DenseRecognition of Spoken Languages

Jaybrata Chakraborty
MAKAUT, West Bengal
Kolkata, India
jaybrata1411@gmail.com

Bappaditya Chakraborty
MAKAUT, West Bengal
Kolkata, India
bappa.chakraborty84@gmail.com

Ujjwal Bhattacharya
CVPR Unit, Indian Statistical Institute
Kolkata, India
ujjwal@isical.ac.in

Problems related to Language Identification

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

Language Corpora

1. 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.

2. Linguistic Data Consortium (LDC) Telephonic Speech Corpus

- Recordings of telephonic conversations in 5 Indian languages.

Characteristics

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

Preprocessing

Audio signals that are recorded in an uncontrolled environment often require several preprocessing stages in order to remove unwanted noises and/or enhance power. In this experiment we use noise filtering to discard low energy frames from a speech sample to minimize the silence, moderately noised zones and low voices present in the original sample. The method is carried out by choosing a sliding window of length 1s and computing the energy of that 1s segment.

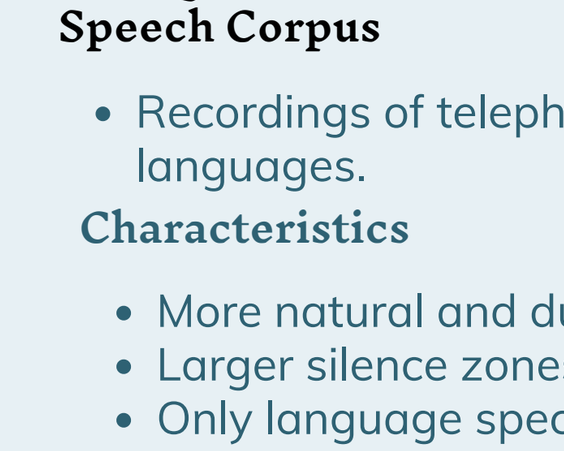


Figure 1A

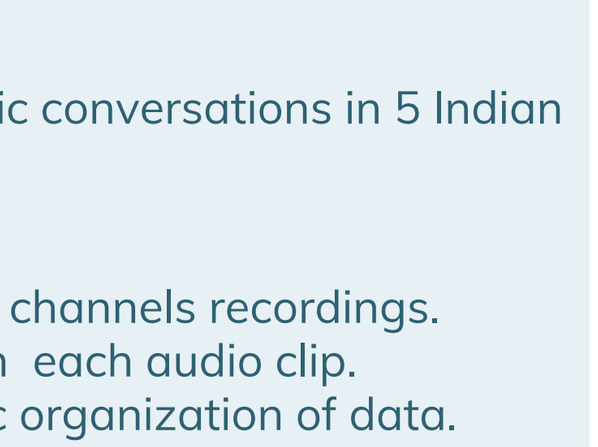


Figure 1B

Figs. 1A and 1B show respectively the amplitude spectrum of a speech segment before and after preprocessing operation

Feature Extraction

- Two sets of features are studied - (i)traditional handcrafted features and (ii) convolutional neural networks (CNN) based features

MFCC, Delta and Delta-Delta Coefficients

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

Mel-Spectrogram

- HANN window based power spectrogram is computed.
- Mel-scale filter banks are applied.
- The Mel-spectrogram is used to feed into Dense CNN architecture.

Other Acoustic and Prosodic features

- Along with 13 MFCCs we have used other prosodic features such as Energy, Energy Entropy, Spectral centroid, Spectral spread,Spectral entropy, Spectral flux and Spectral Rolloff.

