5th International Conference on Inventive Computation Technologies - ICICT 2022

Automatic Speech Recognition for the Nepali Language using CNN, bidirectional LSTM, and ResNet

Organized by
Tribhuvan University, Nepal

Manish Dhakal
Arman Chhetri
Prabin Lamichhane
Aman Kumar Gupta
Suraj Pandey
Prof. Dr. Subarna Shakya

Objective

- Design and train automatic speech recognition (ASR) model that can transcribe the spoken audio to Devnagari texts with fewer errors

$$D_1 D_2 ... D_N = F(\theta, X)$$
$$D_t \in Devanagari\ Token\ Set, \{\text{क},....\text{ज्ञ},..\}$$
$$\theta = Learned\ parameters\ of\ model$$
$$X = Audio\ Features$$

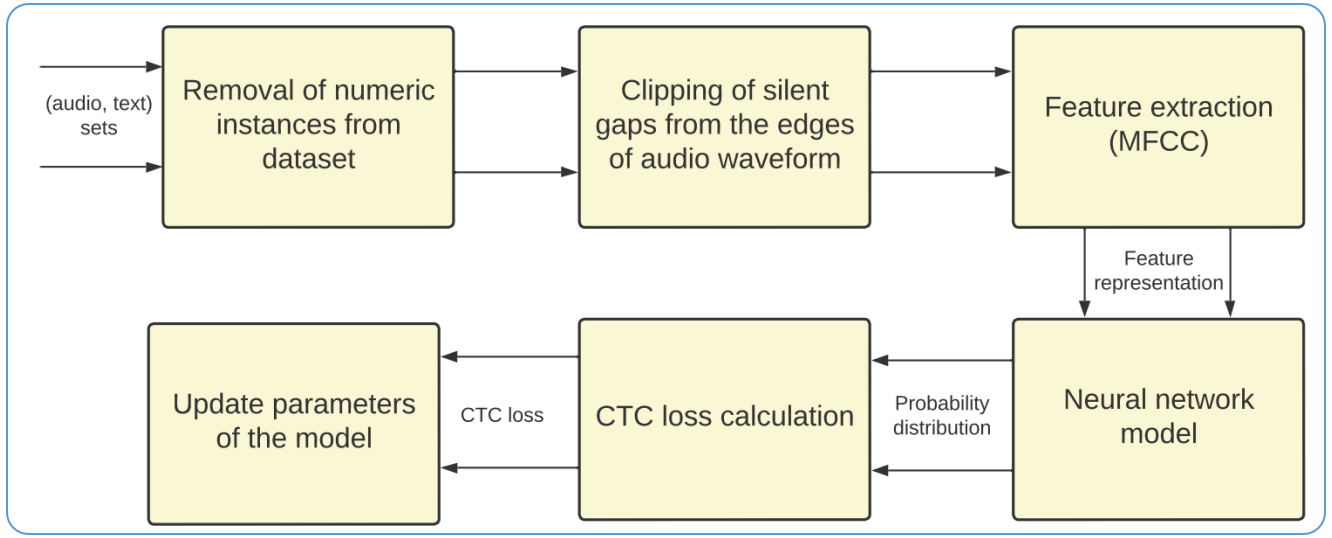
Motivation

- Lack of extensive research in Nepali ASR
- Applications in the field of home automation, banking, education, etc.

