Estimating the age of a person from their speak data

19th November review with Prof Bhiksha Raj

Data processing

We shall extract mel spectrograms using short time fourier transforms. The following parameters will be used:

- Get SRE and Fisher datasets in addition to Timit we already have

For TIMIT:

- 64 mel filters
- Bandwidth: 100 - 8000

For SRE:

- 40 Mel filters
- Bandwidth 150-3800 Hz

In both cases:

- Window size shall be 100ms
- 90 ms overlap/10 ms shift

Baseline experiment 1

The baseline experiment 1 involves predicting the age in years.

The model: 1 TDNN layer with kernel size ranging from 2 to 5 and a couple of LSTM layers stack on top.

The task for this experiment would be to minimize L2(Least Square) and L1(Least Absolute Deviations) errors for the age. This would be a mixed weighted combination of both losses.

On the test data, we would find the root mean squared error i.e. finding the square root of the l2 error and MAE

Default classifier: predict everyone's age to be the average age of the training set.

Baseline experiment 2

Experiment two would involve partitioning the data by gender; that is assuming the speaker's gender is known. The motivation for this experiment is that the manner in which pitch changes is gender specific.

Therefore for our model we predict two outputs, we predict both gender and the age. The loss we would calculate would be a combined KL divergence and l2 divergence(for the age loss). This would help segregate gender from the data we learn.

baseline model on the test data

we will take the average age of the female voices in the training set. if the test instance is a female we predict it as the average age of female in the training set.

Experiment 3

This experiment involves replacing the CNN with a transformer.

Experiment 4 (using a source filter)

The speech of humans is a combination of sounds from lungs and in the mouth. The human mouth is a resonance chamber that acts as a filter to sound that comes out of the lungs. Basically one can talk when they breath out. Talking while breathing in is practically impossible. However certain countries have fewer sounds they make while breathing in.

Therefore for the final experiment, the aim is to separate the sounds that come from the lungs and that produced in the mouth. The separation would be done with a source filter model. The model would separate excitation from the lungs from the filter(modification that occurs in the mouth). The output would be 2 different spectrograms (spectro characteristics from signals that came from the mouth and that of the lungs).

We will use these two spectrograms in parallel with our models. Moreover, we will use cross attention to extra resonances from the mouth and that from the lungs.