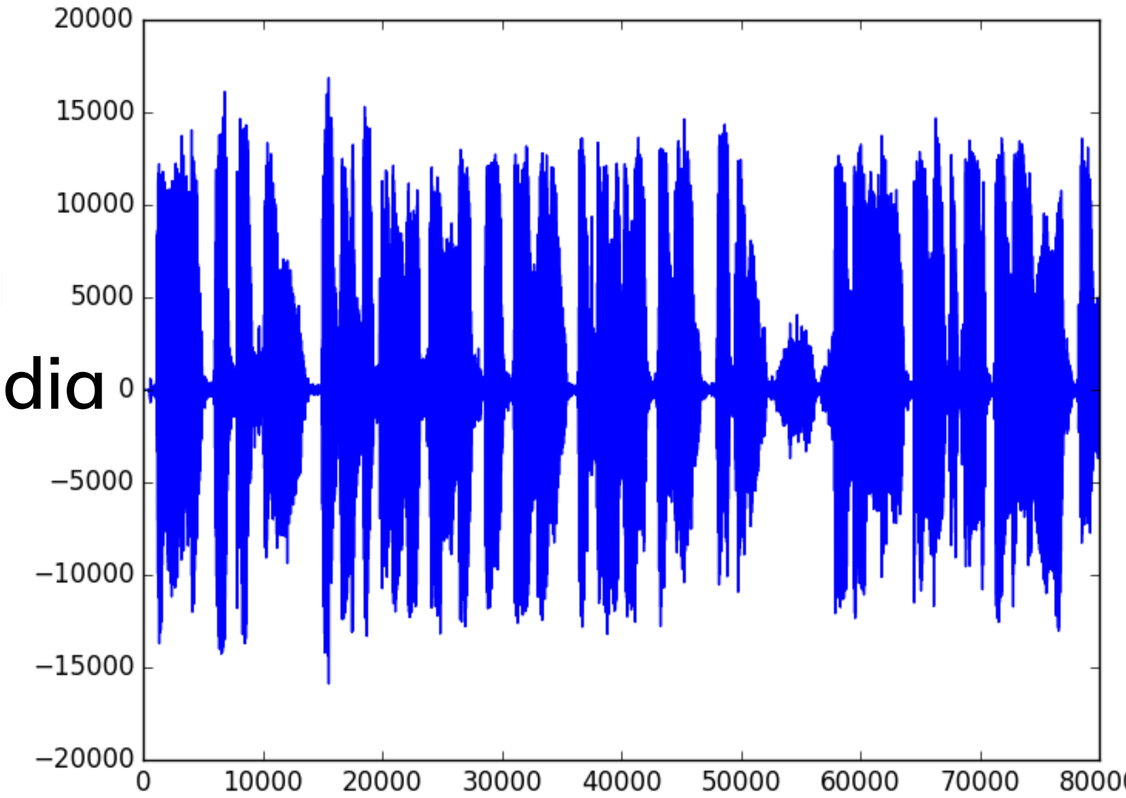
DenseRecognition of Spoken Languages

Jaybrata Chakraborty, MAKAUT WB, India
Bappaditya Chakraborty, MAKAUT WB, India
Ujjwal Bhattacharya, ISI Kolkata, India



Problems Related to Language Identification

- Understanding the language from the spoken utterances.
- Large class classification of Indian languages having significant pronunciation similarities.
- Presence of noise and silence zones in the speech signal.
- Limitations of handcrafted features.

Motivation

- Articulatory parameters , prosody , phonotactic and lexical knowledge are used as features in many literatures around the world.
- Identification of Indian languages cannot produce results with high accuracy due to their phonetic closeness with a few others.
- Hand-crafted features may not be able to encode enough discriminatory characteristics of speech signal for efficient automatic classification.
- Automatic extraction of efficient features for speaker independent recognition of language from its speech segments in the presence of noise & phonetic similarities.

