Modelling Phoneme Similarity in Varieties of English for Human Language Technologies

Thesis by
Emma O'Neill
Student Number: 17203917

This thesis is submitted to University College Dublin
in fulfillment of the requirements for the degree of
Doctor of Philosophy

Principal Supervisor:	Prof. Julie Carson-Berndsen
Co-supervisor:	Assoc. Prof. Anthony Ventresque
Head of School:	Prof. Neil Hurley
RSP Panel:	Prof. Mark Keane
	Dr. João Cabral

School of Computer Science
University College Dublin
May 2023

Table of Contents

List of Figures

List of Tables

Glossary

AAE African American English. 19

acoustic Relating to the auditory signal of speech. xi, 4, 5, 7, 9, 20, 23, 27, 33–36, 40, 90, 106, 109, 113

affricate Relating to sounds produced with a plosive-like closure and a fricative-like release. 24

allophone One of the possible realisations of a specific phoneme. 9, 99

alveolar Relating to sounds articulated with the tongue positioned on the ridge of the roof of the mouth just behind the teeth. 9, 11, 13, 24, 31, 50, 51, 57, 61, 62, 67, 74, 91, 92, 95, 97, 99

apico Relating to sounds produced with the tip of the tongue. 74, 95

approximant Relating to sounds produced with the articulators approaching each other but not touching. 26, 95

ASR Automatic Speech Recognition. xi, 1, 4, 6–8, 10, 20, 21, 33, 71–75, 77, 78, 83, 85, 87–90, 92, 94–99, 101–107, 109, 112–114

assimilation A process by which a speech sound takes on characteristics of neighbouring sounds. 2, 96

CBOW Continuous Bag Of Words. 29

coda The final consonant(s) of a syllable occurring after the central vowel or nucleus. 28, 31, 95

dental Relating to sounds articulated with the tongue in contact with the teeth. 24, 51, 57, 67

devoicing A process by which a typically voiced sound is produced without vibration of the vocal folds. 67, 95, 97, 98

EFL English as a Foreign Language. 98

elision The omission of one or more sounds from a word or phrase. 40, 91, 93

embedding A numerical representation of an object. xi, 14–16, 23, 24, 27–30, 33–36, 109, 110

epenthesis The insertion of one or more sounds into a word or phrase. 40, 93, 94, 101, 106

flap Relating to speech sounds articulate with a single contact of articulators but with no release burst. 92, 95

fortition A phonological process by which a consonant is strengthened, increasing the degree of stricture. 51, 57, 67

fricative Relating to sounds where airflow is forced through a narrow space in the vocal tract. 11, 12, 24, 26, 32, 40, 51, 57, 67, 92, 95–97, 99

glide Relating to sounds produced with vowel-like properties but which are non-syllabic. 24

glottal Relating to sounds articulated with the glotis or vocal folds. 9, 24

grapheme The smallest functional unit of a language's writing system. 5, 15, 17, 28, 37, 39, 50–52, 60, 68, 69, 74, 87, 92, 94, 97, 98, 101, 110, 112

IPA International Phonetic Alphabet. 1, 9, 25

L1 First Language. xi, 16, 17, 95, 99–101, 103, 104, 112

L2 Second Language. 16, 17, 89, 95, 112

labial Relating to sounds articulated with the lips. 24, 31

language transference Where linguistic features of one language influence the production and/or perception of another. 1, 6, 9, 16, 26, 75, 89, 107, 112

lateral Relating to sounds produced with the tip of the tongue raised to the roof of the mouth such that the airflow is directed along the sides of the tongue. 24

lenition A process by which a consonant is weakened, making them more sonorous. 67

MAE Mainstream American English. 19

nasal Relating to sounds produced with a lowered velum such that air flows through the nose. 12, 24, 26, 31, 33, 50, 62, 67, 94

nucleus The central vowel of a syllable. 13, 34

obstruent Relating to speech sounds produced with obstructions in the airflow e.g. plosives and fricatives. 24, 25, 32, 33, 94–96, 98

onset The initial consonant(s) of a syllable occurring before the central vowel or nucleus. 13, 28, 31, 68

orthographic Relating to the written form of language. xi, 5, 6, 12, 17, 18, 37–39, 43, 45, 49, 50, 52, 61, 62, 71, 87, 97, 110, 111

palatal Relating to sounds articulated with the tongue in contact with the palate. 24, 26, 32, 33

PAM Perceptual Assimilation Model. 16

PER Phone Error Rate. 75, 78, 85, 87

phone The acoustic realisation of a phoneme. 20

phoneme An abstract label for the smallest unit of sound within a language that can distinguish one word from another. xi, 1–3, 5–9, 11–16, 18, 23–42, 49–53, 55–62, 65, 67–69, 71–78, 80, 85, 87–92, 94–97, 99, 101–104, 106, 109–114

phonics A method of teaching early literacy which involves mapping sounds to their orthographic encodings. 3, 18, 37

phonological inventory The set of phonemes of a particular language or variety. 2, 9, 94, 96, 97

phonotactic Relating to phonotactics, the branch of linguistics concerned with restrictions and permissible combinations of phonemes within a language. 28, 29, 31, 32, 110

plosive Relating to sounds where airflow through the vocal tract is obstructed then released in a burst. 9, 11, 12, 24, 26, 32, 34, 40, 51, 57, 61, 67, 91, 92, 94, 95, 97, 99

pretonic Denoting the syllable occuring immediately before that which holds primary stress. 93

realisation The physical production of a phoneme. xi, 1, 5, 7, 9, 16, 23, 27, 28, 34–36, 50, 51, 57, 61, 67, 68, 71, 72, 87, 88, 93–96, 98, 99, 103–105, 109, 111

retroflex Relating to sounds produced with the tip of the tongue curled back towards the hard palate. 13, 94, 104

rhotic Relating to /R/-like sounds or the property of a spoken variety to retain the production of an /R/-like sound when it occurs in a syllable rime. 24, 41, 52, 67, 92, 94, 95

schwa A central, unstressed vowel sound. 92–94, 96–98

semantic Relating to the linguistic meaning of a word or phrase. 14, 27–29, 34

sibilant Relating to fricative sounds produced with the tip of the tongue towards the teeth. 12, 67

SLM Speech Learning Model. 17

sonorant Relating to speech sounds produced with a continuous and non-obstructed airflow e.g. liquids and nasals. 24–26, 31, 33, 34, 92, 96

syllabification The process of splitting a word into its constituent syllables. 28, 34

transformer A particular deep learning model architecture capable of converting an input sequence into an output sequence using self-attention to learn important relationships within the sequences. 4, 20, 21, 35, 39, 87, 101, 110

trill Relating to speech sounds articulated with a vibration between the active and passive articulators. 95

typographic In relation to typographic errors, a misspelling produced in the process of producing typed text. 3, 17, 37, 40, 43–45, 52, 69, 111

variety A style of language used by a group of people. xi, 1–8, 10, 19, 55, 56, 58, 59, 61, 62, 64, 65, 67, 69, 71–73, 77, 78, 80, 83–85, 87–89, 93–97, 106, 107, 111–113

velar Relating to sounds articulated with the tongue against the soft palate or velum. 11, 24, 50, 62, 67

voiced Relating to sounds produced with vibration of the vocal folds. 2, 11, 12, 16, 24–26, 31, 32, 40, 92, 95–99

voiceless Relating to sounds produced with no vibration of the vocal folds. 2, 11, 12, 16, 24–26, 31, 32, 40, 67, 94, 97, 99

WER Word Error Rate. 4, 103

Abstract

At its core, this thesis endeavours to model similarity judgements of phoneme categories in different varieties of English with a focus on simplicity and interpretability. These models can be incorporated into adapted language technologies in order to benefit users with underrepresented spoken varieties and can be examined to provide insight into the pronunciation features of an individual speaker or variety.

First, it is shown that speaker judgements of phoneme similarity are not fully predictable based solely on traditional phonological features. Similarity hierarchies of the phonemes of English are constructed from three different phoneme embedding approaches; perception based, feature based, and distribution based. Through qualitative comparison of these three hierarchies it is demonstrated that some elements of speaker perceptions, which were not explainable in terms of phonological features, appear to be influenced by the environments in which phonemes typically occur in English. As a result of this finding, a model of general English phoneme similarity is constructed based on the distributive properties of phonemes and the acoustic properties of their realisations.

A practical application of this similarity model is then explored in the form of a spelling correction method. It is demonstrated that a spellchecker based on comparing the phonemic similarities between a misspelling and potential real-word corrections is better suited to the phonetic writing of children than traditional character based systems. This interaction between spoken words and written forms prompts further work investigating how regional pronunciation variation might influence orthographic encoding. An adapted spelling correction tool, developed by fine-tuning the similarity model to Irish Accented English, exhibits better performance on misspellings from Irish school children. Furthermore, the resulting tuned model is interpretable and captures phonetic and phonological features of the English variety.

Furthering this investigation into the interplay between spoken variation and phoneme similarity, models of individuals' English varieties are constructed for speakers from different regions and with different First Languages (L1s). Erroneous Automatic Speech Recognition (ASR) output is leveraged to construct these models by, again, considering similarity as a function of confusability (this time on the part of the recogniser). It is then demonstrated that speakers with similar varieties produce similar representations. Additionally, similarity models at the variety level are analysed to discover which pronunciation features are detected and, as a result, lead to recognition errors and indicate a weakness in the ASR system. The variant pronunciations captured by the models are shown to align with those which arise from human annotations and with existing literature on the specific English varieties.

The work presented in this thesis draws from many areas across Computer Science, Linguistics, and Education. These include machine learning, human language technologies, phonetics and phonology, sociolinguistic variation, and children's literacy acquisition. It is hoped that, above all else, this thesis highlights the benefits of multidisciplinary approaches to these topics and will promote collaboration between the fields.

Declaration

I hereby certify that the submitted work is my own work, was completed while registered as a candidate for the degree stated on the Title Page, and I have not obtained a degree elsewhere on the basis of the research presented in this submitted work.

Emma O'Neill

Acknowledgements

There are a number of people and institutions who deserve thanks for their input and influence on my PhD journey.

To my Principal Supervisor Professor Julie Carson-Berndsen and Co-Supervisor Associate Professor Anthony Ventresque. Thank you for your expertise, guidance, and endless patience. I am extremely grateful for the opportunities I have been given throughout the PhD journey and for the support you have both provided.

To University College Dublin and the ADAPT Research Centre, for hosting my studies, funding my research, and providing a supportive and stimulating environment to carry out and share my work.

To SoapBox, for taking me on as an intern during my studies and for inviting me back into the team. Thank you for giving me the opportunity to apply my skills and experience outside of academia.

To my godfather, Eddie, for your support, encouragement, and for reading every one of my papers - published or otherwise. Thank you for taking an interest.

To my parents, Denise and Mark, who always taught me the importance of a good education and then complained when I spent over two decades of my life in school. Thank you for making my seemingly endless studies possible.

To my siblings, who have made the sound decision not to follow in my footsteps. You all keep me grounded and remind me that life is too short not to do the things that make you happy.

To my fellow PhD students in Computer Science at UCD who provided invaluable conversation, insights and company in the student bar on Friday nights.

To the INF-YT for tolerating my incessant word polls and pronunciation questions. Thank you for such words of encouragement as “don’t summon the word witch”, “stop analysing me”, and “why are you like this?”

To my old Edinburgh flatmates, who have never doubted I would make it to the end and who have been my biggest cheerleaders throughout this process.

To my Glaswegian friends back home. I’m incredibly grateful that we can always pick up where we left off despite the distance and long periods without contact.

And finally, to my partner Jakim, for your support, understanding, and encouragement. I don’t think I would have made it to the end in one piece without you. Thank you for always keeping our home stocked with snacks and forcing me to take breaks when I needed them.

This research was conducted with the financial support of the School of Computer Science at the ADAPT SFI Research Centre at University College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme

Publications and Presentations

O'Neill, Emma and Julie Carson-Berndsen (in preparation). *Leveraging Erroneous ASR Output to Build Interpretable Speaker Models of Pronunciation Variation in English*. Journal paper in preparation.

O'Neill, Emma and Julie Carson-Berndsen (2023). "Investigating the Sensitivity of Automatic Speech Recognition Systems to Phonetic Variation in L2 Englishes". In: *University of Pennsylvania Working Papers in Linguistics*. Vol. 29.2.

O'Neill, Emma and Julie Carson-Berndsen (2022a). "Investigating the Sensitivity of Automatic Speech Recognition Systems to Phonetic Variation in L2 Englishes". In: *New Ways of Analyzing Variation 50*. [Abstract].

O'Neill, Emma and Julie Carson-Berndsen (2022b). "Modelling Pronunciation Variation in Different Spoken Englishes". In: *UK Speech Conference 2022*. [Abstract].

O'Neill, Emma, Joe Kenny, Anthony Ventresque, and Julie Carson-Berndsen (2021). "The Influence of Regional Pronunciation Variation on Children's Spelling and the Potential Benefits of Accent Adapted Spellcheckers". In: *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 674–683.

O'Neill, Emma, Robert Young, Elsa Thiaville, Muireann MacCarthy, Julie Carson-Berndsen, and Anthony Ventresque (2020). "S-capade: Spelling correction aimed at particularly deviant errors". In: *International Conference on Statistical Language and Speech Processing*. Springer, pp. 85–96.

O'Neill, Emma and Julie Carson-Berndsen (2019). "The Effect of Phoneme Distribution on Perceptual Similarity in English". In: *INTERSPEECH 2019*, pp. 1941–1945.

O'Neill, Emma, Mark Kane, and Julie Carson-Berndsen (2018). "Two Data-Driven Perspectives on Phonetic Similarity". In: *UK Speech Conference 2018*. [Abstract].

Chapter 1

Introduction

At its core, this thesis examines the concept of phoneme similarity. Specifically, how intuitions regarding the similarity relationships between phoneme categories might differ between speakers and across varieties of English. Differences in usage, variant phonetic realisations, and, for multi-lingual speakers, language transference effects appear to influence the perception and production of phonemes. This, in turn, impacts how similar or dissimilar two phonemes are considered to be by a speaker. Several approaches to modelling these similarity relationships are explored throughout this thesis. Firstly, through an investigation into which factors affect similarity judgements and then by considering both the spoken and written form of the language variety. In all cases, these models rely on the central idea that similarity is a function of confusability and that the likelihood of mistaking one phoneme for another is directly correlated with how similar they are believed to be. These confusions are considered from both speaker perceptions and productions as well as from the performance of automatic speech recognition (ASR) systems.

The potential advantages of such models of similarity are also a central concern of this thesis. Investigation of the models reveals insight into both the pronunciation features of a spoken variety and the salience of such features amongst speakers. Furthermore, the benefits of applying similarity models to the development and evaluation of language technologies are demonstrated. This includes their incorporation into an automatic spelling correction method and their use in probing the sensitivities of an ASR system to patterns of variation.

This thesis is multi-disciplinary in nature and lies at the intersection between Linguistics, Computer Science, and Education. The following sections provide a brief introduction to the areas of *phonology and phoneme similarity*, *children's literacy and spelling correction methods*, and *spoken variation and automatic speech recognition* with the aim of providing context and outlining the gaps in existing research which motivate the work presented here. The five central research questions are then detailed followed by an overview of the chapter structure of the thesis. It is important to note here that, when discussing phonemes, this thesis predominantly uses ARPAbet notation but, where more fine-grained phonetic detail is required, International Phonetic Alphabet (IPA) symbols may be used (see Section 2.1). As per standard convention, phonemes are enclosed between slashes whilst phonetic realisations are presented in square brackets. Moreover, the term 'variety' is opted for when discussing a specific language style shared by speakers rather than enter the debate regarding what constitutes a distinct language, accent or dialect.

1.1 Phonology and Phoneme Similarity

Phonemes are the smallest meaningful units of speech in that they are contrastive within a spoken language. The complete set of phonemes, or phonological inventory, differs across languages and can also vary depending on the dialect or accent of a language variety. If two sounds are phonemically distinct in a variety, then they can form minimal pairs; two words which are identical in all sounds but one. For example, /P/ and /B/ are considered distinct phonemes in English. The word 'pat' is made up of the phonemes /P/, /AE/, and /T/ whilst the word 'bat' consists of /B/, /AE/, and /T/. Since replacing the /P/ sound with a /B/ sound produces a new word, 'pat' and 'bat' are considered a minimal pair and /P/ and /B/ are shown to be phonologically distinct. /P/ and /B/ differ only in their voicing property, being voiceless and voiced respectively. Voicing is considered a phonological, or distinctive, feature of English. Phonological features are the qualities of a phoneme which differentiate it from others. English consonantal phonemes are distinguished by features such as *voicing*, *manner of articulation*, and *place of articulation*.

Typically, phoneme similarity is discussed in terms of the phonological features two phonemes have in common or the natural classes formed by grouping together phonemes with shared features (Chomsky and Halle, 1968; Wheeler, 1972). For instance, two phonemes differing in only one feature, /P/ vs /B/ which differ only in *voicing*, would be considered more similar than phonemes with a number of differing features, /P/ vs /N/ for example, which differ in *voicing*, *place*, and *manner*. The concept of similarity amongst phonemes is integral to a number of phonological theories such as Generative Phonology (Kenstowicz and Kisseberth, 1979) or Optimality Theory (Smolensky and Prince, 1993). Both make use of the idea that the realised spoken form of a word, the *surface form*, comes from a more abstract representation, the so-called *underlying form*, and that the two forms must be in some ways similar. Likewise, the phonological process of *assimilation* involves sound segments becoming more similar to adjacent sounds by adopting their properties like, for instance, the place of articulation (Ohala et al., 1990).

However, studies of the similarity judgements of speakers, whether through overt questioning, an examination of half-rhyme acceptability, or research into phoneme confusability, reveal that phonological features are insufficient in capturing perceived similarity (Gallagher and Graff, 2012). Often the results of such studies are not predictable based on distinctive features alone. In other words, there exists a disparity between speaker intuitions regarding phoneme similarity and similarity as defined in terms of phonological features. Prompted by experiments which demonstrate that speakers' perceptions of phoneme categories are adjustable and can be influenced by exposure and context, for example that of Scharenborg et al. (2018), this thesis seeks to investigate additional factors, specifically the distributive properties of phonemes, which might be of influence. It is the speaker intuitions and similarity judgements which the models generated are intended to capture.

1.2 Children's Literacy and Spelling Correction Methods

Children's attempts at spelling have long been considered rooted in phonetics (Read, 2018). Nowadays, modern literacy education approaches involve the teaching of phonics; mappings between sounds and letters. For example, children may be taught that the letter 'a' represents an /AE/-like sound. In terms of spelling, children are encouraged to first identify the sound sequence of a word and then encode those sounds using the appropriate letters which represent them. The word 'cat', for instance, is broken down into the component phonemes /K/, /AE/, and /T/ which are, respectively, denoted by the letters 'c', 'a', and 't'. However, English orthography is complex in nature and there is not an exclusively one-to-one mapping between sounds and letters. The /F/ phoneme might be written using the letter 'f' or with 'ph'. The letter 'c' could represent /K/ or /S/. As a result, as children are acquiring written literacy and attempting to spell words they may not be familiar with, this 'sound-it-out' approach often leads to misspellings which deviate substantially from the canonical spelling.

Such spelling errors produced by children tend to prove difficult for conventional spelling correction tools. Traditional automatic spellcheckers are typically designed to correct typographic errors - those which differ from the target spelling by only one or two characters (Kukich, 1992). They are much less effective when tasked with correcting the phonetic misspellings of children which are far removed from the standard form. These sorts of errors require a spelling correction approach that considers sound similarity between a misspelling and its intended target. Whilst there are some spellcheckers which incorporate predefined mappings between letters and sounds, the area of spelling correction for children provides an interesting application for a model of phoneme similarity within a sound based method of correcting errors.

Moreover, it stands to reason that if spelling attempts are based on sound sequences and identifying the constituent sounds of a word, then a child's individual pronunciation would impact this process. Whilst there are many studies into the interaction between spoken variation and literacy acquisition in general, there is much less research regarding spelling specifically. Snell and Andrews (2017) surmised that regional dialect has only a minor impact on written literacy but, throughout their review, difficulties with spelling were not considered related to pronunciation variation and instead merely a product of the complexities found in English orthography. Despite this, over the course of the research carried out for this thesis, examples of spelling errors which appeared to capture pronunciation features of specific spoken varieties were observed. This potential influence of the spoken form on spelling efforts was identified as a gap in the literature to which this thesis hopes to contribute.

If it is indeed the case that pronunciation influences spelling, it can be assumed that speakers of a specific variety will face particular difficulties related to that variety. As Terry (2006) states, "while all children must learn to negotiate mismatches between speech and print in

order to become good readers and writers, this process may be particularly problematic for children whose spoken language differs substantially from standard written forms". It could be the case that certain groups of children are at a greater disadvantage when acquiring literacy skills as a direct result of their spoken language and that the spelling correction tools available to them are less beneficial in the aid they can provide. In an effort to address such a problem, this thesis examines the potential benefits of adapting language technologies to a specific spoken variety. By tailoring the technology to the variety, it is hoped that the user will experience greater success when using such an adapted spellchecker.

1.3 Automatic Speech Recognition and Variation

Automatic speech recognition is the computational approach to transcribing speech. Traditional methods comprised separate components, namely, an acoustic model, a pronunciation model, and a language model. However, recently, end-to-end models for ASR have been rapidly gaining popularity and outperforming traditional hybrid models. The self-attention based transformer model architecture is one such model used for ASR which has exhibited state-of-the-art performance on various speech recognition benchmarks (Vaswani et al., 2017). In this thesis, experimentation involving ASR uses a transformer based model known as wav2vec (Baevski et al., 2020).

It is generally agreed that, whilst ASR systems are capable of delivering ceiling-level performance on so-called "standard" speech, when it comes to spoken varieties that are underrepresented or absent in the training data this performance deteriorates (Hinsvark et al., 2021). With the complex network architectures and "black box" nature of many high performing ASR systems, it is often difficult to pinpoint exactly which features of a spoken language variety prompt such performance deterioration beyond just deviation from the expected "norm". This negatively effects already marginalised groups like minority speakers, those who use minority languages, and those with non-native accents.

Koenecke et al. (2020) investigated the commercial ASR systems of Amazon, Apple, Google, IBM, and Microsoft and found significantly higher Word Error Rates (WERs) for African American varieties of English. This deterioration of performance was attributed to underrepresentation and the authors call for data collection to include "nonstandard varieties [...] including those with regional and nonnative-English accents". Race and dialect were also shown to impact the accuracy of the automatic YouTube captions in work by Tatman and Kasten (2017) where it was noted that the best performance was achieved on white speakers with a General American (non-regional) dialect. The work presented in this thesis seeks to explore exactly which features of "nonstandard" varieties lead to speech recognition errors and how this information might be used for target training data collection with the end goal of improving ASR performance on these varieties.

1.4 Research Questions

Here, the five main research questions which this thesis seeks to address are defined.

Research Question 1: *What factors influence a speaker's intuitions regarding phoneme similarity?*

A speaker's similarity judgements regarding phonemes can affect their production and perception of spoken language. Since this can vary between speakers, it is not the case that such intuitions are based purely on phonological features as are traditionally used to distinguish between sound categories. Instead, it is hypothesised that there must be some influence stemming from how a phoneme is typically realised within a spoken variety and how it is used in the language. The influence of the distributive properties of English phonemes on native speakers' perceptions is investigated. Through the adaptation of a word embedding approach, vector representations of phonemes are generated based on the environments in which they occur in English usage. Phonemes which occur in similar environments have corresponding vector representations which are in close proximity in the multi-dimensional space. Examination of the resulting similarity hierarchy and comparison with similar feature based and perception based representations obtained from existing literature demonstrates that distribution influences perception. A model of phonemic similarity in English is then constructed in the form of a phoneme distance matrix using both distributive features of phonemes and the acoustic properties of their realisations. This work is discussed in Chapter 3 and has been published in part in O'Neill and Carson-Berndsen (2019) and presented in O'Neill et al. (2018).

Research Question 2: *Can a model of phoneme similarity be applied to the development of language technologies; specifically in a spelling correction tool for children?*

Children's spelling tends to be phonetic and relies on their ability to identify and encode sounds orthographically. An effective spelling correction tool for children needs to work on a sound-based level rather than character-based. The phoneme similarity model constructed previously is used as a basis for a collaborative project developing a spelling correction method specifically for children's writing. Misspellings are converted to sequences of phonemes using a grapheme-to-phoneme tool and potential real-word corrections are suggested based on their phonemic similarity to the original error. The tool is shown to outperform standard spellcheckers when tested on children's spelling. It is also capable of correcting misspellings which deviate extensively from the canonical form and which even the best performing tools struggle with. This work is discussed in Chapter 4 and was published in O'Neill et al. (2020).

Research Question 3: *How might the pronunciation variation present in a child's English impact the misspellings they produce and can a spelling correction tool be adapted to better perform with a specific variety?*

If a child must first identify the sounds of a word before attempting to encode them orthographically, it stands to reason that their individual pronunciation would influence this process. The potential benefits of accent adaption in language technology are investigated with an Irish Accented English spelling correction tool developed by fine-tuning the phoneme similarity matrix of the previously discussed system. This adapted model is shown to perform better on misspellings from Irish children than a similar system tuned to British English or the original baseline system. This work is discussed in Chapter 5 and was published in O'Neill et al. (2021).

Research Question 4: *Can an individual's variety of English be modelled based on phoneme confusability and do speakers with similar varieties produce similar representations?*

The success of the previous Irish Accented English tuned model of phoneme similarity and its correspondence with the existing literature regarding specific pronunciation features prompt further research into modelling spoken varieties. Erroneous ASR output is leveraged to construct speaker models of phoneme confusability in the form of confusion matrices by comparing the output and original text prompt at the phonemic level. The resulting matrices benefit from being both simple and interpretable: they require minimal annotated data to build, can be reshaped into a multi-dimensional vector for various computational tasks, and can be examined in order to extract and analyse specific patterns of pronunciation variation exhibited by the speaker. It is also then demonstrated that speakers of the same region tend to cluster together in the multi-dimensional space, suggesting that the ASR performs consistently on similar spoken varieties and that they produce similar representations. This is further confirmed through the success of a classification task using these representations which demonstrates that relatively little annotated data is required for representative models of a speaker. This work is discussed in Chapter 6, was presented in O'Neill and Carson-Berndsen (2022b) and features in O'Neill and Carson-Berndsen (in preparation).