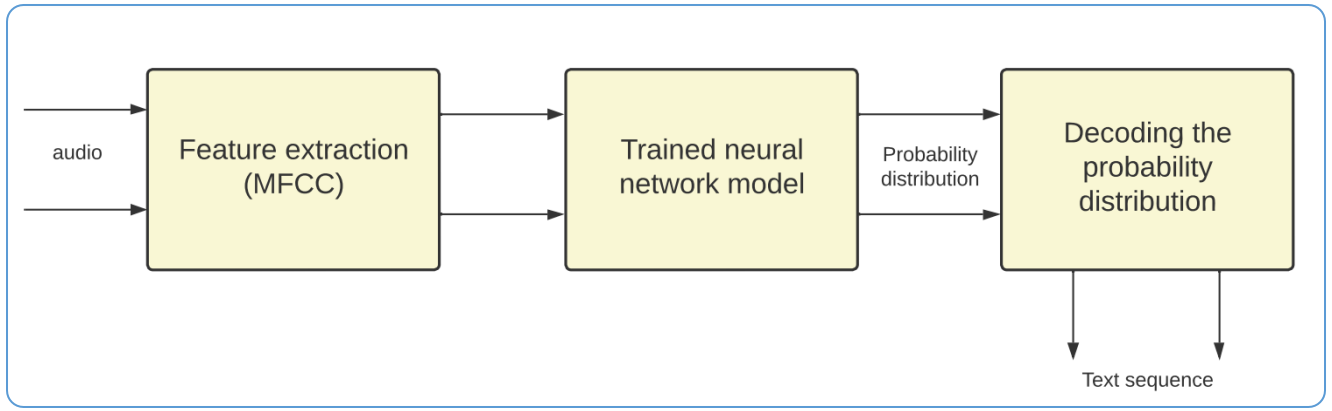
Roadmap

- Flow of methodology
- Dataset
- Architecture of the model
- Experiment
- Analysis

