



School of Informatics, Computing and Engineering

ENGR-E533: Project Presentation

Speech Recognition using Deep Neural Network

Taslima Akter, Khandokar Md. Nayem

SECTION 1

Overview

Speech Recognition

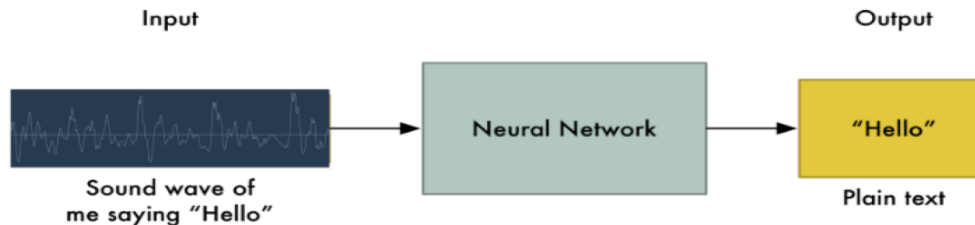
1. Automatic **recognition** and **translation** of spoken language into text.



2. For training, use extensive datasets which are not free.
3. Struggles in noisy environments, in recognizing accented speech, or speaking styles and languages (limited training data available).

Our Speech Recognition

1. Simple idea, feed sound recordings into a neural network and train it to produce text (architecture: Deep Speech).



2. Big problem: speech varies in **speed**.
3. Need to **align** audio files of various lengths to a fixed-length piece of text.

Corpus

1. Trained on **Timit** corpus (6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major US dialect regions).
2. For robustness, synthetically increased the volume of the dataset.
 - 10 types of noises (Babble, Cafe, Car, Factory, Machine gun, Plane, Restaurant, SSN, Tank, White Noise)
 - 5 Mixture levels (-3dB, 0dB, 3dB, 6dB, 9dB)
3. Total dataset size, $6300 \times 10 \times 5 = 3,15,000$



Feature Extraction

1. Python **LibROSA** feature extraction methods are used.
 - **Spectrogram**: Time-Frequency representation of the speech signal.
 - **Mel-frequency cepstral coefficients (MFCC)** : The coefficient that collectively make up the short term power spectrum of a sound.
2. Currently using **MFCC** feature as baseline performance.
3. No manual feature extraction (simply **Spectrogram**) is the ultimate goal.



SECTION 2

Methodology

Connectionist Temporal Classification (CTS)

1. One more unit than there are labels in L .
2. The activations of the first $|L|$ units are interpreted as the probabilities of observing the corresponding labels at particular times.
3. The activation of the extra unit is the probability of observing a **<blank>**, or no label.
4. Together, these outputs define the probabilities of all possible ways of aligning all possible label sequences with the input sequence.

$$p(\pi|\mathbf{x}) = \prod_{t=1}^T y_{\pi_t}^t, \bigvee \pi \in L'^T, L' = L \cup \{blank\} \quad \text{where, } \mathbf{x} \text{ is input sequence and } \pi \text{ is label sequence.}$$



Connectionist Temporal Classification (CTS)

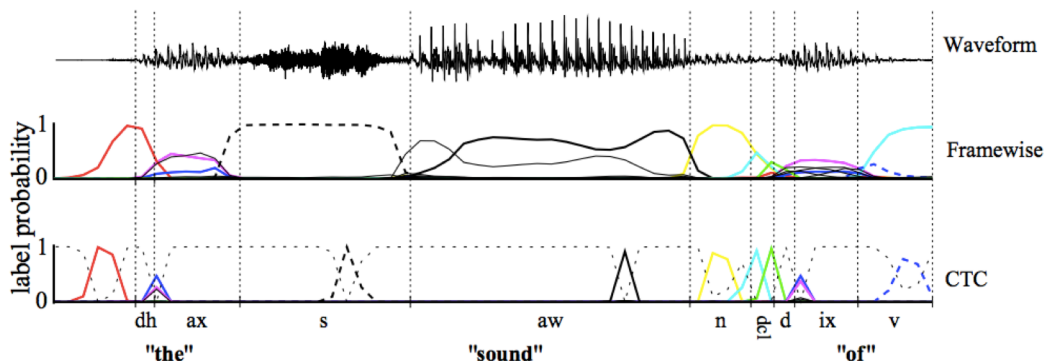


Figure 1. Frame-wise and CTC networks classifying a speech signal. The shaded lines are the output activations, corresponding to the probabilities of observing phonemes at particular times. The CTC network predicts only the sequence of phonemes (typically as a series of spikes, separated by 'blanks', or null predictions), while the frame-wise network attempts to align them with the manual segmentation (vertical lines). The frame-wise network receives an error for misaligning the segment boundaries, even if it predicts the correct phoneme (e.g. 'dh'). When one phoneme always occurs beside another (e.g. the closure 'dɪl' with the stop 'd'), CTC tends to predict them together in a double spike. The choice of labelling can be read directly from the CTC outputs (follow the spikes), whereas the predictions of the frame-wise network must be post-processed before use.