Research Question 5: *How does the information captured by these models compare to existing knowledge of a spoken variety and what can they tell us about the sensitivity or robustness of ASR systems to pronunciation variation?*

Region based models, constructed by pooling together speakers of the same region, are analysed in order to investigate whether such representations capture known pronunciation features of a particular English variety and language transference effects from other languages spoken in the region. The characteristic traits of each region are extracted and explained in relation to existing research and literature on phonetic variation. Additionally, to understand the feasibility of using an ASR system to automate the annotation or de-

tection processes in variation research, it is important to investigate how the ASR output and resulting speaker models described previously compare with expert annotations. It is demonstrated that the patterns of variation, including problematic phonemes and their most common substitutes align well with the judgements of human annotators with some discrepancies. This analysis thus allows for evaluation of the robustness or sensitivity of an ASR system in terms of specific variant pronunciations. Different ASR applications will require different sensitivities, for example ASR for variation research or Computer Aided Pronunciation Training (CAPT) would need to be highly sensitive to pronunciation differences whereas more general voice controlled technology would want to be robust to variation and perform similarly for all language varieties. This method allows a detailed analysis of a specific ASR system's handling of variation. This work is discussed in Chapter 7, was presented in part in O'Neill and Carson-Berndsen (2022a), will be published in part in O'Neill and Carson-Berndsen (in preparation), and features in O'Neill and Carson-Berndsen (2023).

1.5 Overview of the Thesis

Having introduced the research areas within which this work is positioned and outlining the problems it seeks to address, the rest of this thesis is structured as follows. Chapter 2 provides relevant background information regarding the different fields of study related to this work. This includes phonology and phoneme similarity, studies of pronunciation variation, spelling correction and children's literacy, and ASR for different spoken varieties. A critical analysis of the work in each of these areas is presented alongside discussion of the theories and frameworks that this thesis builds on.

An investigation into the sources of perceptual similarity and the construction of a general English model of phoneme similarity is then presented in Chapter 3. Here, it is demonstrated that the distribution of phonemes influences speakers' perception of similarity and a phoneme distance matrix is generated by combining both the acoustic properties of phoneme realisations with the distributive properties based on the environments in which they occur in English.

Chapter 4 then focuses on an application of this similarity model through its incorporation into a sound-based spellchecker for children. By using the similarities between phonemes in an algorithm which ranks potential candidate corrections of a misspelling, the spelling correction method is shown to handle those difficult-to-correct errors which deviate drastically from the target spelling.

In Chapter 5, this spellchecker is then fine-tuned to a specific spoken variety of English, namely Irish Accented English, to investigate whether systematic pronunciation variation influences children's spelling and to explore the potential benefits of accent adaptation in

language technologies. This adapted approach achieved better performance on the spelling of Irish children than a similar system tuned on British English data. This demonstrates both that commonalities in the pronunciation features of a spoken variety manifest in writing, and that children would see benefits from literacy tools designed for their particular spoken variety.

A new method for modelling a speaker's variety based on phoneme similarity is then presented in Chapter 6. By analysing the phonemic sequences of an ASR system's output, speaker level similarity models are constructed based on the idea of phoneme confusability. In order to demonstrate that speakers with similar varieties produce similar representations, these models are first shown to cluster by region in multidimensional space and then exhibit success in their use in a region classification task.

Chapter 7 then considers what can be learned through the analysis of the previous speaker models and region based profiles. Pooling together representations from speakers of the same region allows for the identification of documented pronunciation features of the English variety common in that region. Furthermore, it is demonstrated that the pronunciation features captured by the speaker models often align with expert judgements and phoneme level annotations and give insight into the sensitivities of the particular ASR system. This has implications for the use of ASR in automating annotation for variation research and in adopting targeted approaches to training data collection for the improvement of ASR performance on underrepresented varieties.

Finally, the thesis concludes in Chapter 8 with a summary of the contributions of this work and the implications it has on the future direction of research in this multidisciplinary field.

Chapter 2

Background

This chapter provides the relevant context within which this thesis is situated across the various disciplines. This includes background information on phonology, modelling phonological similarity, variation and language transference effects, spelling correction, early literacy acquisition, and automatic speech recognition. The primary aim is to provide an overview of the different areas and highlight the key themes and concepts that will be addressed throughout this thesis.

2.1 Phonology

Phonology is the study of the individual sound segments that make up the words in our language. These units of speech, known as phonemes, are fundamental to the study of spoken language and are the most basic contrastive units of speech. This means that if two distinct words in a language differ only in one sound segment then the sounds in question are considered distinct phonemes since they form *minimal pairs*. For example, the words “bat" and “pat" are a minimal pair since they differ in only one phoneme but are considered different words. This demonstrates that /B/ and /P/ are distinct phonemes in English. The set of phonemes, or phonological inventory, varies across languages. Phonemes are theoretical categories used to discuss the contrastive sound segments of a language. They are a level of abstraction away from the actual spoken production known as their *realisation*. For example, the phoneme /T/ might be produced as an alveolar plosive [t], an alveolar tap [ɾ] or as a glottal stop [ʔ] but these are all allophones of the /T/ phoneme. Given the tendency for researchers in the speech technology field to conflate the term ‘phoneme’ with the acoustic sound (Moore and Skidmore, 2019), it is endeavoured, throughout this thesis, to maintain a clear distinction between the two. In line with standard practice, phonemic transcriptions are given between slashes whilst phonetic transcriptions appear within square brackets.

There are a number of notations for labelling phoneme categories. The IPA provides symbols and various diacritics for referring to the sounds of languages. These can be seen in Figures 2.1 and 2.2 taken from International Phonetic Association and others (1999). A more machine readable set of transcription codes is ARPAbet notation. The set of ARPAbet symbols used throughout this thesis, their IPA symbol counterparts and example words are given in Table 2.1. ARPAbet notation also features in the CMU Pronouncing Dictionary (Weide, 1998), a phonetic dictionary of over 100,000 words of English and their pronunciations (based on

an American variety of English). The CMU Pronouncing Dictionary is used in a number of different applications in the field of speech technology from ASR to text-to-speech. It is also used extensively throughout this thesis.

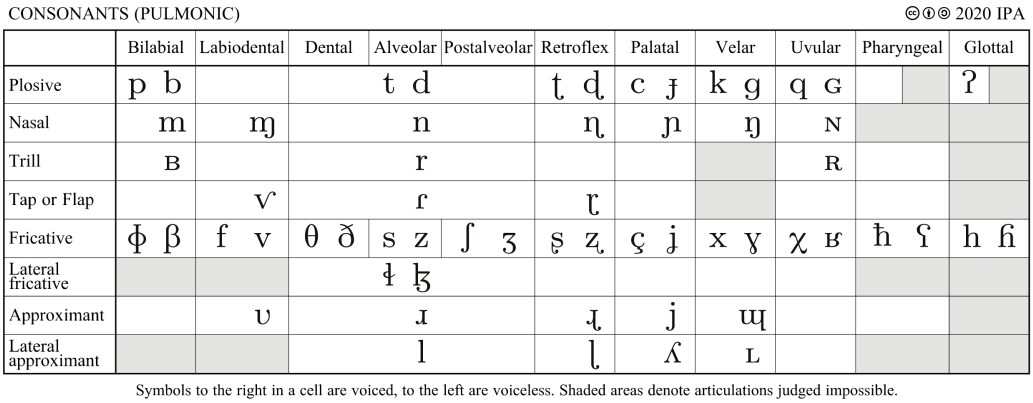


Figure 2.1: International Phonetic Alphabet notation for the consonantal sounds.

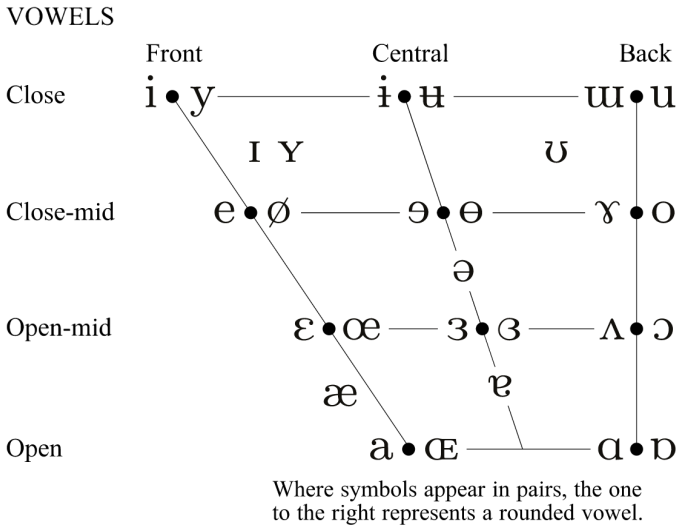


Figure 2.2: International Phonetic Alphabet notation for the vowel sounds.

ARPAbet	IPA	Example	ARPAbet	IPA	Example
AA	ɑ	p**a**lm	G	g	**g**ive
AE	æ	tr**a**p	HH	h	**h**and
AH	ʌ	str**u**t	JH	ʤ	**j**oin
AO	ɔ	l**o**t	K	k	**k**ite
AW	aʊ	m**ou**th	L	l	**l**oud
AY	aɪ	pr**i**ce	M	m	**m**ost
EH	ɛ	dr**e**ss	N	n	**n**ever
ER	ɝ	n**ur**se	NG	ŋ	so**ng**
EY	eɪ	f**a**ce	P	p	**p**lay
IH	ɪ	k**i**t	R	ɹ	**r**un
IY	i	fl**ee**ce	S	s	**s**un
OW	oʊ	g**oa**t	SH	ʃ	**sh**ock
OY	ɔɪ	ch**oi**ce	T	t	**t**ie
UH	ʊ	f**oo**t	TH	θ	**th**ink
UW	u	g**oo**se	V	v	**v**ery
B	b	**b**ell	W	w	**w**ait
CH	ʧ	**ch**oose	Y	j	**y**oung
D	d	**d**og	Z	z	bu**zz**
DH	ð	**th**en	ZH	ʒ	vi**s**ion
F	f	**f**rog			

Table 2.1: A comparison of ARPAbet and IPA notation with examples.

Standard phonological theory typically refers to the various properties which define and distinguish the individual phonemes in a language. These properties are known as distinctive features. For example, consonantal phonemes might be described in terms of;

- *Place of articulation*: The location of the tongue when the phoneme is realised. For example /T/ is alveolar whilst /K/ is velar.
- *Manner of articulation*: How the air stream is affected by the articulators. For example /T/ is a plosive whilst /S/ is a fricative.
- *Voicing*: Whether or not the vocal folds vibrate during production. For example /T/ is voiceless whilst /D/ is voiced.

Whilst for vowels, a different set of features could be used;

- *Height*: The vertical position of the tongue when the phoneme is realised. For example /IY/ is a high vowel whilst /AE/ is a low vowel.
- *Backness*: The horizontal position of the tongue when the phoneme is realised. For example /IY/ is a front vowel whilst /UW/ is a back vowel.
- *Tensing*: The tensing of the tongue when the phoneme is realised. For example /IY/ is a tense vowel whilst /IH/ is a lax vowel.

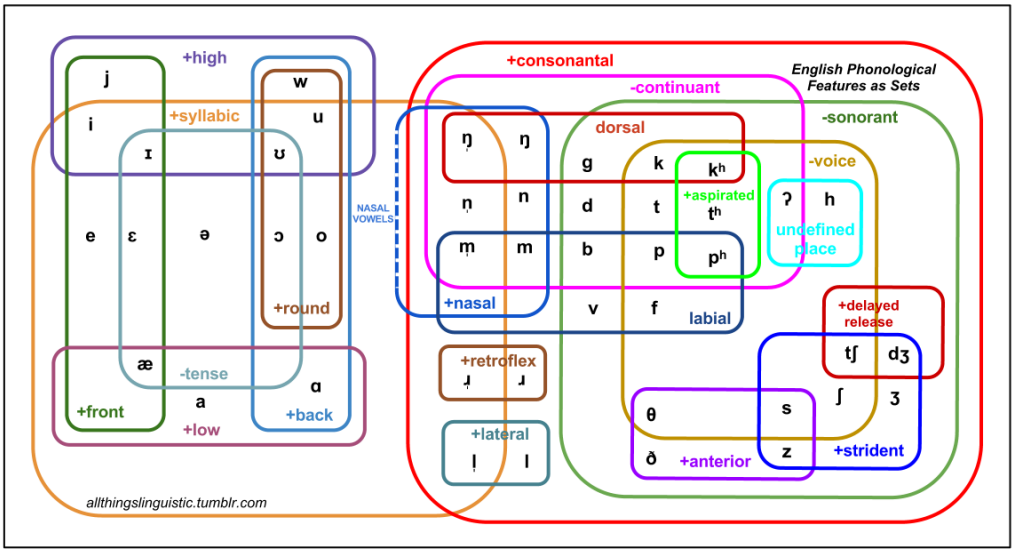


Figure 2.3: Natural classes of phonemes based on their shared phonological features (McCulloch, 2013).

It is common for groups of phonemes with a particular feature or features in common to be treated as a single group called a natural class (for details see Giegerich et al. (1992)). These classes allow for generalisation across a group of phonemes that behave in the same way in various linguistic processes. For example, for the pluralisation of words in English there are 3 possible affixes; namely /S/, /Z/, and /IH Z/. /S/ is used when the last phoneme in the singular word is *voiceless* as in "cat**s**". /Z/ is used when the last phoneme is *voiced* as in "dog**s**". /IH Z/ is used when the last phoneme is a *sibilant* as in "hors**es**". (Note that the phonemes used when the word is spoken do not necessarily match with the orthographic spelling).

Essentially, in order to explain how to pluralise a word in English, it is necessary to be able to group phonemes in terms of their features: e.g. *voiced*, *voiceless*, *sibilant*. Typically, phonological theories are based on and refer to the distinctive features of phonemes and the natural classes into which they can be grouped (Chomsky and Halle, 1968; Giegerich et al., 1992). However there is no one singular feature specification used throughout the field. In some cases, the minimal number of features which can adequately differentiate between all of the phonemes in a given language is considered optimal whereas others prefer a rich and complex system of features (Jakobson et al., 1951; Frisch, 1996). Furthermore, features can be binary or multi-valued. A binary feature relates to the presence or absence of a single phonological property, [+voice] for example, whilst multi-valued features can have many values, like the feature 'manner' with possible values including 'plosive', 'fricative', 'nasal' etc. (Chomsky and Halle, 1968; Bailey and Hahn, 2005). Figure 2.3, taken from McCulloch (2013), illustrates an arrangement of phonemes into natural classes based on shared phonological features.

2.2 Modelling Phonological Similarity

Phonological features have permeated the field of phonology for decades. Thus, the concept of similarity between phonemes is traditionally measured in terms of the features they do or do not have in common. For instance, two phonemes differing in only one feature, /P/ vs /B/ which differ only in *voicing*, would be considered more similar than phonemes with a number of differing features, /P/ vs /N/ which differ in *voicing*, *place*, and *manner*. The concept of similarity amongst phonemes is integral to a number of phonological theories such as Generative Phonology (Kenstowicz and Kisseberth, 1979) and Optimality Theory (Smolensky and Prince, 1993). As Oostendorp (2004) explains, both make use of the idea that the realised spoken form of a word, the *surface form* comes from a more abstract representation, the so-called *underlying form*, and that the two forms must be in some ways similar.

However, it has been noted that a purely feature-based description of similarity between phonemes can contrast with observed phonological processes which are thought to be driven by physiological and perceptual factors. Johnsen (2012) reports how a particular phonological process (Norwegian retroflexion) that applies to the alveolar consonants is obligatory for some phonemes (/T/, /D/, and /N/) but optional for another (/S/) and states that “the likelihood of retroflexion correlates with its perceptual properties". Despite the process involving a change in the same feature across all the alveolar consonants, the application to /S/ results in a sound that is perceived as being too dissimilar to the underlying phoneme in certain contexts. These apparent inconsistencies are the result of a speaker’s mental representation of the phonemes of their language and the similarities between them, or *perceptual similarity*. Furthermore, Gallagher and Graff (2012) state that perceptual similarity “does not necessarily coincide with natural classes or feature values”.

Perceptual similarity relates to a speaker’s intuition regarding the sound system of their language. Modelling this relies on examining how speakers produce or recognise these sounds. One particular approach, used in a number of studies, looks to what is deemed acceptable in the context of rhyming words (Zwicky, 1976; Kawahara, 2007). A syllable can be broken down into its onset, nucleus, and coda; the consonants occurring before the central vowel, the vowel itself, and the consonants following respectively. Figure 2.4 shows an example structure of a syllable.

The vowel and following consonant cluster is known as the rhyme and when two words end with the same sequence of sounds they are considered perfect rhymes (*line* and *mine* for example). However, in the world of poetry or musical lyrics, writers often employ half-rhymes; words with similar but not identical rhymes in the final syllable (*line* and *time* for instance). Research into the use of half-rhyme suggests that there exists a gradient of how acceptable a half-rhyme is based on how similar the phoneme components are, and this

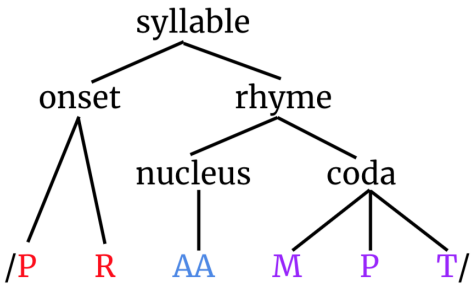


Figure 2.4: The syllable structure for the word ‘prompt’.

correlates with the frequency with which it is used (Kaplan and Woodmansee, 2018). For example, *line* and *light* constitute a half rhyme where the only difference between the words is the use of /N/ or /T/. Similarly, the half-rhyme pair *line* and *lime* differ in one phoneme, /N/ or /M/. However most English speakers agree that the latter is a more acceptable half-rhyme suggesting that /N/ and /M/ are perceptually more similar than /N/ and /T/. This point is confirmed in the literature where /N/-/M/ half-rhymes are the most commonly observed (Zwicky, 1976; Kaplan and Woodmansee, 2018).

Another way of gathering similarity judgements lies in the idea that similarity is a function of confusability (Gallagher and Graff, 2012). Phonemes that are more perceptually similar to the speaker are more likely to be confused for each other in a classification task. Thus, the frequency with which a phoneme is mistakenly identified as another correlates to how perceptually similar the two phonemes are. Perceptual confusability experiments have a long history in the field with early examples including that of Miller and Nicely (1955) to more modern studies like that of Cutler et al. (2004). These experiments elicit errors in phoneme identification, typically through the use of increased background noise or distortion of the speech signal, in order to gather information regarding which phonemes are more often confused and thus perceptually similar.

Over time, phonological research has shifted from the traditional theoretical approach to a computational machine learning one. Jarosz (2019) notes that “computational modelling has become an essential tool of modern phonological research" and the use of large data sets to analyse phonological phenomena has prompted the emergence of the term ‘Corpus Phonetics’ (Liberman, 2019). This is the landscape for early work carried out as part of this thesis whereby phoneme similarity is modelled by generating phoneme embeddings from observed data using machine learning techniques (see Chapter 3).

The use of embeddings has seen a surge in popularity in the field of Computational Linguistics in the last decade due to the success of the *word2vec* model in capturing the semantic meaning of words based on their observed contexts (Mikolov et al., 2013). Leveraging the words of John Firth, “you shall know a word by the company it keeps", words are represented as vectors based on the context in which they occur. Their proximity to each other in the

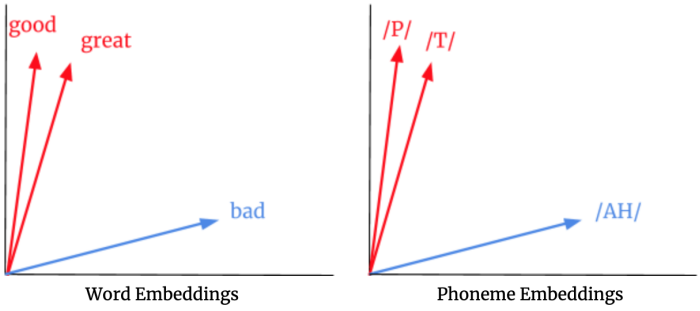


Figure 2.5: An illustration of possible vector representations capturing *semantic similarity* in word embeddings (left) and *phonemic similarity* in phoneme embeddings (right).

vector space correlates to their similarity in meaning; vectors that are very close together depict words that are used in the same way and tend to be synonymous. In a similar vein, this research involves the generation of phoneme embeddings where the context in which a phoneme appears is used to generate vector representations with the hypothesis that the proximity of the vectors captures perceived phonemic similarity. This idea is illustrated in Figure 2.5 using 2-dimensional vectors although most embedding approaches work in a higher multi-dimensional space.

The concept of embeddings built from phonological data has been gaining traction in recent literature. Parrish (2017) sought to generate word embeddings based on phonological feature bigrams within words. This paper demonstrated the effectiveness of such embeddings in predicting similarity judgements of native speakers and demonstrated their use in poetic sound symbolism. In work by Ma et al. (2016), phonemic n-gram embeddings were generated from child-directed speech and were shown to improve performance of a language acquisition model in a word segmentation task. Evidently, not only are phonologically based embeddings feasible, but they have been shown to be influential in a number of tasks.

With regards to similarity, an investigation of various phonological distance measures in a tonal language was carried out by Do and Lai (2019). They compared the predicted similarities of phonemes using a number of distance metrics and feature-based models to native speaker judgements of similarity. Their work displays the success of considering factors outside of the traditional phonological features, in this case tone, when modelling similarity. Furthermore, the word2vec algorithm has been adapted for spoken language analysis in an approach coined Speech2Vec (Glass, 2018) which demonstrates both that speech data can effectively be used to generate embeddings and that the resulting model outperforms that of a standard word2vec model.

Silfverberg et al. (2018) generate phoneme embeddings from graphemes (the written forms) for the purpose of investigating sound analogies. This is a technique often used to examine the effectiveness of word embeddings where vector geometry is used to test whether an

analogy, like *king* is to *queen* as *man* is to *woman*, holds true. Similarly, these phoneme embeddings are shown to be effective in their ability to capture distinctive feature analogies like /P/ is to /B/ as /T/ is to /D/ (a change from voiceless to voiced). It is also noted that the analogies arising from the embeddings do not always coincide with what would be expected from distinctive features, and that the "flexibility to learn a vector space representation that does not always strictly conform to distinctive features is then an advantage of the representations" (Silfverberg et al., 2018). Chapter 3 explores the generation of phoneme embeddings in an effort to capture perceptual similarity in a data-driven manner.

2.3 Variation and Language Transference Effects

Language variation is a key concept in the field of sociolinguistics. It refers to the way language production differs, usually across different social groups, geographical locations, or time periods. Labov (1986) carried out a sociolinguistic study on the presence or absence of [ɹ] in post vocalic position in the speech of department store workers in New York City. The study discusses the [ɹ] variable as a social differentiator with its usage directly related to perceived social class. Similarly, differences in vowel productions by speakers across North America were collated in Labov et al. (2008) demonstrating the regional variation that exists across America and Canada and the active sound shifts taking place. Variation can be observed at the syntactic or morphological levels but this thesis focuses specifically on phonetic variation and differences in pronunciation seen in speakers of different spoken varieties. One source of spoken variation which features prominently in the later chapters of this thesis is that of *language transference* in multi-lingual speakers. This refers to the variation (phonetic or otherwise) which stems from the influence of a speaker's native language (L1) on the perception and production of an Second Language (L2).

The Perceptual Assimilation Model (PAM) (Best et al., 1994; Best et al., 2007) is a framework for understanding how listeners perceive and categorise non-native sounds. Essentially, upon exposure to a novel speech sound, a listener will either categorise this sound as an example of an existing category or of a non-native sound category. PAM considers the ease with which a listener can perceive certain sound contrasts in a non-native language based on the contrasts of their native language. It suggests that phoneme contrasts in the L2 which also exist in the L1 are easily perceived whilst those that do not are discernible based on how similar the realisations are to instances of native categories. The realisations of a phoneme contrast in the non-native language might be perceived as instances of a single phoneme category in the native language and are likely to be assimilated into the existing native phoneme category. This makes the contrast difficult to perceive and produce for the listener. Conversely, L2 realisations which are particularly deviant from any L1 category are more likely to result in the formation of a new phoneme category which is more easily perceived.

Another prominent model relating to the production of non-native speech sounds is that of Flege and Davidian (1984) known as the Speech Learning Model (SLM). Under this framework, sounds in the L2 are mapped to the closest (most similar) sound in the L1. The greater the perceived difference between an L2 sound and the L1 sound, the more likely it is that the phonetic differences will be discerned, and a new category will form for the L2 sound. Category formation, however may be blocked if the L2 sound is perceived as equivalent to an L1 sound in which case it is much harder for the speaker to perceive or produce the distinction.

Through these models, difficulties faced by L2 speakers in the perception or production of specific phonemic contrasts can be predicted and systematic patterns of variation can be explained in relation to the phonetic and phonological properties of the L1. As such, speakers with the same L1 will likely have similar spoken L2 varieties and exhibit the same pronunciation features. This concept is central to the speaker modelling and subsequent interpretations presented in Chapters 6 and 7.

2.4 Spelling Correction

Kukich (1992) describes spelling errors as belonging to one of two types; typographic or cognitive. The former occurs when the writer knows the correct spelling of a word but makes an error when producing it. For example, entering a different character than intended by mistakenly pressing an adjacent key on the keyboard. Early spelling correction algorithms use character edit distances between misspellings and real-word corrections in order to suggest a correction that is orthographically similar to the error. They rely on the finding that the majority of misspellings differ by a single edit operation (insertion, deletion, substitution, or transposition) (Damerau, 1964).

By contrast, cognitive errors result from a lack of knowledge of how to correctly spell a word. Misspellings that result from an effort to capture the sound sequence of a word fall under a subset of cognitive errors labelled phonetic errors and they typically deviate substantially from the target word (Kukich, 1992). Improved performance on more deviant errors is seen with the use of noisy-channel models which allow for multiple edit operations (Church and Gale, 1991). In particular, Brill and Moore (2000) demonstrated significant performance improvements to the noisy channel model by calculating the probabilities of string-to-string edits and combining these when comparing a misspelling to real-word candidate corrections.

The correction of phonetic errors in particular has been tackled by incorporating pronunciation information as opposed to just orthographic representations. Veronis (1988) uses a weighted edit-distance algorithm where the costs of edit operations are based on the phonetic similarity between graphemes. It is also common to convert words from their

orthographic form to one which captures the phonetic features. For example, Soundex, described in Kukich (1992) and patented in Russell and Odell (1918), maps words to a fixed length alpha-numeric code based on their characters. Numeric values are assigned to groups of letters that are phonetically similar. Thus words which are pronounced similarly will have the same encoding (e.g. 'sure' and 'shore' both have encoding S600). Edit-distance algorithms can be applied to these encodings to find real-word candidate corrections that are phonetically similar to a misspelling.

