Machine Learning for Feature Extraction and Classification of English-language Accents in Ireland

[Speech Processing, Machine Learning]

Peter Nolan
x22154116
MSCDATOP — Research in Computing
National College of Ireland

12 August 2024

Abstract

Pronunciation in the English language in Ireland is of strong interest to the Irish public and the research community, both as a marker of identity and in considering interactions with modern automated speech processing tools. Qualitative linguistic research has consistently shown that the pronunciation within Ireland of the English language has shown significant differences by geographic origin among the world's English speakers, within the US and Britain, and between these traditional mother-tongue speakers and accents in India, Australia and elsewhere. area and over time. Recent data analysis reinforces the distinctness of Irish-English from forms of English spoken in mainland Britain. Using speech samples from the wide survey published in Hickey's (2004) 'Sound Atlas of Irish English', we estimate models to classify the Belfast and Dublin accents of Irish English using logistic regression, neural networks, convolutional neural network and large audio models. Evaluation by accuracy, confusion matrix and ROC curve methods showed strong classification ability for these regression and neural network models. However, performance using recent transformer-based large audio models was poor. Overall, this research points to continued future data-gathering and more modelling work while preserving privacy as promising avenues for future research, leading to greater socio-linguistic understanding and to reduce bias impacting consumers.

Keywords: Accent Classification, Socio-linguistics, Ireland

Contents

1 Introduction

This research describes a contribution to research in linguistics on English-language speech in Ireland (known as 'Irish English' to linguists and from hereafter) by an innovative application of data analytics applied to speech recording data to understand and to classify the accents used in two major urban regions, Dublin city and county and Belfast together with its hinterland in the surrounding county of Antrim.

1.1 Research Question

The research question examined here can be summarised as: How can we classify some selected regional accents in Ireland using data analysis techniques based on the features within the sound of their speech and on demographic characteristics of the speakers? This points to three closely-linked research questions in turn:

First, what datasets may be available that can support data analysis of sub-populations within Ireland? This requires a search of the available datasets of Irish speech to discover those that might be useful for such a data analysis. The answer may be that no useful data is available, or we may discover an accessible dataset that can be used for data analysis for this research project and for future work in the field.

Second, what features of the sound signals and what data on personal characteristics can be used as inputs for attempting classification of a sound sample, ascribing it to a particular geographic location within Ireland? The target here is to use available data to build a model that uses speech recordings as an input, perhaps after some processing to extract features from the speech data, and then outputs an accent classification by regional geography. If available, there may be some personal or group characteristics but not just simiply a personal identification of the speaker, as an additional input, to such a classification model.

Third, how would data analysis models, as used in similar research in the published literature on automated classification of accents, perform in classification of Irish speech samples by location? One way of answering this question is to compare the performance using the common metrics of classification models that we can build with similar techniques used in studies in different geographical areas. With the availability now of many machine learning libraries, models through API access or by using published code for models makes adaptation of existing models to new data much easier.

Note the difference to the CA1 and CA2 outlines from the research question as specified here in this section. The proposal in CA1 to collect original speech samples from schools in Ireland in CA1 has been dropped from the current project scope in CA2 owing to the pressure of the deadline. Then, much of the effort following CA2 went into searching for a useable dataset as several candidates were unavailable or unsuitable, one key dataset being replaced by another as the main source, as discussed in section 2.1 in the Literature Review below. Furthermore, a request for data to Mozilla about CommonVoice, a crowd-sourcing project for speech samples, for more data, has so far gone unanswered.

1.2 Motivation for the Literature Review

The starting point was in textbooks covering the basics of linguistics, such as those by (Hornsby 2014), so as to understand basic terms.

An accent is a 'way of pronouncing a language that is distinctive to a country, area, social class, or individual' *Oxford English Dictionary* (2023). The lines between an accent and a dialect can be blurred: A dialect is understood by linguists as the distinctive sounds, but also with grammar and vocabulary in a spoken language, whereas an accent instead refers to the identifiable differences in pronunciation of the words and sentences among a group (Hughes et al. 2012, chapter 1).

Identifying, understanding and using an accent is an important social behaviour, the process understanding and, if necessary choosing, how one speaks. Accents are identifying factors for individuals, which can indicate gender, nationality, socio-economic status, regional or local origin, membership of a neighbourhood or other identities. This can create trust and supportive relationships but might also trigger hostility, discrimination or even violence.

Everybody who speaks does so with an accent, although the term might only be used to describe speech by any person with a relatively lower status than some standard accent, typically associated with the 'supraregional' accent, such as the spoken English of Dublin 4 associated with RTÉ) broadcasts or the English accent that has been dominant historically on the BBC Hickey (2007), Markl & Lai (2023).

Statistical analysis of speech shows that variance between accents might be found in the length, stresses and tones used in pronunciation Zuluaga-Gomez et al. (2023). Analysis of accent and other patterns in speech might look at the phonemes, the basic sounds of speech, typically mapping to a syllable and fully observable in a short period, perhaps under a second: Other distinctive patterns may be visible over longer time horizons, corresponding to clauses or complete sentences Jiao et al. (2016).

Irish literature and media has long taken an interest in accents, their location, how they change over time and their social roles. Nationality is one such role. (Dolan 1991) quotes James Joyce's classic novel **Portrait of the Artist as a Young Man** on an encounter between the Irish protaganist and an Englishman, 'An Irishman's pronunciation of these words classifies him as a speaker of non-Standard English, because his vowels and consonants, as well as his pronunciation of the endings of the words Christ and master, all demonstrate that he is using Irish sounds for English letters.' The distinctive Irish English, the Irish form of the English language, defiantly held, was cultivated in Irish literature, especially Joyce's masterwork and Ireland's most important novel, **Ulysses**.

Linguistics research has consistently documented a number of distinctive regional accents all over the island of Ireland. Hickey maps the accents in Ireland as shown in the map in figure1, taken from his 'Sound Atlas of Irish English' (hereafter 'SAIE'). Another such atlas is famous monologue by the actor Niall Tóibín in which he speaks in and comments on a variety of Irish-English accents[1].

Dublin's early adoption of English and close links, as a major port, across the Irish Sea has lead to distinctive accents, closer to English pronunciation. From the 'nineties onwards, younger generations in Dublin often use a less-structured New Dublin English accent with a distinctive pronunciation (Hickey 2007, chapter 5). The Irish public seem fascinated by observing accents and the social stratification they signal, as shown in 25 years of best-selling novels, podcasts and weekly newspaper column satirising the stereotypical affluent south Dublin rugby player, Ross O'Carroll-Kelly, who exhibits this 'DORT' accent, how he refers to the area's DART commuter railway[2].

Belfast's accents have formed by historic migrations from lowland Scotland. The accent influenced by the Scots-Ulster pronunciation remains and from the southern and western part of the current Northern Ireland and County Donegal Hickey (2007). Besides the status hierarchy common to all urban areas, Northern Ireland also experiences the persistent ethnic/religious divide between Irish Catholic and British Protestant populations that was violent in the past and remains pertinent today. Like Dublin, Belfast is today also the largest city by population in their areas and both are the centre for broadcast media within their regions.

By its interaction with automated speech processing tools, the accents of Irish-English have an observable effect. Indeed, it can significantly impact usability in how accurately a person can be understood by machines for voice input, such as the Amazon Alexa or Apple Siri. Recent empirical work Markl (2022) shows how, among recordings of speech in the IViE dataset described below fed into transcription tools offered by Google, that Belfast accents were distinctive in their significantly higher error rate than Dublin or any British accents.

1.3 Findings of the Literature Review

Beginning by drawing on no prior background in either digitial signal processing or in linguistics, the research began by surveying the literature and selecting books, articles, datasets and software libraries. This was based on availability within the library at NCI as well in the extensive collections at the Trinity College Dublin libraries. Bearing this in mind, the literature selected fell into a number of topics. While all worked are related, those within each subset show contrasting approaches, so the choices made for carrying out the research are explained in detail.