Prior Work

- Only a few studies of Indian language recognition from speech signal are found.
- A major bottleneck of pursuing effective studies on Indian language recognition was unavailability of necessary speech corpus for major languages until 2012.
- IITKGP-MLILSC speech (news) corpus of 27 Indian languages published in 2012.
- Authors of IITKGP-MLILSC corpus studied two distinct systems on a subset of its languages, for speaker dependent and speaker independent scenarios.
- Existing recognition studies include a few statistical methods using handcrafted features.
- Performance of existing approaches in speaker independent scenario are limited.
- Proposed DenseNet based approach have significantly outperformed the state-of-the-art recognition results of Indian languages.

Datasets Used

IITKGP-MLILSC Corpus

Recordings of news clips in 27 Indian languages.

Characteristics

- Mainly noise free.
- Smaller silence zones in individual audio clips.
- Language and gender specific organization of data.

Linguistic Data Consortium Corpus

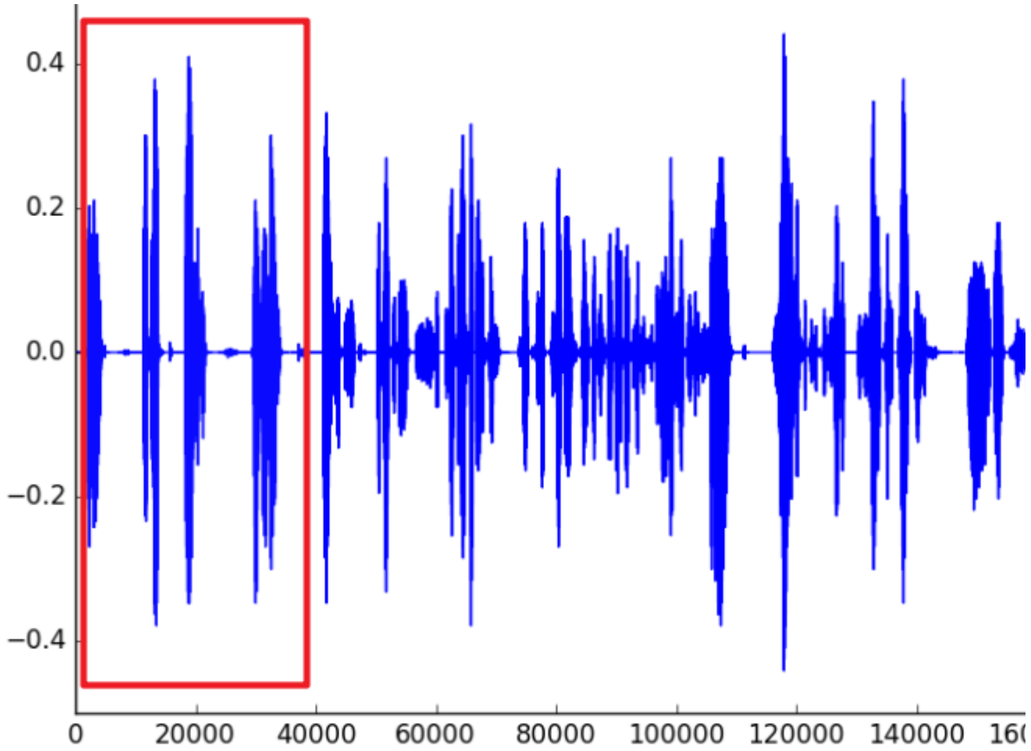
Recordings of telephonic conversations in 5 Indian languages.