However, Soundex has been criticised as being too general given its limited permutations (Hodge and Austin, 2001; Mendonça Almeida et al., 2016). Thus, phonetic transformation rules, determined by linguistic knowledge of the target language, are used before encoding in approaches like those of Philips (2000) and Hodge and Austin (2001). Alternatively, phonemic forms can be used directly by transforming a misspelling to its corresponding phoneme sequence using letter-to-sound-rules as in Fisher (1999), Toutanova and Moore (2002), Khoury (2015), and Mendonça Almeida et al. (2016).

Other approaches to spelling correction include tackling the problem as one of Machine Translation (Aw et al., 2006; Silfverberg et al., 2016) or as a synthesis/recognition task (Stüker et al., 2011). Spelling correction tools targeted specifically towards children have also arisen more recently. Downs et al. (2020) released Kidspell: a child-oriented, rule-based, phonetic spellchecker. Their system makes use of the phonetic rules of English to map letters to keys which aim to capture accurate phonetic representations. Candidate suggestions for spelling correction are generated by identifying words with matching or similar phonetic keys. The incorporation of phoneme similarity information into a spelling correction tool and the benefits it exhibits for children's misspellings is the subject of Chapter 4.

2.5 Early Literacy Acquisition

Nowadays, a popular focus in early literacy education is the teaching of phonics with schools across the English speaking world incorporating the method into their curricula (National Council of Curriculum and Assessment, 2019; Bowers and Bowers, 2017). Phonics based approaches to reading and writing involve teaching the relationship between letters and sounds. For example, the word 'cat' can be broken down into the letters 'c', 'a', and 't' and the corresponding sounds of the phonemes /K/, /AE/, and /T/. When tackling the spelling of an unfamiliar word, children are then encouraged to adopt a "sound it out" approach by identifying the phonetic sequence of the word and the letters which represent these sounds. This approach is heavily relied upon by low achieving spellers (Daffern and Critten, 2019).

Through phonics education, reading and writing development become intrinsically linked to the spoken language which raises the question of how spoken variation impacts the acqui-

sition process. Early research into factors affecting children's literacy acquisition surmised that speaking a "non-standard" or "non-mainstream" language variety hindered reading and writing performance with Schwartz (1982) coining the term 'dialect interference'. More recently, an alternative (though not mutually exclusive) explanation for the association between non-mainstream productions and literacy achievement has become prevalent in the literature. This is the idea that children who produce more non-mainstream forms, particularly in contexts where this would not be appropriate (i.e. in the classroom), potentially have less linguistic awareness in general. The lack of sufficient awareness and flexibility required to code-switch likely extends to other aspects including phonological awareness which is regarded as integral to literacy acquisition (Terry and Scarborough, 2011). Thus, it is not the case that the use of "non-mainstream" forms or "non-standard" dialect usage in general negatively impacts reading and writing skills. Rather, a high frequency of usage is an indicator of an underlying linguistic weakness that also impacts literacy.

Much of the existing research on the interaction between spoken variation and literacy acquisition typically focuses on children with an African American English (AAE))dialect and the deviation from what is referred to as Mainstream American English (MAE) (Charity et al., 2004; Terry, 2006; Connor and Craig, 2006; Terry and Scarborough, 2011; Terry, 2012; Terry and Connor, 2010). This research confirms that children who frequently use the AAE variant typically have more trouble with learning to read and write. In particular, Terry and Connor (2010) demonstrated that words with dialect sensitive features caused spelling issues in both struggling and typically achieving readers. Beyond the examination of AAE, similar supporting research on spoken language and education has been carried out concerning language variants of Dutch, Arabic, Greek, and Caribbean creoles (Driessen and Withagen, 1999; Siegel, 1999; Saiegh-Haddad, 2003; Yiakoumetti, 2007). This is evidently an issue not limited to the AAE variety or even to varieties of English.

In their systematic review, Snell and Andrews (2017) suggest that there is insufficient research on the relationship between regional accent or dialect and written English literacy in England. From the works reviewed, they surmise that there is no straightforward relationship between literacy achievement and language background and state that regional dialect has only a minor impact on writing. However, throughout the review, difficulties with spelling were not considered related to pronunciation variation and instead a result of the complexity between English orthography and pronunciation that affects all children regardless of their accent. Terry (2006) proposes that "while all children must learn to negotiate mismatches between speech and print [...] this process may be particularly problematic for children whose spoken language differs substantially from standard written forms". In much the same way as one might employ a targeted approach to teaching a foreign language based on the learner's native language, it could prove beneficial to children with regional language variants to be taught reading and writing skills with such variation in mind and to have access to tools designed specifically to handle their variant. Chapter 5 investigates this idea further.

2.6 Automatic Speech Recognition

Automatic speech recognition is the process of transcribing spoken language into text using computational methods. For decades, ASR systems typically consisted of three component parts; an acoustic model, a pronunciation model, and a language model. Acoustic models traditionally used *Hidden Markov Models* (HMMs) and *Gaussian Mixture Models* (GMMs) to model the relationship between the acoustic features and phones in the speech signal. Hinton et al. (2012) then introduced the use of a deep learning model, a *Recursive Neural Network* (RNN) to replace the GMM of the acoustic model. This HMM/RNN hybrid model outperformed the traditional models and was seen as a significant breakthrough in the field.

More recently, another breakthrough has reshaped the field of speech recognition. End-to-end models are fast gaining popularity and exhibiting state of the art performance on a number of benchmarks (Li et al., 2022). These models directly translate the input speech signal into an output sequence using a single network thus removing the need for the separate modelling components of traditional systems. One popular end-to-end ASR approach is known as the *transformer* (Vaswani et al., 2017). Transformer based ASR models make use of self-attention mechanisms (Bahdanau et al., 2014) which allows the model to learn which parts of an input sequence are most relevant to predicting the required output.

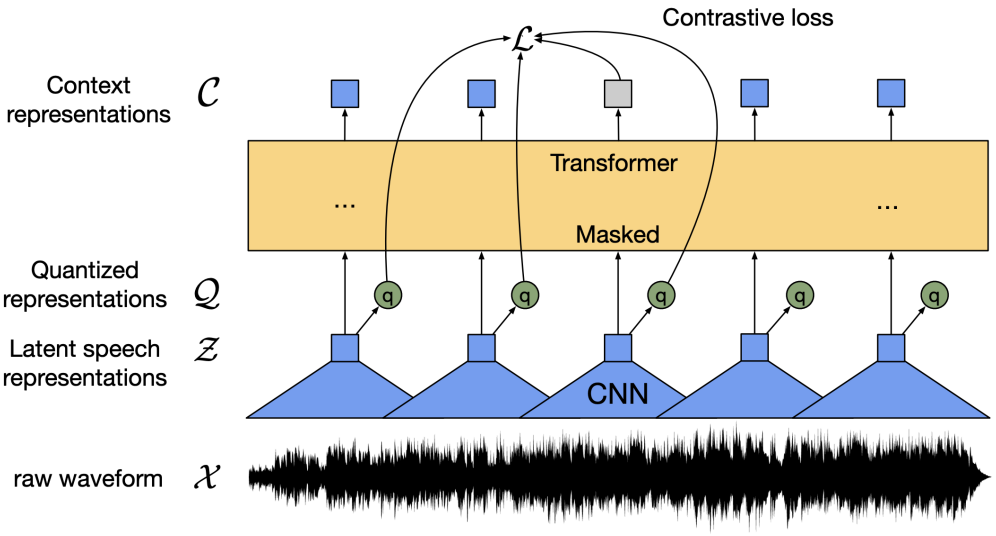


Figure 2.6: Architecture of the wav2vec 2.0 model (Baevski et al., 2020).

The ASR component used in this thesis is the wav2vec 2.0 model (Baevski et al., 2020). This ASR model uses an unsupervised learning approach to encode unlabelled speech sounds from audio during training. The model can then be fine tuned using a small proportion of labelled data for specific downstream tasks. The network architecture of the wav2vec model is given in Figure 2.6 (taken directly from Baevski et al. (2020)). The feature encoder

first takes the raw audio data and produces the latent speech representations. These are then passed to the transformer component which outputs contextualised representations. For the self-supervised training phase, the latent speech representations are discretised to a finite set of speech representations via quantisation. Model training then uses unlabelled speech data in a contrastive learning approach where the true quantized representation for a masked time step must be chosen from amongst a set of distractors. In the fine tuning phase, a comparatively smaller proportion of labelled speech data is used to train the system using *Connectionist Temporal Classification* (Graves et al., 2006). The model learns a linear projection between the context representations output by the transformer component and the required textual output. Interestingly, this output is not constrained to real English words and transcripts can contain non-words which appear to capture the phonetic properties of the speech. This is further explored in Chapter 6.

ASR technology has a number of everyday use-cases, from real-time captioning to voice controlled assistants. One particular application examined within this thesis is the use of ASR to aid in the annotation of large scale speech corpora for the purpose of linguistic research. In an investigation into the use of ASR for the purposes of linguistic annotation, Markl (2022) notes the potential time-saving benefits to incorporating ASR and the feasibility of working with much larger corpora. However, they outline the limitations of such tools and the challenges faced in applying ASR to diverse speech corpora containing various accents, topics, and recording environments. It seems clear, then, that the incorporation of ASR technology could be of great benefit to large-scale linguistic research but the deterioration of performance when used with “non-standard" speech is a barrier for many researchers. Even when choosing an ASR solution best suited for the speech data to be annotated, the manual correction of any errors in the output can still be a difficult task when the behaviour of the system is not fully understood. Investigating ASR sensitivity to spoken variation is a focal point of Chapter 7.

Chapter 3

Phoneme Similarity in English

The idea of measuring phoneme similarity is a complex one since, by definition, phonemes are abstract categories. From a theoretical perspective, phoneme similarity is usually expressed in terms of phonological features. Studies into the sound patterns of languages often refer to natural classes where phonemes form groupings based on shared features. Some phonological theories rely on judgements regarding the faithfulness (or similarity) of an underlying form to its surface form (Kenstowicz and Kisseberth, 1979; Smolensky and Prince, 1993). However, it has been noted that a purely feature-based description of similarity between phonemes can contradict observed phoneme groupings which are thought to be driven by physiological and perceptual factors (Mielke, 2012). Furthermore, Gallagher and Graff (2012) state that perceptual similarity "does not necessarily coincide with natural classes or feature values". In other words, there exists a disparity between native speakers' perceptions of phoneme similarity and similarity as defined in terms of phonological features. This inconsistency is thus the focus of this chapter which seeks to address Research Question 1: *What factors influence a speaker's intuitions regarding phoneme similarity?*

To explore this question, phoneme similarity models based on both phonological features and perceptual experiments are constructed and compared. Through this comparison, it is established that there are aspects of the resulting similarity hierarchies which differ. Thus suggesting that phonological features alone are not sufficient in adequately modelling perceptual similarity. Then, prompted by the hypothesis that usage influences perception, a similarity model based on phonological distribution in English is constructed using a phoneme embedding approach similar to that of *word2vec* (Mikolov et al., 2013). It is demonstrated that this model appears to capture characteristics of the perception based model that are not adequately explained by the feature based model of similarity. Finally, a similarity matrix of the English phonemes is generated in a data-driven manner using these distributive embeddings combined with the acoustic properties of the phoneme realisations.

The main contributions presented in this chapter include an investigation into the effect of phoneme distribution on perceptual similarity in English. The results indicate that the frequency and contextual environments of phonemes, as they are typically used in the language, influence a speaker's intuitions regarding similarity and confusability. A further contribution is the generation of a purely data-driven model of phoneme similarity in English which combines acoustic and distributional properties of phonemes. This results in a model of similarity which more closely resembles speaker perceptions and phoneme confusability and which acts as the foundation of many of the following chapters in this thesis. This work

has previously been presented in part in O'Neill et al. (2018) and published in O'Neill and Carson-Berndsen (2019). The outline of this chapter is as follows. In Section 3.1 a similarity hierarchy of English consonantal phonemes is generated based on work by Bailey and Hahn (2005). A second hierarchy based instead on speaker perceptions and phoneme confusability stemming from Cutler et al. (2004) is then presented in Section 3.2. Section 3.3 details the method for generating distribution based phoneme embeddings, and the resulting similarity hierarchy and a comparison of the three models is given in Section 3.4. The limitations of this investigation are discussed in Section 3.6 and the chapter concludes with a summary of the work in Section 3.7.

3.1 Modelling Similarity through Phonological Features

Classifying and categorising phonemes in terms of distinctive phonological features is a standard practice. However, there are differing schools of thought when it comes to agreeing on the set of features which should be used to describe the phonemes of a language. Some would argue that the minimal number of features that can adequately differentiate between all phonemes is optimal whilst others feel a rich and complex system of features can be more informative (Jakobson et al., 1951; Frisch, 1996). Some believe that a distinctive feature should be a binary property whilst others use multi-valued features (Chomsky and Halle, 1968; Bailey and Hahn, 2005). Thus, before exploring similarity in the context of phonological features, an appropriate feature specification must be chosen.

Bailey and Hahn (2005) investigate a number of similarity models based on their ability to predict human judgements of similarity. Participants performed an auditory, two-alternative, forced choice task where they were asked to pick which of two words was most similar to a given target. Choice words differed from the target word in a single phoneme (for example, /B AH S P/ - /P AH S P/ and /B AH S P/ - /G AH S P/). In the first experiment, choice A differed from the target word by a single phonological feature whilst choice B differed by two. The second experiment used a random sample of all possible phoneme contrasts. Both feature based and empirically driven models of English phoneme similarity were investigated for their ability to predict the favoured choice in these experiments. It was concluded that the best predictor is a minimal and multi-valued feature model based on "simple counts of the number of features [...] in which the two phonemes fail to match". As such, this is the model chosen for analysis in this chapter. The model uses 4 equally weighted, multi-valued, distinctive phonological features;

- *Place of articulation*: labial, dental, alveolar, palatal, velar, or glottal.
- *Manner of articulation*: plosive, fricative, nasal, glide, lateral, rhotic or affricate.
- *Sonority*: sonorant or obstruent.
- *Voicing*: voiced or voiceless.

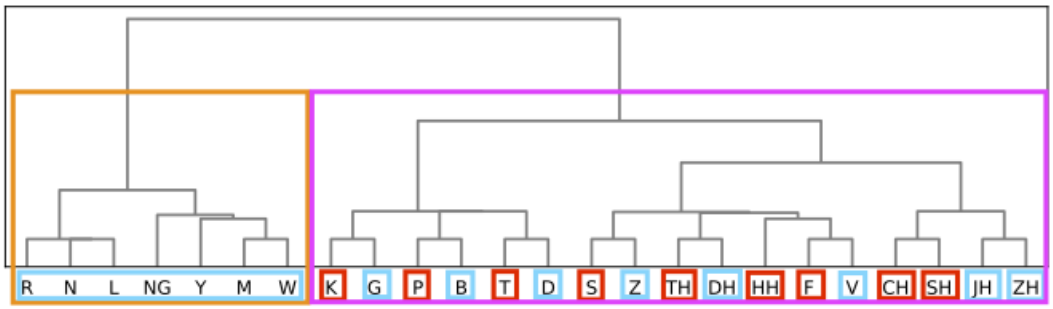


Figure 3.1: The similarity hierarchy generated from the phonological feature based model.

Note that only consonantal phonemes were considered in the original experiments and so only consonants are modelled. Using the distance matrix presented in Bailey and Hahn (2005), the Ward clustering algorithm (Ward, 1963) is applied to produce hierarchical clusters which could then be visualised as a dendrogram as seen in Figure 3.1. Ward's method tends to create small and even sized clusters and is widely used in linguistics, for instance in the areas of corpus linguistics and dialectometry (Szmrecsanyi, 2013).

Here, the high level major distinction is seen to be that of sonority with a clear separation of *sonorants* (highlighted in orange) and *obstruents* (pink). At the lowest level the majority of pairwise clusters group phonemes that differ only in their voicing property (*voiced* sounds in blue and *voiceless* in red). This similarity hierarchy is in agreement with most theoretically-based descriptions of phonemes and resembles the groupings seen in the IPA chart (International Phonetic Association and others, 1999).

3.2 Modelling Similarity through Speaker Perception

Modelling phoneme similarity through speaker perception relies on examining how speakers produce and recognise the sounds of their language. The forced choice task of Bailey and Hahn (2005) detailed previously in Section 3.1 is one such method of doing so. Another approach looks to what is deemed acceptable in the context of rhyme and half-rhyme (Zwicky, 1976; Kawahara, 2007; Johnsen, 2011). However, a significant theory regarding perceptual similarity, and one which is foundational to the work in this chapter and the thesis as a whole, is the idea that "similarity is a function of perceptual confusability" (Gallagher and Graff, 2012). That is, phoneme similarity is directly correlated with the likelihood of a speaker confusing one phoneme for another through either perception or production.

Experimentation carried out by Cutler et al. (2004) asked both native and non-native English speakers to identify phonemes in consonant-vowel or vowel-consonant pairings. Varying levels of background noise were used to invoke errors in the identifications, and the frequencies with which phonemes were recognised incorrectly were recorded in a series of confusion

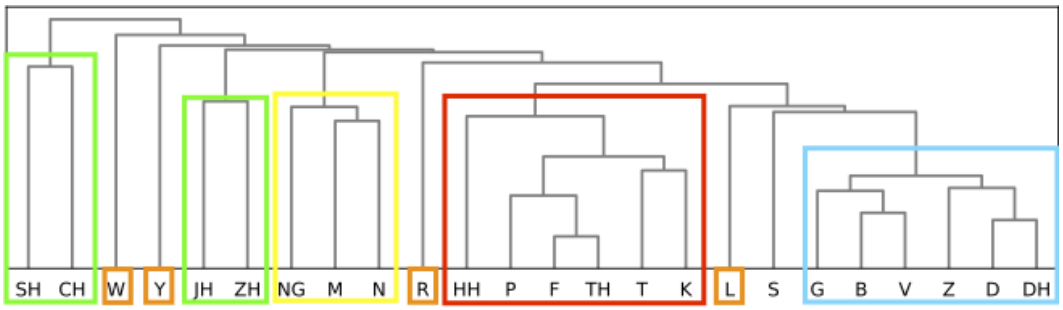


Figure 3.2: The similarity hierarchy generated from the speaker perception based model.

matrices. For the purposes of this exploration into phoneme similarity from the perspective of speaker perceptions, the confusion matrices corresponding to native speaker identifications of consonants in both word-initial and word-final position were combined by summing their values. In this way, potential language transference effects from the Dutch speaking participants are avoided. Since the intention was to make a comparison with the feature based similarity hierarchy generated previously in Section 3.1 where only consonants were considered, again, just the results related to consonantal phonemes were used. The resultant confusion matrix is taken from the original paper and, again, the Ward clustering algorithm is used to produce hierarchical clusters which were visualised in the dendrogram shown in Figure 3.2.

Whilst it is noted that a single confusability experiment is unlikely to give rise to a perfect model of perceptual similarity, it is believed that the resulting hierarchy is sufficient in highlighting general areas of significance in how phoneme similarity is judged by native speakers. In contrast to the feature-based model, the perception-based model exhibits a clear distinction between the *voiced* and *voiceless* plosives and fricatives which form two separate clusters in the hierarchy (highlighted in blue and red respectively). Furthermore, it suggests the *palatal* sounds [/SH/, /CH/, /JH/, /ZH/] are more weakly connected than a feature-based model might suggest (green), *nasal* sounds [/M/, /N/, /NG/] form their own distinct cluster (yellow), and the other sonorants [/W/, /Y/, /R/, /L/] (orange) known as the *approximants* do not.

3.3 Modelling Similarity through Phoneme Distribution

Whilst there has been some level of investigation into the differences between perceptual and feature-based models of similarity and their respective performances at predicting observed phonological usage (Bailey and Hahn, 2005; Johnsen, 2011), there has been little study into what causes this difference. A particular novelty of this PhD research lies in its examination of a potential source of this discrepancy, namely, how the distributional patterns of phonemes might influence a speaker's judgement of phoneme similarity.

3.3.1 Motivations

There exists evidence to suggest that perception and experience are linked. Perceptual learning experiments have demonstrated that an individual's mental representation of phoneme categories are malleable and can adapt with exposure to new inputs (Norris et al., 2003; McQueen et al., 2006; Scharenborg and Janse, 2013). In these studies, participants first engage in a training phase where an ambiguous sound is produced in the context of a word where the phoneme target is obvious. For example, a sound halfway between the typical realisations of /S/ and /F/ (henceforth denoted as /SF/) would be heard in either an /F/-context as in the word 'giraffe' or an /S/-context as in the word 'dress'. Note that neither *'girass' nor *'dreff' are words of English. Each participant is trained in one context using words where only the target phoneme would constitute a real word in their language. In the second phase, all participants are then exposed to new lexical items where either of the phonemes could constitute a real word. For example, /N AY SF/ could be interpreted as /N AY S/ 'nice' or as /N AY F/ 'knife'. Participants who are trained in the /F/-context are more likely to perceive this ambiguous item as 'knife' whereas those trained in the /S/-context typically hear 'nice'. This leads to the conclusion that an individual's mental representation of a phoneme category can change based on experience. This category adaptation is not dependent solely on the acoustic property of the sound since in this case all participants heard the same ambiguous sound. It is also impacted by the environment in which the sound is heard, as this context is what triggers the adaptation of either the /F/ or /S/ category. These studies prompted the hypothesis that usage and exposure might influence not just the mental representation of a single phoneme, but the relationships between phoneme categories. Intuitions regarding the similarity between phonemes might be impacted not just by the acoustic properties of the sounds (which are often linked to their phonological features), but also by the environments in which they are observed to occur in everyday use.

3.3.2 Phoneme Embeddings

Phoneme embeddings have emerged as an area of interest for speech segmentation and for the determination of sound analogies (Ma et al., 2016; Parrish, 2017; Silfverberg et al., 2018; Kolachina and Magyar, 2019). It has also been demonstrated that the use of speech data and factors outside of traditional phonological features result in embeddings that better capture phonological similarity (Glass, 2018; Do and Lai, 2019). In order to investigate this idea further, phonemes were modelled based on their distribution within English using a word embedding approach - specifically by adapting the word2vec algorithm (Mikolov et al., 2013). Semantic similarity between words has been shown to be reliably captured through the modelling of words based on their distribution. This same idea is applied at the phoneme level to determine whether phoneme similarity can be modelled based on the environments in which they occur and whether this model of similarity can explain the differences between the previously discussed feature based and perception based models.

To model distributional similarity phoneme embeddings are generated from their observed environments. Rather than modelling words based on the sentences in which they occur and working under the assumption that semantically similar words appear in similar contexts, phonemes are modelled based on their syllabic environments assuming that distributionally similar phonemes will occur in similar contexts.

3.3.3 Data Preprocessing

The modelling of phonemic similarity from the distributional properties of phonemes is tackled in a purely data-driven manner by adapting the typical word2vec approach. The data used comes from the Brown Corpus, a corpus of American English texts from a variety of sources and covering a variety of topics (Francis and Kučer, 1964). Each word in the corpus is translated to its equivalent sequence of phonemes using the CMU pronunciation dictionary (Weide, 1998) resulting in approximately 1 million phoneme sequences. For words with multiple possible pronunciations one is chosen at random. By doing so, the over exposure to words with many pronunciation possibilities is avoided as this would impact the frequency effects. For words which do not appear in the dictionary, a grapheme-to-phoneme tool trained on the CMU pronunciation dictionary is used to predict its corresponding phoneme sequence (CMUSphinx, 2016).

The environment considered for phoneme embeddings differs to that of word embeddings. In the case of word embeddings, a word is supplied to the model in context using the words occurring before and after the target word. Whilst the exact window size is specified as a hyperparameter (typically 5 words either side as default), the context words are never taken from across a sentence boundary. For phonemes, embeddings are generated using the context of the surrounding contextual phonemes but never across a syllable boundary. The syllabic environment is chosen since the phonotactic rules of English, which constrain the environments in which a phoneme can occur, typically apply at the syllable level. Across syllable boundaries there are minimal restrictions on phoneme sequences (although there are effects on phonetic realisations which are outwith the focus of this work).

To achieve syllabic phoneme sequences, a syllabification tool is developed in order to split the word-level phoneme sequences into syllables. Syllable boundaries in words are assigned based on the phonotactic rules of English and linguistic theories. These include the sonority sequencing principle (Selkirk, 1984), where sounds at the syllable edges tend to be less sonorous than those in the middle, and the maximum onset principle (Kahn, 1980), where the preferred syllable structure is one that maximises the number of consonants in the syllable onset rather than the coda. This resulted in approximately 1.5 million syllables with 4 million phoneme tokens, 2.5 million of which were consonants. An example of the phoneme translation and syllabification process can be seen in Figure 3.3. The syllabification tool is available at the link in Appendix A.2.1.

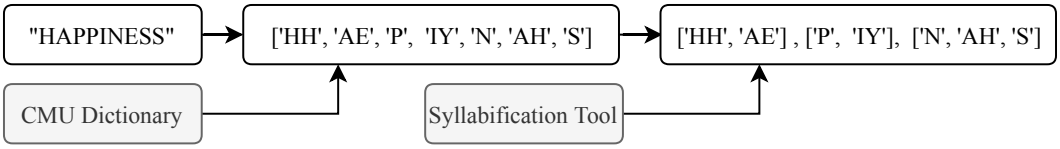


Figure 3.3: An example of the syllabification process for the word “happiness".

3.3.4 Model Parameters

To generate the phoneme embeddings the *gensim* word2vec implementation is used (Řehůřek and Sojka, 2010). Where a typical word embedding model takes individual sentences as inputs, for the purposes of phoneme embeddings the inputs are syllabic phoneme sequences. There are two possible model architectures for generating embeddings using word2vec: the Continuous Bag Of Words (CBOW) model and the skipgram model. A CBOW model seeks to maximise the likelihood of the target given the context whereas a skipgram model seeks to maximise the likelihood of the context given the target. Both model architectures were tested in generating phoneme embeddings with a qualitative analysis of the outputs suggesting that the CBOW model is more appropriate for this task. As stated by Levy et al. (2015) the effectiveness of word2vec lies mainly in the choice of hyperparameters. These parameters include vector dimensionality, window size, the use of subsampling, and the use of negative sampling.

