Deep Learning based multi-task speech classification of gender and accent

Sai Satya Vamsi Karthik Bhamidipati
200420608

Abstract

This paper presents how different deep learning based techniques work for multi-task speech classification of speaker characteristics such as gender and accent. In the training phase, the models are trained to classify both gender and accent of an english speaker using multi-task learning. The dataset used for training and evaluation is the Common voice corpus [1]. Experiments have shown a few models achieving high performance. The code implemented has been released under an open-source license and is available in GitHub.[1]

1 Introduction

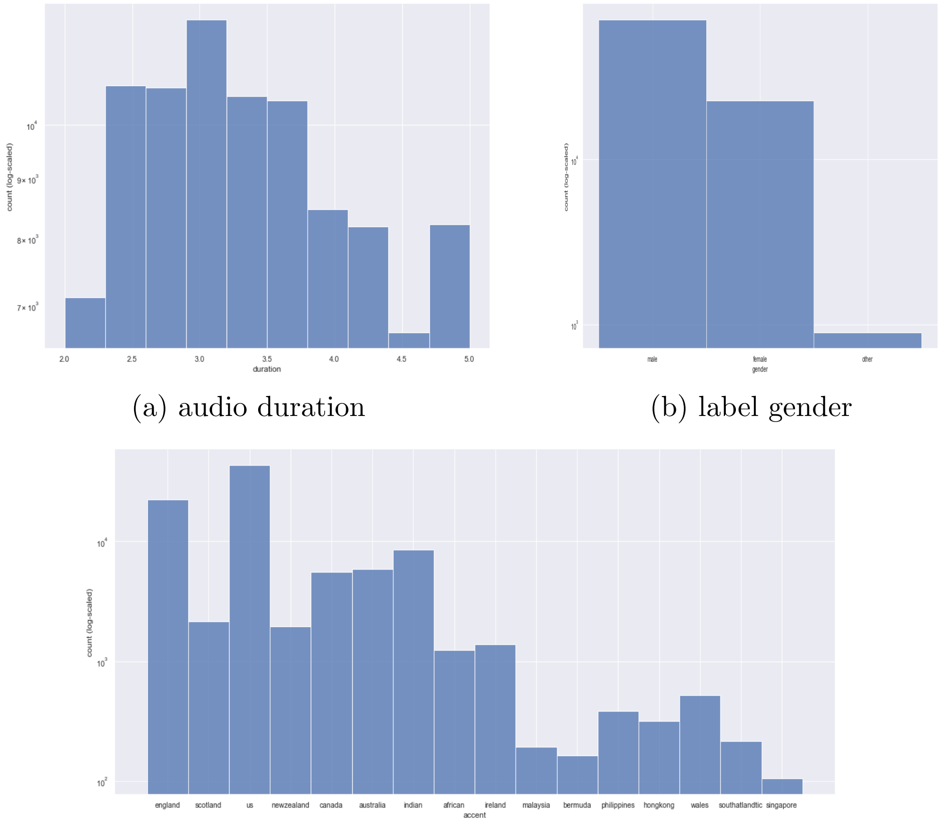
Speech is considered to be one of the most important part of human communication. Humans are extremely good at identifying characteristics like accent, gender in a speakers' speech. Although it's very easy for humans, it has traditionally been a challenge for speech processing systems. Identifying speaker characteristics is a important task in several speech related applications like E-Language learning system between U.S. Department of Education and the Chinese Ministry of Education [2]. In this paper, I have presented different deep learning approaches like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for multi-task speech classification of gender and speech of english speakers. These models have been trained and evaluated against the Common voice corpus [1].

[1] https://github.com/karthikbhamidipati/multi-task-speech-classification

2 Dataset

The Common voice corpus is a massively multilingual collection of transcribed speech [1], that is intended for speech technology research and development. It is created to make the speech data open and decentralized. It is multilingual consisting of 17 languages. However for this paper, speech data consisting of only english speakers has been considered. It consists of around 380k audio files and metadata with fields such as audio transcription, up votes, down votes, age, gender, accent. For this paper, only gender and accent are used as the labels to train the model.