Network Architectures

BLSTM Based Architecture

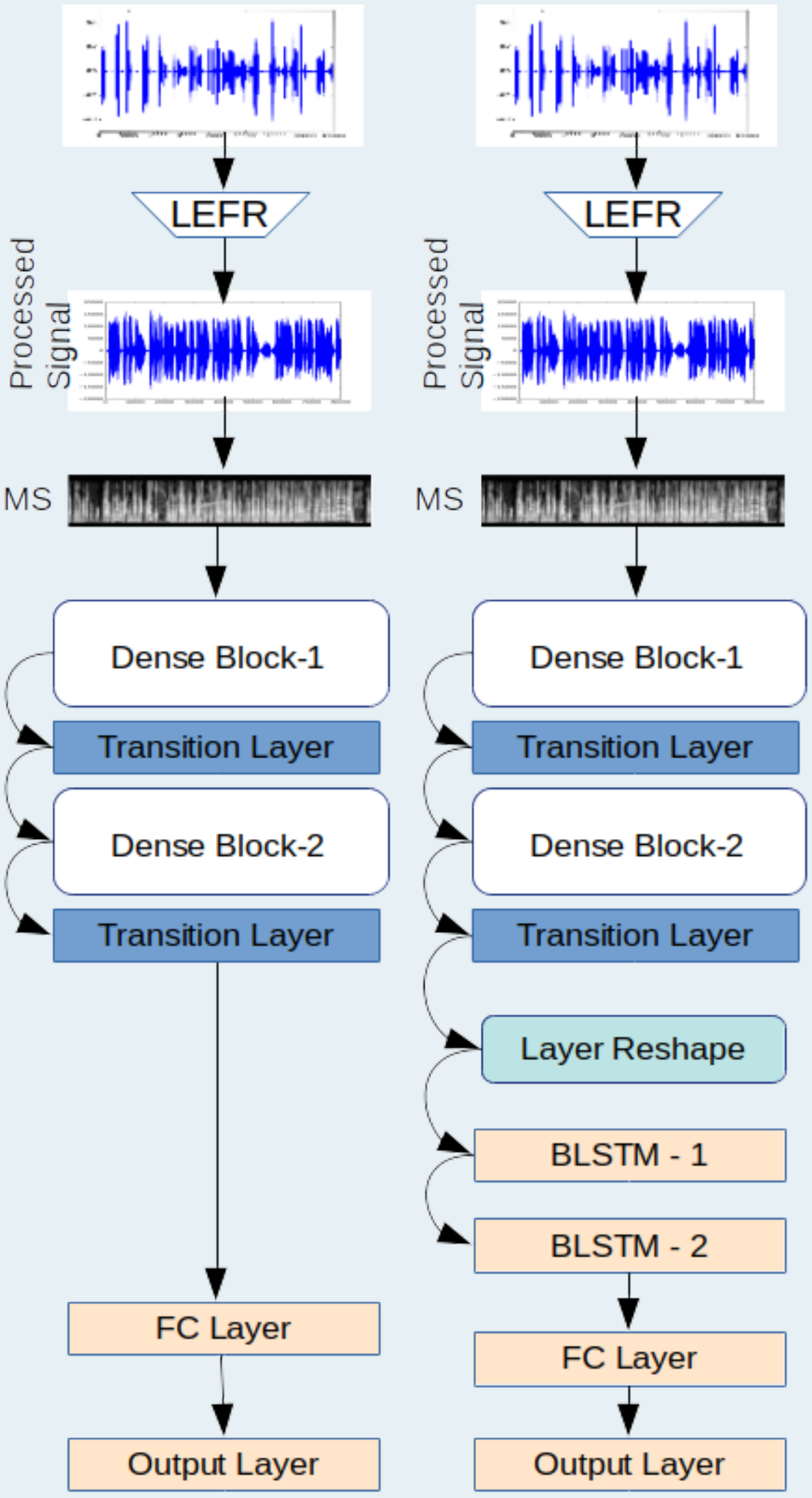
MFCC + Delta + Delta-Delta features (39 F)
Other acoustic and phonetic features (34 F)
Mel-Spectrograms with DenseNet-BLSTM hybrid.
50 ms window strides with 50% overlap generating 399 and 199 frames respectively for 10s and 5s speech segments.
BLSTM network fed separately with two sets of handcrafted features and tested for both the datasets.

DenseNet Based Architecture

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

DenseNet based approach have provided better recognition performance over all other architectures compared in this study.

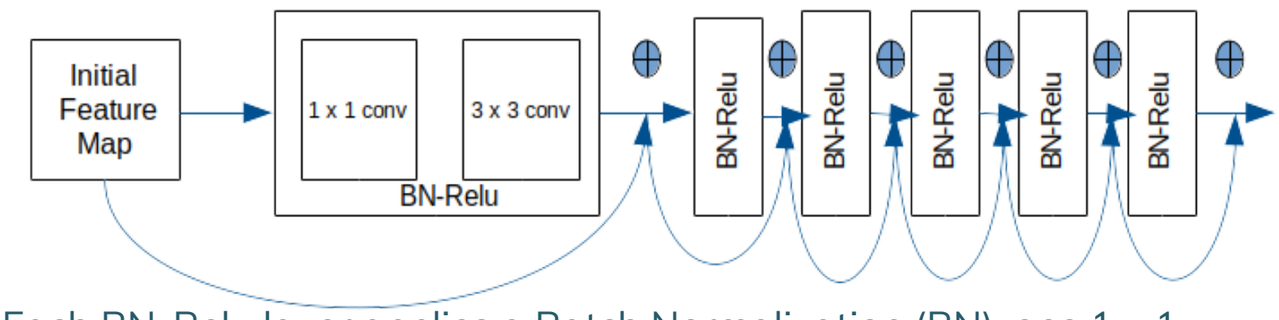
Proposed DenseNet Based Recognition Framework