Care has been taken in selecting the material for this literature review according to criteria of scholarly reputation, of practicality in implementation and of ethical correctness. As previously justified in CA2, priority was given to those papers with higher citation counts as indicators of impact and quality. For publications highlighted below, the citations were counted whereever available on the Scopus database and on Google Scholar. Being part of proceedings, papers were evaluated with reference to the CORE and h5 rankings of conferences. The Conference of the International Speech Communication Association, commonly known as INTERSPEECH, is particularly highly-regarded among industry practitioners and academic researchers, and papers in those proceedings were a priority. Papers from the Advances in Neural Information Processing Systems conference ('NEURIPS', previously 'NIPS') was also particularly

[1] Irish regional accents: https://www.youtube.com/watch?v=EhLdKJnY194

[2] For example, this extract from a recent Ross O'Carroll-Kelly audiobook is available online: https://www.youtube.com/watch?v=vlobTz-9_Fc

valued, as was work authored by Turing Prize-winning pioneers of deep learning, Geoffrey Hinton and Yoshua Bengio.

Some case-studies were selected as pointers for the data analysis in the project. These were chosen to show useful results using the data analysis techniques in current use at the time of publication. (Sheng & Edmund 2017) is one such case, to build classification models English language speech with different accents and getting increasingly good performance from using decision trees, multi-layer perceptron and convolutional neural networks model. (Schuller et al. 2016) uses SVM models for classification of accents with a datasets shared in an INTERSPEECH competition. (Jiao et al. 2016) builds unsupervised learning and then convolutional models that perform significantly better with the same dataset. Similarly, (Abdel-Hamid et al. 2013) uses unsupervised learning and feeds output into a convolutional neural network. Maas et al. (2017)[3] and (Lesnichaia et al. 2022)[4] showed that more hidden layers, random initialisation of network weights, newer optimisers and Dropout pruning all have improved deep learning classification performance on accent classification. Finally, a recent INTERSPEECH paper Zuluaga-Gomez et al. (2023) which uses variants of the **wav2vec2** models using *wavenet*, transformer and masking models, cross-training on non-English language data produces the highest classification accuracy of any of these examples at over 97%.

1.4 Outline of this Report

This report follows the outline given in the *MSCDATOP - RIC Module Assessment Guide - 24 Jan 2024.pdf*. The material expands and develops from the papers submitted previously as **CA1** and **CA2**. Any differences between this report and the earlier documents are highlighted and justified here.

The literature review is presented in the next section, organised by topic and presenting the arguments for choices about what research literature to include in our analysis. It proceeds examine the topics to draw on for the research design, one in each subjection. One reference is typically broken down by subject area in the subsections. The Literature Review begins with a survey of available datasets and justifications for choosing among the candidates. It continues to consider the preprocessing activities needed to get the available data in form suitable for the data analysis models. Models available are then considered for their suitability for answering our research question. The evaluation suitable for these models is then considered, common classification metrics being an obvious choice,

The report goes on to the Research Methods and Specification Section. This begins by justifying the choice of the models selected for the research, namely logistic regression, k-means clustering, neural networks with the Multi-Layer Perceptron architecture (or 'MLP'), convolutional neural networks ('CNNs') and finally, a likely Large Audio Model (LAM) utilising recent advances in architecture and showing the highest level of performance available. Following on, the models and actions are described in detail, together with the results from running these on our data.

The technical setup for the models is then described in following subsection. Progress in the project since CA2 was submitted is then compared to the project plan and the issues that delayed or prevented tasks being completed described, in particular in setup for the Large Audio Model.

Finally, in the Conclusion, the report is summarised and the results of the research evaluated for their likely benefits and contribution to the research literature and significance to stakeholdes. With strong performance on the classsication tasks by the logistic regressions, MLP and CNN, the choice of the SAIE dataset is vindicated. With this novel methods applied to this dataset, it can be expected that the linguistics research community will take on the valuable future work of further data analysis projects on Irish English speech and of gathering more data.

2 Literature Review

In this section, a review of selected literature relevant to the research question, is presented. This is organised by topic, one per subsection. Research referred to will be cited in different sections where its contents contribute to those questions.

2.1 Datasets

Searching for candidate data then meant moving on to look at available datasets, whether international ones such as CommonVoice Ardila et al. (2020) or those with an Irish focus such as IViE Nolan & Post

[3]Scopus=95,GS=179 citations
[4]Scopus=5, GS=6 citations

(2014) and the Sound Atlas of Irish English (hereafter 'SAIE') Hickey (2004).

As soon as sound recording became technically possible more than a century ago, this was then used to collect samples of speech in Irish and in English from people throughout the country(Hickey 2007). With our research question here focusing here on Irish English speech recordings, what datasets may be available to use in a data analysisof Ireland's accents?

Nowadays, we might expect that the largest repositories of such voice recordings would belong to the large technology companies offering voice assistants, who gather it in the course of interactions with customers. However rich these might be in insights, this data is not often released for outside researchers, although the results of studies using the data by their own staff have been published. Hinton et al. (2012) estimates models based on data available to those co-authors working at Microsoft, Google and IBM, Abdel-Hamid et al. (2013) uses other data from Microsoft. Despite requesting, making such data avaialble for this study was refused by a number of such businesses in Dublin, citing both data protection and competitive necessity as justificaton. So, as much as possible, the research should use data publicly-available, to aid replication and future work, and, if possible, add to what is available.

(Markl 2022) used the public transcription tools offered by Amazon and Google to evaluate their accuracy with a set of recordings from published sources. One, the **Speech Accent Archive** Weinberger & Kunath (2011)[5] has recordings of speakers reading scripted sentences in English over 15-25 seconds from around the world. However, only 5 were from named districts in Northern Ireland and 12 more in the Republic. Besides the small number of speakers, gender imbalance was severe, with only 3 female speakers among these 17 records.

The second dataset examined was the Intonation Variation in English (**IViE**) dataset, maintained by the University of Oxford Linguistics Laboratory and described in more detail in Nolan & Post (2014)[6]. The data was surveyed from school pupils across the UK and Ireland, including both Belfast and Dublin. Six male and six female readers provided samples in each city, with 260 scripted sentences and 65 longer passages recorded. With more observations and greater gender balance than SAA, this looked more promising as a basis for research. Managers of some schools in Dublin and Belfast were approached during February 2024 to ask if they would consider adding more samples using the same scripts, but none would commit to doing so within the project timeline.

A third candidate dataset was that described by (Hickey 2004) in his 'Sound Atlas of Irish English' (hereafter 'SAIE')[7]. This contained 1500 speech samples recorded all over the island of Ireland, some 266 in total from Dublin and Belfast cities and the surrounding counties of Dublin and Antrim respectively. The recordings and other files were provided as a CD accompanying the book, which was retrieved at the Trinity College Dublin Library only in early April. Manual checking abstracts for all citations on Google Scholar to date of this dataset showed that no data analytics had been used, that the work involved the traditional linguists' method of manual replay of recorded speech and analysis by human capabilities alone. His mapping of accents across Irealand is shown in figure1 below.

Selecting a dataset for the study here from among available candidates, the quantity of data and especially the number of speakers in a dataset seemed vital. In matching voice characteristics, it was certainly necessary, as one study argued that the training and the test datasets be created without including any one speaker in both, otherwise the data analysis risked simply matching the voice characteristics of that individual speaker and not matching the features across a whole population Ahamad et al. (2020). With only six speakers of each gender for each location in the IViE dataset, an 80% train to test data split would likely leave only a single speaker in a test dataset for any estimation, which did not appear a sound approach. Therefore, the preference during the research was for the the SAIE database.

Also of interest was the **CommonVoice** (Ardila et al. 2020)[8] public dataset of scripted speech sampled. This was collected by crowd-sourcing. Volunteers among the public may deposit a voice sample through Mozilla's website or iPhone app. Donors self-report their accent on a country-level, gender and age. The pronunciation and the self-reported accent is verified by other users' manually verifying it. Already by the 2017 initial dataset, CommonVoice had 981 samples labelled as Irish accents out of over 67,000 in English, but only as a national-level identifier i.e. as Irish rather than in regional detail as with IViE and SAIE. This dataset is also used as the input for accent detection by the claimed State Of The Art (SOTA) model in (Zuluaga-Gomez et al. 2023) and would be useful if we are to try to reproduce or update that model.