Figure 2.1: Data Distributions



(a) audio duration

(b) label gender

(c) label accent

As part of pre-processing, the following steps were applied. 1. The audio files with no annotations of gender or accent have been filtered out. This reduced the number of audio files to around 125k. 2. The audio files with duration less than 2 seconds and greater than 5 seconds were also filtered out. This reduced the number of audio files to around 95k. After the preprocessing, the reduced audio data and the average duration are around 90 hours and 3.5 seconds respectively as shown in the Figure 2.1a.

The data distributions[2] of gender and accent shown in Figure 2.1b and Figure 2.1c respectively suggest that there's a class imbalance in both the annotated labels. The number of classes for accent and gender are 16 and 3 respectively. To account for the class imbalance, the data is split using stratified sampling into 65% training data, 15% validation data and 20% test data.

The Mel Frequency Cepstral Coefficents (MFCCs) were extracted from the audio with a sampling rate of 22.05 kHz, window length of 1024, and hop length of 512, number of mfcc filters of 64. Additionally the MFCC is transposed and padding is applied along the time dimension resulting in an output tensor of shape (256, 64). MFCCs were chosen for audio preprocessing because it takes into account human perception for sensitivity at appropriate frequencies which is very useful in applications dealing with speech. The different deep learning methods shown later in Section 3 will train using MFCC outputs as the input.

3 Models

The audio representations (MFCCs) have been used to train and predict the user characteristics such as gender and accent using multi-task learning. The architecture consists of two parts: 1. A feature extractor such as a CNN (Section 3.1) or an RNN (Section 3.2) is used to extract linear 1-dimensional features of length 512 from the 2-dimensional MFCC output. 2. The linear features from the feature extractor is then used by two different sub-networks of dense layers for classifying gender and accent [3].

As shown in Figure 3.1, there's a hard parameter sharing of the feature extractor network between the gender classification network and the accent classification network. So, the output of the feature extractor will be a shared representation of both accent and gender. This helps the model to learn simultaneously to classify both the gender and the accent. The gender

[2] counts in histogram have been logarithmically scaled

classification network consists of a Dropout layer with a probability of 0.1, and a fully connected layer with number of input nodes, output nodes as 512 and 3 respectively. The accent classification network is similar to the gender classification network with the only difference being the number of output nodes in the fully connected layer equal to 16. The dropout is used here to avoid overfitting.

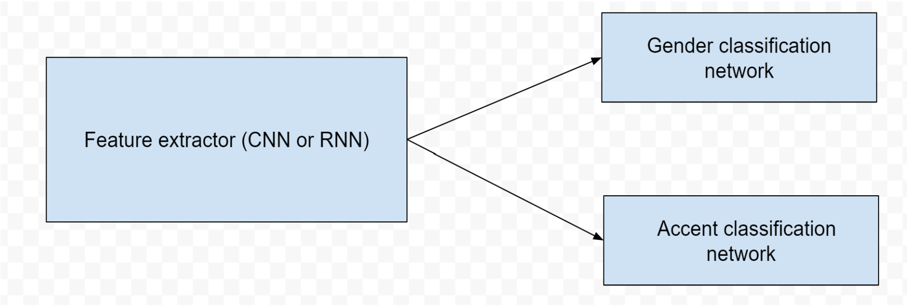

Feature extractor (CNN or RNN)
Gender classification network
Accent classification network