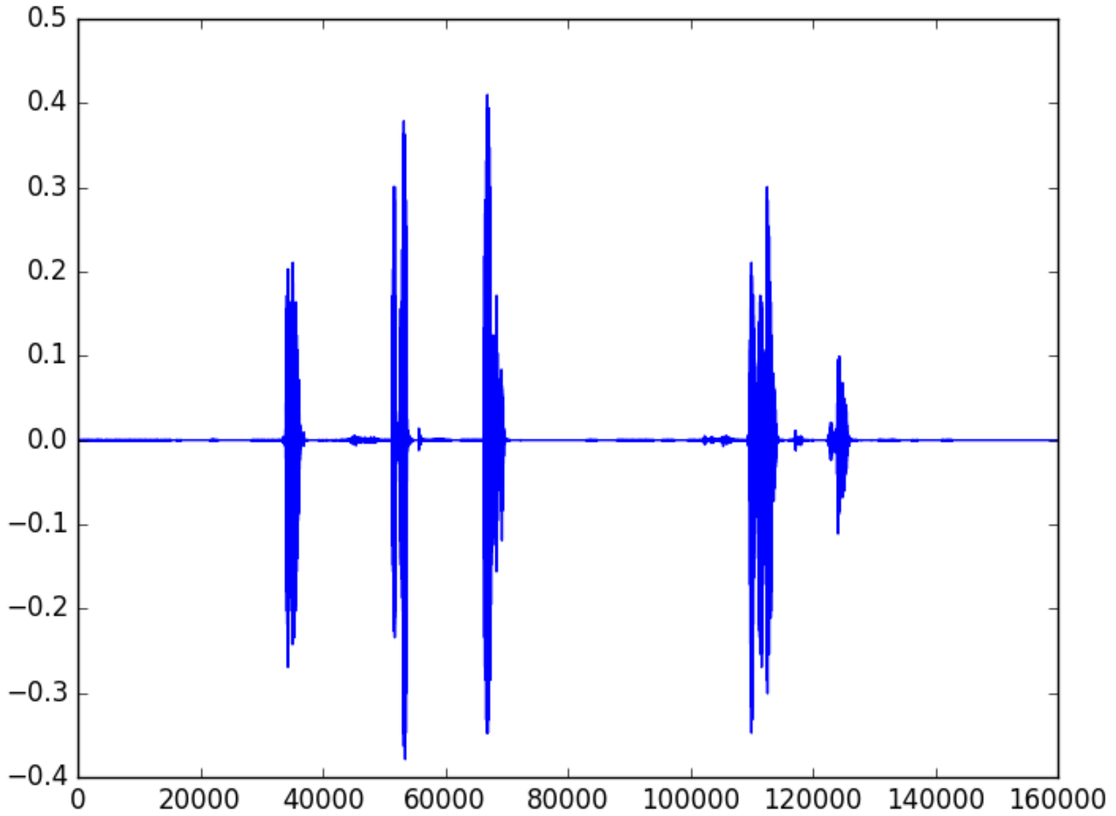
Characteristics

- More natural and dual channels.
- Larger silence zones in each audio clip.
- Only language specific organization of data.

Preprocessing

Certain sliding window based strategy removes noisy zones from input raw audio signals.





Low energy frames of an input speech sample are discarded to minimize the silent and moderately noisy zones present in the input sample.

Feature Extraction

Two types of features have been studied -- (i) traditional handcrafted features and (ii) CNN based automatic features

Mel-Spectrogram

- HANN window based power spectrogram is computed.
- Mel-scale filter banks are applied.
- The Mel-spectrogram is fed into the input layer of Dense CNN architecture.

MFCC, Delta and Delta-Delta Coefficients

- Thirteen MFCC coefficients represent local spectral features of short utterances.
- Delta and Delta-Delta coefficients represent velocity and acceleration of computed MFCC.
- Total no. of features: 39 per frame.

Other Prosodic features

- Other prosodic features include Energy, Energy Entropy, Spectral centroid, Spectral spread, Spectral entropy, Spectral flux and Spectral Rolloff.