For word-embeddings the recommended vector dimension is somewhere between 100 and 1000 (Mikolov et al., 2013). However, given the reduced vocabulary size of the fixed set of all phonemes (39 unique tokens) compared to the ever growing set of all words, a significantly smaller dimension of 20 is chosen. The window size parameter determines how far removed from the target a context token can be located and still be considered relevant context. As stated previously, the phonotactic rules of English that affect the environment in which a phoneme can occur apply at the syllable level. However these constraints are more heavily effected by the phonemes immediately before and after the target. Thus a window size of 1 token either side of the target is chosen. Unlike with phoneme embeddings, when generating word embeddings in an effort to capture semantic similarity, it is desirable to disregard frequency effects. For instance, two words like “good" and “wonderful" would be considered synonyms and as such their vector representations should ideally be close together in the vector space. However, “good" appears in general usages with much greater frequency than “wonderful" and this can negatively impact the embeddings. To overcome such a problem it is customary to use subsampling (where instances of very common words are ignored so as to lessen the effect of their frequency) and/or negative sampling (where randomly generated ’incorrect’ target-context pairs are generated to push the embeddings away from contexts where they would not occur). However, the frequency of a phoneme in typical usage is potentially an influential property that should be preserved for the purpose of

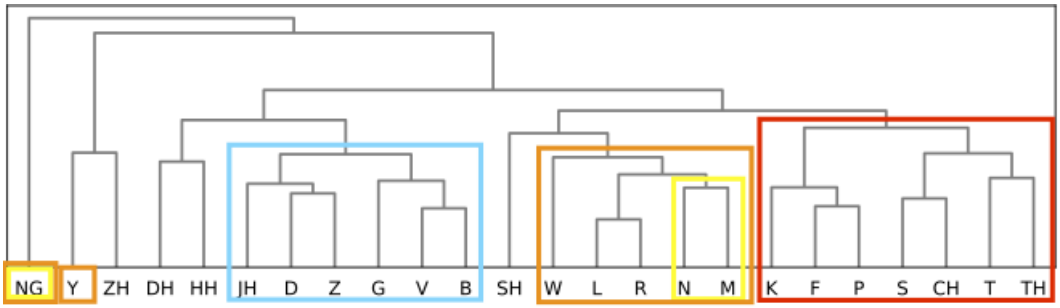


Figure 3.4: The similarity hierarchy generated from the phoneme distribution based model.

these embeddings. Furthermore, given the limited vocabulary size, it is likely that randomly generated negative samples would overlap with true instances in the data thus having an adverse effect. As such, these techniques are not applied.

3.3.5 The Distribution Based Similarity Model

The distances between the generated phoneme vector representations are calculated using the Euclidean distance metric as shown in Formula 3.1 where $d(x, y)$ is the distance between two vectors x and y, and n is the number of dimensions of the vectors. Typically the cosine distance (see Formula 3.2 again where $d(x, y)$ is the distance between two vectors x and y, and n is the number of dimensions of the vectors) is used to measure distance between word embeddings so as to avoid the effects of vector magnitude which is influenced by a word's frequency. However, in this case, these frequency effects are sought to be preserved. The calculated distances between all phonemes are stored in a distance matrix and hierarchical clustering, using the Ward clustering algorithm (Ward, 1963), is applied. The phoneme embeddings include both vowels and consonants which form two distinct and separate clusters. For the purposes of comparison with the previous models, only the branch containing the consonantal phonemes is examined and can be seen in the similarity hierarchy dendrogram of Figure 3.4.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.1}$$

$$d(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3.2}$$

Here, it can be seen that the voiced and voiceless clusters (highlighted in blue and red respectively) bear a striking resemblance to those seen in the perception-based model and not in the feature-based model. Looking at the nasal sounds (yellow), there appears to be something akin to the nasal cluster of the perception-based model but with an extreme separation of /NG/ from /M/ and /N/. This corresponds to the phonotactic rule of English preventing /NG/ from occurring initially in a syllable and thus behaving differently to /M/ or /N/. Interestingly, there also appears to be a cluster consisting of most of the sonorants (orange) which is more consistent with the feature-based model than the perception-based.

3.4 Comparison of the Similarity Models

In this section the similarity hierarchies generated from each of the previously described models are compared. Focus is given specifically to general areas where the feature based and perception based models differ thus supporting the idea that phonological features alone are insufficient in modelling perceptual similarity. These areas are then examined within the distributive model through which phoneme distribution is demonstrated to be an influential factor in determining similarity perceptions.

3.4.1 The nasal sounds: /M/, /N/, and /NG/

Within the perception based model (Figure 3.2) the clustering of the nasal sounds /M/, /N/ and /NG/ is observed. Of the three sounds, /NG/ appears to be the more dissimilar of the three as evidenced by its leaf node being further away from the other two. This judgement also aligns with half rhyme usage where /M/-/N/ rhymes are much more common than /M/-/NG/ or /N/-/NG/ (Johnsen, 2011). This cluster is absent in the feature based model (Figure 3.1) since these sounds group more closely with phonemes that share their place of articulation rather than their manner of articulation. In this model /M/ is most similar to /W/, another labial sound, whilst /N/ is considered more similar to the other alveolar sonorants /R/ and /L/. If manner of articulation were to be more heavily weighted over the place of articulation a cluster of nasal sounds would likely occur in the feature based similarity hierarchy. However, it would be unlikely to exhibit the slight removal of /NG/ from the other two nasal sounds that is seen in the perception based model and in half-rhyme usage. Conversely, this distinction between /M/ and /N/ on the one hand and /NG/ on the other is magnified in the distribution based model (Figure 3.4). This model judges /M/ and /N/ to be very similar to each other whilst /NG/ is extremely far removed. This is presumably a result of the unique property of /NG/ only occurring in English syllable codas and never in onsets. This distributional property is potentially a source of the greater distance between /NG/ and the other nasal sounds in the perception based model.

3.4.2 The palatal sounds: /SH/, /CH/, /ZH/, and /JH/

Within the feature based model (Figure 3.1) there is a branch containing the pairings of palatal sounds: /CH/ and /SH/, and /JH/ and /ZH/. These pairings are also seen in the perception based model (Figure 3.2) but the degree of similarity between the phonemes in each pair and between the pairs themselves is much less (as evidenced by the length of the branches connecting them). In the perception based model the palatal sounds appear not to constitute a distinct clustering but instead are rather dissimilar phonemes with respect to the entire vocabulary. The average rate of occurrence in the corpus for a consonant is 4%. However, the palatal sounds fall significantly below this value with each of them making up less than 1% of the total number of phonemes in the corpus. The frequency with which a particular phoneme is encountered appears to impact the strength of mental associations between it and other phonemes. The palatal sounds being relatively 'rare' to a speaker seem to result in the similarity between them and their closest neighbours being much weaker than, say, that of the more frequent /T/ and /K/.

3.4.3 The voicing contrast

Perhaps the most prominent characteristic of the perception based model (Figure 3.2) is the distinct separation between voiced and voiceless obstruents. The two main clusters referred to in this subsection are the branch containing the voiced sounds /G/, /B/, /V/, /Z/, /D/, and /DH/ and the branch containing the voiceless sounds /HH/, /P/, /F/, /TH/, /T/, and /K/. These consist of the plosives and fricatives of English minus the palatal fricatives discussed previously in Section 3.4.2. In the feature based model (Figure 3.1) voicing is typically the most fine-grained distinguishing feature between a pair of phonemes that have all other features in common. As a result a phoneme's most similar neighbour is typically its voicing counterpart: /P/ and /B/, /K/ and /G/, /F/ and /V/ etc. The perception based model, on the other hand, would suggest that a difference in voicing is much more distinguishing than is captured by the feature based model and this difference is also seen in the distribution based model (Figure 3.4). Here, two clusters are observed which are almost identical to the two seen in the perception based model. In particular there is the branch containing voiced sounds /JH/, /D/, /Z/, /G/, /V/, and /B/ and the branch containing voiceless sounds /K/, /F/, /P/, /S/, /CH/, /T/, and /HH/. The voicing of plosives and fricatives is evidently an important aspect of the sound in terms of its phonemic environment and the phonotactic rules surrounding it. In turn this has led to a significant perceptual distinction between these groups of sounds.

3.5 A Data-Driven Model of Phoneme Similarity in English

Whilst it is clear that the distribution-based model captures some salient features of perception that are absent in a feature-based model, on its own it does not appear to be objectively 'better'. Both the feature and distribution based models have their merits and it would seem that a combination of the two would be a closer step towards modelling native speaker judgements. However, the data driven nature of the distribution based model is a favourable attribute due to the analogous learning style to that of a speaker. The average speaker makes judgements of similarity without explicitly being told about phonological features. Their perceptions are generated purely through exposure to the language and as such a model of these perceptions would ideally learn in a similar fashion.

To overcome this issue, a data-driven approach to modelling phoneme similarity in a way that captures the same type of information as the knowledge-driven feature based model is sought. Kane and Carson-Berndsen (2016) presented work whereby an enhanced confusion matrix was produced from acoustic speech data. In essence, an acoustic model for a particular phoneme within an ASR system was excluded from the recognition process in an iterative manner. This resulted in forced errors which were used to supplement a standard confusion matrix of ASR results. The similarities captured by this enhanced confusion matrix are remarkably similar to those of the feature-based model. Using this confusion data and the Euclidean distances between the previously generated phoneme embeddings, a distance matrix is constructed with similarity values calculated as in Equation 3.3 where S_{PQ} is the similarity between phoneme P and Q, C_{PQ} is the rate of confusion from the acoustic modelling experiment where P was the target phoneme and Q was the forced confusion, and D_{PQ} is the Euclidean distance between the vector representations of phonemes P and Q from the phoneme embedding approach.

$$S_{PQ} = \frac{(1 - C_{PQ}) + (D_{PQ})}{2} \tag{3.3}$$

Using this similarity matrix, hierarchical clustering was again performed using the Ward clustering method (Ward, 1963) and the resultant similarity hierarchy can be seen in the dendrogram of Figure 3.5. Again, the vowels and consonants form two distinct clusters so only the consonantal branch is shown here for comparison with the previous models.

The similarity hierarchy of the acoustic-distributional model reveals a number of salient points. The sonorant cluster (in orange) is still evident but the nasals (yellow) resemble those in the perception-based model. Some of the palatal sounds (in green) also retain the higher level of dissimilarity noted in the perception-based model. Interestingly, there is also the separation of the obstruents into two clusters based on place of articulation, specifically the coronals, (purple), and the non-coronals (dark blue).

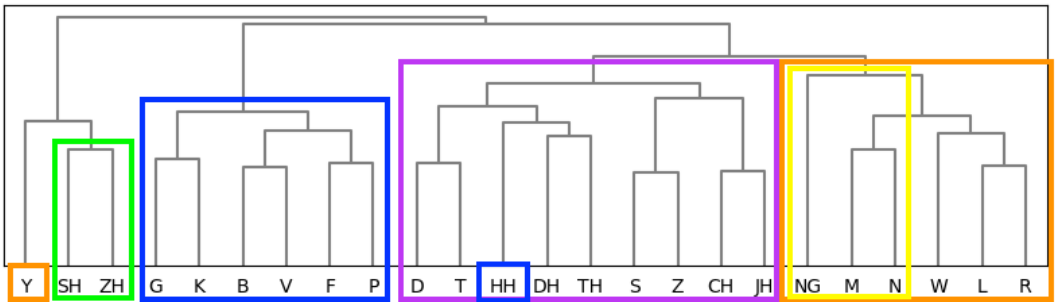


Figure 3.5: The similarity hierarchy generated from the acoustic and distributional based model.

This particular model of phoneme similarity, in the form of the distance based similarity matrix constructed from acoustic and distributional properties of phonemes, plays a significant role in the following chapters of this thesis. It is visualised as a heatmap in Figure 3.6 with phonemes along the y-axis relating to those target phonemes removed during acoustic modelling and phonemes along the x-axis relating to those which were recognised instead. Darker cells of the matrix represent a higher degree of similarity between the phonemes in terms of their acoustic realisations and their distribution.

3.6 Limitations

This chapter demonstrates that the distribution of English phonemes in everyday usage appears to influence speaker perceptions of phoneme similarity in ways which are not adequately explained by phonological features alone. However, the approach is not without its limitations. Firstly, the perceptual model was based on a single confusability experiment which cannot be considered to be representative of the intuitions of every English speaker. Whilst it is not suggested that such a model is a ground truth of human perception, additional studies could be examined and their results pooled to determine the shared observations.

Furthermore, the word2vec algorithm was originally intended to model words and the semantic similarity between them based on the contexts in which they occur. In terms of the context window, words further removed from the target word typically have less weighting during modelling but the exact word order is lost. This is typically not a major issue in modelling the meaning of a word. Phonemes, however, are much more constrained in relation to the sequence in which they occur. The sonority sequencing principal suggests that less sonorant sounds, like plosives, typically occur at syllable boundaries whilst more sonorant sounds, like liquids, occur nearer the syllable nucleus. This ordering of phonemes is lost through the word2vec approach and as a result syllabification must be carried out prior to generating embeddings and the window size must be constricted to a single phoneme either side of the target. It is possible, then, that an encoding approach that makes use of attention,

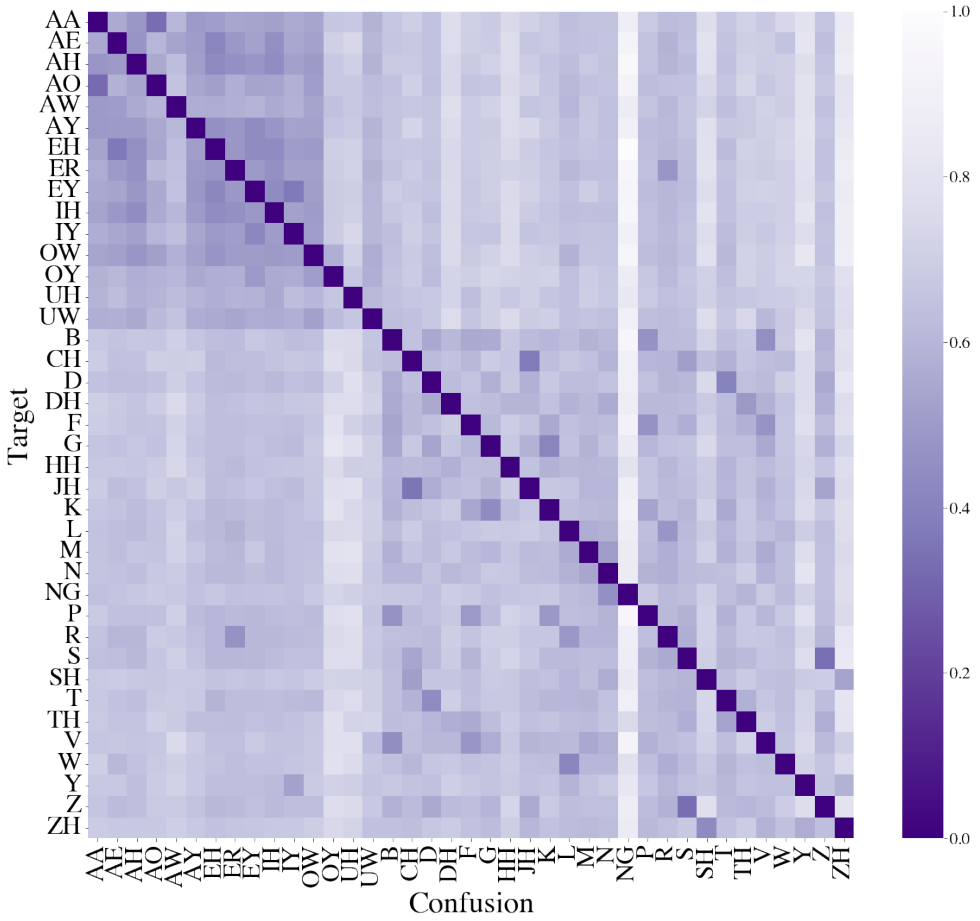


Figure 3.6: The phoneme similarity matrix based on phoneme distribution and the acoustic properties of their realisations visualised as a heatmap.

like the encoder component of transformer models, might capture more environmental influences on the phonemes and produce improved phoneme embeddings for the purposes of measuring their distributional similarity.

Finally, a data-driven model of phoneme similarity was constructed based on phoneme distribution and the acoustic properties of their realisations. However, this model was not tested for its ability to predict human perceptions in the likes of forced choice tasks or confusability experiments. In the following chapters, this model is applied to the task of spelling correction and shown to be beneficial in this domain. However, future work could investigate more directly how the similarities captured in this model compare to human intuition in a more quantitative and empirical manner.

3.7 Summary

This chapter sought to answer Research Question 1: *What factors influence a speaker's intuitions regarding phoneme similarity?* Prompted by perceptual learning experiments which demonstrate how an individual's mental representation of an abstract phoneme category can change and adapt based on exposure, the usage and distribution of phonemes in English was investigated for its potential influence on similarity judgements. By adapting the word2vec word embedding algorithm and applying it at the phoneme level, phoneme embeddings were generated based on the contexts in which specific phonemes occur. Similarity between phonemes could then be examined by comparing the distances between the vector representations in the multi-dimensional space. Hierarchical clustering was performed on three different models of phoneme similarity (feature based, perception based, and distribution based) and it was demonstrated how aspects of perceptual similarity that were not explained by phonological features appeared to be impacted by the distributional properties of the phonemes. Finally, an overarching model of phoneme similarity was constructed in a purely data-driven manner by combining both the distributional properties of the phonemes and the acoustic properties of their realisations.

Chapter 4

Children's Spelling Correction

In this chapter, a practical application of the previously constructed phoneme distance matrix is explored in the form of a spelling correction tool designed specifically to target the phonetic spelling errors produced by children. Traditional spellcheckers typically focus on typographic errors where the source of the misspelling lies not in a lack of knowledge of the spelling of a word but in its production. For instance, hitting an adjacent key on a keyboard to the one intended whilst typing. Since typographic errors tend to deviate from the target spelling by only one or two characters (Damerau, 1964), spelling correction tools intended to correct such misspellings work best when the misspelling and target are similar on an orthographic or character level. In contrast, cognitive errors stem from not knowing how to spell a word correctly and are common in the writings of children still acquiring written literacy. The popular phonics based approaches to teaching encourage children to 'sound out' words they are unsure how to spell by identifying the individual sounds which make up a word and then encoding them using letters which represent those sounds. As a result, phonetic misspellings are often produced which endeavour to capture the sound sequence of a word but which can deviate drastically from the intended target on an orthographic level. Such misspellings prove difficult for traditional typographic spellcheckers to correct. The core aim of this chapter lies in demonstrating how applying knowledge of phoneme similarity in English can aid in targeting these types of misspellings. It seeks to answer Research Question 2: *Can a model of phoneme similarity be applied to and benefit the development of language technologies; specifically in a spelling correction tool for children?*

This chapter introduces S-capade (Spelling Correction Aimed at Particularly Deviant Errors), a spelling correction method which suggests potential corrections for misspellings based on phoneme similarity. Misspellings are converted to sequences of phonemes using a grapheme-to-phoneme tool. These sequences are then aligned with those of potential real-word corrections by a weighted edit distance algorithm. This algorithm uses the values from the similarity matrix constructed in Chapter 3 as edit operation weights. S-capade is available publicly and can be found at the link in Appendix A.2.2. The strengths of the S-capade tool are then demonstrated by evaluating its performance on various corpora of misspellings compared with other spellcheckers. A further analysis of the more phonetic misspellings across the corpora discusses the potential sources of error and how they relate to aspects of spoken English. Overall, it is shown that the S-capade tool, through its use of a phoneme similarity approach, is capable of correcting phonetic misspellings which, on an orthographic level, are much further removed from the intended real-word target than other spellcheckers are equipped to handle.

The main contributions of the thesis presented in this chapter are the incorporation of a phoneme similarity model of English into an automatic spelling correction framework and an evaluation of its effectiveness on phonetic misspellings produced by children which deviate heavily from the real-word target. This work has been previously published in O'Neill et al. (2020). The rest of the chapter is structured as follows. First, Section 4.1 describes the development of S-capade with a focus on how the previous similarity matrix allows for potential real-word corrections to be ranked in terms of how phonemically similar they are to a misspelling. The system is then tested against other existing spellcheckers on corpora of misspellings from various sources in Section 4.2. Section 4.3 dives deeper into the phonetic misspellings of the corpora to demonstrate the influence of spoken English on the acquisition of written literacy. The limitations of the S-capade method and evaluation are considered in Section 4.4 followed by a final summary of the work presented in this chapter in Section 4.5.

4.1 The S-capade Method

Despite the prevalence of phonetic errors in children's writing, conventional spelling correction tools are not fully capable of correcting these types of misspellings and, as such, exhibit poorer performance on children's spelling. When a child uses a 'sounding out' approach to spelling they are approximating the sounds they perceive in the target word with letters they believe represent those sounds. As such, deviations from the correct spelling occur both as a result of incorrect phonemes being perceived, e.g. the phoneme /V/ being perceived as /F/ resulting in the misspelling *'gif' ('give'), and of incorrect letters being chosen, e.g. representing the /K/ phoneme with a 'c' and the /IY/ phoneme with 'ea' in the misspelling *'ceap' ('keep')[1]. The majority of misspellings resulting from the latter case encode the same sequence of sounds as the intended target and so are more easily corrected by simply converting a misspelling to its phonemic form (see Section 4.1.1) and finding a real word target with a matching phonemic form. However, misspellings of the former variety map to phoneme sequences that are similar to that of the correct spelling but not necessarily identical. In these cases some measure of similarity at the phoneme level is required so that, in the case of *'gif', the possible real-word target 'give' is ranked more highly as a correction candidate than, for example, the more orthographically similar word 'gig' as a direct result of the /F/ phoneme encoded in the misspelling being more similar to the /V/ in 'give' than the /G/ in 'gig'. As such, the automatic correction of children's spelling is a fitting task for applying a model of phoneme similarity.

[1] All misspelling examples, unless otherwise stated, are real instances taken directly from various corpora of children's writing.

4.1.1 Predicting Phoneme Sequences

In order to develop a spelling correction approach that works at the phoneme level, misspellings and their potential real word targets must be converted to sequences of phonemes. For real words this is done using the CMU Pronouncing dictionary (Weide, 1998) which contains over 120,000 words and their corresponding phonemic representations (based on North American English pronunciation). For example, upon looking up the word 'situation' in the dictionary the phoneme sequence /S IH CH UW EY SH AH N/ is returned. Since this dictionary is the same one used previously in Chapter 3 to build the similarity matrix, it uses the same symbols to represent the same set of phonemes (based on the ARPAbet symbol set). As such, by using the CMU Pronouncing Dictionary to obtain the phoneme sequences of words, the similarity values can be used directly from the matrix without requiring conversion to any other phoneme symbol set.

However, by their nature, misspellings typically are not found in any dictionary and so a different method is required for predicting the phoneme sequence which they encode. For any words which do not appear in the CMU Pronouncing dictionary, a grapheme-to-phoneme tool was used instead. G2P-seq2seq (CMUSphinx, 2016) is a 3-layer transformer model which converts grapheme forms to their phonemic counterparts. Importantly, the pre-trained model used in this work was trained on the CMU Pronouncing Dictionary and, as such, uses the same phoneme symbol set and produces sequences which are consistent with those obtained from the dictionary for real words. For example, applying the G2P tool on the misspelling 'sichweshen' produces the sequence /S IH CH W EH SH AH N/.

4.1.2 The Weighted Edit Distance Algorithm

Many character-based spellcheckers rank possible corrections of a misspelling based on their orthographic similarity. To do this, the Levenshtein (Levenshtein, 1966) or Damerau–Levenshtein (Damerau, 1964) distance is typically used which measures the edit distance between two words based on the number of operations which are required to convert one word to another. Edit operations include character insertions, deletions, substitutions, and, in the case of the Damerau-Levenshtein distance, transpositions. In order to find the minimal edit distance for a misspelling and a potential correction, the Wagner-Fischer dynamic programming algorithm (Wagner and Fischer, 1974) is employed to find the optimal alignment between two strings. The S-capade approach uses a similar method but, rather than applying the Wagner-Fischer algorithm and Levenshtein edit distance measure at the character level, it instead looks at the alignment and edit operations between strings of phonemes. Table 4.1 shows a comparison between character-level and phoneme-level alignments and edit distances for the misspelling *'sichweshen' and its corresponding target 'situation' in terms of deletions (d), insertions (i) and substitutions (s). As can be seen, comparison at

	Character Sequence										Phoneme Sequence							
Misspelling	S	I	C	H	W	E	S	H	E	N	S	IH	CH	W	EH	SH	AH	N
Real-Word Target	S	I	T	U	A	T	I	O		N	S	IH	CH	UW	EY	SH	AH	N
Edit Operations			s	s	s	s	s	s	d					s	s			
Edit Distance	7										2							

Table 4.1: A comparison of the Levenshtein edit distances and alignments at the character level vs the phoneme level.

a phoneme level shows a stronger degree of similarity between this type of highly deviant phonetic misspelling and its real-word target.

However, the Levenshtein edit distance measure only looks at the total number of edit operations and does not account for the fact that some operations might be more likely to occur than others. Again consider the example of the misspelling *‘gif’ whose target was the phonemically similar ‘give’ rather than another possible correction ‘gig’. Both potential targets differ from the misspelling by a single phoneme substitution and so the current edit distance measure would consider them equally likely. However, the voiceless fricative /F/ of *‘gif’ is much more similar to the voiced fricative /V/ of ‘give’ than the voiced plosive /G/ of ‘gig’ and so ‘give’ should be considered a more likely target for the misspelling. This is assuming the aim is to target phonetic misspellings exclusively rather than typographic ones. Otherwise, ‘gig’ would be a likely correction for ‘gif’ given the single character difference and the fact that ‘g’ and ‘f’ are adjacent on the keyboard. To address this idea of some substitutions being more probable than others, the S-capade approach uses a weighted edit distance measure whereby different operations have different associated weights (or costs) depending on how likely they are to occur. The cost associated with a particular phoneme substitution, if the phonemes involved are similar and thus likely to be confused, is relatively low compared to other substitutions. Thus, the overall edit distance between a real word and a misspelling which encodes a likely phoneme substitution is lower.

4.1.3 Edit Operation Weights

In order to apply a weighted edit distance measure to rank potential real-word corrections of a misspelling, each possible edit operation needs to be assigned a weight value. These values are based on how likely such an operation is to occur. Since the confusability of phonemes is inherently linked to their similarity, the values in the phoneme similarity matrix constructed previously in Chapter 3 using the acoustic and distributional properties of phonemes, are used directly as substitution weights for this purpose. However, this only provides costs for substitution operations, and weight values for insertions and deletions are also required. Based on existing literature regarding phoneme epenthesis and elision (Collins and Mees, 2013; Fourakis and Port, 1986; Gimson and Ramsaran, 1970; Itô, 1989; Yip, 1987), cost values between 0 and 1 are assigned to phonemes depending on their likelihood of being

inserted or deleted in English. These are added to the similarity matrix by treating them as substitutions involving the empty string (ε). For example, in non-rhotic varieties of spoken English the /R/ sound is not pronounced in post vocalic position and so could potentially result in the deletion of an /R/ phoneme in a misspelling e.g. 'nearly' written as *'nealy'. This is treated as a substitution of /R/ in the target word with ε (nothing) in the misspelling and this substitution is allocated a relatively low cost value. The resulting similarity matrix including insertion and deletion costs, visualised as a heatmap, can be seen in Figure 4.1. The raw values of this matrix can be found at the link in Appendix A.1.1. Notably, the matrix itself is not symmetrical: the cost of substituting phoneme X for phoneme Y is not the same as substituting phoneme Y for phoneme X. This allows for the distinction between a likely phoneme substitution, e.g. the /NG/ in 'anything' being perceived as a /N/ and resulting in the misspelling *'anythin', and the less likely reverse substitution, e.g. the /N/ of 'happen' being perceived as /NG/ and written as *'happeng'[2].