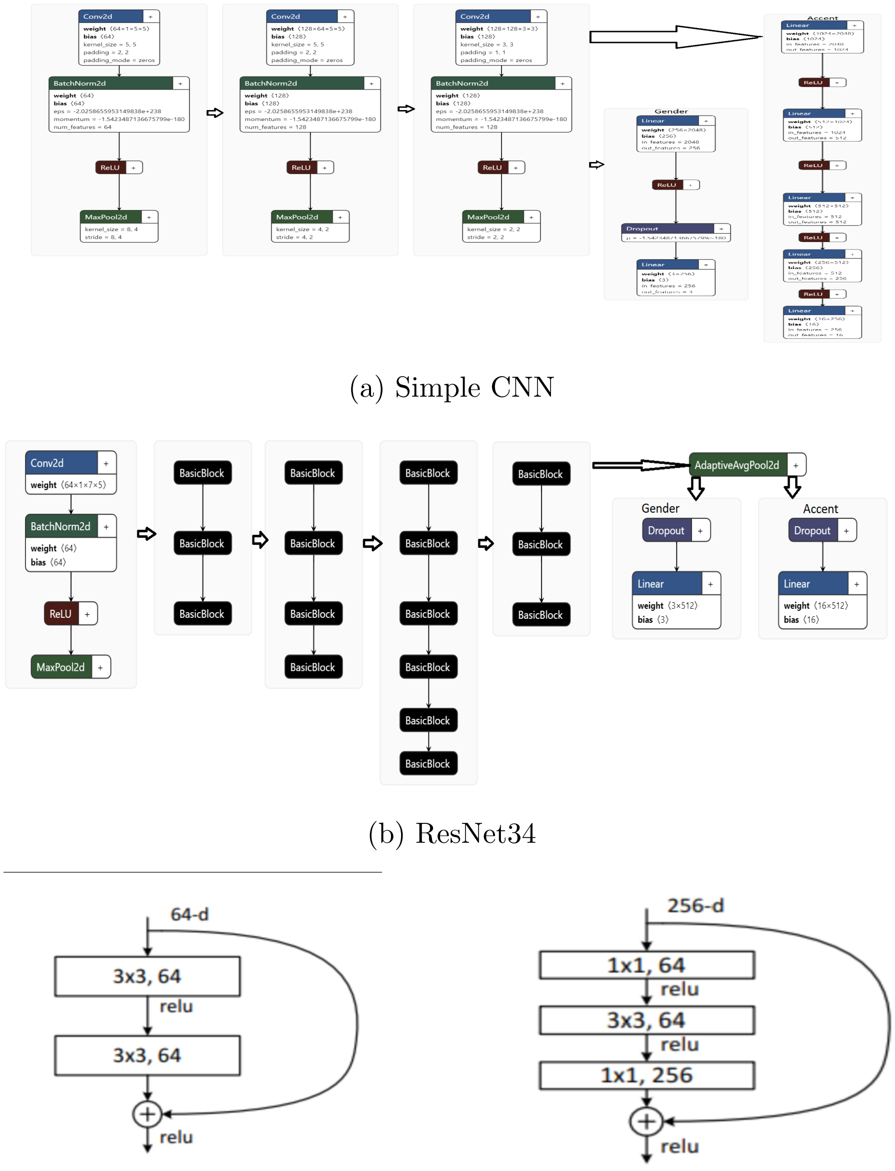
Figure 3.1: High level model architecture

3.1 Convolutional Neural Network (CNN)

Convolutional neural networks exhibit shift-invariance because of their shared-weight architecture represented as kernel filters. This could be very useful in detecting the underlying repeated patterns in an audio representation. Two different types of CNN models have been used for extracting the shared representation.

1. Simple CNN model with 3 CNN blocks with each block consisting of a Conv2d layer, BatchNorm2d, ReLU activation, MaxPool2d. The first block had a convolution layer with 64 out filters, kernel size of 5, padding of 2, and a maxpooling layer with kernel size (8, 4), stride (8, 4). The second block had a convolution layer with 128 out filters, kernel size of 5, padding of 2, and a maxpooling layer with kernel size (4, 2), stride (4, 2). The final block had a convolution layer with 128 out filters, kernel size of 3, padding of 1, and a maxpooling layer with kernel size 2, stride 2. This resulted in 2048 number of flattened features which are then passed onto two different fully-connected networks for gender and accent as shown in Figure 3.2a. The CNN blocks will be able to extract a dense feature representation from the audio representation which can be very useful for the multi-task classification.