(a) (b)

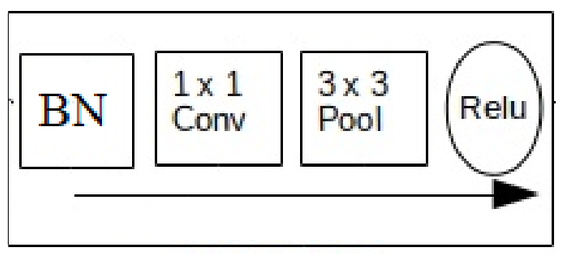
(a) DenseNet based proposed recognition framework. (b) DenseNet-BLSTM hybrid architecture. Mel-spectrogram(MS) of speech signal is fed as input to both the networks (LEFR represents preprocessing operation for removal of low energy frame removal).

Architecture of a Dense Block



Each BN-Relu layer applies a Batch Normalization (BN), one 1 x 1 convolution, one 3 x 3 convolution and a Rectifier Linear Unit (Relu) successively. Input to a BN-Relu layer barring the first is obtained by concatenating the input and output of the preceding BN-Relu layer. The $\oplus$ symbol represents this concatenation operation.

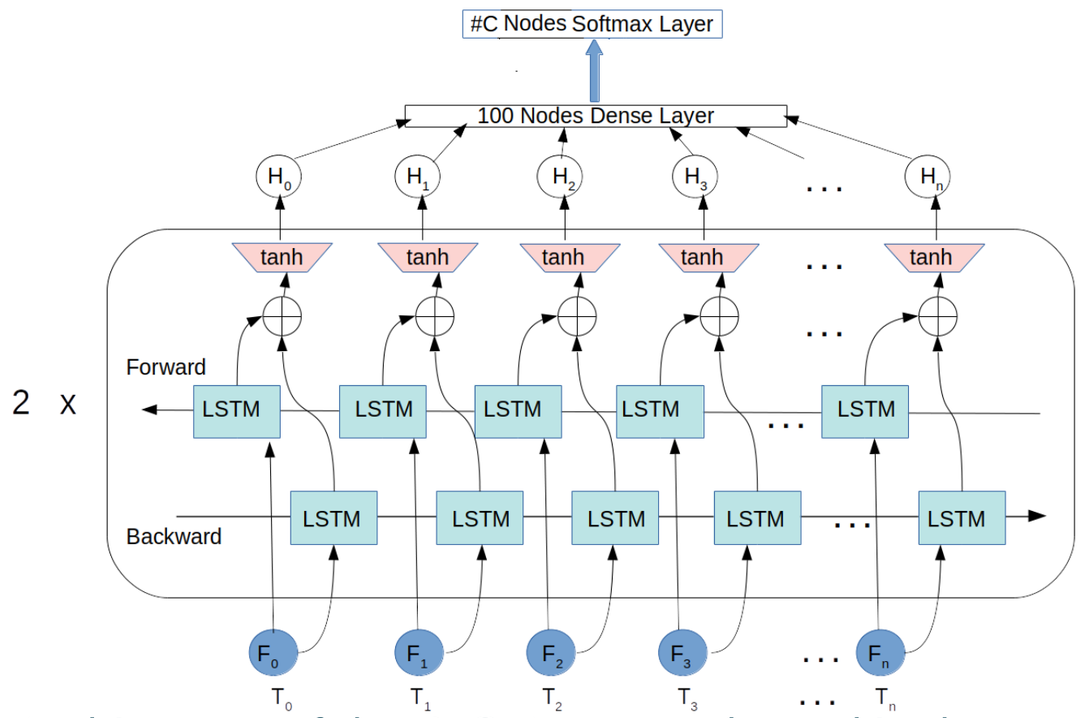
Transition Layer



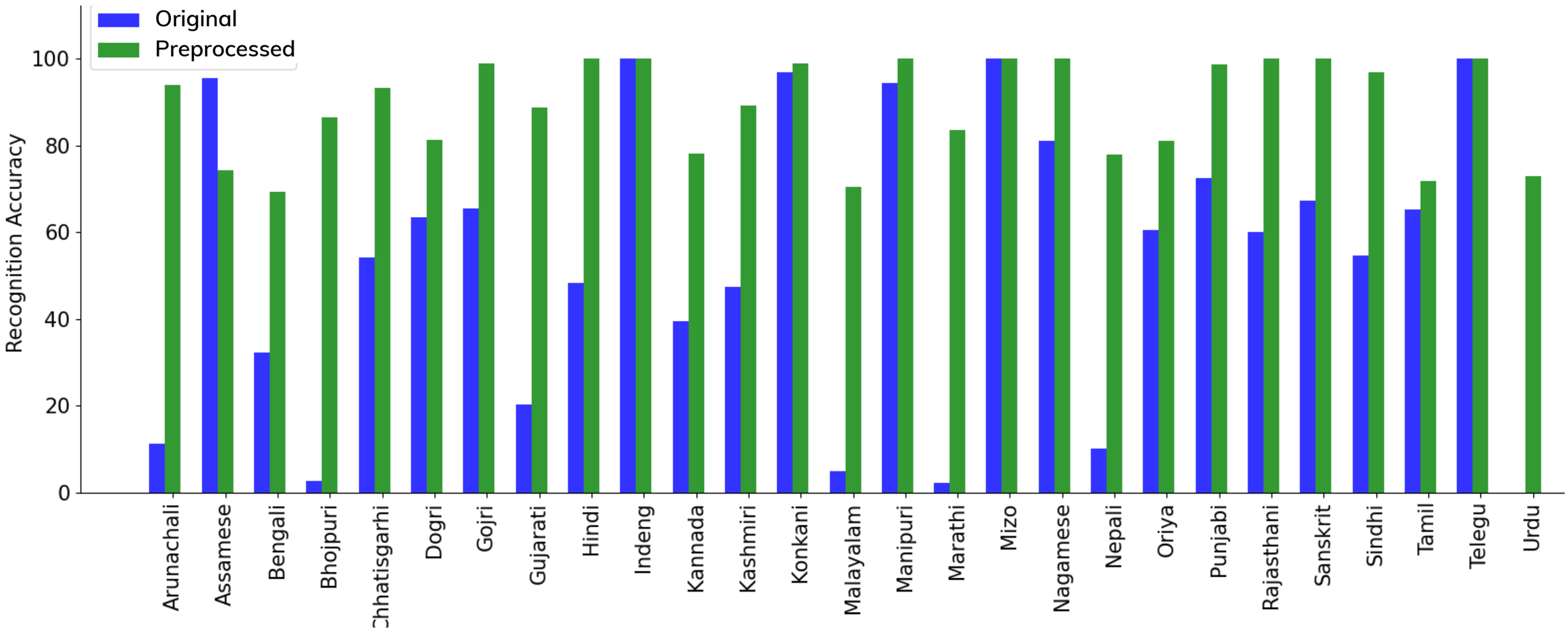
Transition Layer

Each transition layer applies a Batch Normalization (BN), one1×1 convolution, one 3×3 pooling and a Rectifier Linear Unit (Relu) successively.

Architecture of BLSTM network



Architecture of the BLSTM network used in the present study. Handcrafted features are fed as input to thisnetwork. 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.



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

ExperimExperimentation Result (IITKGP-MLILSC Corpus Speaker Dependent)

(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 Result (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 Result (LDC Corpus 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) + ResNet18
5 sec - 83.98 %
10 sec- 91.82 %
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

Conclusion