Deep Speech Model

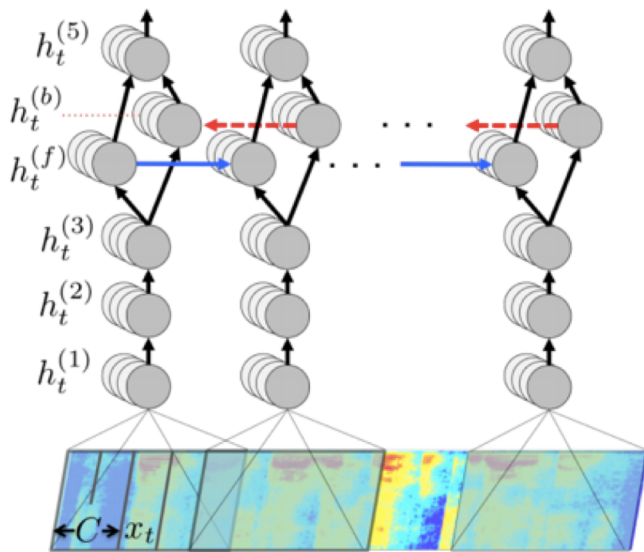


Figure : Structure of Deep Speech model and notation.

Layer h^6 : Softmax output layer

Layer h^5 : $h_t^{(5)} = g(W^{(5)}h_t^{(4)} + b^{(5)})$

Layer h^4 : Bi-directional RNN

Layer h^2, h^3 : $h_t^{(l)} = g(W^{(l)}h_t^{(l-1)} + b^{(l)})$

Layer h^1 : spectrogram frame x^t with context C

Here, $g()$ is a clipped ReLu, $C \in \{5, 7, 9\}$

Our Model (CRNN)

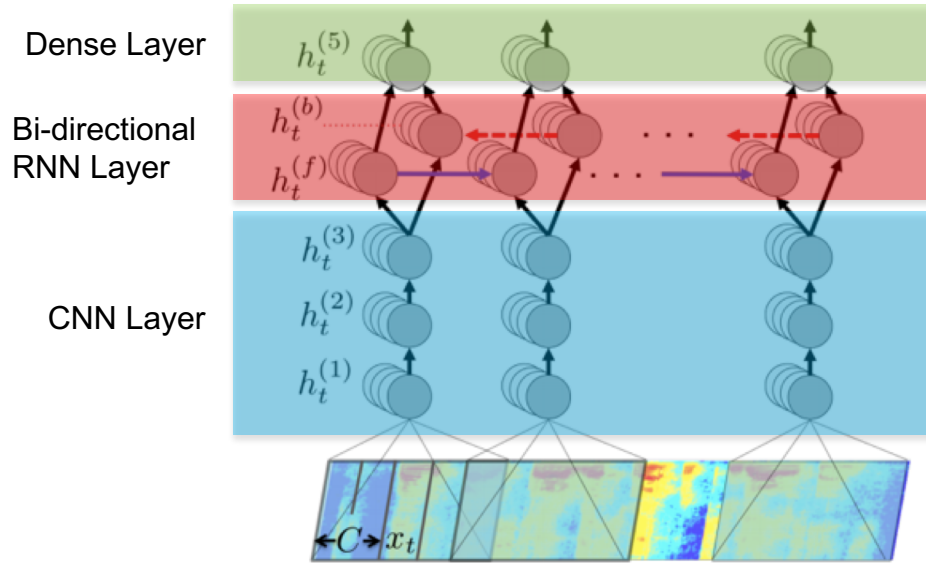


Figure : Structure of DeepSpeech model and notation.

Layer h^6 : Softmax output layer

$$\text{Layer } h^5 : h_t^{(5)} = g(W^{(5)}h_t^{(4)} + b^{(5)})$$

Layer h^4 : Bi-directional RNN

$$\text{Layer } h^2, h^3 : h_t^{(l)} = g(W^{(l)}h_t^{(l-1)} + b^{(l)})$$

Layer h^1 : spectrogram frame x^t with context C

Here, $g()$ is a clipped ReLu, $C \in \{5, 7, 9\}$

Postprocessing (Language Model)

RNN output	Decoded Transcription
what is the weather like in bostin right now prime miniter nerenr modi arther n tickets for the game	what is the weather like in boston right now prime minister narendra modi are there any tickets for the game

Table 1: Examples of transcriptions directly from the RNN (left) with errors that are fixed by addition of a language model (right).

Thank You!



INDIANA UNIVERSITY BLOOMINGTON
FULFILLING *the* PROMISE