Network Architectures

BLSTM Based Architecture

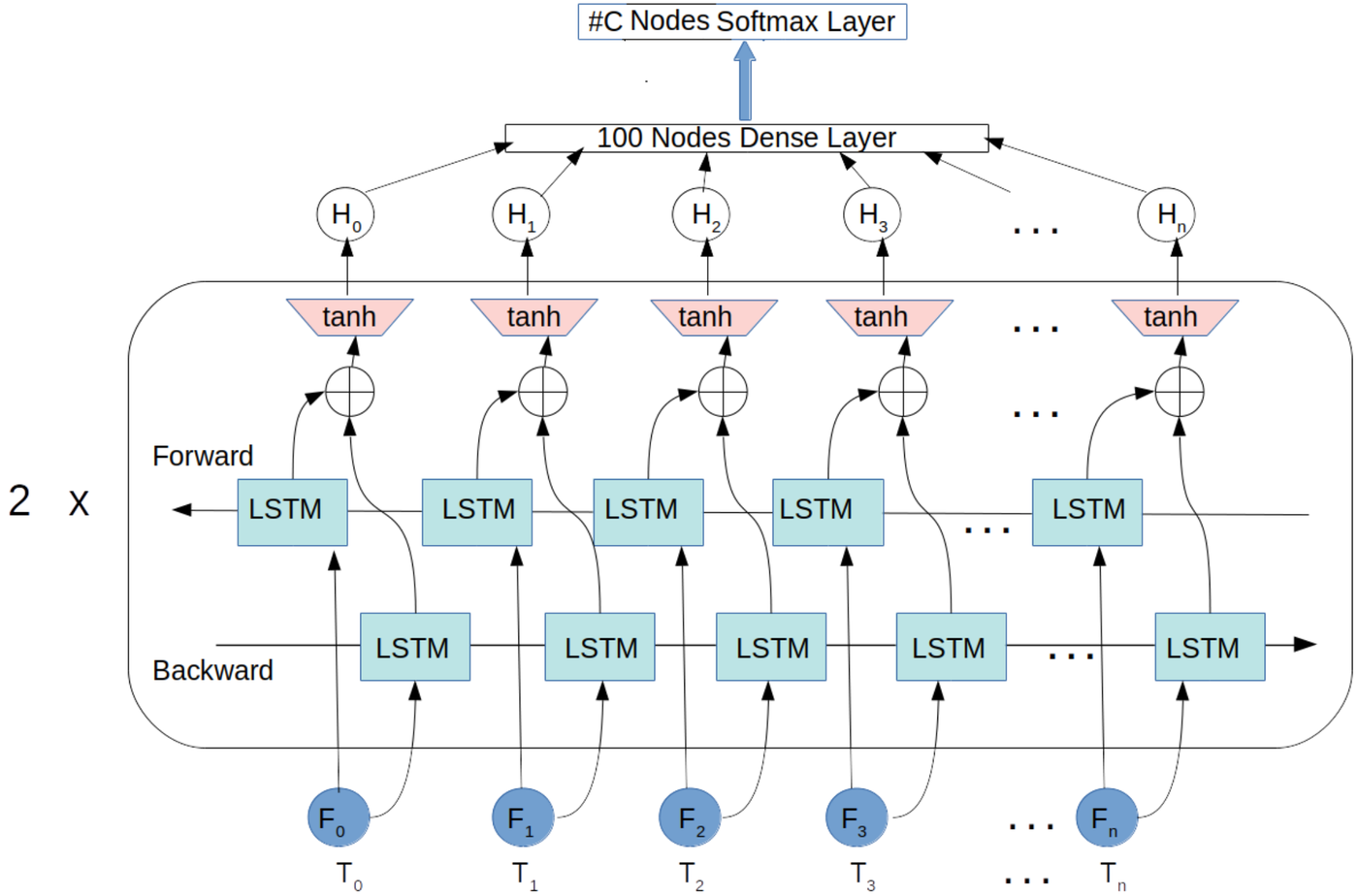
- 39 Features: MFCC + Delta + Delta-Delta
- 34 Features: MFCC + certain acoustic and phonetic features
- Strategy used: 50 ms strides with 50% overlap generating 399 and 199 frames respectively for 10s and 5s speech segments.
- BLSTM Network is fed separately with both sets of handcrafted features and tested on both the datasets.

DenseNet Based Architecture

- Mel-spectrograms are fed as features.
- DenseNet based architecture that is capable of automatic feature extraction.
- Mel-Spectrograms with DenseNet-BLSTM hybrid architecture.

