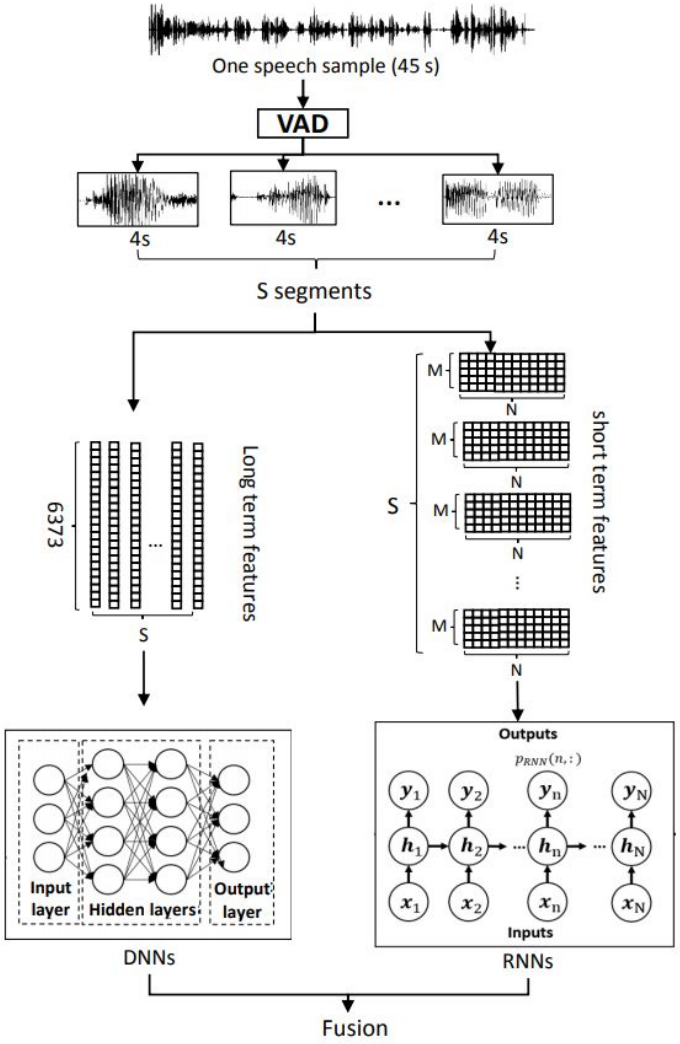
CS753: Automatic Speech Recognition

Course Project: Accent Recognition

By : Rishabh Ramteke
Anshul Tomar



Architecture

Feature Selection

There are two types of features involved in this project:-

1)Long term features : - Extracted via openSmile
2)Short term features:- Extracted by python script

Long term feature selection

These features represent the overall statistics of the wav file.
There are 6373 features overall. These include features like the power of the signal and its uptime.

As there are 2466 audio files (each of length 4s) the total number of long term features are 2466*6373

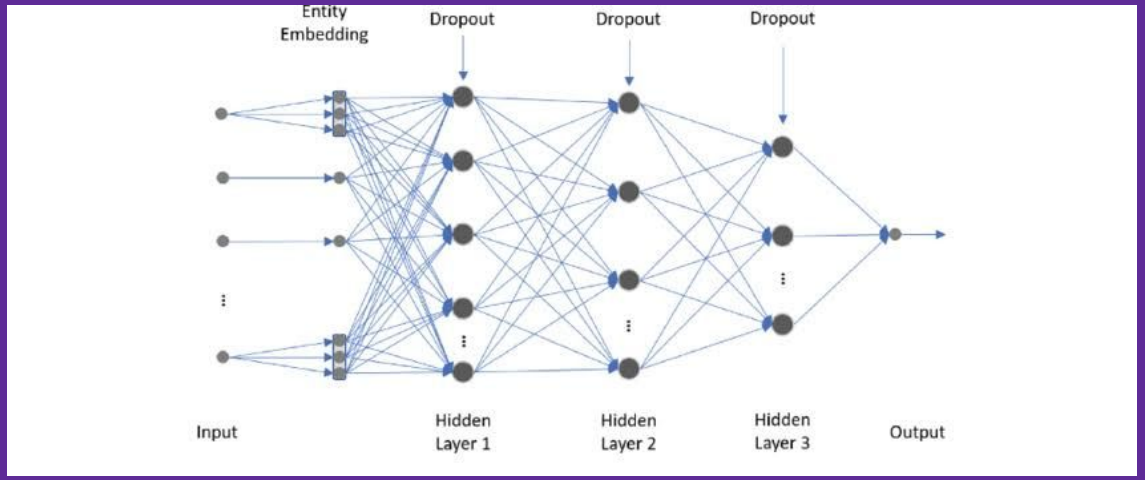
Short term feature selection

These are the regular MFCC features calculated on 25ms samples and with frame gaps of 10ms. Hence there are 4/0.01 -1 = 399 vectors each of size 39.

