Capstone Proposal

Domain Background

Speech recognition has recently seen an unprecedented level of development in the past few decades. Previously, Hidden Markov Models were used and fed through to neural network architectures (Bourlard, 1994). However, the emergence of deep learning has revolutionised the domain of speech recognition (Schmidhuber, 2015). The key problems of historic neural network methods such as gradient diminishing were overcome with increased computing power and the associated ability for big data to be analysed. From 2009 onwards, the availability of computing power has allowed deep learning to be applied to acoustic modelling and speech recognition, which has spawned a host of collaborative research teams from companies such as Google and Microsoft (Hinton et al., 2012).

Using machine learning to analyse audio signals has already been widely applied in everyday life for billions of people across the world. Most famously, Apple's Siri uses sophisticated deep learning techniques to analyse users' voices, transcribe, action, and reply. A recent example of Siri's development has been released by Apple's Machine Learning Journal, where a deep learning network was used to turn Siri on when a user says 'Hey Siri' (Apple, 2017).

My personal motivation stems from my workplace. I work as a (very junior) developer in a consultancy in London. I have built the website and components of the backend, including implementing a binary tree that takes machine learning outputs and decides whether a recorded call from a customer to the company is compliant. My team uses the transcription of calls as feature inputs into models and I wanted to work on audio analysis to try to use other audio features as inputs to machine learning models, rather than just the transcription.

Problem Statement

Industry is increasingly looking to use the data that they capture. One major data source that has been captured but not used is telephone calls with customers. These calls can be analysed to identify areas of risk between customers and a company. For example, if a customer is showing anger, the call is likely to have been below standard for the company and should be investigated. By targeting these calls, companies can target business plans to address issues derived from data, rather than guessing at what needs to be fixed.

The problem I am going to try to solve is to identify the sex and age of a telephone call. A telephone call is an audiofile, which is an encoded binary stream. The required output from the audiofile will be a Boolean answer of whether it is a male or female speaker and a category of age (young, adult, elderly). Audiofiles have many features that can be explored and correlated through a deep neural network with labelled data to attempt to extract some of the relationships between features and the age/gender labels. This problem is quantifiable because each audiofile has an age/gender label, measurable because audio features are quantifiable by nature, and replicable because there is a whole host of open source data that is continuously added to.

Datasets and Inputs

The aim of this project is to be open-source and therefore I have decided to use an open-source dataset. The Common Voice Dataset (Common Voice, 2018) has 12GB of recorded and validated data. This dataset has labels for upvotes and downvotes, which I can use to filter out poor-quality audiofiles (which may be because of a bad transcription or poor-quality voice). Although in future I can use 12GB of data, I shall limit this submission to 500MB as per the course guidelines. I shall choose the subset by ordering the data with the most upvotes, then taking the top that fall within 500MB. This should give the cleanest data. Each file is approximately 50kb, meaning that 10,000 audiofile datapoints can be used (each file is approximately 5 seconds long giving nearly 14 hours of data). The licence for the dataset is CC0 (CC0, 2018), which means that the author of the dataset gives anyone the right to use it.

The audiofiles in this dataset are given as mono .mp3s and a 48,000 sampling rate. This is all sufficient to process the files into any form that is required for machine learning. The audiofiles are one sentence long, which is sufficient for analysis, although longer files would perhaps have made the pre-processing stages easier. Features shall be taken directly from the audiofiles so no other information is required in this respect. I have tested pre-processing some of the files and I was able to use them to train a very simple CNN. The labels for the audiofiles are not as complete as the features themselves. Not all audiofiles have sufficient labels, but about half do. About of the 12GB have labels for the sex and age of the speaker. There is enough labelled data to train, cross-validate, and test.

Wang et al. (2017) used audiofiles to predict the emotion, age, and gender of speakers in Mandarin. They used a similar dataset with 17,408 utterances from a Microsoft spoken dialogue system, which were then judged by a panel of five judges. In addition, an undergraduate research team used web-scraping techniques to get audio data from YouTube (Gill et al., 2017). I believe my open-source library offers as much information as the two examples given here, if not more. It is appropriate because it gives the inputs to pre-processing in order to create features and labels required.

In my project, I shall build two distinct models. The first model will predict age and then another will independently predict the sex of the speaker. The classes will be split into the following:

Model 1

- Male
- Female

Model 2

- Young
- Adult
- Elderly

Although this is the aim, I am taking note that Wang et al. (2017)'s study mixed the classification of sex and gender. This may be because a younger male's audio features are similar to that of an adult female, which could make the sex model inaccurate. I intend to experiment with the outcomes of both to come to a final conclusion, although the original aim is to separately classify sex and age.

The train/test set is already split by Common Voice already. However, I would like to replicate Wang et al.'s (2017) distribution of 60% train, 15% cross-validate, 15% test. The dataset is skewed towards male voices. Of the 73,000 datapoints for training, 75% of the voices are male. The data is split into the following categories for age:

Age (category)	Distribution (%)
Teen	7.5
Twenties	31.6
Thirties	25.2
Forties	15.3
Fifties	13.0
Sixties	6.3
Seventies	2.2
Eighties	0.3

In order to make this more equitable, I shall select more from the categories that are under-represented in order to make a more generalised model.