Network Architectures

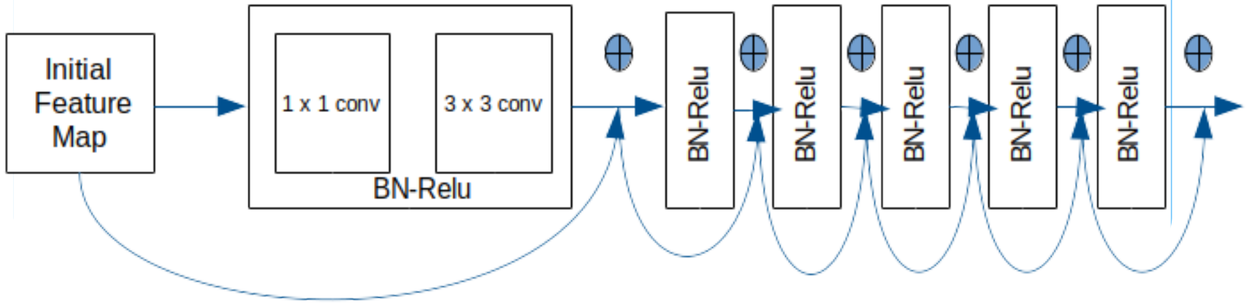
BLSTM Based Architecture



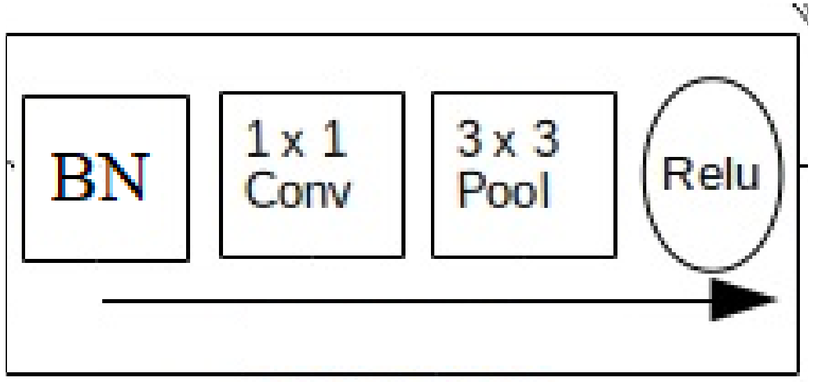
Architecture of the BLSTM network used in the present study. Handcrafted features are fed as input to this network. Output layer consists of #C number of nodes, where #C denotes the number of underlying classes. #C =27 for the IITKGP-MLILSC dataset and #C = 5 for the LDC dataset.

Network Architectures

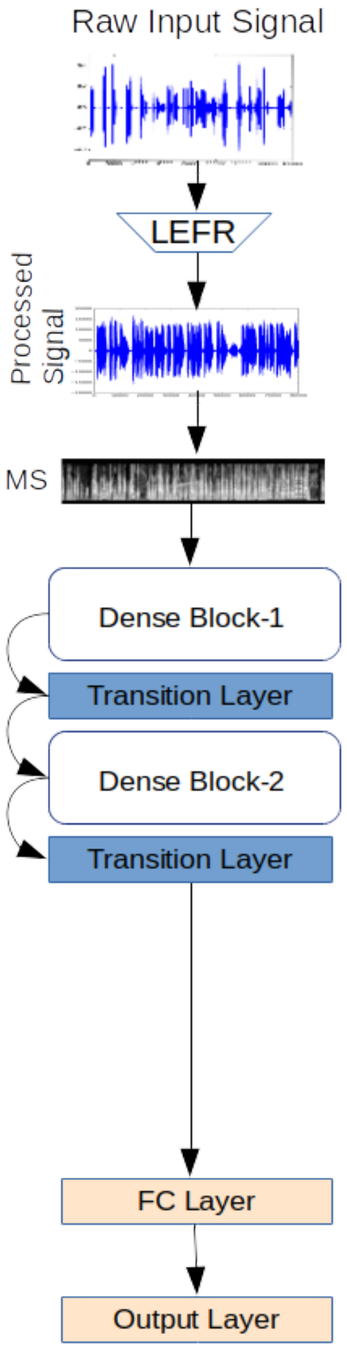
DenseNet Based Architecture



Architecture of a Dense Block



Transition Layer



Proposed recognition framework.

Mel-spectrogram (MS) of speech signal is fed as input to both the networks (LEFR represents preprocessing operation for removal of low energy frame removal).