Figure 3.2: Architecture of CNN Models



(a) Simple CNN

(b) ResNet34

(c) Basic block [4]

(d) Bottleneck [4]

2. ResNet model with the updated first convolution block and the final fully connected layers as per following. The first convolutional block

has been replaced with a Conv2d layer with 64 output filters, kernel size of (7, 5), stride of (2, 1), padding of (3, 2), BatchNorm2d layer, ReLU activation, MaxPool2d layer with kernel size of (5, 3), stride of (2, 1), padding of (2, 1). The final fully connected layer has been replaced with two separate fully connected networks for gender and accent respectively as shown in Figure 3.2b. The blocks inside the ResNet architecture that consists of skip connections i.e., Basic blocks or Bottleneck blocks were unmodified as shown in Figures 3.2b, 3.2c, 3.2d. ResNet18, ResNet34, ResNet50 with the above mentioned modifications. ResNet allowed the possibility of very deep neural networks without the vanishing gradient problem [4]. Since, deeper networks can learn more relationships between features, ResNet have been experimented with.

3.2 Recurrent Neural Network (RNN)

Since Recurrent Neural networks can identify and learn temporal sequences, Long Short Term Memory (LSTM) layers with different configurations have been experimented with.

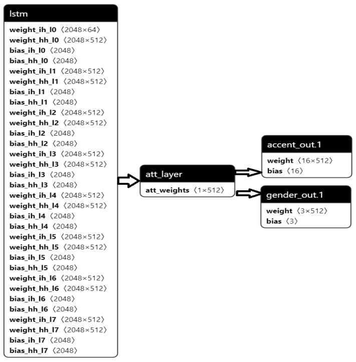
1. A Simple LSTM based model with an LSTM layer having hidden size 512 and 8 stacked layers, average pooling layer, and two separate fully connected networks containing a dropout layer with probability 0.1 and Dense layer for gender and accent respectively.

2. A Bidirectional LSTM based model with a Bidirectional LSTM layer having hidden size 512 and 8 stacked layers, average pooling layer, and two separate fully connected networks containing a dropout layer with probability 0.1 and Dense layer for gender and accent respectively.

3. A LSTM with attention based model with an LSTM layer having hidden size 512 and 8 stacked layers, attention layer, and two separate fully connected networks containing a dropout layer with probability 0.1 and Dense layer for gender and accent respectively.

4. A Bidirectional LSTM with attention based model with a Bidirectional LSTM layer having hidden size 512 and 8 stacked layers, attention layer, and two separate fully connected networks containing a dropout layer with probability 0.1 and Dense layer for gender and accent respectively.