[5]Citations in Scopus = 23, Google Scholar = 53

[6]Citations in Google Scholar = 13 GS, not returned by the Scopus database

[7]Citations on Google Scholar of 175, book not present in Scopus

[8]Scopus = 412, GS=1,124 citations

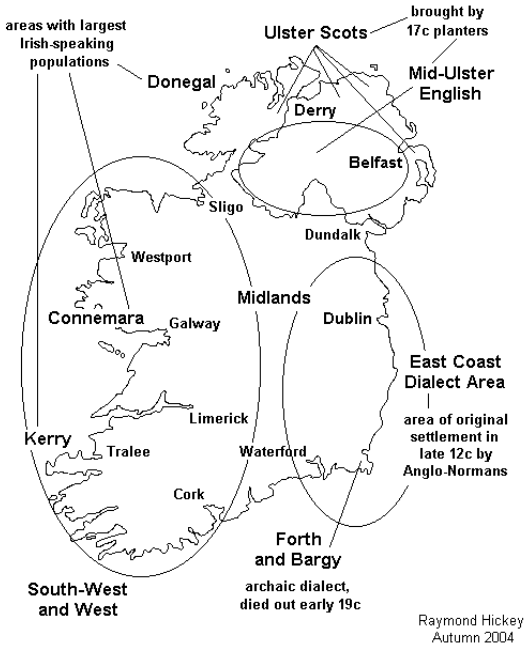

Figure 1: Accent Map of Ireland from Hickey (2004)

2.2 Preprocessing

For deciding on the signal processing of the sound inputs in this research study, the starting point was the treatment of the data in dsicovering the metadata, graphing and transforming the sound files in [Chapter 16]Jurafsky & Martin (2024). These were compared with the the data preprocessing and signal processing used in published data analysis studies that have chosen as suitable precedents.

Sound files are usually made availabe in datasets such as SAIE as **.wav** files, or less often as **.mp3** files. Each dataset described above is documented as being in one format rather than mixing. On loading, these can be graphed showing the signal against time, as shown below in the Exploratory Data Analysis sub-section3.5.

After that, the convention since the early work in the field has been to process the signal to generate the Mel Frequency Cepstrum Coefficients (hereafter 'MFCC', whose calculation was described in CA2) for time periods during the recording Jurafsky & Martin (2024). The MFCC gives a 2-dimensional image of the spectrogram to use as input for data analysis models. Also typically used in the literature are the first and second time differences of the MFCC, Hinton et al. (2012).

Variations around this basic approach with different parameters for initial processing of speech samples to prepare them as inputs to data analysis are used in the literature.

One example is graduate thesis for Stanford's machine learning course CS229 by (Sheng & Edmund 2017)[9] or another later project by (Andrea J Scott and James Thieu and John M Wall 2017), They build classification models for English language speech with different accents and get increasingly good performance as they move from using decision trees to multi-layer perceptron ('MLP') and then convolutional neural networks ('CNN') models.

They begin by cutting up the recording files to segment out the words separately, abandoning the silent periods and keeping the peak energy periods of the sound wave instead. Their code then either padded out with silence or cut down the remaining speech sample to give a 1-second long time window for the MFCC calculation. The simpler approach in this research, to segment each speaker's recording into segments of 1-second matched this time horizon. The **librosa** Python library was then used produce 50 MFCC values, compared with the librosa default value of 20 MFCCs used in the research project, which, with a sampling rate of 22,050 Hertz for the SAIE WAV files generated MFCCs as arrays of MFCC's, each of shape (20,44). This approach and their code seems to have come from a Stanford project for classifying animal calls, which would not have the informaiton content of human speech in

[9]Google Scholar = 14 citations, none on Scopus, but was consistenly the top result when querying Google Search for 'accent classification' together with 'machine learning' or 'deep learning': This might reflect a particular approach popular among Stanford's ecosystem of researchers and start-ups such as `sanas.ai` which provides a tool for changing accents spoken with one accent to synthetic speech with another accentShoichet (2021)

which potentially the whole sequence of data is important, so this approch was not applied in the data analysis here.

The paper used the MFCC length as a hyperparameter, using 30, 50 and 75 per second of sound recording data. A more common and hence easier to compare option would be to 20 per second of speech, drawing on the default values of the MFCC estimator in librosa library. Data was divided up in an 80/20 ratio of training to dev and test observations, which we followed here also.

Preprocessing in a later speech processing data analsis by (Jiao et al. 2016) used MFCCs calculated over 25 milliseconds (ms) of their recordings as inputs. Any silent periods of 300ms or longer were removed from samples.

For data augmentation, (Sheng & Edmund 2017) used Gaussian noise added to the original dataset to double the number of observations. However, SAIE datasets, being recorded quickly and usually in public places, are already noisy and have unpredictable timing gaps, so this technique of adding extra noise was not used in our research.

Preprocessing by Abdel-Hamid et al. (2013)[10] used the MFCC and its first and second deltas on windows of 40 samples of 25ms each one second total.

Lesnichaia, Mikhailava, Bogach, Lezhenin, Blake & Pyshkin (2022)[11] used other features as well as MFCC, and showed improved performance. However, the research here follows the most the most common approach was to use the python **librosa** library, for a faster implementation of the data analysis and with the added benefits of avoiding extra design or coding errors and better comparisons with the published research to evaluate our data and models.

The transformer-based models such as those with the **wav2vec2** and others Mohamed et al. (2022) Zuluaga-Gomez et al. (2023) architecture incorporate the unsupervised learning as the first stage of processing the raw sound waveform data, so no MFCC or other calculated measure is present in the input data, but instead features are calculated by the model.

2.3 Machine Learning Models

Two textbooks in particular gave a detailed account of the evolution of data analysis for automated speech processing applications such as this research. The newest edition of the textbook *Speech and Language Processing* by Jurafsky & Martin (2024) covers methods applied to speech processing in detail starting from the origins of the field in the 'sixties. Complementing this, Geron (2022) outlines the machine learning and deep learning models and the history over time of their applications in different areas, including audio and speech processing.

For accent classification, the earlier less-sophisticated machine-learning models have long been of use as a baseline for comparing performance to deep learning models, so these will be ued here.

Sheng & Edmund (2017) first run ensemble decision-tree methods, gradient boosting and random forests with their MFCC calculated from speech observations as inputs to classify their speech samples into one of three accents, reaching a respectable level of accuracy over 69% compared to 80% and 88% for the MLP and CNN respectively. As they then comment, this is less work to set up and quicker to run than their other methods, and so was considered worthwhile

Schuller, Steidl, Batliner, Hirschberg, Burgoon, Baird, Elkins, Zhang, Coutinho & Evanini (2016)[12] used Support Vector Models (SVM) to classify speech samples among 11 different accents among English as a Second Language learners, with 5132 distinct speakers. The deep learning solution by on the same accent classification task by Jiao, Tu, Berisha & Liss (2016) in contrast outperformed significantly by 45.1% to 52.2%.

With these precedents in mind, the research project also chose to start estimation with a simple and quick method, beginning estimation by using logistic regression to classify the accent in the sample. Besides generating a first-attempt model, this is intended to also help in evaluating potential input features, whether the demographic factors, genders and ages, or the three MFCC features calculated as part of the preprocessing.

2.4 Deep Learning Models

Since they are theoretically capable of estimating almost any functional form, linear or non-linear, including with interactions between the input variables [Chapter 5.1]Bishop (2006), feedforward neural

[10]Scopus = 239, GS = 506 citations
[11]GS = 6, Scopus = 5 citations
[12]Scopus = 248, GS = 371 citations

networks are an obvious choice for analysing speech data.

Important new techniques are described in detail in original papers in the deep learning literature such as (Bengio et al. 2006) and (Hinton et al. 2006) on unsupervised learning and (Hinton et al. 2012) applying this to speech processing, but the searching as part of this literature review points to there being a lag before new techniques in deep learning were applied to data analysis for accent classification.

Hinton, Osindero & Teh (2006)[13] recommended labelling input data for deep learning through unsupervised learning, recommending input data being fed into sequential layers of Restricted Boltzmann machines (RBM) to make up Deep Belief Networks: Once estimated, these layers can then be incorporated and the features elictited into neural network architectures for further estimation such as that needed for classifying accents, a method that reduced or eliminated the need for slow and expensive labelling of input data by humans needed for supervised learning [pp.377-8]Geron (2022). Soon after, Bengio, Lamblin, Popovici & Larochelle (2006) proposed using stacks of autoencoders instead, which has become the common practise and has shown consistently more stable training for the neural nets, with less of the problems with vanishing gradients that bedevilled deep neural network structures previously [pp.377-8]Geron (2022).

One advantage of the autoencoder structures and tools over the RBMs is their being easier to understand for those without a deep mathematical or physics background. A further hope could be that such estimation might find patterns outside the MFCC typically used, perhaps including those not easily perceptible to humans or exploited by linguistics research in the past. A problem, however, might be interpreting the features generated in this process, which may not be easily understood to humans or even observable by them in listening to sound data.

