Supporting communication in children with non-typical speech using Deep Learning

James Toop

Data Science, Capstone Project

Purpose –
The importance of communication

Communication is a basic human right and ensures that people can:

- Express themselves;
- Control their environment;
- Make friends and build relationships;
- Make sense of the world around them.

Source: Foundation for People with Learning Disabilities, National Health Service (NHS) UK and British Institute of Learning Disabilities (BILD)

Purpose –
The importance of communication

Communication has greater importance for a person with a learning disability.

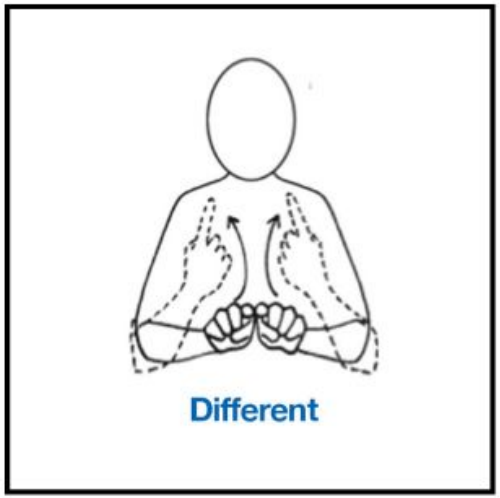
- May not able to interpret their environment as easily as others;
- Inability can cause frustration and behaviours that might be seen as challenging;
- Can lead to being marginalised or excluded by society.

Source: Foundation for People with Learning Disabilities, MENCAP

Purpose – Supporting communication

A variety of techniques have been developed to support children for whom speech is difficult:

- Communication systems such as Makaton or PECS;
- Easy read symbols;
- Speech and language therapy.



Example of Makaton signing

Source: Foundation for People with Learning Disabilities, MENCAP, The Makaton Charity

Purpose – Supporting communication

With supported communication a child with a speech disorder or severe learning disability will be better able to:

- Express their needs, feelings or opinions;
- Make choices;
- Interact with others and form relationships;
- Develop skills which will help them access learning opportunities.


Martha
U
S
R

Objective –

Use Deep Learning techniques to build a model that can identify an audio sample of non-typical speech and suggest possible words as a tool for helping parents support their child’s communication.

Approach – The raw datasets

2 datasets totalling almost 140,000 - one second audio samples of single spoken words

- **Speech Commands:** 105,829 audio samples of 35 spoken keywords;
- **Ultrasuite:** 14,286 audio samples from 86 different children, 28 with speech disorders;
- After processing, 33,800 audio samples of 991 spoken words;

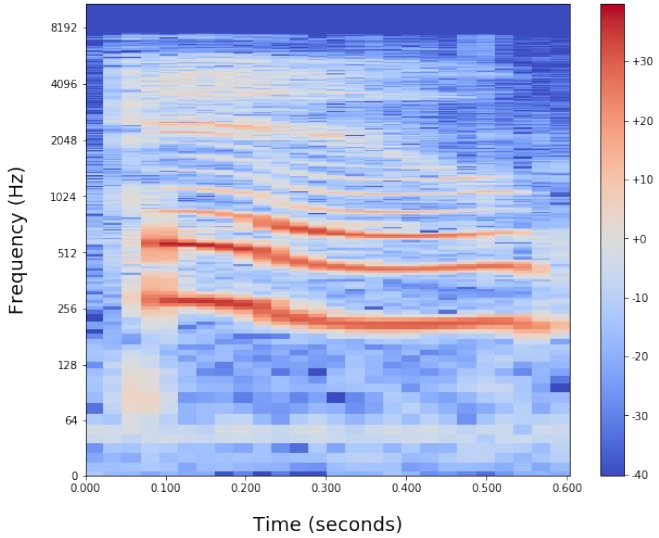
Speech Commands audio example

Ultrasuite audio example

Approach –
Preparing the data and creating the models

Iterating through models, adjusting architecture and variables in response to previous results

- Significant preprocessing to isolate and resample Ultrasuite audio samples then extract spectrograms and other features
- Create simple, baseline model based on reduced dataset as proof-of-concept and observed results.
- Final model uses a Convolutional Neural Network.