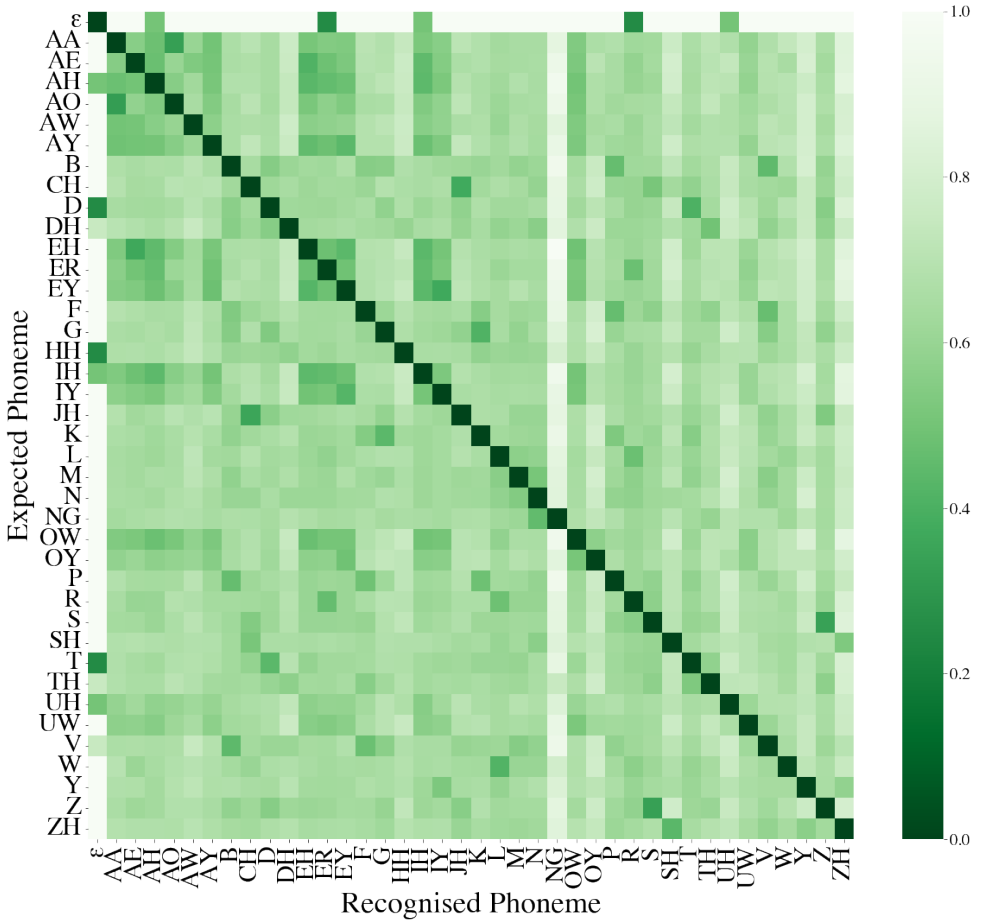


Figure 4.1: The phoneme similarity matrix with added insertion and deletion values visualised as a heatmap.

[2]This example does not occur in the corpora.

4.2 Spellchecker Evaluation

The S-capade approach is not intended as a standalone spellchecker, but, rather, as a component part designed to specifically address phonetic misspellings which deviate heavily from the target word. Its success is, thus, not determined through the overall accuracy or recall of the system on a corpus of misspellings. Instead, it is investigated whether or not the approach is capable of correcting errors where other tools fail by comparing the system's performance against existing spelling correction methods. Furthermore, constructing a corpus consisting solely of confirmed phonetic misspellings for the purpose of testing S-capade would be too time and resource expensive. Instead, the spellcheckers are tested on multiple misspelling corpora which are likely to contain various levels of these phonetic misspellings due to the nature of their sources. This section describes the spelling correction tools and misspelling corpora used for comparison and the metrics used to evaluate the performance of the S-capade approach. It is demonstrated that this phoneme-based method of suggesting real word corrections of misspellings is capable of handling the particularly deviant errors produced by children which lie beyond the scope of conventional spellcheckers.

4.2.1 Spellcheckers for Comparison

Three different spelling correction tools are used for comparison with the S-capade method. For each one the maximum edit distance between a misspelling and potential correction (whether this was measured in terms of characters, metaphone representations (see below), or phoneme sequences) is set to 2. Additionally, all target words in the misspelling corpora used for evaluation (described in Section 4.2.2) are manually added to each of the spellcheckers' inbuilt dictionaries if they are not already present. This includes the internal dictionary of S-capade which uses the CMU Pronouncing Dictionary. The spelling correction approaches considered for comparison are:

- **PySpellChecker**: This automatic spelling correction approach was based on work by Norvig (2007) and is available for download from Barrus (2018). Here, all possible permutations of a misspelling with a Damerau-Levenshtein edit distance of up to 2 are generated. These permutations are then compared to a list of known words in a frequency dictionary to find possible real-word corrections. The candidate corrections are then ranked according to their frequency with more frequently used words being considered more likely to be the true target word of the misspelling.
- **SymSpell**: A spelling correction algorithm developed by Garbe (2012) which again uses the Damerau-Levenshtein edit distance measure and word frequencies to rank potential corrections. However it is much faster due to its use of the symmetric delete algorithm for looking up potential corrections in the dictionary. Rather than generating all possible permutations of a misspelling, only those involving deletions

are considered. Insertions, substitutions, and transpositions are handled by adding similar deletion only permutations of real words as additional entries to the dictionary. By initially generating a much larger dictionary, fewer permutations of the misspelling need to be searched for and so it is much faster to find real word candidate corrections. The S-capade approach also makes use of the symmetric delete algorithm for candidate searching.

- **GNU Aspell**: This is a spellchecker designed to handle both typographic misspelling and phonetic misspellings (Atkinson, 2004). Potential corrections are found by considering real words within one or two character level edit operations of the misspelling much like the previously discussed approaches. However, GNU Aspell also converts misspellings to their 'soundslike' equivalent - using the metaphone representation for English (Philips, 2000). These strings are compared with those of the dictionary entries again in terms of edit operations between the sequences. In this way, similar sounding real word corrections are included in the suggested candidates. These candidates are then ranked in terms of a weighted average between the orthographic and 'soundslike' weighted edit distances where, as with the S-capade approach, each edit operation is assigned its own cost value. The main difference lies in Aspell's use of metaphone sequences which approximate phonetic information using a rule-based method for encoding an orthographic word.

4.2.2 Misspelling Corpora

Each of the spelling correction methods are tested on five separate misspelling corpora. Based on the various sources of these corpora, they are each likely to differ on the proportion of phonetic errors contained within - and particularly on the proportion of those errors produced by early literacy children which heavily deviate from the target word. The aim is to evaluate how the relative behaviour of the S-capade approach differs across the different corpora compared with the other correction methods. Additionally, misspelling corpora are restricted to errors whose target consists of a single word. For example, misspellings like *'alot' ('a lot') were removed since the dictionaries of the spellcheckers evaluated contain single words and so are unable to correct such an error. The misspelling corpora used for evaluation are:

- **Wikipedia Corpus** (2,230 misspellings): A corpus constructed from Wikipedia's 'Lists of common misspellings' (Wikipedia, 2020). These lists are used to correct the typographic errors made by article editors and a misspelling is considered 'common' if it occurs in Wikipedia at least once a year. Since these misspellings are presumably produced by adults using a keyboard, this corpus is the least likely to contain the phonetic errors for which the S-capade approach was designed.

- **Aspell Test Corpus** (515 misspellings): This corpus is used to test the GNU Aspell spellchecker and focuses on “really bad misspellings" (Atkinson, 2002). Since the GNU Aspell method considers corrections that are similar to a misspelling both in terms of their orthography and their ‘soundslike’ metaphone representations, this corpus likely contains a mixture of both typographic and phonetic errors although not necessarily the particularly deviant misspellings common to the writing of children.
- **Irish Schoolchildren Corpus** (232 misspellings): A corpus of misspellings extracted from schoolchildren’s free-text responses to surveys carried out by Irish educational company Zeeko (Everri and Park, 2018). Since these misspellings come from the writing of children, they likely contain a number of phonetic errors including those which deviate heavily from the intended target. However, the surveys were issued electronically and the children’s use of a keyboard means typographic errors are also likely to appear in the corpus moreso than if responses were written by hand.
- **Birkbeck Spelling Error Corpus** (33,887 misspellings): This is a large-scale corpus of English misspellings created from a number of sources but consisting primarily of cognitive errors (Mitton, 1985). It contains misspellings from children and adults, some of whom were receiving specialised literacy tuition or were self described ‘bad spellers’, and the majority of errors were extracted from handwritten text. Thus, this corpus is very likely to contain phonetic errors and creative attempts at encoding the sound sequence of a word. However, many of the misspellings of this corpus were obtained through spelling tests which were far beyond the ability of the writer and, as such, can be so far removed from the target word that there would appear to be no relation between the two. For example, one recorded misspelling of the word ‘according’ was simply *‘o’.
- **Holbrook Passages Corpus** (1,562 misspellings): A corpus of misspellings extracted from the passages presented in Holbrook (1964). These passages consist of the writings of British schoolchildren who were considered low academic achievers. As such, this corpus of originally handwritten misspellings produced by those still acquiring written literacy is likely to contain the largest proportion of the types of difficult-to-correct phonetic errors that the S-capade method was developed to target.

4.2.3 Evaluation Results

Kukich (1992) grouped work on spelling correction into three distinct tasks; detection of errors, isolated error correction, and context dependent error correction. The evaluation presented here concentrates on isolated error correction whereby the different approaches are supplied the misspellings and tasked with suggesting potential corrections based on the error alone and without the original context in which it appears. For each misspelling, the various spelling correction methods returned the top 10 closest real-word correction

candidates ranked in terms of the specific edit distance metric they employed. Both accuracy and recall are considered for evaluation purposes. Accuracy refers to whether or not the closest ranked candidate correction was the intended target of the misspelling. It is thus equivalent to recall at 1 (R@1) where a value of 1 is obtained if the target is the top candidate and 0 otherwise averaged across all misspellings in a corpus. The recall metric, or recall at 10 (R@10), considers whether the target word appears in the top 10 ranked candidates returned by the spellcheckers. Since all methods were restricted to a maximum edit distance of 2, there were some misspellings which returned fewer than ten potential corrections. Similarly to the evaluation carried out by (Hodge and Austin, 2001), the recall metric is favoured over accuracy. This is both due to the nature of typical spelling correction tools whereby the user is often presented with multiple correction options to choose from, and due to the S-capade method itself being envisioned as a component part of a larger spelling correction tool. Thus, being able to generate the true target word as a correction candidate is considered more important than the target being considered the most similar. This is particularly true for the more creative errors where neither an orthographic nor phonemic measure of similarity would consider the target word as the closest correction. The accuracy (R@1) and recall (R@10) results, for each of the spelling correction methods across the different misspelling corpora are presented in Table 4.2 with the best performing approach for each metric and for each corpus indicated in bold.

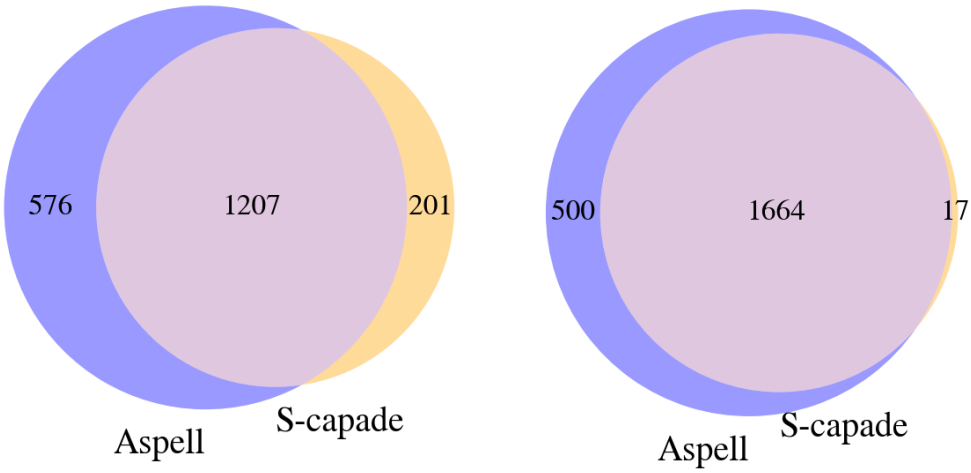
From these results it would appear that the GNU Aspell spellchecker is the overall best performing approach across the corpora and particularly in terms of recall. This is unsurprising since this method is designed to correct both typographic and phonetic errors. What is interesting is how the S-capade approach compares to the other tools based on the corpus it is applied to. Accuracy-wise, despite being the poorest performing approach, S-capade is notably more competitive compared to the other methods on the corpora more likely to contain a greater proportion of phonetic errors. Furthermore, for recall, the S-capade method outperforms the character-based spellcheckers on the Irish Schoolchildren Corpus, the Birkbeck Spelling Error Corpus and the Holbrook Passages Corpus - all of which are considered more likely to contain the creative phonetic misspellings of children for which

	Wikipedia Corpus		Aspell Test Corpus		Irish Schoolchildren Corpus		Birkbeck Spelling Error Corpus		Holbrook Passages Corpus	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
PySpellchecker	0.784	0.885	0.493	0.623	**0.569**	0.724	0.353	0.426	**0.293**	0.421
SymSpell	**0.810**	0.921	0.532	0.672	0.547	0.707	0.347	0.430	0.275	0.425
GNU Aspell	0.800	**0.970**	**0.557**	**0.856**	0.547	**0.866**	**0.399**	**0.660**	0.271	**0.679**
S-capade	0.631	0.778	0.464	0.652	0.496	0.763	0.344	0.515	0.271	0.528

Table 4.2: A summary of the accuracy and recall results for the different spelling correction approaches across the various misspelling corpora.

the method was designed. Indeed, the difference in recall scores between S-capade and the two character-based approaches appears to correlate with the hypothesised proportion of these types of errors. The greatest positive difference is seen with the Holbrook Passages Corpus and the greatest negative difference with the Wikipedia Corpus.

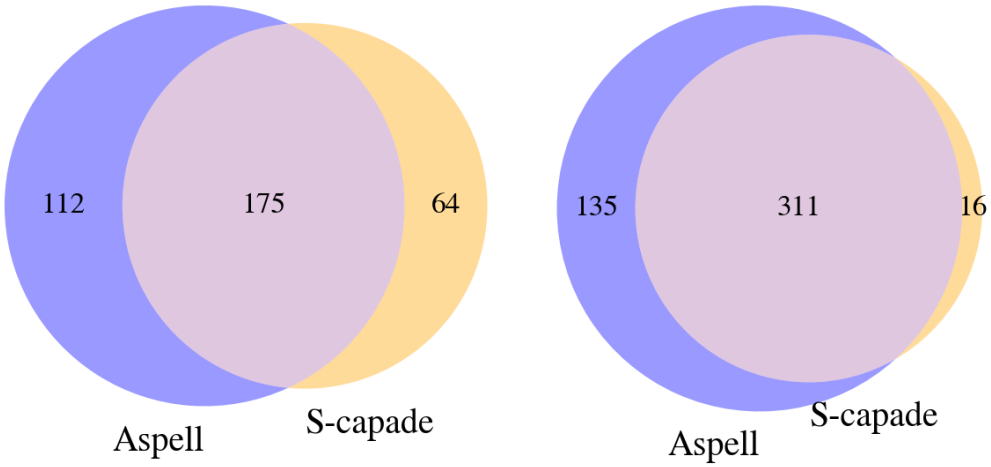
As discussed previously, the main goal of the spelling correction approach discussed in this chapter is to target the misspellings which other methods are unable to correct. That being the case, the overlap of corrected misspellings between the S-capade approach and that of GNU Aspell (the established best performing method of those evaluated) were considered so as to determine whether or not there were specific errors that S-capade was uniquely capable of correcting. Figures 4.2 - 4.6 depict the overlap of corrected misspellings for each of the corpora both in terms of accuracy and recall. These results demonstrate that, particularly for the corpora most likely to contain deviant phonetic misspellings, the S-capade approach is able to handle a not insignificant proportion of errors for which the GNU Aspell spellchecker is inadequate. A particularly interesting case is that of the Holbrook Passages corpus since this was identified as being the most likely to contain the particularly deviant phonetic errors for which S-capade was designed. Indeed, 13.0% of misspellings in the Holbrook corpus had their target word returned as the top candidate by S-capade and not by GNU Aspell. Similarly, S-capade was able to return the target word amongst the top 10 most similar candidates for 7.8% of misspellings in the Holbrook corpus where the GNU Aspell approach did not. This would suggest that, when it comes to the writing of children, a spelling correction approach that incorporates the phonemic similarity method of S-capade would exhibit improved performance over existing tools.



(a) Accuracy

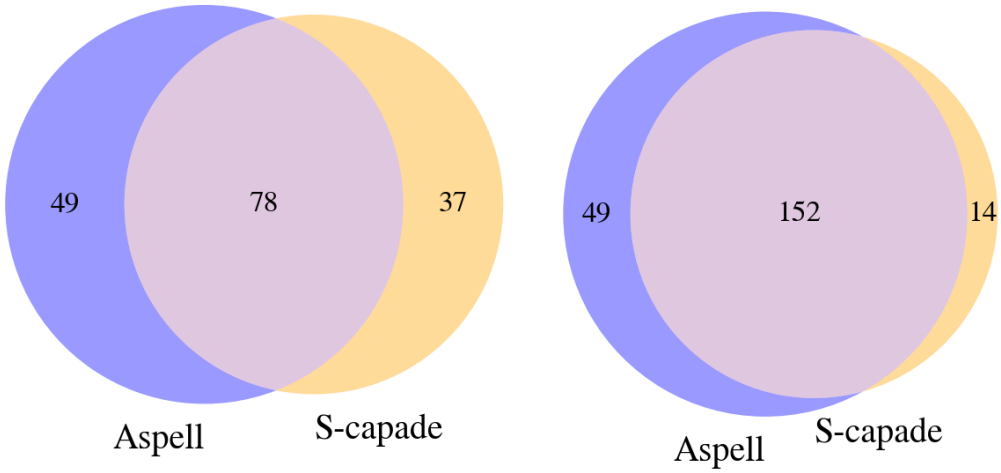
(b) Recall

Figure 4.2: Venn diagram displaying the overlap between Aspell and S-capade of misspellings in the Wikipedia Corpus whose intended target occurred as the top candidate (left) or within the top 10 candidates (right).



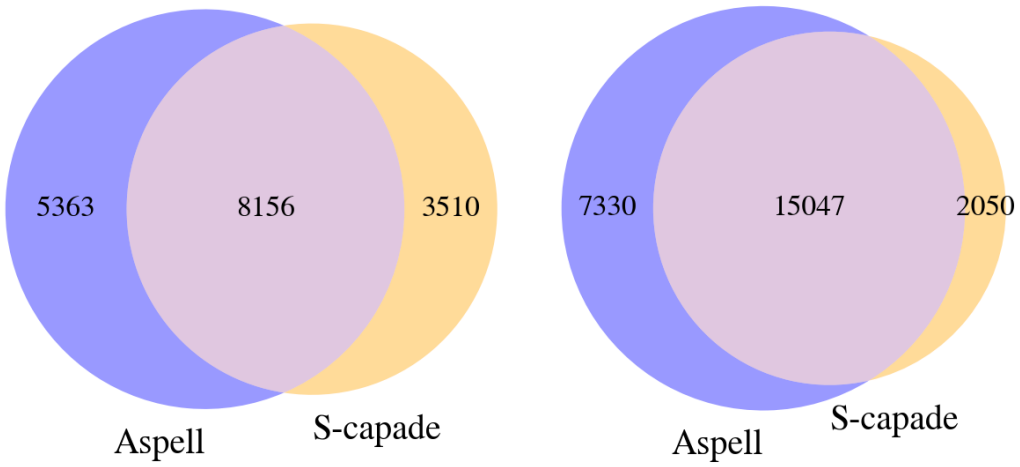
(a) Accuracy (b) Recall

Figure 4.3: Venn diagram displaying the overlap between Aspell and S-capade of misspellings in the Aspell Test Corpus whose intended target occurred as the top candidate (left) or within the top 10 candidates (right).



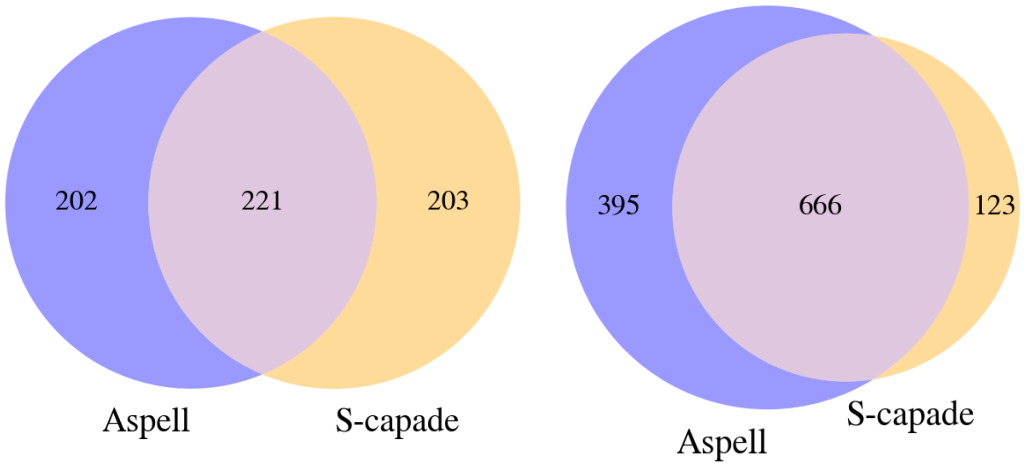
(a) Accuracy (b) Recall

Figure 4.4: Venn diagram displaying the overlap between Aspell and S-capade of misspellings in the Irish Schoolchildren Corpus whose intended target occurred as the top candidate (left) or within the top 10 candidates (right).



(a) Accuracy (b) Recall

Figure 4.5: Venn diagram displaying the overlap between Aspell and S-capade of misspellings in the Birkbeck Spelling Error Corpus whose intended target occurred as the top candidate (left) or within the top 10 candidates (right).



(a) Accuracy (b) Recall

Figure 4.6: Venn diagram displaying the overlap between Aspell and S-capade of misspellings in the Holbrook Passages Corpus whose intended target occurred as the top candidate (left) or within the top 10 candidates (right).

4.3 A Closer Look at Children's Misspellings

The motivation behind the development of the S-capade approach lies in the idea that children produce misspellings through an effort to encode the sounds they identify in a word using characters they know represent those sounds. To further explore how spoken English impacts literacy acquisition, a qualitative analysis is carried out on the more phonetic errors which appear in the corpora containing the writing of children. An examination of these errors, for which the S-capade method of spelling correction is better suited, reveals that these types of misspellings typically stem from either the highly variable nature of English orthography or as a result of pronunciation features of spoken English being encoded in the spelling attempt.

4.3.1 Misspellings and English Orthography

Firstly, standard English spelling is neither totally phonetic nor predictable without knowledge of word etymologies. There is not a one-to-one mapping between the sounds of English and the characters used to denote these sounds. For example, the /F/ phoneme could be written using 'f' as in 'fish' or with 'ph' as in 'photo'. Conversely, the letter 'c' could encode the phoneme /K/ as in 'cat', the phoneme /S/ as in 'cent', or even the phoneme /SH/ as seen in 'ocean'. In addition, many English spellings retain elements of historical pronunciations which are no longer used in modern day spoken English. This is seen in words like 'knight' or 'gnaw' where the now silent initial 'k' and 'g' would have been pronounced in Old English (Kökeritz, 1945). As a result, a number of instances are observed where a misspelling encodes the correct sound sequence of a word but is orthographically too far removed from the canonical spelling to be corrected by traditional character-based spellcheckers. Some examples of these types of errors, which were corrected by the S-capade method but not by Aspell, are given in Table 4.3. In these examples, the phoneme sequence of a misspelling is an exact match to the target word.

Spelling	Phoneme Sequence				
strate	S	T	R	EY	T
straight	S	T	R	EY	T
waid	W	EY	D		
weighed	W	EY	D		

Table 4.3: Misspelling examples whose predicted phoneme sequences match those of their intended targets. Taken from the Holbrook Passages Corpus.

Some further examples are given in Table 4.4 where the predicted phoneme sequence of the misspelling does not match that of the target word. However, it would appear in these cases that the differences between the phoneme representations are the result of the grapheme-to-phoneme process failing to predict the exact sound sequence intended by the writer. For instance, the final ‘y’ of *‘folocify’ was likely intended to encode /IY/ in the same way as the final ‘y’ of the target ‘philosophy’. However, the predicted phoneme sequence of this misspelling ends with the phoneme /AY/ instead which prevents it from being an exact match to the target word. Similarly the mismatches between the predicted phoneme sequence of *‘egicasinol’ and that of ‘educational’ are not necessarily the intention of the writer since ‘s’ could represent the /SH/ phoneme as in ‘tension’ and ‘o’ could denote /AH/ as seen in ‘idol’. As such, these examples are considered as encoding the same phoneme sequence as their targets but failing to match the orthographic form. Again, the true intended targets of these misspellings were suggested as the top candidate by the S-capade approach but not by Aspell.

Spelling	Phoneme Sequence									
folocify	F	AH	L	AA	S	AH	F	AY		
philosophy	F	AH	L	AA	S	AH	F	IY		
egicasinol	EH	JH	AH	K	EY	S	IH	N	AO	L
educational	EH	JH	AH	K	EY	SH	AH	N	AH	L

Table 4.4: Misspelling examples whose predicted phoneme sequences differ from those of their intended targets as a result of the grapheme-to-phoneme process. Taken from the Birkbeck Spelling Error Corpus and the Irish Schoolchildren Corpus (respectively).