Solution Statement

The solution to classifying age and gender from audiofiles is to extract features from the audiofiles and to process through a machine learning engine with labels. The audiofiles will require some feature engineering in order to extract features that will be passed to the model. There are two options that I am looking to explore for features, either using a spectrogram or a MFCC (mel-frequency cepstral coefficients). Both spectrograms and MFCCs have been used in the Wang et al. (2017) and Gill et al. (2017) paper. These features will then be fed into a convolutional neural network. This may use transfer learning by using the CIFAR-10 CNN (Wang et al. 2017). I aim to try some other transfer learning architectures such as VGG19. This will then either by soft-maxed or passed through a more traditional machine learning model to classify. This process will be reproduced for each audiofile so is repeatable.

Benchmark Model

The primary benchmark model would be a very simple architecture like has been demonstrated below. It would use one convolutional layer before flattening through some dense layers to the classification categories. I have used this architecture to process 1,500 practice files. This achieved an accuracy of 73% when attempting to classify males and females. However, 73% of the files were male, so the model was classifying every input as a male. My model must be better than this,

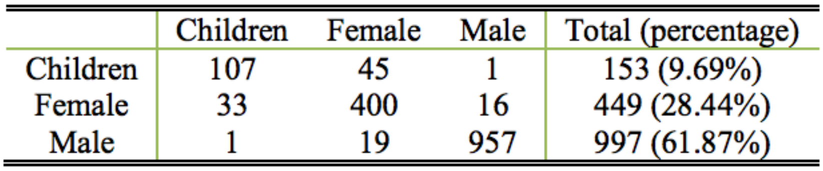
The secondary benchmark model for this project will relate to the two papers that have previously been cited: Wang et al. (2017) and Gill et al. (2017). Gill et al. made a model that was 97% accurate for classifying the gender for three people. Although this is very good, for just three people there may have been a lot of overfitting. I shall be working with thousands of voices and therefore aim to create a far more generalised model. As for age, the Wang et al. paper differentiated between males, females, and children. They achieved a 92% accuracy, which is higher than the 5 judges that were used.

Evaluation Metrics

Due to the simple nature of the classification task, previous studies have been very simple. Gill et al. (2017) uses simple accuracy:

$$accuracy = \frac{correctly\ classified}{total\ classified}$$

Wang et al. (2017) show a confusion matrix to differentiate between male, female, and children:



	Children	Female	Male	Total (percentage)
Children	107	45	1	153 (9.69%)
Female	33	400	16	449 (28.44%)
Male	1	19	957	997 (61.87%)

By creating very simple evaluation metrics, it allows direct comparison with other studies, which also use simple metrics. This gives context to the data and problem because it uses the original labels for the data (age and gender) and then evaluates the solution by comparing the classification with the label. A confusion matrix will be used to see where the misclassifications are occurring, which will provide insight into where the model is going wrong.

However, I shall use more metrics than just accuracy, As shown above in the evaluation metrics section, accuracy can hide the poor performance of a model. An accuracy of 73% may seem good at first glance but the model was predicting the same class every time, which is an extremely poor model. Even though I shall balance the classes, I shall give some other metrics in order to make sure constant classification isn't happening. A ROC curve (and associated AUC) will be able to present the specificity and sensitivity, which will be able to show the predictive power in addition to the accuracy. I shall use both these metrics in order to choose my preferred model.