Training flow of the model



Inference flow of the model

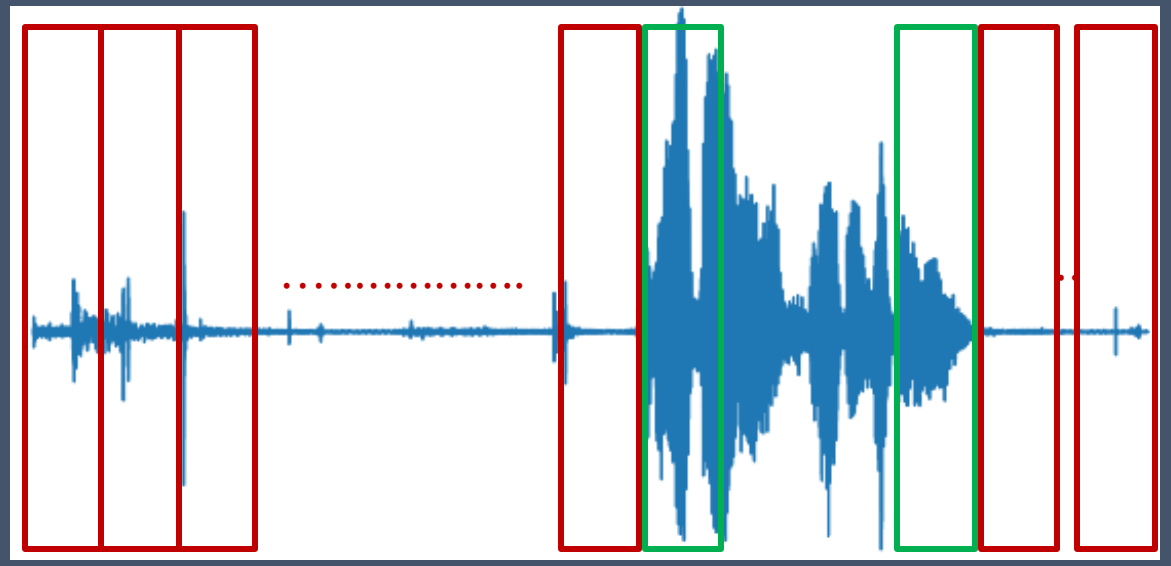


Dataset

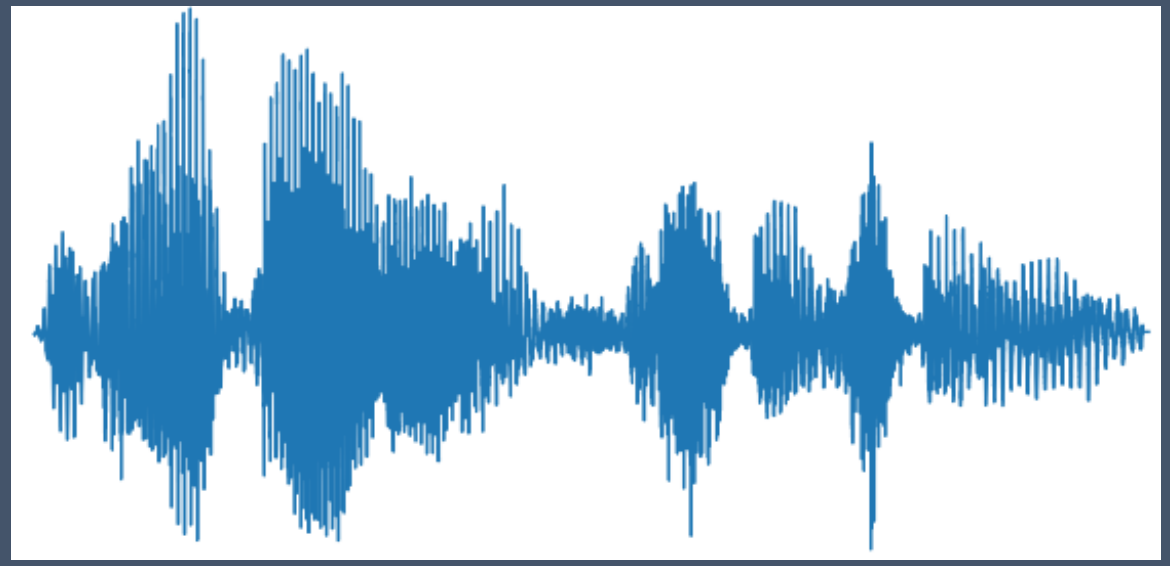
OPENSLR AS DATASET SOURCE

NUMERIC TRANSCRIPTION DISCARDED

Dataset clipping



Audio clip with silent gaps



Audio clip without silent gaps

MFCC as feature extraction

- Mimics the non-linear perception of the sound by human ear
- Discriminative ability to the lower frequencies better than higher ones
- Cosine transform of a log power spectrum on a nonlinear mel scale of frequency
- 13 mel scales for extracting features from human voice

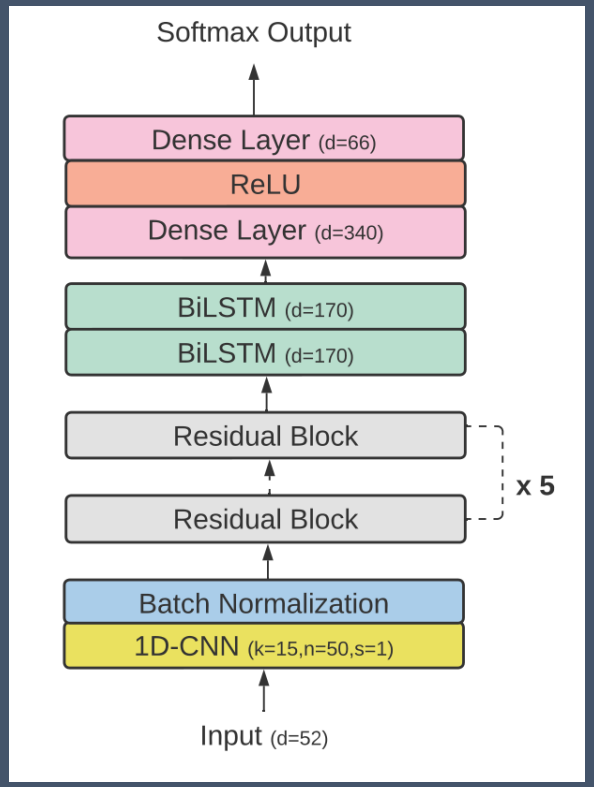
ML Techniques

- CNN
  - Localized features extraction with fewer learnable parameters
- ResNet
  - Shortcut connections in very deep neural network
  - Addresses the problem of larger training error
- RNN (GRU or LSTM)
  - Sequence to sequence mapping between input and output data
  - Preserves the information from past to be used in the current step
  - Bidirectional RNN preserves the contextual information of both future's and past's time step