In a survey by Hinton et al. (2012) done with data and modellers shared by Microsoft, IBM and Google, these unsupervised learning approaches added to architectures led to improvements in 10 to 20% in the Word Error Rate measures on five speech recognition tasks over the the leading HMM solutions. Not surprisingly, this approach has become increasingly common over time, as surveyed by Mohamed et al. (2022), particularly for smaller and less-spoken languages.

Experiments with different neural network architectures by Maas, Qi, Xie, Hannun, Lengerich, Jurafsky & Ng (2017)[14] showed that increased size was associated with greater power to represent input data features with deep learning models typically using 10 to 100 times more parameters than for HMM ones. Furthermore, they reported unsupervised learning not absolutely necessary for good performance. Convolutional neural network structures, they argued, were likely to perform better than MLP because, like in image processing, the models see repeated patterns, with a limited number of basic sounds being repeated in speech. Integrating more techniques from the deep learning literature such as dropout regularisation and momentum optimisation also led to better estimation performance in speech recognition. However, the demands of greater model complexity and training time were an unavoidable drawback, as were their reliance on more powerful processing, typically moving to GPU processing, and memory. However, by this time, the integration of speech processing tasks with the mainstream of deep learning research and the benefits from this, was obvious.

2.5 Transformer and Large Audio Models

In parallel with deep learning models that used increasingly large sequential data series as inputs, the models used for automated speech processing continued to work with larger input samples and longer sequences Mohamed et al. (2022). The best performance on accent classification tasks in recent years has usually involved architectures built on top of **wavenet** models Mohamed et al. (2022), Zuluaga-Gomez et al. (2023) combining a number of recent innovations. The revolutionary transformer architectures introduced by Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin (2017) provide much greater flexibility in estimating relationships at lengths all across an input sequence. **wavenet** introduced the use of stacked CNN layers with dilation, where a convolution is used with some values skipped in the sequence, increasing efficiency in estimation van den Oord et al. (2016), as shown in figure2 below. The original **wavenet** achieved State of The Art results on speech recognition, better than any published before, for their input dataset van den Oord et al. (2016). As was becoming increasingly common by then, inputs used were the raw sound files, without features such as the MFCC calculated during preprocessing van den Oord et al. (2016), Mohamed et al. (2022).

[13] Scopus=12,296, GS=20,955 citations

[14] Scopus = 95, GS = 179 citations

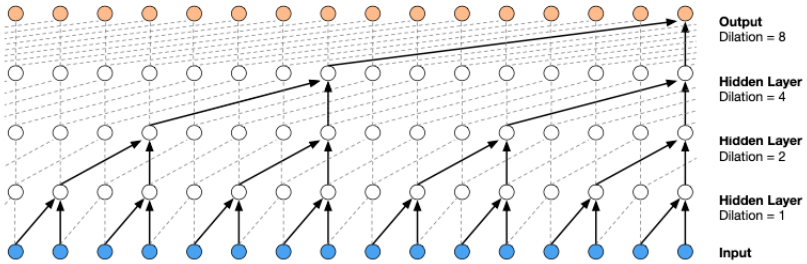

Figure 2: WaveNet architecture showing dilated input (van den Oord et al. 2016, Section 2.1)

Viglino, Motlicek & Cernak (2019)[15] and (Mohamed et al. 2022) provides a survey of datasets, architectures and research on accent classification using these newer deep-learning techniques. The best performance on accent classification tasks in recent years has usually involved architectures built on top of **wavenet** (van den Oord et al. 2016) models using stacked CNN models to provide inputs, combining transformer architectures introduced by Vaswani et al. (2017).

Baevski, Zhou, Mohamed & Auli (2020) adapt this architecture for the speech recognition task to create the **wav2vec2** model. Masking, the selective dropping of items from the input sequence during training [pp.590–3]Geron (2022), is used to improve the generalisation during learning. Connectionist Temporal Classification (CTC) is used in training, maximising the probability of the distribution of outputs, conditional on the inputs rather than measuring performance directly on a task Graves et al. (2006)[16].

The **CommonAccent** model of Zuluaga-Gomez, Ahmed, Visockas & Subakan (2023) uses variants of the **wav2vec2** models in accent classification. As is increasingly common also, the models are increasingly training with multimodal-data, which might be visual data of the speaker talking [§III E] Mohamed et al. (2022), training on multiple languages at once [§2, 4.2] Zuluaga-Gomez et al. (2023) and to train for multiple tasks such as speaker identification with accent classification at the same time [§2.2] Viglino et al. (2019). This variant of the **wav2vec2** model produces what appears to be the produces the highest classification accuracy of any of these examples, at over 97%. While demanding in terms of technical setup, of data required and the resources needed, this model looks like an excellent candidate for use in this accent classification research project.

2.6 Evaluation

While some of the studies above refer to transcription accuracy measured by Word Error Rate, this metric cannot apply to whether an accent on speech, a purely sound-related feature, is correctly classified. In general, the metrics for evaluating the classification models in the research cited did not vary much from those commonly used in the literature e.g. in [Chapter 6](Raschka et al. 2022) such as accuracy, precision, recall and their harmonic mean, the F1. Confusion Matrices were reported and the Receiver Operating Characteristic (ROC) curves plotted and the Area Under the Curve (AOC) measures calculated from them.

Related to the evaluation by the classification model is the question as to how we can evaluate the human performance in determining the accents apart. Some simple tests, by selecting recording files manually at random, researchers found accuracy in classification by humans to be almost perfect, with one attempt in 20 failing. However, this was with a Dublin native familiar with both cities. During the early stages, after CA1 but before CA2, an Amazon Mechanical Turk proposal was put out to hire people in Ireland to do the accent classification manual, but there were no useful responders, those answering being either too aged, 65+ and thus with hearing unavoidably deteriorated or were recent residents of Ireland showing broken English. So, unfortunately, there is no reliable human benchmark in evaluating the modelling in this research project.

[15]Citations Scopus = 50 , GS = 73
[16]Citations Scopus = 1261, GS = 6828

3 Research Method and Specification

3.1 The Research Aims and Objectives

This project aims to contribute to an active body of research into the features and classification of speech samples in English in Ireland and how this speech can be classifed by geographic origin by the properties of the speech audio or data available about the speakers.

Representative speech datasets and accurate analysis of accent data is needed to make services accessible and error-free for all users without misunderstanding their speech. To this end, I will estimate machine-learning and deep-learning models, applying the State Of The Art (SOTA) models used in previous research for accent classification with data from other countries.

Furthermore, analysis of our accents has direct applications in social understanding, a central theme in Irish culture, so highlighting geographic differences, which are often celebrated but also may be a foundation for bias and discrimination.

If differences can be detected by automated data analysis tools, then both the opportunity for self-knowledge and the risk of discrimination are more probable to occur, in the same way that benefits and perils that come with data generated by individual human genetic testing.

3.2 Data Analysis Methods

The data analysis methods described here are designed to translate our research question into models that answer it. The data analysis has 8 stages in total, all falling under the Knowledge Discovery in Databases (KDD) (Fayyad 1996) process that was specified in **CA2** as being best describing this project. The tasks and sequence of the data analysis is described and justified in this section and each activity in the data analysis outlined in detail in the following selections.

