1 Nepali Speech Recognition Using CNN, GRU, and CTC

1.1 Introduction

This paper presents an idea to build the Nepali ASR system to convert the spoken Nepali language to its textual representation using a CNN, GRU, and CTC model. The features in the raw audio are extracted by using the MFCC algorithm. MFCC features are a sequence of Acoustic feature vectors where each vector represents information in a small time window of the signal. CNN is used to capture high-level spatial features from the image. The plot of MFCC can be viewed as a transformed intensity of frequencies over time which resembles images, hence CNN can be used to capture high-level features in the spatial domain. GRU is responsible for constructing the acoustic model. The decoding is carried out using a CTC network. The CTC is based on Bayes' decision theory. It receives output from the softmax function.

1.2 Model Architecture

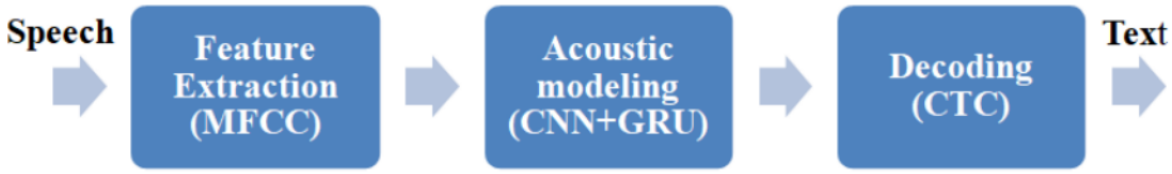

Figure 1: Architecture of proposed ASR system

The experimental setup is carried out on the GPU MX150. For the pre-processing, feature extraction, training, and testing, python and its library have been used. The obtained results from various experiments were as follows:

Experiment	learning rate	batch size	total epochs	WER
1	0.03	100	44	90
2	0.03	300	100	80
3	0.015	50	100	11

Figure 2: Summary of experiments with the results