4.3.2 Misspellings and English Pronunciation

A different source of phonetic error stems from the encoding of pronunciation features. In these cases, shown in Table 4.5, the phoneme sequence of the misspelling deviates from that of the target because the misspelling endeavours to capture the sound sequence of the target word as it occurs in spoken English. For instance, the misspelling *‘wokin’ (‘walking’) appears to capture a process known as /NG/-fronting (or ‘g’-dropping) where the, typically word-final, velar nasal is instead produced as an alveolar nasal like the realisation of /N/. Similarly, the misspelling *‘saur’ (‘saw’) is a potential result of a phenomenon known as intrusive-/R/. For example, an /R/ realisation is often inserted at the end of the word ‘saw’ in connected speech if it is immediately followed by a vowel sound as in “they saw a...". Perceiving this /R/ sound thus results in the target word ‘saw’ being misspelled as *‘saur’. Both the misspellings *‘wokin’ and *‘saur’ can be corrected using the S-capade approach whilst the true intended target is not found in the candidates suggested by Aspell.

Spelling	Phoneme Sequence				
wokin	W	OW	K	IH	N
walking	W	AO	K	IH	NG
saur	S	AO	R		
saw	S	AO			

Table 4.5: Misspelling examples whose predicted phoneme sequences differ from those of their intended targets due to the encoding of general English pronunciation features. Taken from the Birkbeck Spelling Error Corpus.

/NG/-fronting and intrusive-/R/ are common pronunciation features observed in many, though not all, varieties of English. However, there are other features of spoken English which are found in a much smaller subset of varieties and are often related to a particular region or social group. For example, the examples given in Table 4.6 are taken from the Irish Schoolchildren Corpus and would seem to be indicative of an influence of Irish Accented English specifically. For example, the misspelling *'tink' ('think') might stem from the fortition of dental fricatives where the /TH/ phoneme is often realised as an alveolar plosive much like the realisation of /T/. This phenomenon is characteristic of the spoken English varieties used across Ireland and likely takes its influence from the absence of a /TH/ phoneme in Irish Gaelic. Furthermore, Dublin English varieties in particular often exhibit back vowel raising and for many speakers, the /AH/ vowel of 'strut' is produced almost identically to the /UH/ vowel of 'foot'. This is potentially the source of the misspelling *'trost' ('trust') where the letter 'o' is used to denote the perceived /UH/ vowel despite the grapheme-to-phoneme tool predicting a /AA/ vowel in the misspelling.

Spelling	Phoneme Sequence				
tink	T	IH	NG	K	
think	TH	IH	NG	K	
trost	T	R	AA	S	T
trust	T	R	AH	S	T

Table 4.6: Misspelling examples whose predicted phoneme sequences differ from those of their intended targets due to the encoding of accent specific pronunciation features. Taken from the Irish Schoolchildren Corpus.

4.4 Limitations

The phoneme based method of spelling correction presented in this chapter has its merits and has been shown capable of handling those difficult-to-correct phonetic errors it was designed for. However, there are several limitations to the approach which should be considered. Most obviously is the fact that, since it was designed to target a specific type of phonetic misspelling, the S-capade method is less suitable for the correction of other types of errors. Children's writing contains both obscure phonetic misspellings and more typical typographic errors. Ideally a single system could handle both types of errors, combining measures of orthographic similarity and phonemic similarity in order to rank potential corrections. Moreover, the method operates on single words in isolation and so loses the context of a misspelling which could be used in predicting the likelihood of a suggested correction. Essentially, the S-capade approach would need to be incorporated into a larger spelling correction system in order to be of benefit to users.

Within the S-capade method itself, the grapheme-to-phoneme tool is not always adequate at predicting the phoneme sequences encoded by misspellings. To use an earlier example, the error *'ceap' ('keep'), where the initial /K/ phoneme was written with the letter 'c', should have produced the phoneme sequence /K IY P/. However, the grapheme-to-phoneme tool instead treated the initial 'c' as denoting the /S/ phoneme and so produced the phoneme sequence /S IY P/. This is an issue arising from the variability of English orthography. Indeed, the letter 'c' can represent both the phonemes /K/ and /S/ as in the difference between 'call' and 'cell'. As a result, the predicted phoneme sequences of misspellings do not always correspond to those sequences a child tried to encode. This can, in turn, affect the method's ability to compare the phoneme similarity between an error and a potential correction.

Furthermore, the cost values for phoneme edit operations arose from experiments and literature regarding spoken English. However, it is likely that not all observations and processes transfer to attempts at writing. For example, /R/ might be an expected edit due to non-rhotic varieties of English which don't realise /R/ in post vocalic position. However, an examination of the misspelling corpora revealed /R/ deletion to be much less frequent than its counterpart - /R/ insertion. This phenomenon likely stems from an overgeneralisation of non-rhotic patterns or intrusive /R/ instances where an /R/ sound is inserted in spoken English where it does not occur in the underlying phonemic form. Evidently, this pronunciation feature of English is much more salient amongst misspellings. Thus, whilst the values in the matrix allow for a reasonable measure of phoneme similarity, it would be beneficial to obtain cost values based on actual misspelling data and observations regarding which spoken English processes are salient enough to be encoded in the writings of children.

4.5 Summary

The focus of this chapter was on Research Question 2: *Can a model of phoneme similarity be applied to and benefit the development of language technologies; specifically in a spelling correction tool for children?* The values in the phoneme similarity matrix constructed previously in Chapter 3 were used as weights in an alignment algorithm comparing the similarity of phoneme sequences. This was then incorporated into an English spelling correction tool aimed at the phonetic misspellings typically produced by children which are difficult to correct using traditional character based spellcheckers. The performance of this spelling correction approach was compared with three other spellcheckers and tested on five different misspelling corpora expected to contain differing proportions of phonetic-type errors. Further analysis was carried out in the form of an examination of the overlap between misspellings corrected by S-capade and those corrected by Aspell (the best performing spellchecker). This analysis revealed that in each corpus, the S-capade approach was capable of correcting misspellings that Aspell could not and that the proportion of errors uniquely handled by S-capade correlated with the expected proportion of phonetic misspellings within the corpus. For instance, the Holbrook passages corpus, taken from handwritten passages from children considered 'low academic achievers', resulted in the largest proportion of misspellings that could be corrected by only the S-capade approach. This would suggest that this phonetic similarity based method of spelling correction could benefit the development of a larger spelling correction tool designed specifically for children.

Chapter 5

Spoken Variety Adaptation for Spelling Correction Tools

In the previous chapter, it was noted that some of the misspellings appeared to be influenced by the child's spoken variety. In particular, the Irish Schoolchildren corpus contained errors which seemed to encode pronunciation features of Irish Accented English. This would seem to contrast with the review of Snell and Andrews (2017) which concluded that there was insufficient evidence that regional dialect had a major impact on writing. Thus, further investigation of the interaction between pronunciation variation and spelling is the core aim of the work presented in this chapter. In addition, the potential benefits of variety adaptation are explored in the context of language technologies to determine whether a user would experience improved performance of a spelling correction tool designed for their particular spoken variety. As such, this chapter addresses Research Question 3: *How might the pronunciation variation present in a child's English impact the misspellings they produce and can a spelling correction tool be adapted to better perform with a specific variety?*

In this chapter it will be shown how the spellchecker described in Chapter 4 can be adapted to Irish Accented English by fine-tuning the phoneme similarity matrix which determines the rankings of candidate real-word corrections. To do this, the similarity scores are treated as a weight matrix for a single layer neural network performing logistic regression and the values are adjusted so as to minimise the cross entropy loss. The resulting adapted spelling correction tool is then evaluated against the misspellings of the Irish Schoolchildren corpus using a k-fold cross validation approach. The original baseline model and one tuned to British Accented English are also tested for comparison. The greater success of the Irish Accented English model over the others is considered evidence that the spelling errors produced by Irish children have some commonality that exists between them that is not adequately captured by fine-tuning on another spoken variety. Further, analysis of the tuned similarity matrix demonstrates effects that likely directly stem from specific features of Irish Accented English pronunciation.

The main contributions presented in this chapter are an investigation of the hypothesis that regional variation in pronunciation influences misspelling productions in children's writing and a demonstration of how language technology adaptation can benefit users of underrepresented spoken varieties. This work has been published in O'Neill et al. (2021). The chapter is outlined as follows. First, Section 5.1 discusses the variant tuning process where the similarity matrix incorporated into a spelling correction tool in Chapter 4 is

adapted to Irish Accented English. In Section 5.2 the performance of this adapted model is then compared with the baseline S-capade model and one tuned to British Accented English. Testing on misspellings from Irish schoolchildren, it is shown that the Irish Accented English adapted model outperforms the others on accuracy, recall and mean reciprocal rank (MRR) thus demonstrating the improved user experience which would be achieved through the use of adapted language technology. The implications of these results are then discussed in Section 5.3 where misspelling examples are presented which appear to encode specific features of Irish Accented English thus suggesting preliminary evidence of spoken variant influence on spelling. The chapter concludes with a consideration of the limitations of this adaptation process in Section 5.4 followed by a summary in Section 5.5.

5.1 Fine-Tuning the Similarity Matrix

One benefit of the S-capade spelling correction tool over other phonetic based spellcheckers that use letter-to-sound rules lies in its potential for adaptation to a particular spoken variety. The overall similarity between a misspelling and a possible real-word target is calculated by summing the weights associated with each observed phoneme substitution, insertion, or deletion. The resulting similarity scores are used to rank the candidate real-word corrections suggested to the user. Ideally, the true real-word target would be calculated as being very similar to the original misspelling and thus be ranked highly among the potential corrections. To achieve this, the distance values which correspond to the observed substitutions in the true *misspelling:target* pair must generally be lower than those of the false *misspelling:non-target* possibilities. If the spelling errors produced by children are in some way influenced by their pronunciation, it would be expected that speakers of the same variety would produce similar types of misspellings. Thus, adapting these distance values based on misspellings from speakers of a specific variety should improve performance of the spellchecker on unseen misspellings from speakers of the same variety.

5.1.1 The Irish Accented English Corpus

The dataset used in this experiment is that labelled earlier in Chapter 4 as the Irish Schoolchildren corpus but which is henceforth referred to as the Irish Accented English corpus. It is compiled from a collection of surveys of schoolchildren across Ireland carried out by Irish education company Zeeko (Everri and Park, 2018). A subset of these surveys contained open-ended questions of the form “Why do you think...?” or “What did you like about...?” and participants responded in a free-text format. In total, survey responses from 628 students contained these free-text answers and were analysed for non-word spelling errors. The students ranged in age from 7 to 17 years with an average age of 10 years. The full distribution of ages across respondents can be seen in Figure 5.1. Whilst responses were received from

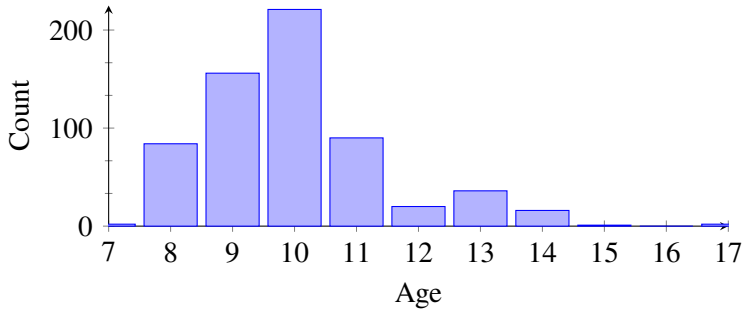


Figure 5.1: The age distribution of survey respondents in the Irish Accented English corpus.

locations across the country, 50.6% of respondents were from schools in County Dublin. As such, whilst the tuned model is referred to as being adapted to Irish Accented English, it is in fact heavily biased towards the specific spoken varieties of Dublin which is incorporated in later analyses.

Non-word misspellings were extracted from the responses and a human annotator judged their corresponding real-word targets. Where a target could not be identified from context, the misspelling was removed. This resulted in a corpus of 232 pairs of misspellings and real-word targets. Examples of some *misspelling:target* pairs can be seen in Table 5.1.

Misspelling	Target
achuly	actually
bekos	because
difrunt	different
egicasinol	educational
mishon	mission
sichweshen	situation

Table 5.1: Example misspellings and real-word targets from the Irish Accented English corpus.

5.1.2 The Similarity Matrix as a Weight Vector

The S-capade spelling correction method discussed previously in Chapter 4 uses a phoneme similarity matrix based on the properties of English phonemes in general. However, different realisations of phonemes in different varieties of English might lead to differences in the degree of perceived similarity between phonemes. For example, consider the /TH/ phoneme which, in Irish Accented English and particularly in Dublin variants, often undergoes fortition and is realised as an alveolar plosive [t] rather than a dental fricative [θ] (Hickey, 2004). In these varieties both /TH/ and /T/ can be realised in similar ways, likely strengthening their perceived similarity and potentially resulting in a child who is adopting a "sound it out" approach encoding the /T/ phoneme instead of /TH/. It is hypothesised that specific

pronunciation features of a variety lead to particular encoding issues and thus similar types of misspellings. In theory, if the similarity matrix reflected these pronunciation features it would exhibit improved performance for users with the spoken variant to which it was adapted.

In order to adapt the similarity matrix to the Irish Accented English variety, the fine-tuning technique used in the training of neural network architectures is explored. By treating the matrix as a weight vector, these weights can be adjusted through backpropagation. Ideally, the similarity matrix should result in low distance values for true *misspelling:target* pairs and comparatively higher distance values for misspellings and potential real-word corrections that are not the intended target. Thus, the likelihood of a *misspelling:target* pair being true or false should be predictable based on its distance score.

$$\hat{y} = f(\sum_{i=1}^{n} x_i w_i + b) \tag{5.1}$$

Consider the equation of a single-layer neural network given in Equation 5.1 with input vector x, weight vector w, bias b and activation function f. By representing a *misspelling:target* pair as the input vector, the similarity matrix as the weight vector, and using a sigmoid activation function, this network can be used to predict the probability that a real-word correction was or was not the intended target of a misspelling. Through a simple reshaping of the 40x40 similarity matrix, a 1600 dimension weight vector is obtained. For the input vectors, each *misspelling:target* pair is first represented by another 40x40 matrix similar to that of the similarity matrix but where the values correspond to the number of times a particular phoneme substitution (or insertion or deletion) was observed in the pair. For example, consider the misspelling 'achuly' and its intended target 'actually' whose phonemic sequence comparison is presented in Table 5.2. The matrix representation for this misspelling:target pair would contain '1' entries in the cells corresponding with the correct encodings of /AE/, /L/, and /IY/ as well as the substitutions /K/:ɛ, /SH/:/CH/, and /AH/:/UW/ with '0' entries everywhere else. Again, this matrix can then be reshaped into a 1600 dimension vector (see Figure 5.2).

The dot product of the resulting input vectors x and the weight vector w, as used in the

Spelling	Phoneme Sequence					
actually	AE	K	SH	AH	L	IY
achuly	AE		CH	UW	L	IY
Edit Operation		del	sub	sub		

Table 5.2: A comparison of the phoneme sequences between the real-word target 'actually' and the misspelling 'achuly' including their optimal phoneme alignment.

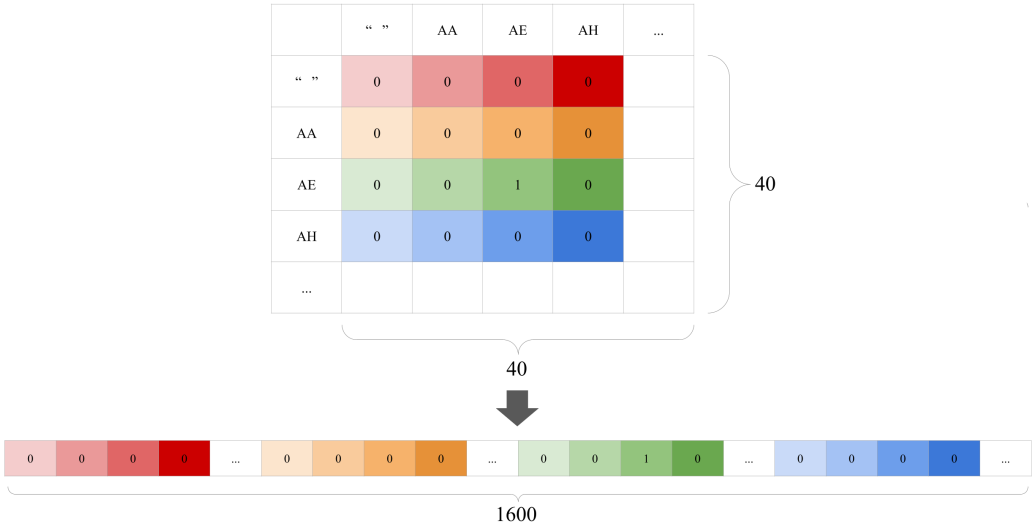


Figure 5.2: A visualisation of transforming a 40x40 matrix of substitution counts into a sparse input vector with 1600 dimensions.

single-layer neural network equation, is equivalent to the similarity score calculated in the S-capade algorithm: the sum of the costs associated with the observed phoneme substitutions in a *misspelling:target* pair. Thus, the network bases its probability predictions on the same value used by the S-capade tool to rank candidate corrections. In theory, a well tuned similarity matrix will result in a ‘0’ or near ‘0’ distance value for true *misspelling:target* pairs. Since higher values indicate greater levels of dissimilarity and result in higher outputs by the single-layer neural network, the probability predicted by the network corresponds to the likelihood that a real-word correction is **not** the intended target of the given misspelling.

5.1.3 Adaptation through Backpropagation

Given the single layer neural network which predicts the probability that a given *misspelling:target* pair does not involve the true target, the weight vector can be adapted to a particular spoken variety by training the network to distinguish between true real-word targets and other non-target potential corrections. Weights associated with observed substitutions in true *misspelling:target* pairs will decrease, resulting in a lower probability prediction by the network, whilst the weights pertaining to the substitutions involving non-targets will increase so as to produce higher predicted probabilities.

Since the Irish Accented English corpus is a relatively small dataset, a k-fold cross validation approach is used to fine-tune the weight vector and evaluate the performance of the resulting adapted spelling correction tool. The corpus is separated randomly into 10 folds and each fold is held out as a test set whilst the other 9 folds are used to train the network. For each misspelling in the training set, the S-capade baseline model is used to generate the 10 most likely candidate corrections based on the calculated similarity measure. As described

previously in Chapter 4, the predicted phoneme sequence of the misspelling is obtained using a grapheme-to-phoneme tool (CMUSphinx, 2016) trained on the CMU Pronouncing Dictionary (Weide, 1998) which itself is used to obtain the phoneme sequences of the real-word candidates. These sequences are then aligned using a weighted edit-distance algorithm and the observed substitutions, insertions and deletions are represented as a 1600 dimension input vector as detailed above.

Each training set has between 208-209 misspellings and each misspelling has 10 generated candidate corrections. If not already present, the real-word target is added to the list of candidates. Some candidate lists contain multiple instances of the target as a result of multiple entries in the dictionary due to variant pronunciations. As such, there are, for each training set, between 250-266 instances of true *misspelling:target* pairs (assigned a '0' label) and between 1856-1883 instances of negative *misspelling:non-target* pairs (assigned a '1' label).

The network is first used to predict the probability of each input pair involving a non-target correction using the original similarity matrix values as initial weights. To fine-tune the weight vector, the performance of the network needs to be measured using a loss function and the weights adjusted to minimise this loss. This is done using a gradient descent approach where the weights are updated in increments of the derivative (or slope) of this loss function. The binary cross-entropy loss is used for this task as defined in Equation 5.2 where N is the number of training inputs, y_n is the true label of input n, and $\hat{y}_n$ is the predicted probability associated with input n.

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}[y_n log\hat{y}_n - (1-y_n)log(1-\hat{y}_n)] \tag{5.2}$$

Given the sigmoid activation function, the derivative of the cross entropy loss with respect to the weight vector $\frac{\partial \mathcal{L}}{\partial w}$ is as shown in Equation 5.3 where N is the number of training inputs, y_n is the true label of input n, $\hat{y}_n$ is the predicted probability associated with input n, and x_n is the input vector n.

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)x_n \tag{5.3}$$

The weights are updated so as to minimise the loss in the manner shown in Equation 5.4 where w' is the updated weight vector, w is the original weight vector, α is the learning rate (here 0.05 was used), and $\frac{\partial \mathcal{L}}{\partial w}$ is the derivative of the loss function with respect to the weights.

$$w' = w + \alpha(\frac{\partial \mathcal{L}}{\partial w}) \tag{5.4}$$

The updated weights are then used to generate new probability predictions for the training inputs and the update process is repeated for 2000 epochs. Both the learning rate and number of epochs were chosen heuristically and are not necessarily the optimal values. However parameter fine-tuning was not the focus of this work. The resulting tuned weight values for each training set are then normalised between 0 and 1 and then the vectors are reshaped back into 40x40 similarity matrices for use with the S-capade spelling correction tool and evaluation on each of the held-out test folds.

5.2 Spelling Correction Model Evaluation

This thesis hypothesises that a child's particular spoken variety influences their phonic recognition skills and their 'sound-it-out' approach to spelling. If that were the case, it would be expected that the spelling errors produced by children with the same spoken variety encode the same pronunciation features that exist in that variety. For example, a word with the /TH/ phoneme written with the letter 't' because of its realisation in Irish Accented English as an alveolar plosive much like the /T/ phoneme. Theoretically, if adapting the spelling correction tool to Irish Accented English misspellings resulted in improved performance on other spelling errors produced by children with this variety, then these misspellings must share something in common and this could be considered evidence of the spoken variety's influence on spelling and literacy acquisition.

It is plausible, however, that any performance gains are not specific to the Irish Accented English variety and, rather, result from adaptation of the similarity matrix to the general task of spelling correction. Recall that the original matrix was developed to represent the confusability and distributive properties of English phonemes in general and did not account for similarity judgements that might be more, or less, likely to result in misspellings. Thus, the adaptation process might result in performance gains unrelated to the specific spoken variety but instead to the general substitutions observed in encoding English sounds orthographically.

In order to test whether or not this is the case, a second adapted spelling correction model is developed by fine-tuning the matrix to British Accented English spelling errors. If the performance gains of the Irish Accented English model are the result of general task specific fine-tuning, then the British Accented English adapted spellchecker should perform equally well on the test set of misspellings by Irish children. Alternatively, if the Irish Accented English spelling correction tool exhibits better performance over that of the British Accented English tool, then this would suggest that, through its adaptation to Irish Accented English, it has captured features in the misspellings that are specific to the Irish Accented English variety.

5.2.1 The British Accented English Model

A second adapted spelling correction tool is constructed by fine-tuning the similarity matrix to a British Accented English variety. To do this, a subset of the Holbrook corpus is collected. The Holbrook corpus, used previously in Chapter 4, contains 1791 misspellings extracted from the writings of academically low achieving British secondary school students (originally appearing in Holbrook (1964) and made available by Mitton (1985)). Given the much larger size of this corpus relative to the Irish Accented English Corpus, 232 misspellings are extracted randomly from the Holbrook corpus to create the British Accented English Corpus. This corpus, now of the same size as the Irish Accented English corpus is the one to which the baseline spelling correction tool is adapted.

The adaptation process is carried out in much the same way as presented previously in Section 5.1. The misspellings are first passed to the baseline S-capade spellchecker to obtain the ten most likely corrections. These correction candidates are then paired with the misspelling and represented as a 1600 dimension vector based on the observed substitutions, insertions, and deletions between the phoneme sequence of the real-word and that of the misspelling. True *misspelling:target* pairs are assigned a '0' label whilst false *misspelling:target* pairs are assigned a '1' label. A single layer neural network using the original similarity matrix values as weights and with a sigmoid activation function is then trained to predict the probability that a given input was a false *misspelling:non-target* pair. During training, the same learning rate and number of epochs are used as with the Irish Accented English fine-tuning. The one difference between this adaptation process and that described previously, is the exception of the k-fold cross validation approach. Since the British Accented English tuned model will be tested on the Irish Accented English data, there is no need to hold out test sets from the corpus. Instead all 232 misspellings are supplied as a single training set. The resultant tuned weights are then reshaped back into a 40x40 matrix to be incorporated into the spelling correction tool method testing.

It is expected that the British Accented English adapted model will outperform the baseline model even when tested on Irish Accented English misspellings. Not only is the tuned similarity matrix better suited to the sorts of phoneme confusions that arise through the orthographic encoding of English sounds, but, also, British Accented English has a number of pronunciation features in common with Irish Accented English. It is likely that the two tuned matrices will share some aspects through their representation of the same spoken English phenomena. For example, /NG/-fronting, the tendency to pronounce word-final /NG/ phonemes not as velar nasals [ŋ] but as alveolar nasals [n] (in the same way /N/ phonemes are typically realised) (Houston, 1985), can lead to these /NG/ phonemes being written with the letter 'n' as in "bein" for "being". Since this pronunciation exists in both varieties of spoken English, it is expected that both of the tuned similarity matrices will capture this feature and both spelling correction tools will be equipped to handle these sorts

of misspellings. Whilst the British Accented English adapted spelling correction tool should thus significantly outperform the baseline S-capade model, it should not reach the same level of performance of the Irish Accented English adapted spellchecker if it is indeed the case that the misspellings produced by Irish children are influenced by the pronunciation features of Irish Accented English.

5.2.2 Evaluation Metrics

As before, in Chapter 4, the evaluation metric deemed most significant in assessing the performance of a spellchecker is recall. The true intended target word needs to be presented to the user in the list of candidate corrections. In addition to this, however, it is preferable that the true target appears at or near the top of the list of possible real-word corrections. As the weights are tuned during training the true *misspelling:target* pairs should produce lower distance scores and thus be ranked more highly in the returned list. If this tuning does indeed capture pronunciation features which, in turn, influence spelling production, then the test set of misspellings should also produce lower distance scores and higher rankings. To determine if this is the case, the spelling correction tools are evaluated on a number of measures:

- **Recall at 1 (R@1)** is equivalent to accuracy, the recall at 1 is '1' if the intended target is the highest ranked correction candidate and 0 otherwise, averaged across all misspellings in the test set.
- **Recall at 3 (R@3)** is 1 if the intended target occurs in the three highest ranked candidates and '0' otherwise, averaged across all misspellings in the test set.
- **Recall at 5 (R@5)** is '1' if the intended target occurs in the five highest ranked candidates and '0' otherwise, averaged across all misspellings in the test set.
- **Recall at 10 (R@10)** is '1' if the intended target occurs in the ten highest ranked candidates and '0' otherwise, averaged across all misspellings in the test set.
- **Mean Reciprocal Rank (MRR)** is a measure of how highly the target typically ranks in the list of candidates (see Equation 5.5) with values closer to one indicating that the target real-words are on average more highly ranked.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{5.5}$$

Using these metrics, more insight is gained into how the fine-tuned similarity matrix impacts the calculated distances between a misspelling and its target and thus the performance of the spellchecker.