However training time of the RNN's model with these feature vectors was around 29 hours for 1 epoch, hence we decided to keep only the 12 cepstral coefficients + 1 power coefficient.(Hence number of input features for RNN are (399*2466)*13

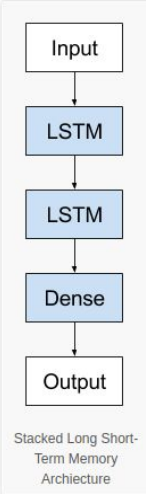
Deep Neural Network Model

3 hidden layers with 256 nodes each



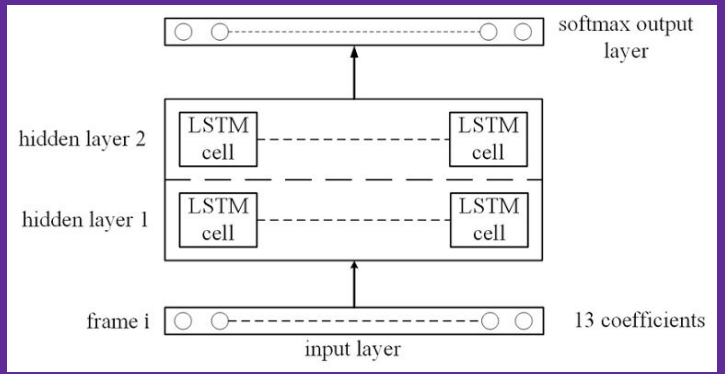
Accuracy of the DNN network

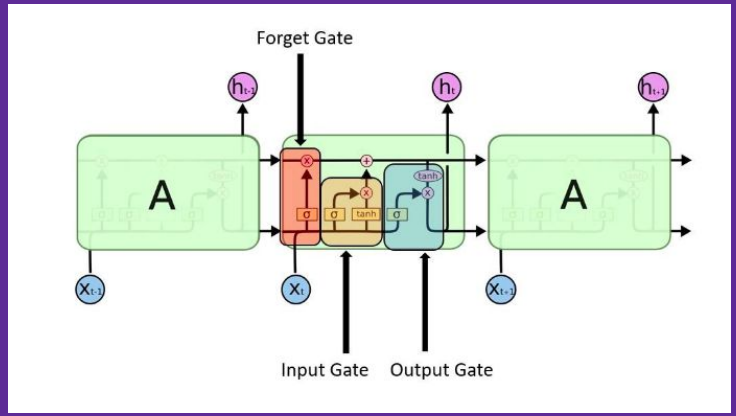
Validation Split	Epochs	Train Accuracy(Avg)	Validation Accuracy(Avg)
0.2	50	53.12%	52.50%
0.2	100	70.56%	53.3%
0.1	100	73.56%	59.00%



Recurrent Neural Network and LSTM Model

2 LSTM layers with 512 nodes each.

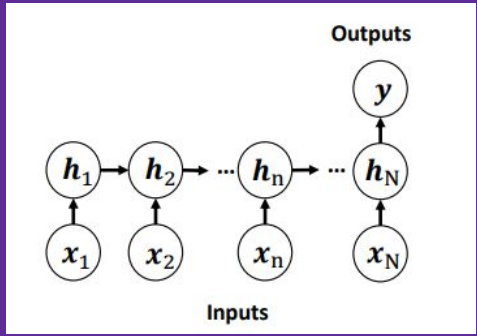




Combination of both models

$$w_{DNN} = \frac{Acc_{DNN}}{Acc_{DNN}+Acc_{RNN}}$$
$$w_{RNN} = 1 - w_{DNN}$$

$$P(j) = \frac{1}{S}\sum_{i=1}^{S}[w_{\mathrm{DNN}}P_{\mathrm{DNN}}(i,j) + w_{\mathrm{RNN}}P_{\mathrm{RNN}}(i,j)].$$



Many-to-one RNN structure used in the method of DNN with RNN(on sequence).

Results and Comparison

Difference between original model and ours : The authors consider 11 classes while we considered only 5. Both the DNN and RNN were trained with the Python neural networks library, Keras

	RNN only	DNN only	DNN with RNN
Original Paper	42.9 %	49.1%	50.2%
Our implementation	55%*	53.3%	54.6%

*The accuracy represented here is on an RNN trained only for 1 epoch and with greatly reduced number of inputs features because it was taking a lot of time.

Reference

Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features

Yishan Jiao , Ming Tu , Visar Berisha , Julie Liss

Thank You