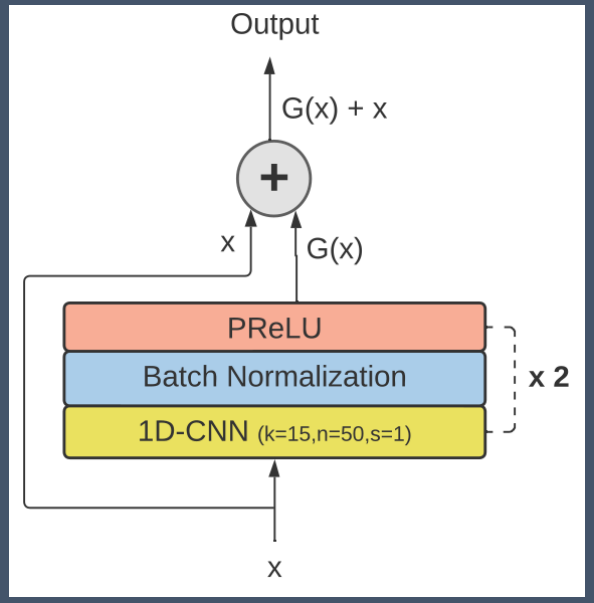
ASR Modelling

- Train multiple models with different combination of the mentioned techniques
- Choose the optimal model based on the evaluation metric of CER

Proposed model architecture



Proposed Optimal Model



Residual Block

CTC loss

- CTC loss for unknown alignment between input audio features and output text
- Alignment-free loss calculation by Introducing the blank token during training

Experimental setup

- Trained (95%) and tested (5%) on the non-numeric OpenSLR dataset
- Adam as the optimization method of the gradient descent
- 20 minutes as the individual epoch training time
- Trained up to 58 epochs in the GPU of the NVIDIA Tesla T4 system.

Evaluation of the models

Model	Test Data CER	# Params
Our trained models		
BiLSTM	19.71%	1.17M
1D-CNN + BiLSTM	24.6%	1.55M
1D-CNN + ResNet + BiGRU	29.6%	1.30M
1D-CNN + ResNet + BiLSTM	**17.06%**	**1.55M**
1D-CNN + ResNet + LSTM	30.27%	0.88M
Other		
1D-CNN + GRU	23.72%	-

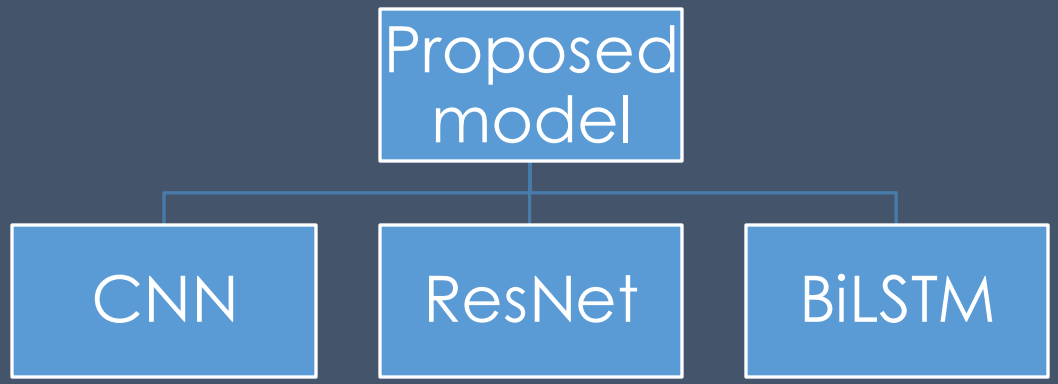
CER = Character Error Rate

Transcriptions from the models

Actual Transcription	Model	Predicted Transcription
मलाई गित गाउन मनपर्छ	BiLSTM	मलाई गीत गाउन भन्पर्छ
	1D-CNN + BiLSTM	मलाई जितगाउन मनुपर्छ
	1D-CNN + ResNet + BiGRU	माल दिनगयाउनु हुन पछ
	1D-CNN + ResNet + BiLSTM	मनाई जीत गाउन मनपछ
	1D-CNN + ResNet + LSTM	मालाई जित ल्ाउनुहुनपर्छ
तिमीलाई ठुलो भए पछि के बन्ने मन छ	BiLSTM	तिमीलाई खुलभएपरछि कय भन्नी भन्छ
	1D-CNN + BiLSTM	तिवीलाईखुनभएपछि को भन्ने मन्छ
	1D-CNN + ResNet + BiGRU	तिमिलाईखलोभयपसित् भन्ने भन्छ
	1D-CNN + ResNet + BiLSTM	तिमीलाई ठुनभएपछि केवनी मन छ
	1D-CNN + ResNet + LSTM	तिमीलाई ठुलभए पछि के मन्ने मन्छ

Summary

- ResNet can solve the problem of early saturation
- Proposed model for ASR is the combination of CNN, ResNet, and BiLSTM



Thank You