5.2.3 Comparison Results

Three spelling correction models are tested and compared. Firstly, the original S-capade model as described previously in Chapter 4 is used as a baseline for comparison. In addition there are two adapted spelling correction models; the Irish Accented English model and the British Accented English model in which the similarity matrix used in the baseline model is replaced with the corresponding tuned matrix as generated in Section 5.1.

For the S-capade Baseline Model and the British Accented English model, the reported evaluation metric scores are obtained by testing on the entirety of the Irish Accented English corpus. For the Irish Accented English model, each of the ten independently tuned matrices are tested on the corresponding held-out test set of misspellings and the reported evaluation metric scores are the averages across these ten folds. The comparison results can be seen in Table 5.3 with the best performance for each metric indicated in bold.

	MRR	R@1	R@3	R@5	R@10
S-capade Baseline Model	0.623	0.543	0.694	0.746	0.789
Irish Accented English Model	**0.693**	**0.629**	**0.746**	0.781	**0.828**
British Accented English Model	0.673	0.599	0.728	**0.784**	0.823

Table 5.3: A summary of the Mean Reciprocal Rank and Mean Recall@K for the three spelling correction tools built using the different phoneme similarity models.

As can be seen, whilst the British Accented English model exhibits substantial performance gains over the baseline model as expected, the Irish Accented English model outperforms the others in almost all metrics. Particularly, it performs notably better on Mean Reciprocal Rank and R@1 (accuracy) meaning that the true target word tends to be ranked more highly under the Irish Accented English model and is more often the top suggested candidate. The Mean Reciprocal Rank score of the Irish Accented English model exhibited an absolute increase of 7.0% (11.2% relative) over the baseline model and 2.0% (3.0% relative) over the British Accented English model. Similarly there was an absolute gain of 8.6% (15.8% relative) in accuracy over the baseline model and 3.0% (5.0% relative) increase over the British Accented English model. In terms of user experience, this means that a child with an Irish Accented English spoken variety using this adapted tool is more likely to find their target word in the suggested candidate list even if relatively few options are presented to them by the system.

5.3 Pronunciation Features of Irish Accented English

The comparative success of the Irish Accented English adapted model over the other spelling correction tools suggests that it is more suited to handling the misspellings produced by Irish children. This can be considered preliminary evidence that not only does there exist a commonality between the misspellings of speakers with the same spoken variety, but that the adapted model is able to capture, in the form of the tuned similarity matrix, the pronunciation features of this variety which influence written productions. To further examine this idea, the resultant tuned matrix is analysed in more detail alongside existing literature regarding the features of Irish Accented English and the observed misspellings in the corpus.

The resultant tuned similarity matrices across each of the ten folds were averaged to give an overall Irish Accented English adapted matrix as presented in Figure 5.3 in the form of a heatmap. Each cell represents the weight associated with substituting a phoneme in the target word (y-axis) with another in the misspelling (x-axis). The empty string ε is used to denote the absence of a phoneme such that substitutions involving ε in the target word indicate phoneme insertions and substitutions involving ε in the misspelling indicate phoneme deletions. Darker squares denote smaller weight values meaning a lower cost associated with the substitution. These lower costs reflect a stronger degree of similarity between the phonemes involved since they are more likely to be confused during encoding. In addition, it is important to consider not just the final values but also how the original values of the similarity matrix change through its adaptation to Irish Accented English. The differences in values between the original matrix used in the baseline model and the tuned matrix pre-normalisation are depicted in Figure 5.4. Here, blue cells indicate a reduction in the original weight associated with a particular substitution and red cells an increase. Accounting for the limited size of the Irish Accented English corpus and the fact that not all pronunciation features will manifest in misspellings, the tuned similarity matrix and the misspellings of the corpus do appear to be influenced by both general English and Irish Accented English phonetics and phonology.

Additionally, as noted previously, the corpus consists heavily of misspellings produced by school children in Dublin and so particular attention is paid to the pronunciation features of Dublin English which appear to manifest in the spelling productions. Specific focus is given to the variety of Dublin English termed *local Dublin English* since this appears to have a strong influence amongst the corpus. *Local Dublin English* is typically associated with the working class vernacular and is considered distinct from *mainstream Dublin English*, used by the middle class and suburban speakers, and *fashionable Dublin English*, used by a smaller group of more socially-mobile speakers who seek to distinguish themselves from the 'low-prestige' of the local variety (Hickey, 2005). Whilst these categories may have evolved somewhat since first described, the features associated with *Local Dublin English* still correspond to the misspellings in the corpus.

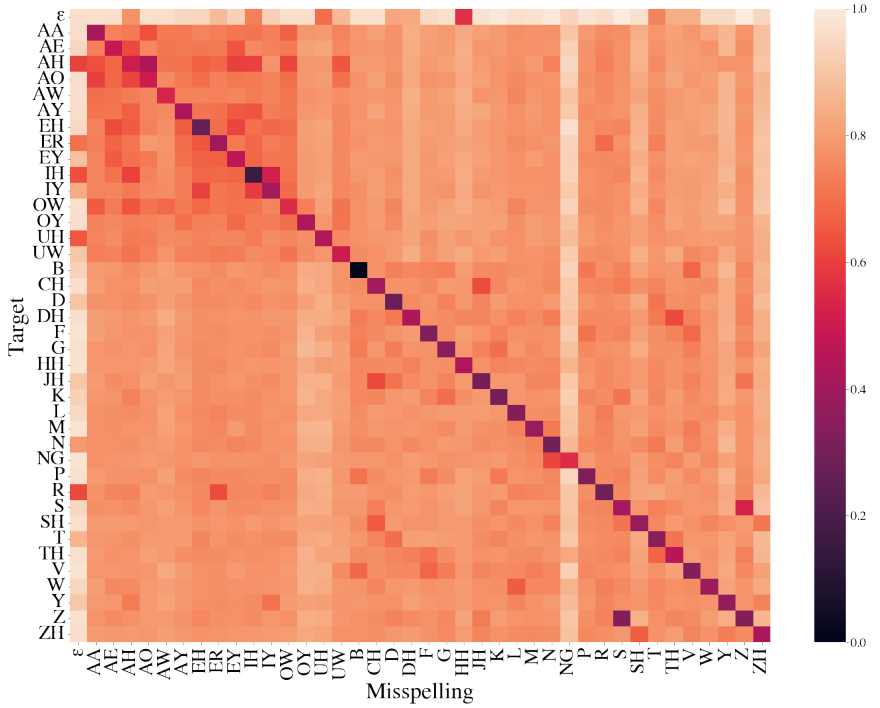


Figure 5.3: The resultant phoneme similarity matrix after adaptation to Irish Accented English visualised as a heatmap.

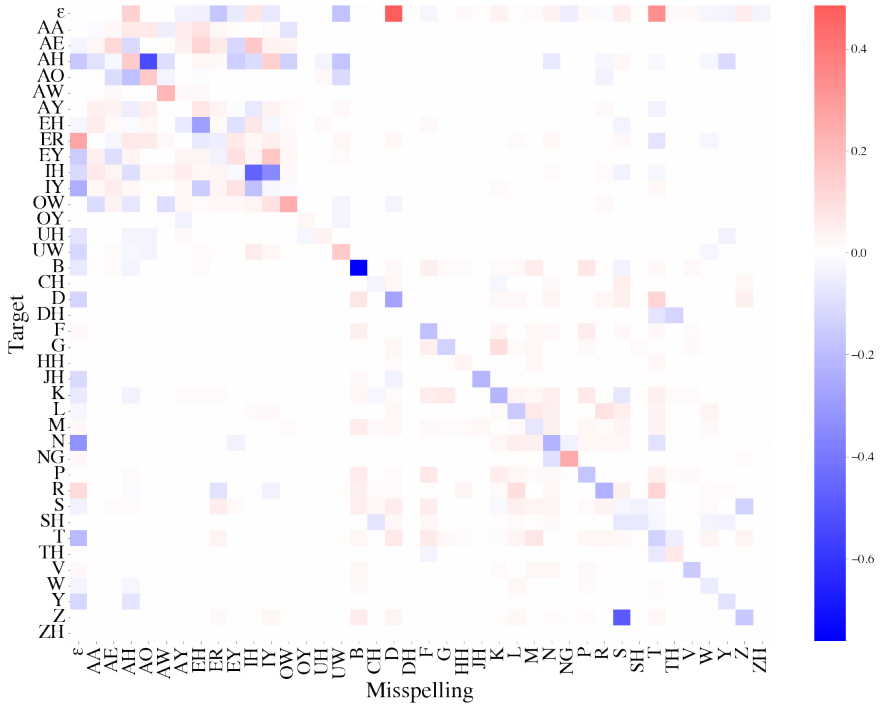


Figure 5.4: A visualisation of the value differences between the baseline phoneme similarity matrix and the Irish Accented English tuned matrix.

5.3.1 Consonants

Perhaps the most salient feature of Irish Accented English is the fortition of the dental fricatives to alveolar plosives (Hickey, 2004). Indeed, the /TH/ phoneme is one of the few consonantal phonemes whose weight associated with its correct encoding in a spelling production[1] increases through adaptation. This suggests that this phoneme was often involved in substitutions with other phonemes. As expected, the most common substitute is /T/ with the weight associated with this substitution showing the largest decrease of any /TH/ substitution. Examples of this fortition can be seen in misspellings such as *'someting' ('something') and *'tink' ('think').

Many Irish Accented English varieties, particularly in the South, exhibit the lenition of /T/ whereby this phoneme is realised as a fricative (Hickey, 2008). As such, it could potentially be mistaken for an instance of the /TH/ phoneme and indeed would explain the decreased cost of substituting a /T/ in the target word for a /TH/ in the misspelling. Potential examples of this process include the misspellings *'tath' ('that') and *'tath's' ('that's'). The large decrease in cost for the deletion of /T/ is also likely an extension of this lenition process.

Hickey (2005) suggests that *Local Dublin English* is non-rhotic whilst Corrigan et al. (2012) argues it is instead weakly rhotic. Since the cost of /R/ deletions increases through adaptation, it would seem likely that this variety is more rhotic in nature than other English varieties like Southern Standard British English. Nevertheless, cases of intrusive /R/ (typically a pronunciation feature associated with non rhotic varieties) are evidenced in the corpus in misspellings such as *'tourt' ('taught') and *'instergram' ('instagram'). These points are reflected in the tuned weight matrix, with the cost associated with deleting /R/ increasing whilst remaining the most likely consonant to undergo deletion, and in the decreased cost of inserting /ER/ and, to a lesser extent /R/.

A more common pronunciation feature found in many varieties of English is that of /NG/-fronting. This is the process by which word-final /NG/ is realised as an alveolar nasal [n] rather than a velar nasal [ŋ] (Houston, 1985). Indeed, the cost of substituting an /NG/ phoneme in the target word with an /N/ phoneme in a misspelling is notably low and examples of this process can be seen in the misspellings *'bein' ('being) and *'cyberbullyin' ('cyberbullying').

Another process which is common to a number of English varieties is word-final devoicing, specifically of the sibilant /Z/ (Docherty, 2011; Smith, 1997). This process results in voiceless productions of /Z/ phonemes much like the typical realisation of /S/. It is likely as a result of this process that misspellings such as *'becus' ('because') and *'bekos' ('because') arise in the corpus and contribute to the reduced cost of substituting /Z/ in the target word with

[1] The weights associated with phonemes being correctly encoded in a spelling production make up the top left to bottom right diagonal of the matrices

/S/ in the misspelling. However, it should be noted that a number of English words ending in the letter ‘s’ elicit a /Z/ pronunciation and so it is not necessarily the case that a devoiced realisation was encoded intentionally.

5.3.2 Vowels

The /AH/-/UH/ (STRUT-FOOT) vowel merger is typical in most varieties of Dublin English and has been observed in varieties across Ireland (Hickey, 2008). /AH/ vowels are often raised and produced as [ʊ] in the same way as the /UH/ vowel. It might be reasonable to think that it is a result of this process that misspellings like *‘trost’ (‘trust’) are observed where the raised /AH/ is denoted with the letter ‘o’. This encoding, however, prompts the grapheme-to-phoneme tool to predict an instance of the /AO/ phoneme and so there is a very significant reduction in the cost of substituting /AH/ in the target word for /AO/ in the misspelling.

Hickey and Amador-Moreno (2020) note the considerable dipthongisation of the GOAT vowel in modern day Irish Accented English. This appears to be reflected in the tuned weight matrix in the form of reduced costs associated with /OW/ being substituted for the low vowels /AA/ and /AH/, which reflect the starting point of the diphthongised /OW/ realisation, and with the /AW/ diphthong. Examples of this process can be seen in the misspellings *‘towled’ (‘told’) and *‘down’t’ (‘don’t’).

One particularly salient feature of *Local Dublin English* is /AY/ onset raising - to the extent that *Fashionable Dublin English* overtly opposes this process by retracting to a low back starting point for this phoneme (Hickey, 2005). Thus it could be expected that such a marked pronunciation would be encoded in some way in the misspellings. The tuned similarity matrix does exhibit a relatively small decrease in costs associated with substituting /AY/ for the short monopthongal /IH/ and /AH/ which might be believed to stem from this pronunciation feature. One misspelling example where the grapheme-to-phoneme tool predicted /IH/ in place of /AY/ is in *‘ciber’ (‘cyber’). However, it is unclear whether or not the child was encoding the raised nature of /AY/ through the use of the letter ‘i’. Indeed a similar instance of an /AY/ -> /IH/ substitution arises from the predicted phoneme sequence of the misspelling *‘niss’ (‘nice’). Here, the use of the letter ‘i’ is not incorrect despite it being predicted as denoting the /IH/ phoneme. This example illustrates the difficulty in investigating the pronunciation features which influence spelling without suitable clarification of the child’s intentions.

5.4 Limitations

This chapter is intended as a preliminary investigation into the idea that regional pronunciation variation manifests in the spelling errors of children and that an adapted spelling correction model is both feasible and beneficial to speakers of such variants. Whilst initial results support the original hypothesis, it is important to note the limitations of the approach presented here.

The Irish Accented English corpus used for tuning and testing is limited in size and scope. It is a relatively small sample of misspellings and is biased towards Irish Accented English speakers from Dublin. With so few misspelling items, not all pronunciation features can be expected to arise. Additionally, for many of the misspellings which appear to encode pronunciation features the differences between the misspelling and the target word are relatively minor. For example, *'tink'('think'). It cannot be confirmed whether this is indeed a cognitive error or merely a typographic error although the strongest argument for the former lies in the fact that, in many cases, multiple instances of the same encoding error across different words were produced by the same child.

In addition, the weighted edit distance approach employed by S-capade only accounts for one-to-one phoneme mappings and is context independent. Thus, potentially influential features of Irish Accented English may not be adequately captured. For example, in Dublin English, short vowels are typically lengthened when they occur before /R/ (Hickey, 2004). This context dependent feature might be encoded in misspellings but the current model has no way of distinguishing the specific _/R/ environment. Should this prove to be a significant shortcoming, a more complex method of calculating similarity would need to be applied.

Finally, as discussed previously in Chapter 4, the grapheme-to-phoneme tool used to predict the phoneme sequences of misspellings does not always accurately represent the child's intended phoneme sequence. Whilst, in the case of the general English spelling correction approach, this could potentially lead to poorer rankings of candidate corrections, in the case of the adapted model, this could be a significant loss of information in terms of investigating the salient pronunciation features which are encoded in children's misspellings.

5.5 Summary

This chapter dealt with Research Question 3: *How might the pronunciation variation present in a child's English impact the misspellings they produce and can a spelling correction tool be adapted to better perform with a specific variety?* The possibility of adaptation was demonstrated through the fine tuning of the weight matrix component of an English spellchecker developed previously in Chapter 4. The performance of this Irish Accented English adapted spellchecker was then tested on writings from Irish Accented English

speakers and compared against a British Accented English model and the baseline untuned model. The comparative success of the Irish Accented English model over the others indicates that adaptation to the specific spoken variety of the writer results in improved spelling correction performance. As such, language technology adaptation can benefit users with specific language varieties. The results also support the hypothesis that misspellings are influenced by the pronunciation features in a child's spoken language. A further qualitative analysis of the misspellings in the Irish Accented English corpus demonstrated the encoding of phonetic variants specific to Irish Accented English and particularly *local Dublin English*.

Chapter 6

Modelling Pronunciation Variation in Individual Speakers

The influence of specific realisations and usage on the perceived similarity between phoneme categories was established previously in Chapter 3. In Chapter 5, it was also shown that specific pronunciation features of a spoken variety could be captured based on the observed substitutions, deletions, and insertions which arose from the orthographic encoding of words in children's phonetic spelling. In this way, a model of Irish Accented English was developed based on the similarity relationships between phonemes which was shown to capture known features of the variety. However, it was noted that not all variant pronunciations would be as likely to feature in these misspellings. For this reason, it is then a natural progression to investigate whether variation could be examined in a similar way using an individual's speech. This is the core aim of this chapter which focuses on Research Question 4: *Can an individual's variety of English be modelled based on phoneme confusability and do speakers with similar varieties produce similar representations?*

To answer this question, speaker models of phoneme similarity are constructed by leveraging erroneous ASR output and considering similarity as a function of confusability. It will be demonstrated how individuals can be represented by a phoneme substitution matrix whose values correspond to observed substitutions between a read text prompt and the output of an ASR system. This includes insertions and deletions which are treated as substitutions involving the empty string (ε) as discussed in previous chapters. These substitution matrices contrast with the previous distance matrices in that higher values correspond to a higher degree of similarity between phonemes. However, the models maintain the same level of simplicity and interpretability of those generated in previous chapters and are capable of capturing features of pronunciation variation in the same way. In fact, models of speakers of the same region, who might be expected to use the same spoken variety of English, are shown to produce similar representations even from relatively little data. This would suggest that such models of phoneme similarity contain pronunciation information which generalises across speakers of a particular variety.

This chapter's main contributions include a process for generating speaker-level models of phoneme similarity which require minimal annotated data and a demonstration of how these models capture pronunciation features common to speakers with the same spoken variety through the success of a region classification task. This work has been presented in O'Neill and Carson-Berndsen (2022b) and is included in O'Neill and Carson-Berndsen

(in preparation). The outline of the chapter is as follows. The method of building speaker models from ASR output is first described in Section 6.1 and the resulting representations are visualised to determine whether speakers from the same region generate similar models. In Section 6.2, the commonalities between speakers of the same region are then further explored by evaluating the models' use in a region classification task. The relative success of this task suggests that the speaker models indeed capture pronunciation features which are common to the English variety of a region. Furthermore, it is then demonstrated that descriptive models can be generated from relatively minimal annotated data by testing the classification accuracy per speaker utterance. The sources of classification errors were then investigated in Section 6.3 and it was determined that the misclassification of a speaker's region is not necessarily a result of a weakness of the model but, rather, due to the natural variation in the speech of an individual who might exhibit features more associated with a region other than their own. Finally, the limitations of the modelling approach are discussed in Section 6.4 followed by a summary of the work presented in this chapter in Section 6.5.

6.1 Modelling a Speaker

The aim of this chapter is to model the pronunciation features of a speaker's variety of English and to determine whether those with similar spoken varieties have similar representations. If this is indeed the case, then analysis of the models would provide information about not only the individual speakers but also general information regarding the pronunciation features of the English variety used in a region. Here, the similarity between phonemes is determined through the misrecognition of phoneme categories by the ASR system as a result of specific phonetic realisations to which the system has not been sufficiently exposed during training. These misrecognitions are leveraged to construct speaker models whilst prioritising simplicity and interpretability. It is then demonstrated that these representations do appear to capture common features of a region's English variety since speakers of the same region tend to cluster together in the multi-dimensional space.

6.1.1 The Corpus

Audio data and text prompts are taken from a corpus of speech data from Datatang, originally used in The Accented English Speech Recognition Challenge at Interspeech 2020 (Shi et al., 2021) and henceforth referred to as the Accented English corpus. In total, this corpus contains approximately 200 hours of spoken English and corresponding text prompts collected from speakers in 10 different regions across the globe. A breakdown of these regions, the amount of speech data available for each, and the number of distinct speakers recorded can be found in Table 6.1. It is important to note, however, that whilst the corpus labels speakers as, for example, "American" or "Spanish", the available metadata does not confirm the nationalities

Region	Hours of Speech	Total Speakers	Training Speakers	Test Speakers	Total Utterances	Training Utterances	Test Utterances
American	20.06	70	49	21	18311	12935	5376
British	21.83	92	64	28	19523	13012	6511
Canadian	20.05	44	31	13	18403	13106	5297
Chinese	20.61	50	35	15	15073	10318	4755
Indian	20.00	42	29	13	14665	10024	4641
Japanese	20.19	46	32	14	16785	11538	5247
Korean	20.01	46	32	14	17652	12220	5432
Portuguese	20.29	53	37	16	17959	13074	4885
Russian	20.04	41	29	12	16561	12031	4530
Spanish	20.02	44	31	13	17675	12422	5253
Total	203.10	528	369	159	172607	120680	51927

Table 6.1: A summary of the Accented English dataset including the number of speakers, hours of speech, and number of utterances for each region and for the training and testing subsets.

of the speakers, nor their linguistic background, fluency in English, nor native language. As such, when the terms “American speakers” or “Spanish speakers” etc. are used throughout this thesis, it is solely to refer to the region in which the speaker was recorded. Whilst an individual speaker might not be representative of the English variety used in a region, it is assumed that any pronunciation feature demonstrated as being common across speakers of a region does pertain to the variety as a whole.

6.1.2 Automatic Speech Recognition and Phoneme Alignment

For the ASR component, the wav2vec 2.0 semi-supervised model fine tuned on 100 hours of the LibriSpeech Corpus (Baevski et al., 2020) is used[1]. The wav2vec 2.0 model aims to learn discrete speech units from unlabelled data and then, using annotated data, is fine-tuned to represent those units through text. Its use of a character level vocabulary, rather than word level, means that the output of this system often includes non-words which capture the phonetic nature of the speech. For example, one Japanese speaker read the prompt;

“I want to sleep, please adjust the car window"

and the ASR output;

“I want to sreep praise a just as a cow windo"

The likely influence of Japanese on the speaker’s English productions is seen in the words ‘sreep’ and ‘praise’. The /R/ and /L/ contrast of English is not distinctive in Japanese phonology which has just one liquid phoneme (Strange and Dittmann, 1984). As such,

[1] Available `https://huggingface.co/facebook/wav2vec2-base-100h`

Prompt	AY	W	AA	N	T	UW		S		L	IY	P	L	IY	Z	AH	JH	AH	S	T		DH	AH	K	AA	R	W	IH	N	D	OW
ASR	AY	W	AA	N	T	UW	SH	R	IY	P	R	EY	Z	AH	JH	AH	S	T	AE		Z	AH	K	AW		W	IH	N	D	OW

Table 6.2: An example of how the phoneme sequence from the prompt text and that of the ASR output are aligned to identify insertions, deletions, and substitutions.

Japanese accented English can lack clear distinction between these sounds, often using the Japanese liquid which is typically realised as an apico-alveolar tap [ɾ]. In this instance, it results in the target /L/ being recognised as an /R/ phoneme by the ASR. Thus, the character level output of the wav2vec 2.0 model allows for the detection of variant pronunciation, the examination of its source, and how it affects the ASR performance.

The ASR is run on the audio files from the Accented English corpus and the outputs, along with the text prompts, are converted to sequences of phonemes in ARPAbet notation using the CMU Pronouncing dictionary (Weide, 1998) for real words and the grapheme-to-phoneme tool (CMUSphinx, 2016) trained on this dictionary for non-words. The phoneme strings of each prompt and corresponding ASR output are then aligned using the method described in Chapter 4 Section 4.1.2. This approach ensures an optimised alignment between phonemes based on their similarity rather than basing it solely on the number of edit operations. For instance, consider the phoneme alignment of our previous example as depicted in Table 6.2. The section ‘the car’ in the text prompt was converted to the sequence /DH AH K AA R/, whilst the recognised ‘as a cow’ was converted to /AE Z AH K AW/. There are two possible alignments for these sequences that have the same number of edit operations as shown in Table 6.3. However, there is a lower cost associated with treating the recognised /AE/ as an insertion and the /Z/ as a substitution for the /DH/ in the text prompt since the /DH/ and /Z/ phonemes are considered more similar than /DH/ and /AE/. A standard non-weighted edit distance approach would have considered equally an alignment where the /AE/ was a substitute for /DH/ and the /Z/ was an insertion which would not have yielded as useful confusions for analysis purposes. Instead, using the weighted approach, the substitutions on which the speaker models are built are more representative of the variant pronunciations observed in the speech from that region.

Alignment 1						Alignment 2					
	DH	AH	K	AA	R	DH		AH	K	AA	R
AE	Z	AH	K	AW		AE	Z	AH	K	AW	
i	s			s	d	s	i			s	d

Table 6.3: Two different potential alignments of the phoneme sequences from the prompt “the car" and the erroneous ASR output “as a cow".

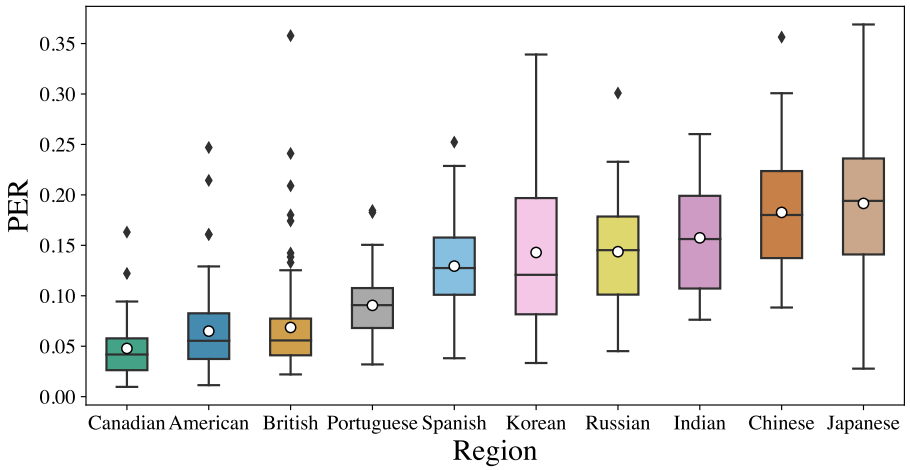


Figure 6.1: Boxplot of Phoneme Error Rates (PER) for each region in the Accented English dataset based on the wav2vec 2.0 output.