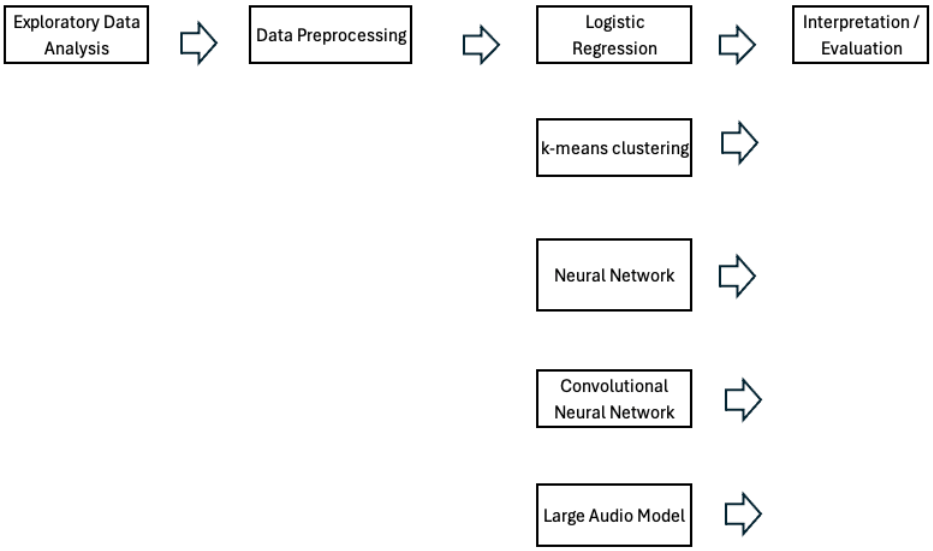

Figure 3: Research Project Data Analysis Process

Our target variable, the dependent variable, is to be the a binary flag, one value for Dublin city and county, another for Belfast for that city and its surrounding county. This flag by the county variable merges cities with the surrounding county, Dublin city locations are included in the Dublin county examples. Belfast is included together with Antrim. In the models this is coded as both a county code, either DUB and ANTBEL in the variable *county* and a binary numerical variable, 0 and 1 respectively in the variable y in the models, as described in more detail below. If a common regional accent exists and is observable from the speech data, this should be present in the speech recordings taken in one location across all speakers, but not in those recordings outside that county and city geographic region.

The task of the classification models here are to explain the differences in this target variable by the candidate independent variables are those that are available. Our sound signal is taken from the SAIE dataset, so either the raw signal or values calculated from it can be used as independent variables. X variables can include the MFCC calculated, as the case studies above do in almost every case. The

time-differenced values, which are labelled as $MFCCdelta$ and $MFCCdelta2$, are also used in our data analysis.

Other variables are tied to each of the speech sample recordings also, coded as an item in the file name Hickey (2004). These include a speaker identity in the form of a number, a narrower geographic identifier at the town and neighbourhood level, an urban/rural binary flag, the gender and approximate age of the speaker. Since these are available from the SAIE dataset, they can be considered also as candidates for inputs as independent variables in our models.

The dataset is described more fully in §3.5 below, in the Exploratory Data Analysis, the first stage of the KDD process, Selection.

The second stage of the process, how the raw dataset processed and then converted into the inputs used for the data analysis models is then described in the following section §3.4, relating to the second stage of KDD, Selection and Transformation.

The third stage, the Data-Mining, involves estimating and evaluating a series of models. These are described here in a sequence roughtly chronological with the case studies from the published literature, going from the earlier and simpler models to the later and more complex ones, but could be run in any order.

The first data analysis method applied was logistic regression, in a series of models incorporating different variables, both the MFCC measures and the age and gender bands, with the location as the dependent y variable. Classification metrics including an accuracy, precision and recall and the F1 statistic, were calculated, and a Confusion Matrix populated. The Receiver Operating Characteristic (ROC) curve, was also estimated and plotted for the logistic regression models, with the Area Under the Curve, its AUC statistic, calculated also.

Logistic regression were chosen as among the simplest and quickest of the classification models to be run and were reported in some case studies above such as Lesnichaia et al. (2022).

Also, the repeated trials with different variables were intended to help with choosing from among the features by observing the impact of adding the variables on model classification performance, a step-wise selection like those done in [Chapter 6.1] James et al. (2021): The logic is working from simpler models with fewer input independent variables, then adding more with later models where justified, a forwards approach done manually.

Handling, mainly by reshaping, casting and flattening the MFCC data, with the (20,44) array for each 1-second sample file among a total dataset of over 21,000 such observations, was not an intuitive process and proved the most time-consuming and error-prone element in this and the other data analysis tasks in the project. With the form of the MFCC data, not easily comprehensible in numerical or graphic form, a key advantage of decision-tree or ensemble methods, the ability to elicit human-interpretable rules from the data, becomes much less likely, so these models were not estimated for this data analysis.

A second stage of data analysis is on the same notebook as the logistic regression, where the MFCC data is calculated and graphed on 2, 3 and 5 k-means clustering to detect commonalities among the speech data by geographic location.

On the other hand, feedforward neural networks are theoretically capable of estimating almost any functional form, linear or non-linear, including with interactions between the input variables [Chapter 5.1]Bishop (2006). They were the obvious approach to follow next in analysing our data, as used in some later case-studies in the literature such as Sheng & Edmund (2017).

Further beyond that, Convolutional Neural Networks (CNN), given the repetition at different times and different scales of the graphic patterns associated with different units of speech, were then applied and evaluated.

Finally, to explore the most recent models similar to those in the Large Language Models performing well in other domains such as the OpenAI (2023) family, a Large Audio Model Zuluaga-Gomez et al. (2023) claiming the current State of the Art (SOTA) in performance in accent classification tasks was attempted. First, the pretrained models were to accessed through the `HuggingFace.co` API[17].

With so many parameters and such large datasets, CommonAccent using the Ardila, Branson, Davis, Henretty, Kohler, Meyer, Morais, Saunders, Tyers & Weber (2020) dataset that is over 88 GB in size, training of the full models was expected to be much slower and more difficult, but was also attempted.

[17] `https://huggingface.co/Jzuluaga/accent-id-commonaccent_ecapa`
`https://huggingface.co/Jzuluaga/accent-id-commonaccent_xlsr-en-english`

3.3 Project Design

3.4 Preprocessing

Causality is specified here to flow from the MFCC sound, age and gender to the accent location indicator. Estimation for the logistic regression, neural networks, convolutional neural networks is a supervised learning problem, as the true values for our classification are already known. The clustering models are unsupervised learning to aid our understanding of the dataset. Our Large Audio Model will take the raw sound data and elicit features through unsupervised and semi-supervised learning before applying supervised learning as previously.

Two Jupyter notebooks named **PRE 0_73 DUB.ipynb** and **PRE 0_73 ANTBEL.ipynb** are interactive scripts for the data preprocessing and should be run first, in order to check the SAIE input and provide the one-second sample input data for executing the EDA data analysis.

The files require the **librosa** and other Python libraries as shown in the **requirements.txt** for the kernel specified in the file and shared with this project.

Users need to set three global variables. These are DIR_PATH, the full explicit path for input files, the recordings files for each speaker for one of the counties from the SAIE dataset, then $DATASET\_NAME$, the region the data relates to, so either 'DUB' OR 'ANTBEL'. Finally, $SAMPLE\_LENGTH$ is the user-specified length in seconds for the sample files into which we split the recordings by time, usually set here as 1 second.

librosa appears to have undocumented bugs in that it will not return the duration of the input files correctly. Also, some files show up in from tjhe SAIE dataset **.mp3** format and not **.wav** and need conversion. Other data relating to the sound recording is needed for calculations in the models, such as the Sampling Rate, the number of data values per second of sound recording, measured in Hertz (Hz), which was 22,050 for every recording: This was preserved in creating the speech samples.

The script retrieves the explicit file paths and names, saved as $filename$, extracts *name* the file names, then extracts the *county*, *gender* and *age* which are saved down for each recording.

The Urban/rural flag is included in the SAIE recording file title names in only some cases to describe rural areas. This is not used, as the simplfying assumption is madethat all accents in the county are part of the larger urban accent rather than a distinct one relating to a locality. Size is sometimes populated for these rural records also, but was not seen as usable.

If present, the *town*, *urbanrural* flag and *size* are retrieved, although there is no use for these in the models. Location is available at the *town* level also, used for smaller towns, a district of the city, or the city name itself, based on where the recording was made. These will be correlated perfectly with the county labels, so that Tallaght for example, a southwestern Dublin suburb, will always be assocated with Dublin county and no other county, so there is no additional information likely to be had from adding this variable unless there was enough local speech data to allow meaningful modelling: there are only 10 recordings from Dalkey, the highest number for any town value within the Dublin dataset, which we judged unlikely to be enough samples to generate meaningful insights, so county was the lowest level of aggregation of the data analysis.

The sound wave graph can be printed for the full recording file, along with a calculated spectrogram for the power and the $MFCC$, $MFCC\_delta$ and $MFCC\_delta\_2$. One of each for the first sample file is shown at the bottom of each notebook.

As discussed in the preprocessing section of the literature review, the analytics are usually calculated on small chunks of speech of a second or less, with each MFCC calculated for betwen 25 and 75 ms. In this script, that logic is implemented with sample files that are created by cutting lengths of the speech recording of $SAMPLE\_LENGTH$ seconds, discarding a surplus at the end of less than that length, then calculating the 3 MFCC features with this subsample sound data, then assigning the other data fields' values taken from the full recording to the sample also, such as *age* and *gender*.