Project Design

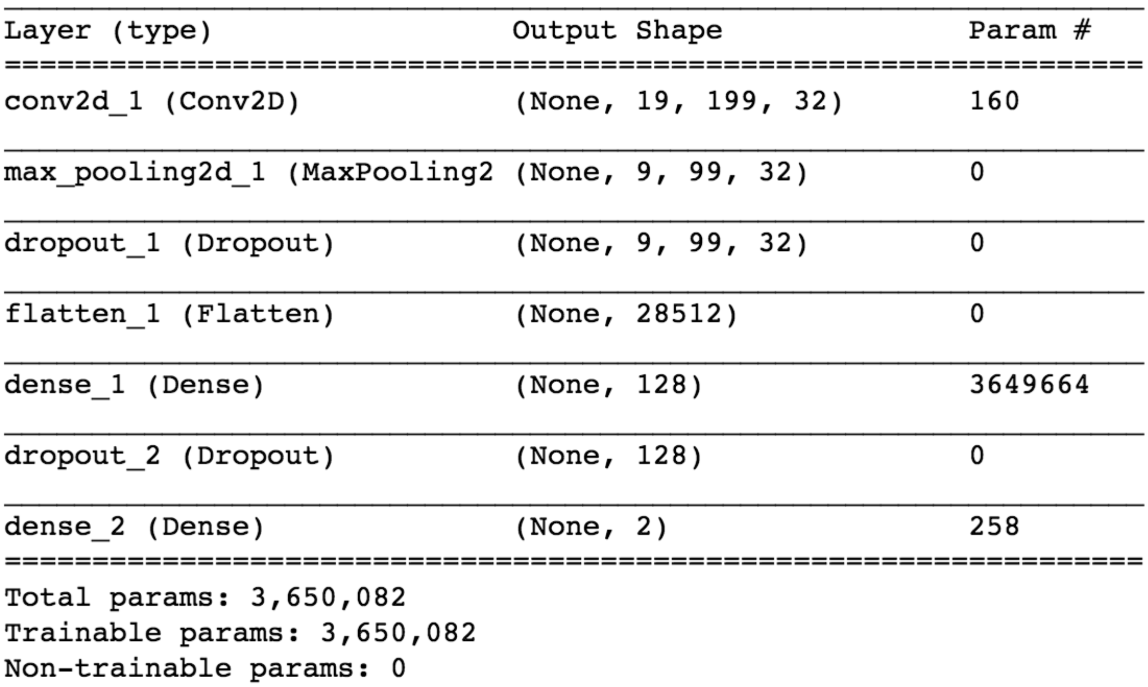
Data Pre-processing

The dataset has mp3s as input and labels in a csv. The following processing will occur:

- Convert mp3 to wav
- Split into segments of constant length
- Filter out segments below an average amplitude threshold
- Create spectrogram
- Create MFCC
- Add extra dimension to MFCC to mimic a picture

Training

At first, a simple CNN model will be built from scratch to analyse the spectrograms and MFCCs. This will be a Keras model with an architecture something like the following:



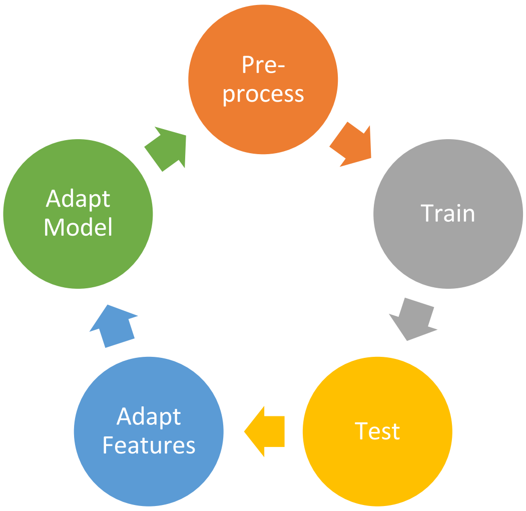
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 19, 199, 32)       160
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 9, 99, 32)         0
_________________________________________________________________
dropout_1 (Dropout)          (None, 9, 99, 32)         0
_________________________________________________________________
flatten_1 (Flatten)          (None, 28512)             0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               3649664
_________________________________________________________________
dropout_2 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 258
=================================================================
Total params: 3,650,082
Trainable params: 3,650,082
Non-trainable params: 0
```

The input shape of the audiofile for the MFCC will be (<MFCC dims>, <duration>, 1). The third dimension here will be the added dimension to mimic a picture. This architecture will be adapted as results are found. In addition, I may use a traditional machine learning model as an output from the CNN to further learn from the features, as presented in Gill et al. (2017).

Also presented in Gill et al. (2017) is the use of transfer learning. They advise using VGG-19 to analyse spectrograms. My final architecture will depend on what shows the greatest success out of:

- Only using a CNN with a final dense layer
- Using a CNN into a traditional machine learning model
- Using transfer learning CNN with a dense layer
- Using transfer learning CNN into a traditional machine learning model

If I have time, I shall look to augment the data in order to get a more generalised feature set. This would include using libraries such as Muda (McFee, 2015), which has been shown to improve models. However, as I have a lot of labelled data, this is not a priority.



References

Apple. (2017). *Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant.* Available: https://machinelearning.apple.com/2017/10/01/hey-siri.html. Last accessed 13th July 2018.

Bourlard, H and Morgan, N (1994). *Connectionist Speech Recognition: A Hybrid Approach*. Boston: Kluwer Academic Publishers.
CC0. (2018). *CC0.* Available: https://creativecommons.org/choose/zero/. Last accessed 13th July 2018.

Gill, C. (2017). *Automatic Speaker Recognition using Transfer Learning.* Available: https://towardsdatascience.com/automatic-speaker-recognition-using-transfer-learning-6fab63e34e74. Last accessed 13th July 2018.

Hinton, G et al.. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*. 29 (6), 82–97.

Mozilla. (2018). Common Voice Dataset. Available: https://voice.mozilla.org/en. Last accessed 13th July 2018.

McFee, B et al. (2015). *A Software Framework For Musical Data Augmentation*. Available: http://bmcfee.github.io/papers/ismir2015_augmentation.pdf. Last accessed 15th July 2018.

Schmidhube, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*. 61 (1), 85–117.

Wang, Z et al.. (2017). *Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks.* Available: https://www.researchgate.net/profile/Ivan_Tashev/publication/316009054_Learning_utterance-level_representations_for_speech_emotion_and_agegender_recognition_using_deep_neural_networks/links/5a29e14eac. Last accessed 13th July 2018.