms and Log Mel Spectrogram are prepared for 23 Major Indian languages[1] and 3 Foreign languages[21] respectively. The SLID task is designed with three classifiers using CNN classifier, ResNet50 classifier and InceptionV3 classifier. SLID performances of various existing methods and features used are compared with the proposed ResNet50 using Log Mel Spetrogram features tabulated in Table 7.

Table 7: Performance of various existing models with proposed model

Work's done	features	Model	Accuracy %
Sreedhar[1]	MFCC	GMM-UBM	41.6%
Jyothilaxmi[25]	MFCC, MFCC Δ MFCC ΔΔ	GMM HMM and ANN	80.56%
Rangan P[22]	Log Mel	CNN LSTM	79.02%
Vander[14]	Log Mel	CRNN Resnet50	89%
Montavon [23]	MFCC	CNN TDNN	91.2%
Sreedhar P[7]	SDC_MFCCΔ	RNN	90.66%
Proposed work	Log Mel Spectrogram	ResNet50	96.7%

REFERENCES

1. Sreedhar Potla and B. V. Vardhan, **Spoken language identification using Gaussian mixture model-universal background model in Indian context**, *International Journal of Applied Engineering Research*, Vol. 13, pp. 2694–2700, 2018.
2. Gundeep Singh, Sahil Sharma, Vijay Kumar, Manjit Kaur, Mohammed Baz, Mehedi Masud, **Spoken Language Identification Using Deep Learning**, *Computational Intelligence and Neuroscience*, Vol.2021, 12 Pages, 2021.
3. Jiang B, Song Y, Wei S, Liu J-H, McLoughlin IV, Dai L-R (2014) **Deep Bottleneck Features for Spoken Language Identification**. *PLoS ONE* 9(7): e100795.
4. Mary, L., and Yegnanarayana, B.: **Extraction and representation of prosodic features for language and speaker recognition**. *Speech communication*, 50(10), 2008, pp. 782–796.
5. Jothilakshmi, S., Palanivel, S., & Ramalingam, V. (2012). **A hierarchical language identification system for Indian languages**. *Digital Signal Processing*, 22(3), 544–553.
6. Reddy, V. R., Maity, S., & Rao, K. S. (2013). **Identification of Indian languages using multi-level spectral and prosodic features**. *International Journal of Speech Technology*, 16(4), 489–511.
7. Sreedhar Potla and Dr. B. V. Vardhan, **Spoken language identification using RNN with SDC features in Indian context**, *GIS SCIENCE JOURNAL*, Vol. 9, Issue 7, 2022, pp. 773-780.
8. Sarmah K, Bhattacharjee U. **GMM based Language Identification using MFCC and SDC Features**. *International Journal of Computer Applications*. 2014 Jan 1;85(5).
9. Hinton G, Deng L, Yu D, Dahl GE, Mohamed Ar, **Deep neural networks for acoustic modeling in speech recognition**: *The shared views of four research groups. IEEE Signal Processing Mag* 29: 2012, pp. 82–97
10. Meng, Hao, Tianhao Yan, Fei Yuan and Hongwei Wei. **Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network**, *IEEE Access* 7 (2019): 125868-125881.
11. Wang S, Wan L, Yu Y, Moreno IL. **Signal combination for language identification.** *arXiv preprint arXiv:*1910.09687. 2019 Oct 21.
12. Celano, G. G. (2021, June). **A resnet-50-based convolutional neural network model for language id identification from speech recordings**. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP* , 2021 June, pp. 136-144.
13. Salesky, E., Abdullah, B. M., Mielke, S. J., Klyachko, E., Serikov, O., Ponti, E., ... & Vylomova, E., **SIGTYP 2021 shared task: robust spoken language identification**. *arXiv preprint arXiv:2106.03895*., 2001.
14. Van der Merwe, R. **Triplet entropy loss: improving the generalization of short speech language identification systems**. *arXiv preprint arXiv:2012.03775*., 2020.
15. Abdullah, B. M., Kudera, J., Avgustinova, T., Möbius, B., & Klakow, D. **Rediscovering the slavic continuum in representations emerging from neural models of spoken language identification.** *arXiv preprint arXiv:2010.11973*. 2020.
16. Rangan, P., Teki, S., & Misra, H. **Exploiting spectral augmentation for code-switched spoken language identification.** *arXiv preprint arXiv:2010.07130*., 2020.
17. Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. **Specaugment: A simple data augmentation method for automatic speech recognition.** *arXiv preprint arXiv:1904.08779*. 2019.
18. Guha, S., Das, A., Singh, P. K., Ahmadian, A., Senu, N., & Sarkar, R.,. **Hybrid feature selection method based on harmony search and naked mole-rat algorithms for spoken language identification from audio signals**. *IEEE Access*, *8*, 182868-182887, 2020.
19. https://superkogito.github.io/blog/2020/01/25/signal_framing.html

20. https://cloud.google.com/tpu/docs/inception-v3-advanced
21. Spoken Language Identification | Kaggle, https://www.kaggle.com/toponowicz/spoken-language-identification, 2021
22. Rangan, P., Teki, S., & Misra, H. **Exploiting spectral augmentation for code-switched spoken language identification.** *arXiv preprint arXiv:2010.07130*. 2020.
23. Montavon, G. **Deep learning for spoken language identification**. In *NIPS Workshop on deep learning for speech recognition and related applications* Vol. 49, No. 10, pp. 911-914, 2009, December.
24. F. Allen, E. Ambikairajah and J. Epps, "Language identification using warping and the shifted delta cepstrum", *Proc. IEEE Workshop on Multimedia Signal Processing*, pp. 1-4, 2005.
25. Jothilakshmi, S., Palanivel, S., & Ramalingam, V. **A hierarchical language identification system for Indian languages**. *Digital Signal Processing*, 22(3), 544–553. 2012.