All the variable values are saved down to the *df* Pandas DataFrame, which is then exported as a pickle (**.pkl**) file for later use in the data analysis, named with the script runtime and the database name variables.

The samples sound data is saved down as a **.wav** file using the original recording name plus a sequenced recording number unique to each recording, plus a sequenced sample number unique to each sample.

The sample sound can be heard by playing the audio player, and also a soundwave graph, the 3 MFCC feature spectrograms and the power spectrograph is displayed at the end of each notebook.

3.5 Exploratory Data Analysis (EDA)

Jupyter notebook **EDA.ipynb** holds calculations and graphics related to Exploratory Data Analysis for the SAIE dataset.

This requires the output of the preprocessing, namely the one second samples in **.wav** format in the directory **sample_output_directory** and the DataFrame pickle file **.pkl** with the preprocessed data and metadata.

The count of recordings by geography is noticeably unbalanced in favour of DUB records, at over 72% of the total.

Age is defined in a set of age bands, making the survey quicker. Overwhelmingly, the most common age band is 20 for all geographies.

Gender is close to balanced for the set of records used in the data analysis, but shows more men than women in the ANTBEL and more women in DUB records.

3.6 Logistic Regression

A Jupyter notebook containing the logistic regression and clustering models named **clustering12.ipynb** is incorporated in the project package.

To begin with, the pickle files are loaded according to names entered in cell 16 of the Jupyter notebook. These are retrieved from the $INPUT\_DIRECTORY$ variable value, which is normally one subdirectory of the script directory named 'sample_output_directory' and contains the sample files and the pickle files.

Preparing to merge the two regional datasets, we load the pickle files, count the recordings in the first DataFrame and renumber the recordings in the 'recording num' field in the second DataFrame so that each value will be unique. The two are then merged into the combined **all_df** dataframe.

In the 'counties' field, all values of 'ANT' or 'BEL' are replaced with a common value of 'ANTBEL', so that only two binary values remain, 'DUB' and 'ANTBEL' remain. Counting the records in each shows how the two classes are not present in equal numbers and so we need to account for imbalanced classes when estimating our models.

As discussed in §2.1 above, best practise will be to split out the training and test datasets with no speaker in common between them, so that the common accent characteristics of speech are classified rather than any data analysis instead simply match the same speaker in both datasets. The datasets are shuffled based on an 80% split by speaker numbers in this way and $train\_df$ and $test\_df$ DataFrames created to hold these training and test datasets. The target independent variable was split out as y with a binary 0 or 1 value for the Dublin or other region. However, a problem with implementing this has been that random seed seems not to set the **np.random.shuffle** to return the same results each time it is run: This remains an outstanding issue at the time of writing.

Two DataFrames were created, $Xtrain$ and $Xtest$, from the values of the $MFCC$, $MFCCdelta$ and $MFCCdelta2$ variables. Also, the gender and age categories were one-hot encoded and added to these DataFrames.

The first logistic regression model was run with the county binary variable, y as the dependent and the one-hot gender and age variables as the dependent variables. We would expect gender to have a distinct effect on speech, with women generally speaking with higher frequencies and men lower. Similarly, accents might be expected to be stronger in those of school age, weakening when exposed to the workplace, or be different among different age cohorts in Dublin as Hickey (2007) describes.

Running the logistic regression from **scikit learn**, with parameters including 'class_weight=balanced', we see a high percentage of those samples classified as 0, Dublin, are correctly so, with accuracy over both classes of 82.7%. From the confusion matrix, we observe the classification success for the Belfast area samples is much worse, with more than half of samples being incorrectly labelled by this classifier. Overall, the ROC AUC is just 69%, so there is certainly a chance to improve classification by adding more variables.

In model 3, we add run the logistic regression just on the first simple MFCC variable observations as the independent X variables. The observations are flattened from (40,22) to (880,) and scaled before the regression is run. This returned slightly worse accuracy from the Dublin samples, but much better classification results for the Belfast region, with 890 of 1072 being correctly classified. AUC was also better for model 3, at 89.7%.

Model 4 uses both MFCC and the categorical variables as independent variables, improving AUC to 91.6%, suggesting an improvement by adding the extra variables.

Model 5 uses only $MFCCdelta$ with age and gender, but results worsen, particularly for Belfast classification, with AUC falllign to 75.9%.

Model 6 uses $MFCCdelta2$ only but with age and gender. Results are among the worst of any of the models, with AUC of 76%.

Model 7 throws in gender and age and all three MFCC measures, performance coming close to the best accuracy and AUC measures seen before on of model 4 at 91.4%.

From this, one might conclude that both the MFCC, the age and gender variables have significant explanatory power. Adding the $MFCCdelta$ and $MFCCdelta2$ seems not to improve performance much, so that these variables may add little in classification power in return for what they demand of preparation and processing time. Comparing to simple case studies such as the Support Vector Machine approach of Schuller, Steidl, Batliner, Hirschberg, Burgoon, Baird, Elkins, Zhang, Coutinho & Evanini (2016), with accuracy under 50% and the accuracy reported for decision-tree and ensemble methods around 69% by Sheng & Edmund (2017), these logistic regression models look seem intuitively correct in the choice of independent variables and show themselves better performing.

3.7 Clustering

In the same Jupyter notebook, Models 8 and 9 applied k-means clustering from **scikit learn** to see if there are any clear partitions in the MFCC (8) and MFCC plus age and gender (9). No clear pattern of clustering was observed using either dataset with either 2 or 3 centroids as shown in figure 4.

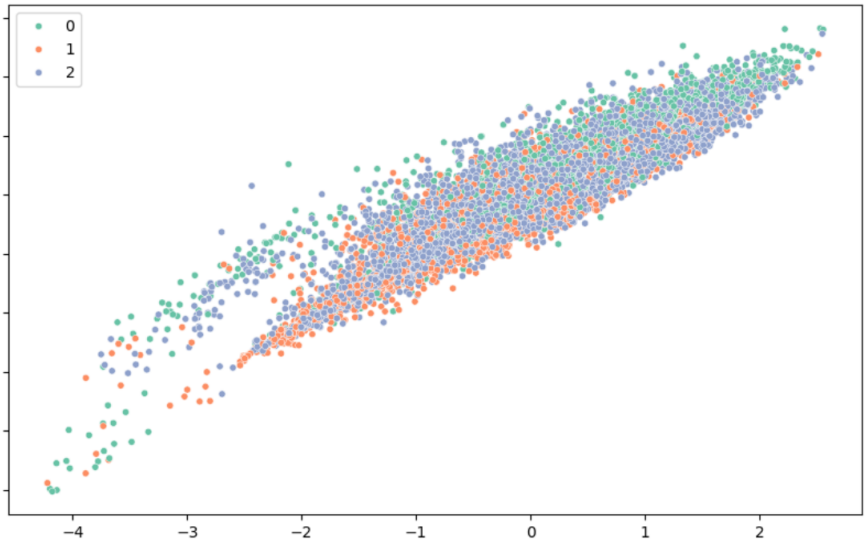

Figure 4: k-means, n=3, MFCC+age+gender

3.8 Neural Networks - Multi-Layer Perceptrons (MLPs)

The next stage of the data analysis was estimation of some relatively simple neural network models. These were implemented in Keras and Tensor Flow and the models and results are contained in the Jupyter notebook **TF8 MLP.ipynb**.

Again the DataFrame pickle files were loaded in the script. Preprocessing continued as in the logistic regression analysis in §3.6 above, with the datasets merged, the recordings count reset, a train/test split done on the recording number so as not to split speaker sample.

Then, drawing from the analysis used in model 9 above, all 3 MFCC related input variables were included, all being flattened as above. One-hot representations of the age and gender features were again added.

Following the examples in the Keras documentation[18] the Keras Tuner was set up to iterate by random search through hyperparameters of a three layer network. ReLU activation units were used between layers except for the logistic activation in the final layer, with loss being binary cross-entropy to support the binary classification task. The learning rate for the Adam optimiser also chosen randomly. An optional 4th layer also open to inclusion by random choice by the Keras Tuner. Some model variants were also tried with just 3 layers, some others with droput layers were also included.