Figure 3.3: Architecture of LSTM Model



4 Training

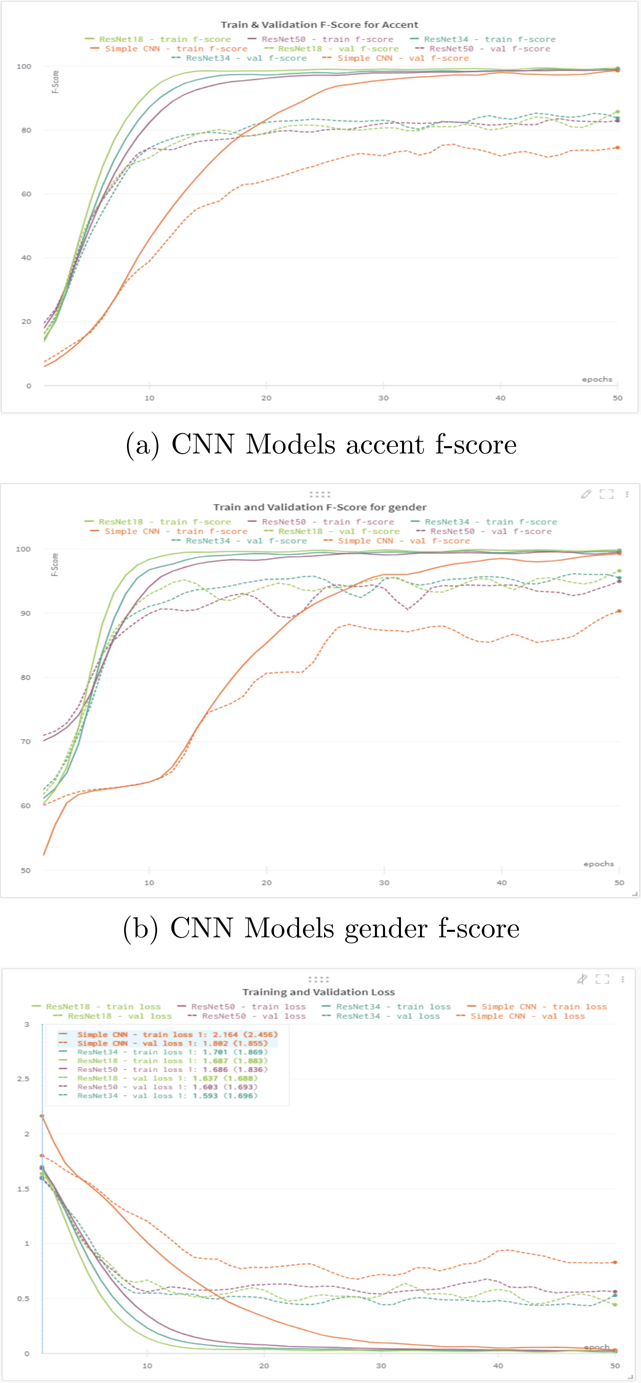
The training configuration of the model is as follows: 1. The loss function is the weighted sum of categorical cross-entropy loss on gender and accent respectively. The weights assigned for the gender and accent loss are 1 and 1 respectively. 2. Adam optimizer is used with a learning rate 0.001. 3. The batch size is 512. 4. Number of epochs are 50. 5. The models are trained using Pytorch's distributed data parallel framework on 4 RTX 2070 GPUs and an Intel Xeon E5 CPU with 64GB ram.

5 Experimental Results

5.1 CNN models

As presented in the Figures 5.1a, 5.1b, 5.1c, 1. All the CNN models converged close to 100 percent F-Score in both accent and gender by 50 epochs during training as shown in Table 1. 2. ResNet18 had the best validation score of 87% and 97% for accent and gender labels respectively amongst all the models as shown in Table 1. 3. With a deep network like ResNet, the improvement in F-Score was around 10% in both accent and gender and has helped the model achieve really high F-Score in validation as shown in Table 1. 4. ResNet18 was the best amongst all the CNN based models as shown in Table 1.

Figure 5.1: CNN Experimental results



(a) CNN Models accent f-score

(b) CNN Models gender f-score

(c) CNN Models loss

Table 1: Validation F-Score for the trained models

Model	**Trainable Params**	**Accent**	**Gender**
ResNet18	11.2M	87%	97%
ResNet34	21.3M	86%	97%
ResNet50	23.5M	86%	96%
Simple CNN	3.9M	79%	90%
BiLSTM + Attention	46.5M	65%	85%
BiLSTM	46.5M	35%	77%
LSTM + Attention	15.9M	6%	31%
LSTM	15.9M	5%	30%