Proposed DenseNet based approach using Mel-spectrogram features has shown significantly improved language recognition performance over the state-of-the-art LID systems. Experimentation performed on IITKGP-MLILSC and LDC datasets has shown higher misclassification rates within a few groups of two or more phonetically similar languages. On the other hand,recognition accuracy on test samples of IITKGP-MLILSC dataset is higher than the same of LDC dataset. This later observation is justified by the fact that the spoken language samples of LDC dataset consist of real-lifeconversation over noisy telephonic channel whereas the samples of IITKGP-MLILSC dataset consist 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, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell,and K. Keutzer, "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," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 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. arandekar, 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, p.841–844.*
- *T. J. Hazen and V. Zue, "Automatic language identification using a segment-based approach," EUROSPEECH, 1993.*
- *D. Matrouf, M. Adda-Decker, L. Lamel, and J.-L. Gauvain,"Language identification incorporating lexical information." ICSLP, vol. 2, 1998, pp. 181–185.*
- *C. Bartz, T. Herold, H. Yang, and C. Meinel, "Language identification using deep convolutional recurrent neural networks," Int. Conf. on Neural Information Processing. Springer, 2017, pp. 880–889.*
- *S. Shukla, G. Mittalet al., "Spoken language identification using ConvNets," European Conference on Ambient Intelligence. Springer, 2019, pp. 252–265.*