[18] e.g. `https://keras.io/guides/keras_tuner/getting_started/`

The best models after training tended to produce training set accuracy percentages in the nineties and test accuracy from 82 to 89 %, better than prior expectations and better compared to prior case-studies with the MLP e.g. around 80% for Sheng & Edmund (2017). AUC near 92% was a typical value. Again, these point to the MLP with the 3 MFCC variables and the age and gender variables as providing good classification performance on our SAIE dataset.

3.9 Convolutional Neural Networks (CNNs)

Data analysis proceeded with estimation of convolutional network models. As with MLP, these were implemented in Keras and Tensor Flow and the models and results are contained in the Jupyter notebook **TF10TK CNN.ipynb**.

The data preprocessing setup for convolutional neural network models was much the same as for the MLP, with the three MFCC measures used as inputs to the classification model. However, problems arose in attempting to incorporate the categorical age and gender variables as inputs also, which seemed to require a custom solution using the Keras API, but this was not run successfully without execution errors within a work period of almost a complete two-week sprint, so the specification was scaled-down, with considerable regret.

Data analysis proceeded with the CNN model using an architecture with three pairs of 2-d convolutional layers as a starting point, following from that used by Ahamad et al. (2020). Performance matched or exceeded that on the MLP structures, reaching test accuracy percentages in the low nineties and AUC of over 96%, so performance coming close to the claimed State of the Art (SOTA) of Zuluaga-Gomez et al. (2023).

3.10 Large Audio Models - CommonAccent

The final model targeted for implmentation during the Data Analysis was the CommonAccent model of Zuluaga-Gomez, Ahmed, Visockas & Subakan (2023), which incorporates many of the newer features of the best-performing architectures for speech accent classification.

However, the most obstructive and difficult engineering problem encountered during the project has been that the CommonAccent installation was not achieved within the project deadline. In repeated attempts over more than the fortnight period of a project sprint, installation of the model libraries, so as to train it on the most recent CommonVoice Ardila et al. (2020) datasets or use it in transfer learning, was not successful. Whether on Google Colab, either version of Windows 11 or Debian or Ubuntu running in the Windows Sub-system for Linux (WSL), repeated Python library incompatabiilties or bugs prevented installation, regardless of the environment and the computer used. As a fallback, the Common Accent API on HuggingFace was sought out instead as the candidate transformer-based architecture for the classification, but this looks to be both broken and unsupported[19]

As an alternative, the HuggingFace API for another model architecture used in the CommonAccent paper was accessed[20] instead. The model is implemented in the Jupyter notebook named **CommonAccent3 Classifier.ipynb**. The model had been trained using the CommonVoice data and queries could be made using an API with a slim installation from **HuggingFace**. The installation was tricky and, unusually, require an Anaconda Explorer environment running with administrator privileges in Windows. This was done and the supporting libraries installed. A number of speech samples come with the model and these were tested with the API and were classifed correctly.

The next step was to use the project speech files as inputs for the API classification model. A logical test would be to have this model classify the sample files used in our data analysis. However, on feeding the full set of over 21,500 samples from the Belfast and Dublin regions from the SAIE dataset, the model returned a classification as Irish on only some 651. So, barely 3% of the samples of Irish speech were recognised by the model as Irish, a surprising and an extremely poor result for this model. This may be explainable by differences in the sound quality between SAIE's recordings in public places and CommonVoice samples gathered by phone or headset in a quiet environment, which is apparent on listening and comparing recordings from both. With the result and the difficulty in executing the API model, further work on them promises little and another candidate Large Audio Model should be used for data analysis in future instead.

[19] `https://huggingface.co/Jzuluaga/accent-id-commonaccent_xlsr-en-english#limitations` .
[20] `https://huggingface.co/Jzuluaga/accent-id-commonaccent_ecapa`

3.11 Ethics

Two ethics forms have been submitted previously, each relating to the datasets that were the main focus of effort, the IViE Nolan & Post (2014) and the SAIE Hickey (2007). As described there, the research and the data had already been published in the linguistics literature some years ago and were available online and in university libraries, with published guidance that these were available for the use of non-commercial research. In addition, the lead author for both datasets had been contacted and gave permission and guidance on the project. Ethical factors and the impact of current Irish data protection regulation were already described in CA1 and CA2 also, in particular the extreme difficulty of using the speech samples to reidentify the original speakers given that their location and identity was concealed by the inexact data available on each, namely town, age and gender in the SAIE dataset. None of the data analysis in this research project would contribute towards deanonymisation.

No identification — names, photographs, or other such personal details -– was taken from the speakers contributing voice samples to either dataset. This and the time elapsed since the samples were taken suggests that the risks to privacy by identifying the speakers or by reusing the sample data or synthetic versions of it that could be used for fake speech output or in breaking security tests based on voice, is likely minimal.

For these reasons, no ethical risks were expected to obstruct the project or cause problems for any stakeholders in it, including the test speakers.

3.12 Technical Setup

Technical setup for the data analysis project involved choices over the tools to use.

Python was an obvious choice to use, given the huge choice of libraries for data analysis. A quick search at the start of the literature on speech processing gave the impression that **R** was comparatively rarely used and that Python dominated published research in the field. Given the requirements for handling and processing large data volumes, the use of scripts rather than GUI tools and of specialised libraries such as **scikit-learn**, **NumPy** and **Pandas** was considered more suitable. As yet, there seems no decisive majority in the literature using one the two main deep learning libraries over another, **Keras** with **TensorFlow** or **PyTorch**. Given past familiarity, TensorFlow was preferred. The **librosa** library is very widely-used in speech processing research, much more so that TorchAudio. With **TensorFlow** already preferred, the choice in favour of **librosa** flowed from that.

Coding and analysis was carried out mainly on two machines. One was a Dell Inspiron with a Intel Core i5-12400 2.50 GHz and 8 GB of RAM running Windows 11 Home. The other was a custom build PC with an Intel Core i9-14900KF and 32 GB of RAM and a NVIDIA GeForce 4090 GPU, running Windows 11 Pro and Debian on bootup and in WSL.

3.13 Project Management and Project Implementation

As described in **CA2** this project was understood to be run under a project plan based on an Agile *The Manifesto for Agile Software Development* (2001) approach, with sprints starting and ending on each fortnightly meeting with the project supervisor.

Unfortunately, the estimate and timing efforts for the tasks and the sprints made beforehand were very inaccurate throughout the project with repeated issues running unsolved over sprints.

Getting access to the datasets was later than anticipated. While IViE was available during the preparation of CA2, the SAIE recordings CDs were missing from the TCD library until Prof. Hickey was contacted and shared the data.

There were repeated problems with the sound libraries, **torchaudio** throwing warnings or proving incompatible with Python libraries. **librosa** had a number of undocumented issues, being unable to retrieve the sampling rate or duration on the SAIE **.wav** files and with them being read as **.mp3** and needing conversion.

As discussed above, there were issues during implementation which remain outstanding. These failures largely those of coding to support the data engineering, as with incorporating the age and gender variables into the CNN models and in successfully implementing the CommonAccent models.

3.14 Conclusions and Future Work

In this report, we repeated our research question i.e. How can we classify some selected regional accents in Ireland using data analysis techniques based on the features within the sound of their speech and on

demographic characteristics of the speakers?

The Literature Review above surveyed the research published prior to now to evaluate how this question could be answered. In the Research Method Section above, these choices were applied in practical applications of data analysis in three areas.

First, what datasets may be available that can support data analysis of sub-populations within Ireland? The evidence here, where the SAIE dataset has been explored in depth and preprocessed into a form suitable for model input, shows that this likely provides a solid foundation for such models. The extensive documentation on the data-gathering process, its wide use in the linguistics research literature over twenty years, number of samples, the geographic spread all over Ireland, a large number of individual speakers, and the clarity of the speech data in spite of challenging recording conditions, the data protection afforded by anonymity to the speakers and the data's availability for non-commercial and academic research all support this over alternative candidates.

Second, our models performed well on standard classification metrics, using most of variables. Starting with single variables and then adding more, performance improved, pointing to our calculated MFCC values (although less so the time difference values) and the age and gender variables as useful independent variables in classifying our speech samples by their geographic origin.

Third, most of our candidate models, logistic regression, MLPs and CNNs, performed better than expected and as well or better as case-studies in the Literature Review.

As a result, the research provides a significant contribution to the linguistics research literature in Ireland, pointing at new ways of using a popular existing dataset. The relative simplicity of the models has the unplanned effect also of making the research more accessible to those outside of the field of deep learning.

Moreover, with some relatively simple models having set a benchmark for performance with the data used, the agenda is set for future research to bring the most advanced techniques to bear on the Research Question.