Experimentation Results (IITKGP-MLILSC Corpus Speaker Dependent Recognition)

(MFCC + Delta + Delta-Delta)+ BLSTM
5 sec - 93.82 %
10 sec- 94.35 %
recognition rate

(MFCC + Additional Features) + BLSTM
5 sec - 90.47 %
10 sec- 93.05 %
recognition rate

(MFCC + Delta + Delta-Delta) + CNN
5 sec - 91.65 %
10 sec- 95.74 %
recognition rate

Mel-Spectrogram (MS) + CNN
5 sec - 93.51 %
10 sec- 96.68 %
recognition rate

(MS) + CNN+ BLSTM
5 sec - 89.13 %
10 sec- 92.19 %
recognition rate

(MS) + ResNet10
5 sec - 92.85 %
10 sec- 93.57 %
recognition rate

(MS) + ResNet18
5 sec - 93.05 %
10 sec- 94.17 %
recognition rate

(MS) + DenseNet-BLSTM
5 sec - 79.5 %
10 sec- 82.39 %
recognition rate

Our Approach
MS+DenseNet
5 sec- 94.44 %
10 sec- 97.07 %
recognition rate

Experimentation Results (IITKGP-MLILSC Corpus Speaker Independent Recognition)

(MFCC + Delta + Delta-Delta)+ BLSTM
5 sec - 65.54 %
10 sec- 66.35 % recognition rate

(MFCC + Additional Features) + BLSTM
5 sec -64.39 %
10 sec- 68.57 % recognition rate

(MFCC + Delta + Delta-Delta) + CNN
5 sec - 70.01 %
10 sec- 69.49 % recognition rate

Mel-Spectrogram(MS) + CNN
5 sec - 72.2 %
10 sec- 76.39 % recognition rate

(MS) + CNN+ BLSTM
5 sec - 62.4 %
10 sec- 67.19 % recognition rate

(MS) + ResNet10
5 sec - 71.25 %
10 sec- 73.05 % recognition rate

(MS) + ResNet18
5 sec - 71.25 %
10 sec- 74.38 % recognition rate

(MS) + DenseNet-BLSTM
5 sec - 80.2 %
10 sec- 82.19 % recognition rate

Our Approach
MS+DenseNet
5 sec- 84.24%
10 sec- 89.07 % recognition rate

Experimentation Results (LDC Corpus Data Speaker Independent Recognition)

(MFCC + Delta + Delta-Delta)+ BLSTM
5 sec - 81.24 %
10 sec- 85.05 %
recognition rate

(MFCC + Additional Features) + BLSTM
5 sec - 78.65 %
10 sec- 84.51 %
recognition rate

(MFCC + Delta + Delta-Delta) + CNN
5 sec - 79.38 %
10 sec- 86.42 %
recognition rate