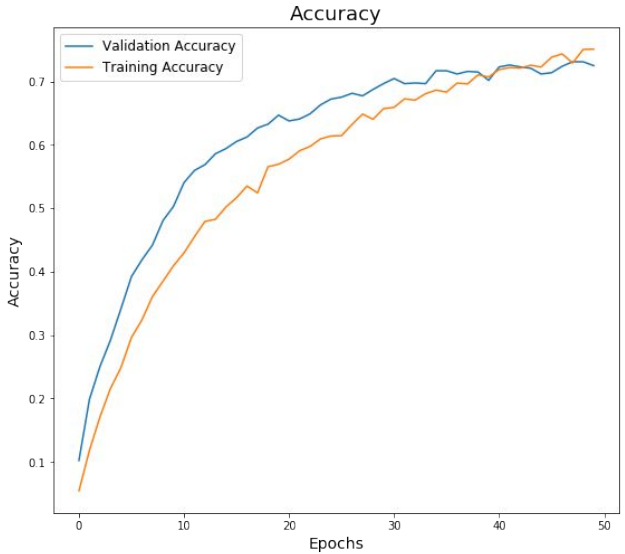


Example spectrogram for “Girl” audio sample

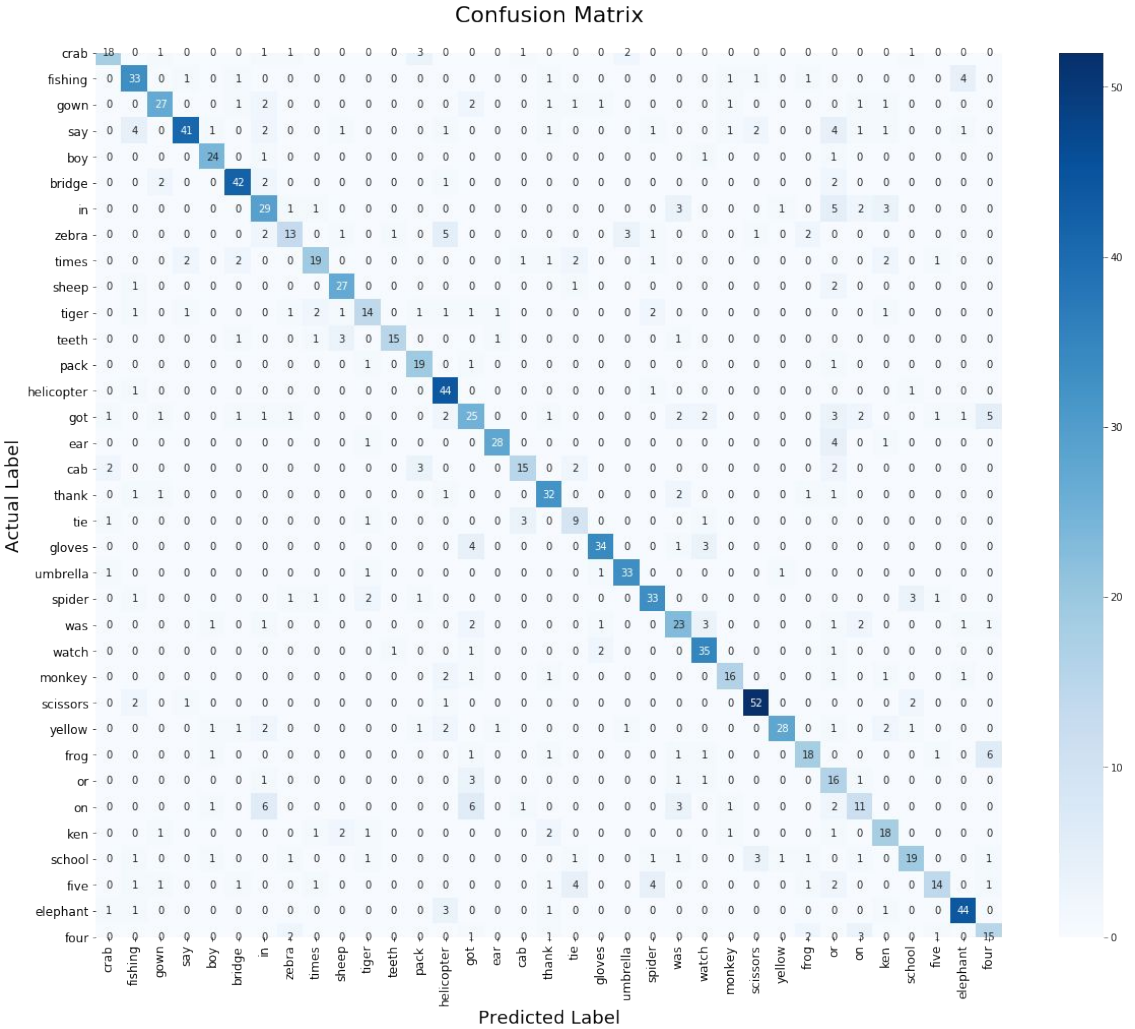
Conclusion –
Lower accuracy, minimised overfitting

Lower accuracy not a hindrance to usefulness of the app

- Most models did not generalise well, tending to overfit the training data. Likely because of the nature of speech impairments.
- Final model produced the most balanced results correctly identifying the spoken keyword for 70% of the “unseen” audio samples.



“Final” model accuracy during training


Confusion Matrix
Actual Label
Predicted Label
crab
fishing
gown
say
boy
bridge
in
zebra
times
sheep
tiger
teeth
pack
helicopter
got
ear
cab
thank
tie
gloves
umbrella
spider
was
watch
monkey
scissors
yellow
frog
or
on
ken
school
five
elephant
four
0
10
20
30
40
50

Conclusion – Potential next steps

Recommendations –

- Continue to use other model architectures and enable to accept longer audio samples as these are more representative of atypical speech patterns.
- Source additional data in the form of further audio samples potentially even using the app as a means for gathering additional samples and improving the model.
- Use data augmentation when training the model, in particular the MixSpeech method that a weighted combination of features such as Mel Spectrograms with MFCCs.

Thank you.

Any questions?

Appendix –
Model results highlights

Model Iteration	Model Details	Training Accuracy	Validation Accuracy
1	Baseline model with single layer	81%	67%
2	Baseline model varying learning rate and batch size	94%	65%
3	Deeper network model	96%	72%
4	Deeper network model but adding dropout layers	85%	70%
5	CNN	83%	70%
6	CNN #2 (descending number of neurons over the first 3 layers)	93%	70%
7	CNN #3 (change dropout layer from 0.3 to 0.5)	94%	69%
8	CNN #4 (Neuron tweaks etc.)	90%	70%
9	CNN #5 (changes to epochs and batch size.) - FINAL	90%	71%

Appendix – Making a prediction

```
# Run inference on the unseen audio file
mfccs = librosa.feature.mfcc(audio_sample, sr, n_mfcc=13, n_fft=2048, hop_length=512)
mfccs = mfccs.T
mfccs = mfccs[np.newaxis, ..., np.newaxis]

prediction = us_cnn_model.predict(mfccs)
predicted_index = np.argmax(prediction)

predicted_keyword = us_keywords[predicted_index]
print('Martha says...',predicted_keyword, '!')
```

```
Martha says... frog !
```