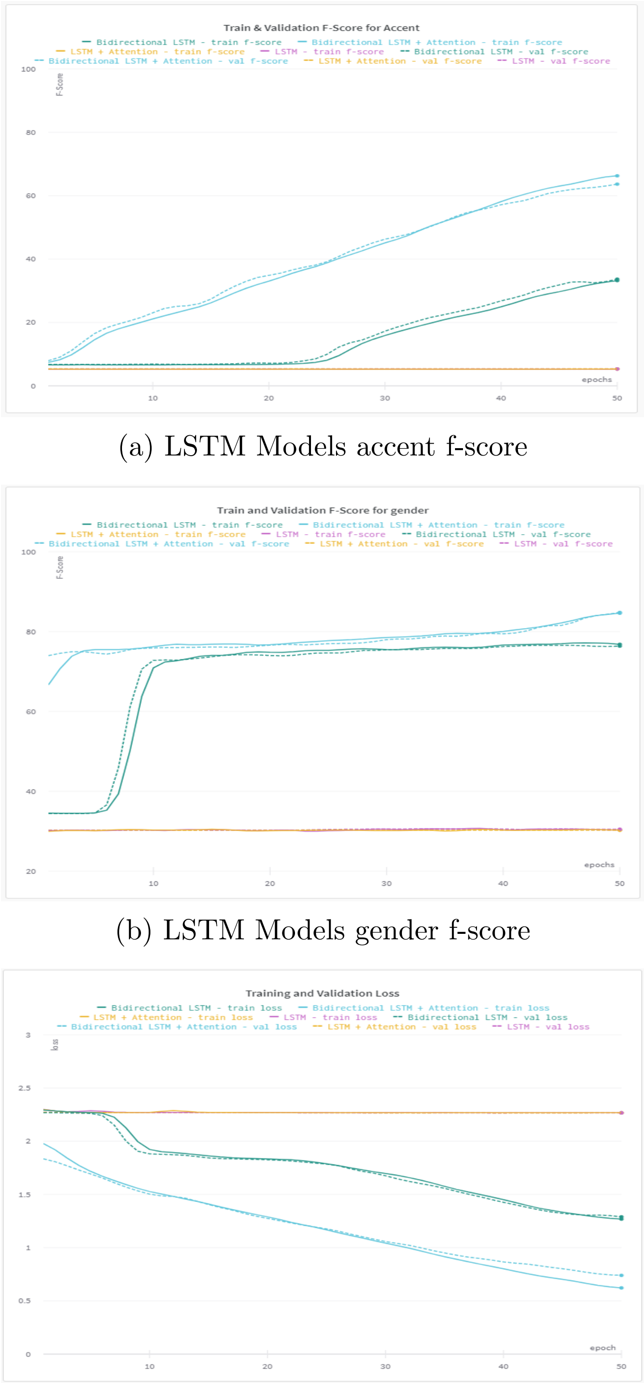
5.2 RNN Models

As presented in Figures 5.2a, 5.2b, 5.2c, 1. The LSTM models didn't improve at all even with attention. The F-Score was stuck at 30 % and 5% for gender and accent respectively. This could be because of the multi-task learning setup of the model. 2. The BiLSTM models were able to learn and improve upon the original LSTM models. They reached an F-Score of 77% and 35% for gender and accent respectively as shown in Table 1. 3. Attention helped the BiLSTM models even more by improving the F-Scores by 30% and 18% for accent and gender respectively as shown in Table 1. 4. BiLSTM with Attention was the best model amongst the RNN based models as shown in Table 1.

5.3 Comparison

Comparing the CNN and RNN based models as per Table 1, we can observe that all the CNN models are outperforming the best of RNN models. ResNet18 seems to be best model amongst all the models. It achieved the highest validation f-score for both accent and gender. ResNet34, ResNet50 are closer to ResNet18 in terms of validation f-score, but ResNet18 was chosen as the best model because it's achieving higher performance with lower number of parameters. The ResNet model achieved an F-Score of 95% and 84% on the test set for gender and accent respectively as shown in Table 2. From the confusion matrix shown below for gender and accent in Figures 5.3a 5.3b, it can be inferred that the model is performing well for the minority classes as well.