Mel-Spectrogram (MS) + CNN
5 sec - 84.34 %
10 sec- 92.42 %
recognition rate

(MS) + CNN+ BLSTM
5 sec - 78.4 %
10 sec- 81.13 %
recognition rate

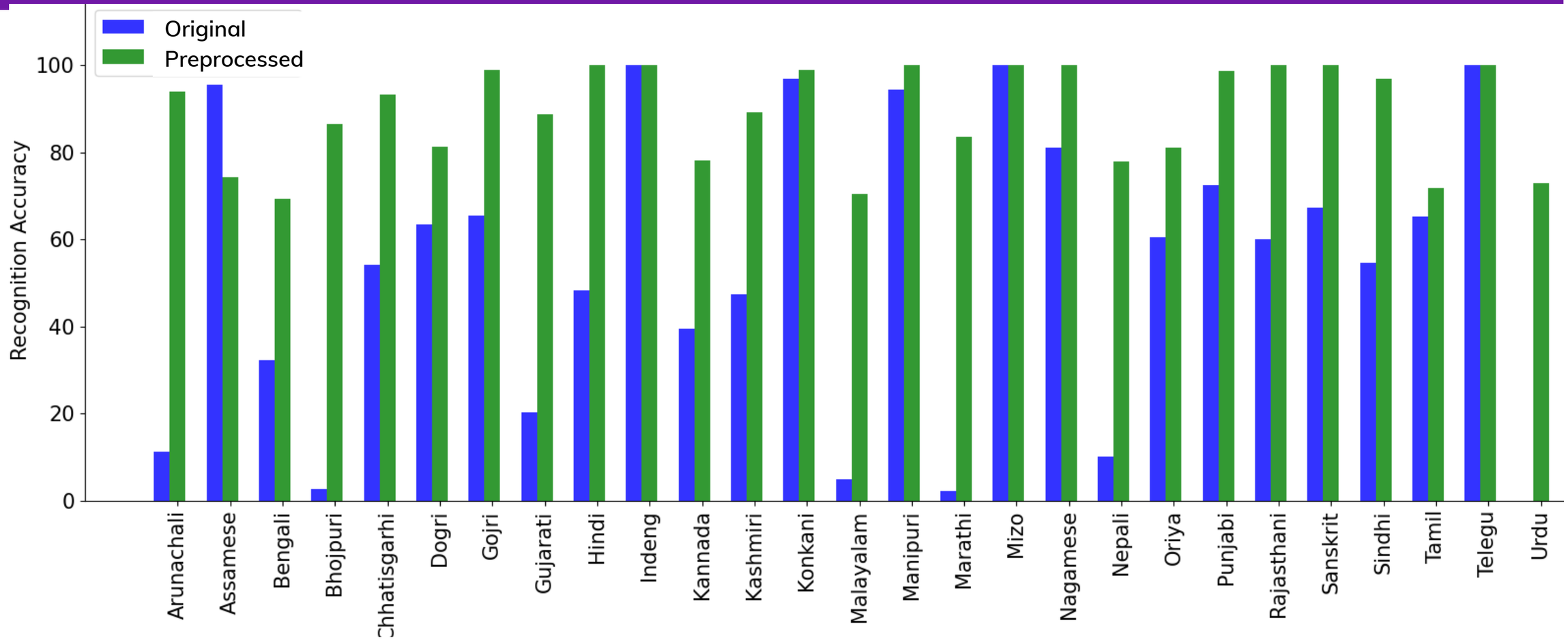
(MS) + ResNet10
5 sec - 82.82 %
10 sec- 91.2 %
recognition rate

(MS) + ResNet18
5 sec - 83.98 %
10 sec- 91.82 %
recognition rate

(MS) + DenseNet-BLSTM
5 sec - 87.5 %
10 sec- 91.19 %
recognition rate

Our Approach
MS+DenseNet
5 sec- 90.24 %
10 sec- 94.06 %
recognition rate

Recognition on Preprocessed vs. Original Signal



Comparative recognition results of the proposed framework on samples of IITKGP-MLILSC Corpus for each individual language with and without preprocessing.

Conclusions

- Proposed approach using Mel-spectrogram features has shown significantly improved recognition performance over the state-of-the-art LID systems.
- Experimentation on IITKGP-MLILSC and LDC datasets has shown higher misclassification rates within a few groups of phonetically similar languages.
- Recognition accuracy with IITKGP-MLILSC is higher as LDC dataset consist of real-life conversations over noisy telephonic channel whereas samples of IITKGP-MLILSC dataset consists of comparatively less noisy and uniformly spoken samples collected from either TV or radio broadcasts.
- Proper representations of natural variations of speech samples with respect to pronunciation, pitch, rates of speech etc. in the training set should lead to better recognition performance of the proposed approach.

Some Relevant References

- *K. S. Rao, V. R. Reddy, and S. Maity, Language Identification Using Spectral and Prosodic Features. Springer, 2015.*
- *K. S. Rao and D. Nandi, Language Identification Using Excitation Source Features. Springer, 2015.*
- *A. Lozano-Diez, O. Plchot, P. Matejka, and J. Gonzalez-Rodriguez, "DNN based embeddings for language recognition," ICASSP. 2018, pp. 5184–5188.*
- *F. Iandola, et al., "DenseNet: Implementing efficient ConvNet descriptor pyramids," arXiv preprint arXiv:1404.1869, 2014*
- *P. Shen, X. Lu, S. Li, and H. Kawai, "Interactive learning of teacher-student model for short utterance spoken language identification," ICASSP. 2019, pp. 5981–5985.*
- *T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech Communication, vol. 52, no. 1, pp. 12–40, 2010.*
- *K. Kirchhoff, S. Parandekar, and J. Bilmes, "Mixed-memory Markov models for automatic language identification," Proc. ICASSP, vol. 1, 2002, pp. I–761.*
- *A. G. Adami and H. Hermansky, "Segmentation of speech for speaker and language recognition." INTERSPEECH, 2003, pp. 841–844.*

Thank You