Finally, while good for its time and proven useful here, the SAIE dataset stands in contrast to the much larger datasets gathered through online crowd-sourcing such as those of Ardila et al. (2020) gathered

Acknowledgements

The research question, choice of literature, modelling and conclusions are the sole responsiblity of the author.

Course director Dr. Catherine Mulwa provided guidance on the research and writing process that was comprehensive, clear and extremely practical. Guidance from supervisor Jorge Basilio was encouraging, informed and useful for this research.

Advice and access from the NCI Library staff was invaluable, as was the assistance in finding relevant resources in linguistics the library staff at Trinity College Dublin, including the sound records used extensively here. The author hopes that the work presented here has justified the assistance given by all of them.

Members of the research community provided valuable feedback also. Dr. Des Ryan of Apple Computer, a graduate of the M.Sc. Data Analytics in 2018 discussed data collection and ethics and he supported the choice of topic. Professor Francis Nolan of the Phonetics Laboratory of Cambridge University confirmed access and offered guidance on the IViE corpus, which he had originally published. Colin Flynn, Professor of Applied Linguistics at Trinity College Dublin commented on the research approach. A particular debt is owed to Professor Raymond Hickey of the University of Limerick, who very generously gave his time to discuss his dataset, the research approach used here and to suggest that this project was a valuable contribution to give back to the linguistics research community.

References

Abdel-Hamid, O., Deng, L. & Yu, D. (2013), Exploring convolutional neural network structures and optimization techniques for speech recognition, *in* 'Proc. Interspeech 2013', pp. 3366–3370.

Ahamad, A., Anand, A. & Bhargava, P. (2020), Accentdb: A database of non-native english accents to assist neural speech recognition, *in* 'Proceedings of The 12th Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 5353–5360.
URL: *https://www.aclweb.org/anthology/2020.lrec-1.659*

Andrea J Scott and James Thieu and John M Wall (2017), CS230 - convolutional neural networks for american accented english region localization and analysis", Master's thesis, Deptartment of Computer Science, Stanford University, Stanford, CA.
URL: *https://cs229.stanford.edu/proj2017/final-reports/5244230.pdf*

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M. & Weber, G. (2020), 'Common Voice: A massively-multilingual speech corpus', *arXiv preprint arXiv:1912.06670* .
URL: *https://commonvoice.mozilla.org/en*

Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. (2020), 'wav2vec 2.0: A framework for self-supervised learning of speech representations', *Advances in neural information processing systems* **33**, 12449–12460.
URL: *https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf*

Bengio, Y., Lamblin, P., Popovici, D. & Larochelle, H. (2006), 'Greedy layer-wise training of deep networks', *Advances in neural information processing systems* **19**.
URL: *http://papers.neurips.cc/paper/3048-greedy-layer-wise-training-of-deep-networks.pdf*

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer.
URL: *https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf*

Dolan, T. P. (1991), Language in Ulysses, *in* J. Genet & E. Hellegouarc'h, eds, 'Studies on Joyce's Ulysses', G.D.R. d'Etudes anglo-irlandaises, Université de Caen, pp. 131—-142.

Fayyad, U. (1996), From Data Mining to Knowledge Discovery: An Overview, *in* 'International Conference of Soft Computing and Pattern Recognition (SoCPaR)', AAAI Press / The MIT Press.

Geron, A. (2022), *Hands-On Machine Learning with Scikit–Learn, Keras and TensorFlow*, 3 edn, O'Reilly Media Inc.

Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. (2006), Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *in* 'Proceedings of the 23rd international conference on Machine learning', pp. 369–376.
URL: *https://www.cs.toronto.edu/ graves/icml_2006.pdf*

Hickey, R. (2004), *A Sound Atlas of Irish English*, number 48 *in* 'Topics in English Linguistics', Walter de Gruyter: Berlin.

Hickey, R. (2007), *Irish English: History and Present-Day Forms*, Studies in English Language, Cambridge University Press: Cambridge.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. 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.

Hinton, G. E., Osindero, S. & Teh, Y.-W. (2006), 'A fast learning algorithm for deep belief nets', *Neural computation* **18**(7), 1527–1554.

Hornsby, D. (2014), *Linguistics: A Complete Introduction*, Teach Yourself / Hodder & Stoughton.

Hughes, A., Trudgill, P. & Watt, D. (2012), *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles*, 5 edn, Routledge: Abingdon-on-Thames, UK.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021), *An Introduction to Statistical Learning with Applications in R*, Springer Texts in Statistics, 2 edn, Springer, New York, New York.

Jiao, Y., Tu, M., Berisha, V. & Liss, J. (2016), Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features, *in* 'Proc. Interspeech 2016', pp. 2388–2392.

Jurafsky, D. & Martin, J. H. (2024), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 3 edn, Prentice Hall: Hoboken NJ.
URL: *https://web.stanford.edu/ jurafsky/slp3/*

Lesnichaia, M., Mikhailava, V., Bogach, N., Lezhenin, I., Blake, J. & Pyshkin, E. (2022), Classification of accented english using CNN model trained on amplitude mel-spectrograms, *in* 'Proc. Interspeech 2022', pp. 3669–3673.
URL: *https://www.isca-archive.org/interspeech_2022/lesnichaia22_interspeech.html*

Maas, A. L., Qi, P., Xie, Z., Hannun, A. Y., Lengerich, C. T., Jurafsky, D. & Ng, A. Y. (2017), 'Building DNN acoustic models for large vocabulary speech recognition', *Computer Speech & Language* **41**, 195–213.
URL: *https://arxiv.org/abs/1406.7806*

Markl, N. (2022), Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition, *in* 'FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 521—-534.
URL: *https://www.pure.ed.ac.uk/ws/portalfiles/portal/266534040/Language_Variation_MARKL_DOA07042022_AFV.*

Markl, N. & Lai, C. (2023), Everyone has an accent, *in* 'Proc. INTERSPEECH 2023', pp. 4424–4427.

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L. et al. (2022), 'Self-supervised speech representation learning: A review', *IEEE Journal of Selected Topics in Signal Processing* .

Nolan, F. & Post, B. (2014), The IViE corpus, *in* J. Durand, U. Gut & G. Kristoffersen, eds, 'The Oxford Handbook of Corpus Phonology', Oxford University Press.

OpenAI (2023), GPT-4 technical report, Technical report, OpenAI.
URL: *https://cdn.openai.com/papers/gpt-4.pdf*

Oxford English Dictionary (2023).
URL: *https://www.oed.com/search/dictionary/?scope=Entries&q=accent*

Raschka, S., Liu, Y. H. & Mirjalili, V. (2022), *Machine Learning with PyTorch and Scikit-Learn*, Packt Publishing.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E. & Evanini, K. (2016), The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native anguage, *in* 'Proc. Interspeech 2016', pp. 2001–2005.

Sheng, L. & Edmund, M. (2017), CS229: Deep learning approach to accent classification, Master's thesis, Computer Science Department, Stanford University, Stanford, CA.
URL: *https://cs229.stanford.edu/proj2017/final-reports/5244230.pdf*

Shoichet, C. E. (2021), 'These former Stanford students are building an app to change your accent', *CNN* .
URL: *https://edition.cnn.com/2021/12/19/us/sanas-accent-translation-cec/index.html*

The Manifesto for Agile Software Development (2001).
URL: *https://agilemanifesto.org/*

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. (2016), 'Wavenet: A generative model for raw audio'.
URL: *https://arxiv.org/abs/1609.03499*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), Attention is all you need, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper_files/paper/2017/file/f5ee243547dee91fbd053c1c4a845aa-Paper.pdf*

Viglino, T., Motlicek, P. & Cernak, M. (2019), End-to-end accented speech recognition., *in* 'Proc. Interspeech 2019', pp. 2140–2144.
URL: *https://www.isca-archive.org/interspeech_2019/viglino19_interspeech.pdf*

Weinberger, S. H. & Kunath, S. A. (2011), The Speech Accent Archive: towards a typology of English accents, *in* 'Corpus-based studies in language use, language learning, and language documentation', Brill: New York, pp. 265–281.
URL: *https://accent.gmu.edu/*

Zuluaga-Gomez, J., Ahmed, S., Visockas, D. & Subakan, C. (2023), CommonAccent: Exploring large acoustic pretrained models for accent classification based on common voice, *in* 'Proc. INTERSPEECH 2023', pp. 5291–5295.