Figure 5.2: LSTM Experimental results



(a) LSTM Models accent f-score

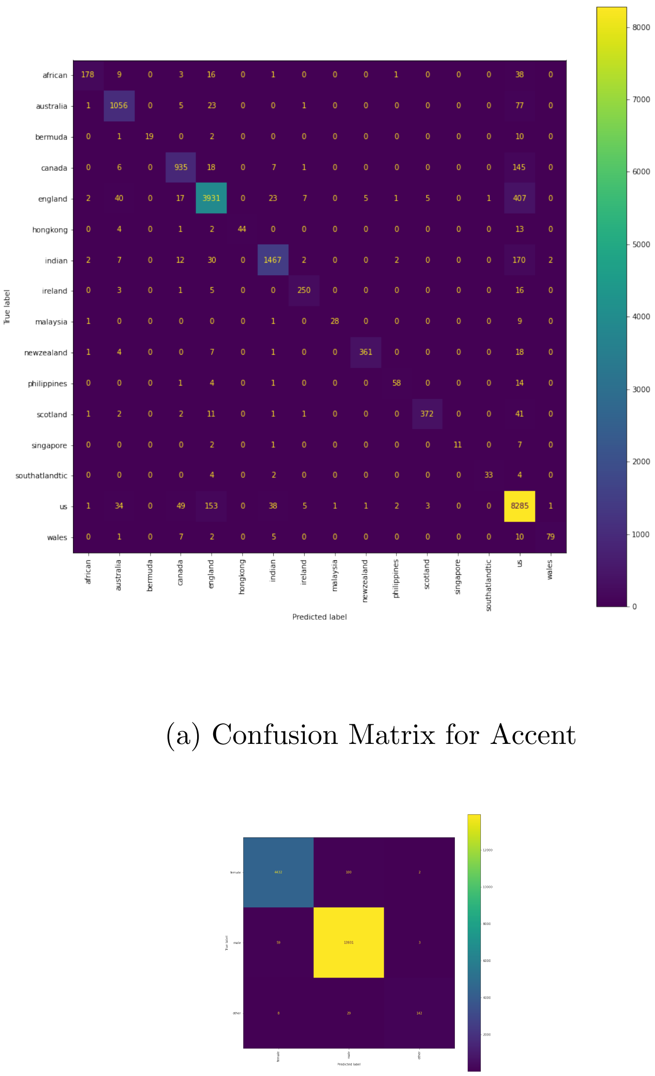
(b) LSTM Models gender f-score

(c) LSTM Models loss

Table 2: Test F-Score for the best model

Model	**Loss**	**Accent**	**Gender**
ResNet18	0.49	84%	95%

Figure 5.3: Testing confusion matrix



(a) Confusion Matrix for Accent

(b) Confusion Matrix for Gender

Table 3: Inference on sample files

filename	**gender**	**gender (pred)**	**accent**	**accent (pred)**
sample-000001.mp3	male	male	scotland	australia
sample-000007.mp3	female	female	us	us
sample-000032.mp3	male	male	ireland	us
sample-000203.mp3	female	female	indian	Indian

5.4 Inference

Inference has been conducted on 4 different audio files as shown in Table 3. As it can be seen from above, the model is getting confused between scotland and australian accent as they sound very similar. The same is also observed between ireland and us accent. Indian accent and us accent has been correctly predicted by model. It can be understood that the class imbalance has caused the model to overfit to majority class accent predictions. However, the model is correctly predicting the gender in all the cases.

6 Conclusion

Speech is considered to be one of the most important part of human communication. Humans are extremely good at identifying characteristics like accent, gender in a speakers' speech. This can be very useful in many applications. The techniques presented in this paper were able to achieve really good performance in identifying characteristics such as gender and accent. The experiments show that CNN based networks achieved really good performance while some LSTM based models struggled to train and the remaining were outperformed by the CNN models. Going deeper by using skip connections like ResNet18 helped the model improve it's performance significantly. However, using even more deeper networks like ResNet34 and ResNet50 didn't improve the performance.

7 Future work

This paper contrasts two different approaches namely using CNN and RNN as feature extractors in multi-task learning. The below are the different experiments/improvements that can be performed as a future work.

1. Weighting the loss of accent more than the loss of gender to improve the validation f-score of accent to go beyond 90%.
2. Adding one more characteristic as the third label like age of the speaker.
3. Improving the LSTM architecture, so that it can reach the accuracy of the CNN based models.
4. Reducing the model size while maintaining similar performance can reduce the training time significantly.

References

[1] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M. and Weber, G., 2019. Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670.

[2] Green, P.J., Sha, M. and Liu, L., 2011. The US-China E-Language Project: A Study of a Gaming Approach to English Language Learning for Middle School Students. RTI International.

[3] Crawshaw, M., 2020. Multi-Task Learning with Deep Neural Networks: A Survey. arXiv preprint arXiv:2009.09796.

[4] He, K., Zhang, X., Ren, S. and Sun, J., 2015. Deep residual learning for image recognition. CoRR abs/1512.03385 (2015).