The distribution of Phone Error Rate (PER)s for speakers in each of the regions is shown in Figure 6.1. This figure depicts how speakers with similar English variants to the speech on which the ASR system was trained (North American varieties) produce more accurate transcriptions. The system exhibits better performance with speakers from Canada and America which achieve average PERs of 4.8% and 6.5% respectively. Conversely, much higher phoneme error rates are observed with speakers in regions where language transference effects might be expected to occur. For instance speakers from China obtain an average PER of 18.3% and speakers from Japan obtain an average PER of 19.2%. The Englishes used in these regions deviate enough from the training data so as to impede recognition. However, these misrecognitions allow for an examination of variant pronunciations and analysis of their source from a phonetic standpoint.

6.1.3 Phoneme Substitution Matrices and Vector Models

Using the aligned prompt and ASR phoneme sequences, a matrix is constructed for each utterance which captures counts of the observed phoneme substitutions. As before, insertions and deletions are also included by treating them as substitutions which involve the empty string (ε). An example of such an utterance matrix (visualised as a heatmap for clarity) for the prompt “I want to sleep, please adjust the car window" and corresponding recognition output “I want to sreep praise a just as a cow windo" can be seen in Figure 6.2. This is a sparse matrix with entries primarily on the diagonal which represent the correctly recognised phonemes. The other darker cells beyond the diagonal capture the observed insertion of /AE/, deletion of /R/, and the various substitutions which occur. As can be seen, the cell representing the substitution of /L/ with /R/ is darker than the other substitution cells since it occurs twice within the prompt.

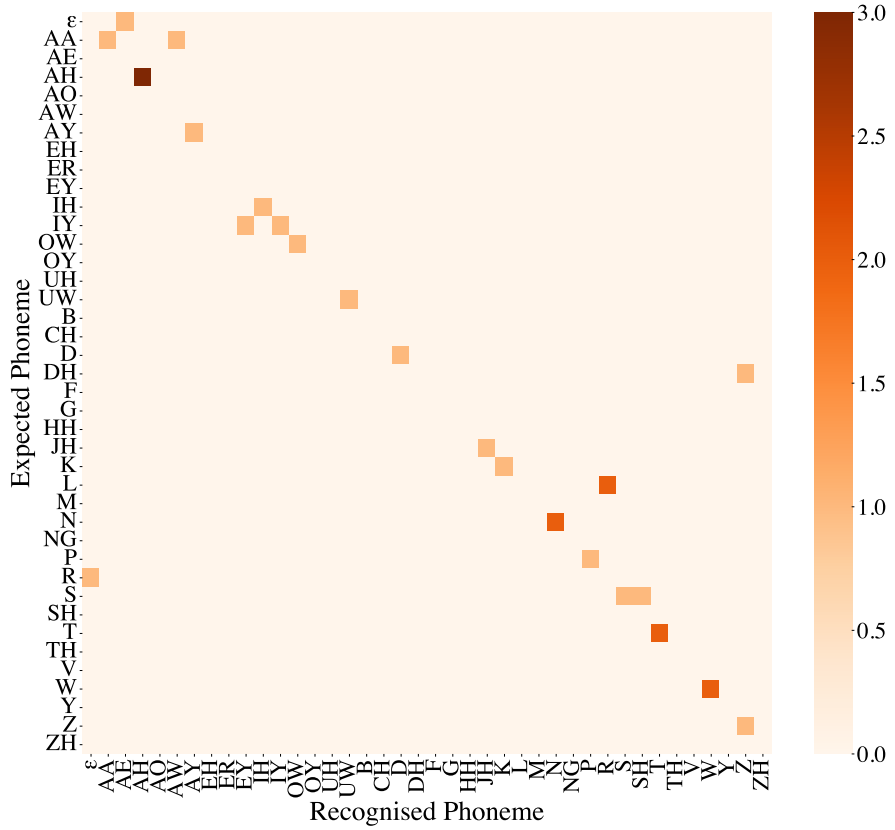


Figure 6.2: The substitution matrix for a single utterance capturing counts of all possible phoneme substitutions visualised as a heatmap.

Speaker-level substitution matrices are then constructed by cumulatively combining the values for each of the utterances recorded by a speaker. Then, the values in the diagonal (correctly recognised phonemes) were set to 0. This moves the emphasis of the representations to the misrecognised phonemes only since the focus is intended to be on the variation that is observed rather than the proportion of speech which exhibits variation. This results in a numerical representation of each speaker in the corpus which can be analysed as a 40x40 matrix or reshaped into a vector of size 1600.

In theory, vectors representing speakers who exhibit similar patterns of spoken variation will be located in close proximity when projected into this 1600-dimensional space. If the captured variation patterns are common to a particular region, it can be expected that the vector representations of speakers from the same region will form clusters in the space. To investigate this hypothesis, a k-means clustering algorithm (k=10) (MacQueen, 1967) is applied to the high-dimensional speaker vectors and then the vectors and cluster centers are reduced to 2 dimensions using tSNE (Van der Maaten and Hinton, 2008) for visualisation purposes. The results of this can be seen in Figure 6.3. In this figure it can be seen that the cluster centres are found in areas surrounded by groups of speakers primarily from the same region. This is despite the information loss when transforming from 1600 dimensions

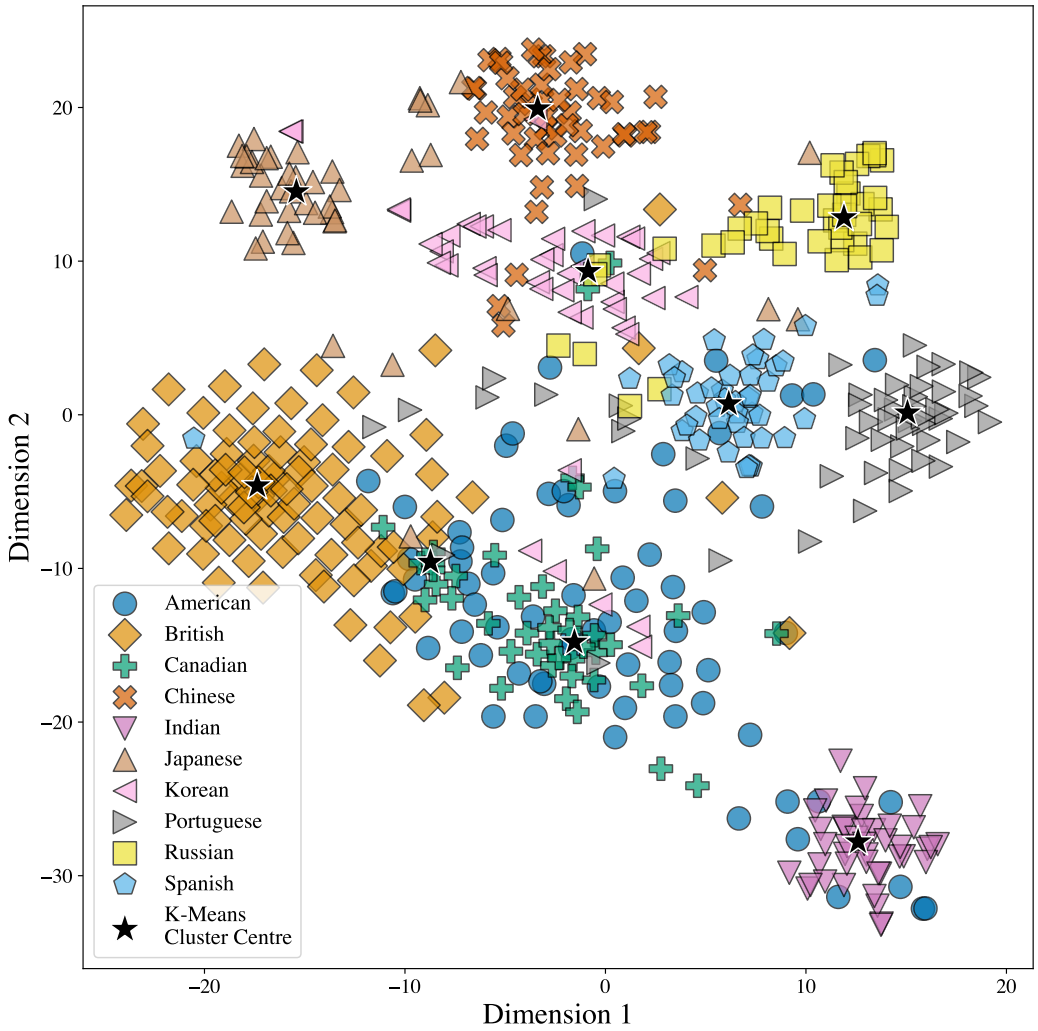


Figure 6.3: A visualisation of the Accented English corpus speaker vectors and the predicted cluster centres from K-means clustering after tSNE dimensionality reduction.

down to only 2. This suggests that there do exist common patterns in the variety of spoken English used in a particular region, that the ASR system is sensitive to these patterns, and that they are captured in the speaker representations constructed from observed phoneme substitutions between the prompts and ASR output.

There are, of course, notable outliers where speakers of one region appear to cluster with those of another. As mentioned previously, the full linguistic background of the speakers is not known. It is plausible that some speakers' English productions are influenced by individual experience and exposure to other varieties and languages thus making their speech more similar to that of a speaker of another region. For example, a high degree of variability and wide distribution amongst American speakers is observed with many located in close proximity to the Spanish cluster centre and to the Indian cluster centre. If it is assumed that the Spanish speakers likely speak Spanish and that their English productions are influenced

by their knowledge and experience of Spanish, then it can be hypothesised that the American speakers located around this cluster also exhibit Spanish influence in their spoken English. This would make sense theoretically given the prevalence of the Spanish language across the US and Latin America. In the same way, the high proportion of Indic languages spoken in the US, for instance Hindi and Bengali, could be the source of American speakers exhibiting pronunciation features in common with the Indian speakers. These assumptions appear to be confirmed after listening to a sample of these outlier speakers although extensive listening experiments and evaluations were outwith the scope of this work. In addition, there is a large overlap between the Canadian and American speakers. Whilst it could be argued that this is a result of the similarity between the Englishes used in these regions, which do share a number of variant pronunciations, it is also important to note that the ASR exhibited the best performance on speakers from these two regions. On average, the ASR demonstrated a PER of 4.76% for Canadian speakers and 6.56% for American speakers (refer to Figure 6.1). Since speakers are represented based on the observed phoneme mismatches between the prompt and ASR, a lower PER means there is considerably less information with which to model these speakers and distinguish between them which could also contribute to the two groups overlapping.

6.2 Region Classification

This section investigates the potential for using the previous vector representations in order to classify the region of a speaker. This is not done with the aim of developing a highly accurate classifier but, rather, to evaluate the effectiveness of the speaker modelling in capturing pronunciation features. If classification can be carried out successfully using only the information pertaining to phoneme substitutions, insertions, and deletions between the prompt and ASR output, this lends credence to such misrecognitions being the result of systematic patterns of variation which are specific to and common across a particular English variety. For use with each of the following classification approaches, the corpus is split into stratified training and test sets by speaker in a 70:30 ratio respectively with proportional numbers of speakers from each region. The breakdown of how many training and test speakers there were from each region can again be seen in Table 6.1.

6.2.1 Classification using KNN

A K-Nearest Neighbours classifier (K=5 chosen as it achieved the best accuracy over other K values tested on the training data) is fit on the 369 unit vectors representing the training set of speakers. The region of each test speaker is then predicted as being that shared with the majority of its 5 closest neighbours in the multi-dimensional space. Classification of the test speakers achieves an overall accuracy of 70.4%. The confusion matrix for the KNN classifier

is depicted in Figure 6.4. From these results, a disparity in the effectiveness of the classifier can be seen. Significantly poorer performance results are observed for the American and Canadian regions with accuracies of 19.0% and 23.1% respectively. Conversely the Chinese, Indian, and Russian test speakers are classified with 100% accuracy. As discussed previously in Section 6.1.3, there is a high degree of variation within the American speakers group which results in a wide distribution of region predictions for these test speakers. As such, comparatively lower classification accuracy is expected for this group. On the other hand, the poor classification results for the Canadian speakers appears to be directly related to the overlap with American speakers rather than the inter-variability between speakers of this region. The Canadian speakers were observed to cluster in the multi-dimensional space, suggesting the majority of speakers exhibit similar patterns of pronunciation. However, many American speakers, of which there are more of in the corpus than Canadian speakers, appear to also produce similar pronunciation variants. This has a negative effect on the classification of the Canadian test speakers with the majority being predicted as American speakers.

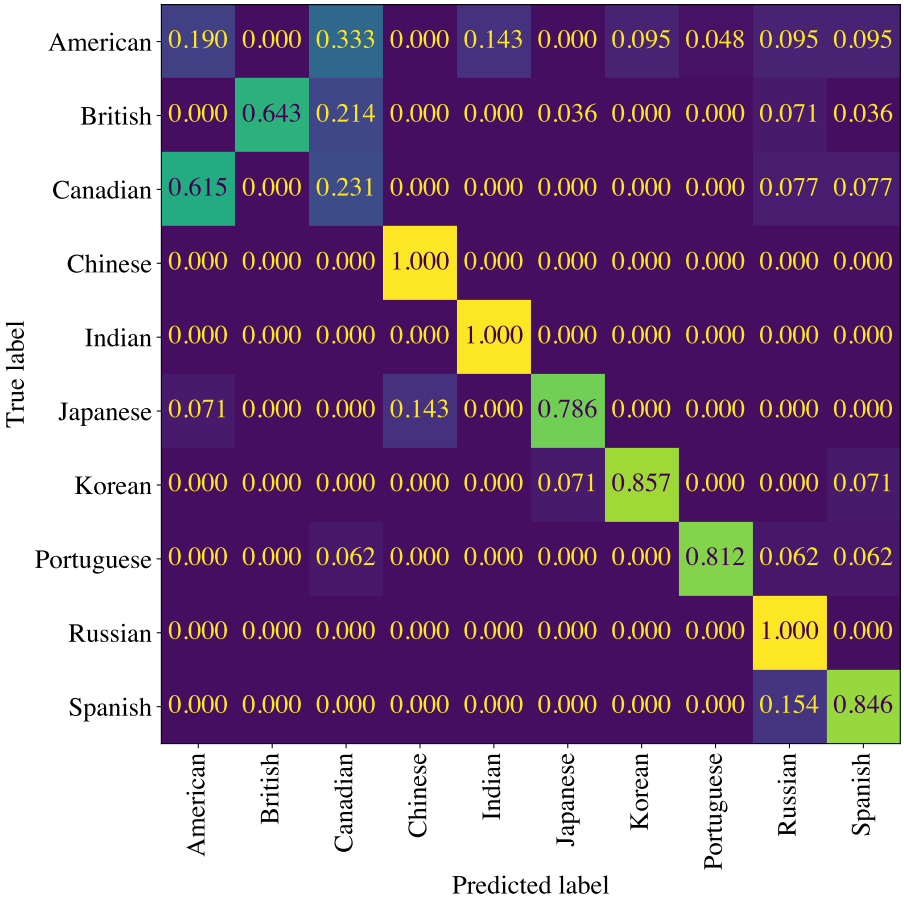


Figure 6.4: The resultant confusion matrix for the KNN classifier.

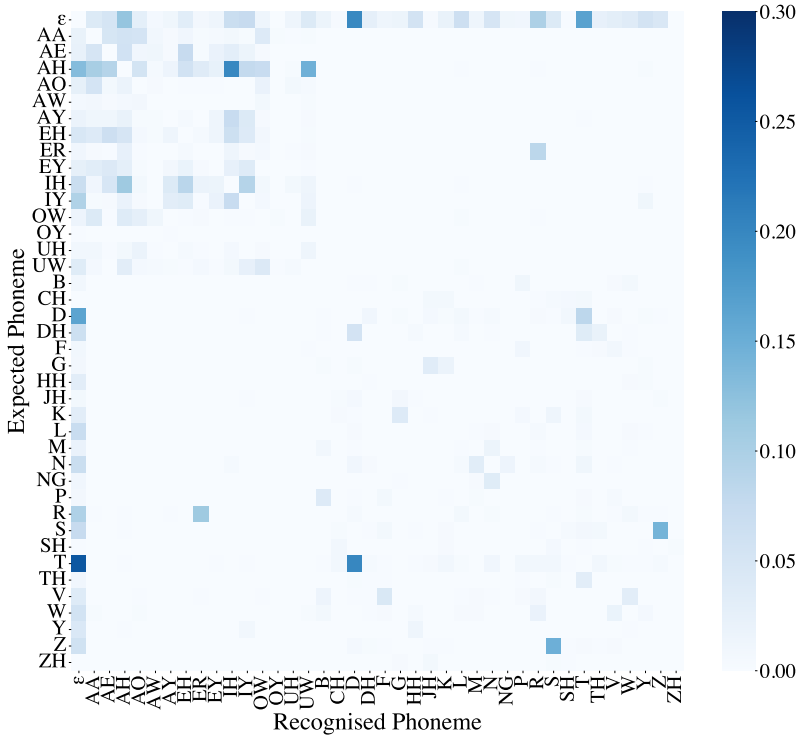
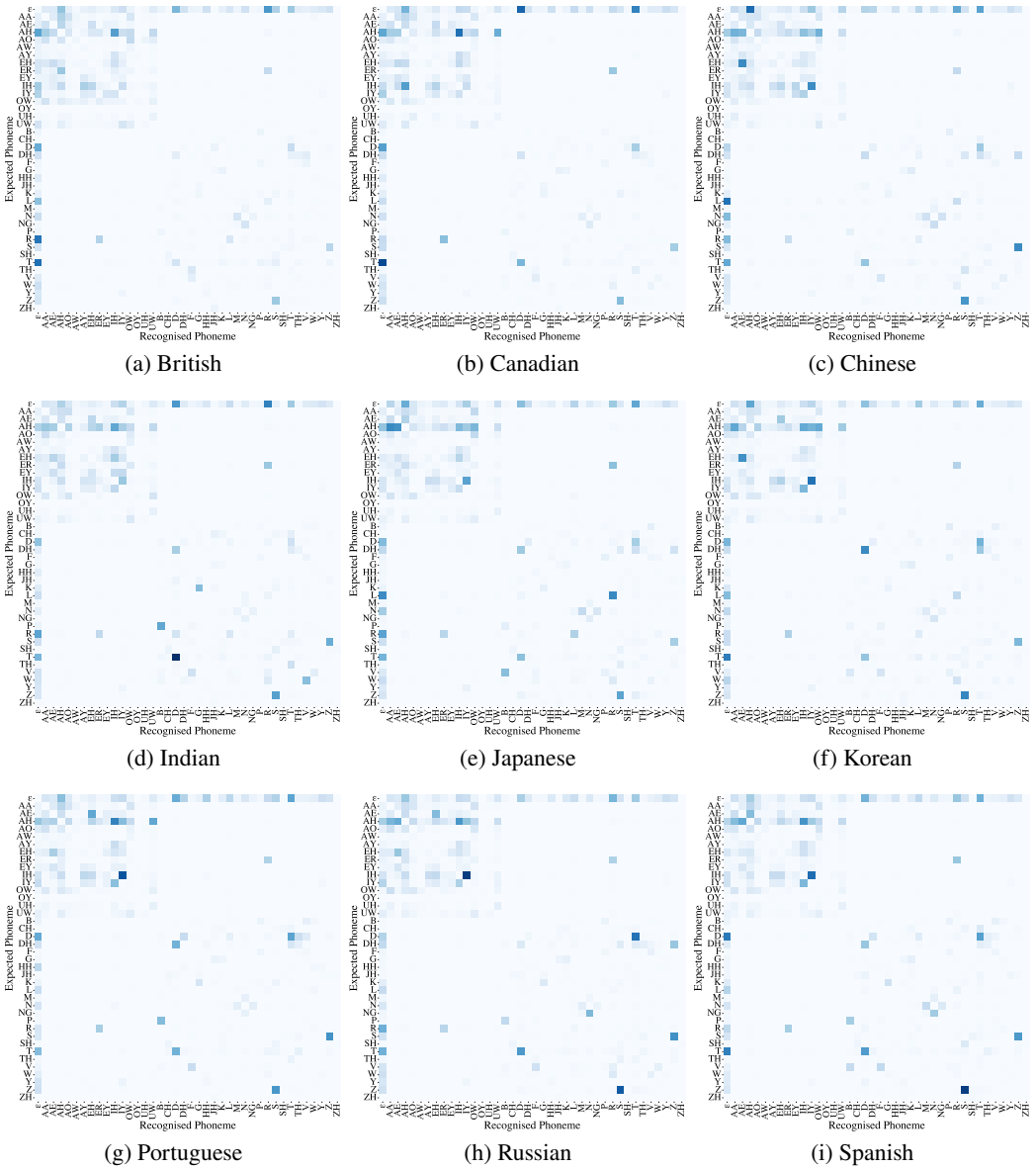


Figure 6.5: The American region profile reconstructed as a 40x40 phoneme substitution matrix visualised as a heatmap.

6.2.2 Classification using Region Profiles

The KNN classifier obtains relatively high classification accuracy. However, as mentioned previously in Section 6.2.1, the method is disadvantaged by the outlier speakers who are labelled as one region but whose English variant appears to more closely resemble that of another. This is especially seen in the American group which appears to have no clear cluster but where many overlap with the Canadian group and negatively impact the classification of the Canadian speakers. This prompts an alternative approach to classifying test speakers by basing the prediction on the similarities between their representations and those of a prototypical speaker of a region. Region profiles are generated by summing the unit vector representations of each training speaker of a region. Thus, the focus is shifted more towards which pronunciation features speakers of a region have in common within their variety. By using the unit vector representations it is ensured that each speaker contributes to the profile equally and speakers whose speech results in higher proportions of substitutions, insertions, and deletions do not bias the representation. Normalisation of the resulting vectors is then carried out purely for the purposes of later examination and comparison of the region profiles since this classification uses a cosine distance measure which is not affected by the magnitude of the vectors. The resulting region profiles, reconstructed as phoneme substitution matrices and visualised as heatmaps, are shown in Figures 6.5 and 6.6. The raw data versions of these profiles can also be found at the link in Appendix A.1.2.



(a) British (b) Canadian (c) Chinese

(d) Indian (e) Japanese (f) Korean

(g) Portuguese (h) Russian (i) Spanish

Figure 6.6: The region profile vectors for the other regions in the Accented English dataset reconstructed as phoneme substitution matrices and visualised as heatmaps.

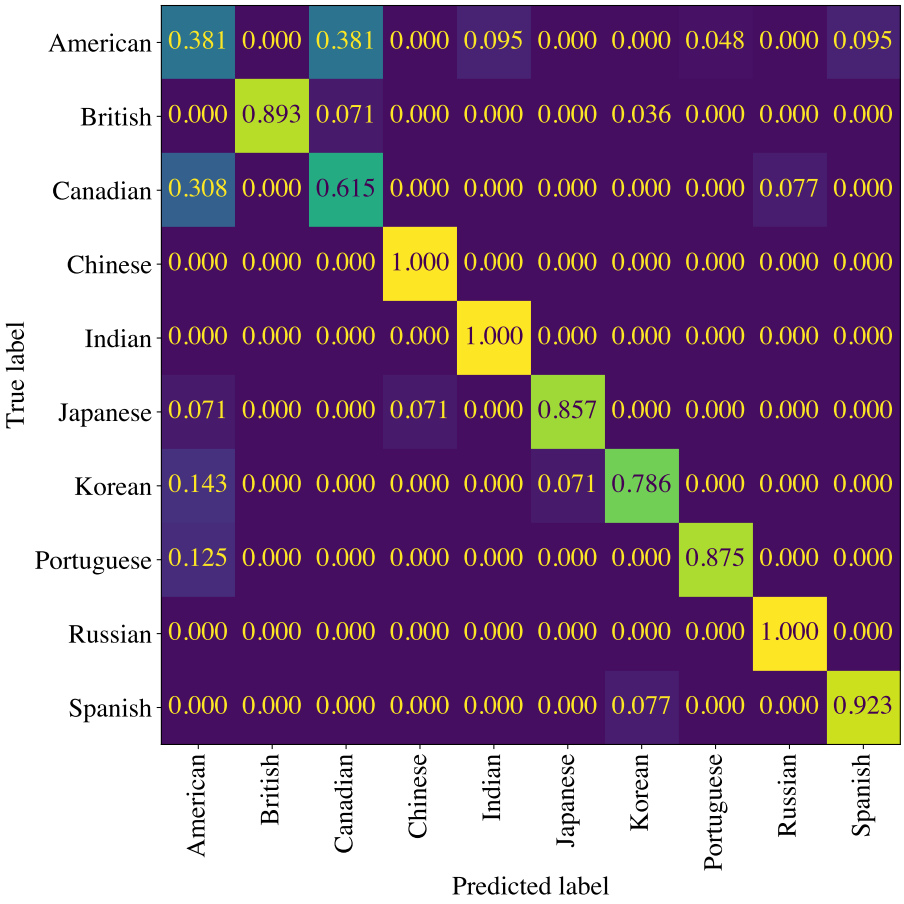


Figure 6.7: The resultant confusion matrix for the Region Profile classifier.

At classification time, the vector representation of a test speaker is compared to the 10 region profile vectors. Regions are ranked in order of cosine distance with the closest being predicted as the test speaker's region. An overall accuracy of 81.8% is achieved and the full breakdown of confusions for each region is shown in Figure 6.7. As can be seen from these results, using average region profiles alleviates confusions between American and Canadian speakers, and improves the classification accuracy for almost all of the regions. Additionally, by using a ranking system, this approach can be evaluated not just on overall accuracy or confusions but on Mean Reciprocal Rank (MRR). This is a measure of how highly the true region label typically ranks in the list of candidates with values closer to 1 indicating higher average rankings. Table 6.4 provides a comparison between the classification accuracies of the KNN model and the region profile model alongside the MRR scores of the region profiles. The MRR values indicate that, even for regions which achieved a lower accuracy, the true region was still ranked highly amongst the other regions. Indeed, the true region occurred in the top 3 profiles for 96% of test speakers.

Region	KNN Accuracy	Region Profiles Accuracy	Region Profiles MRR
American	19.0%	38.1%	0.668
British	64.3%	89.2%	0.935
Canadian	23.1%	61.5%	0.780
Chinese	100.0%	100.0%	1.000
Indian	100.0%	100.0%	1.000
Japanese	78.6%	85.7%	0.900
Korean	85.7%	78.6%	0.857
Portuguese	81.3%	87.5%	0.906
Russian	100.0%	100.0%	1.000
Spanish	84.6%	92.3%	0.942
Total	70.4%	81.8%	0.891

Table 6.4: A summary of the classification accuracy and Mean Reciprocal Rank of the KNN classifier and the Region Profile classifier across each region in the Accented English dataset.

6.2.3 Classification by Utterance

Underrepresentation of a variety in the training of an ASR system negatively impacts speakers of this variety and often stems from a general lack of data. Thus, it is important to examine the quantity of annotated audio recordings required to build adequate speaker models that capture pronunciation variation. In order to do this, multiple models were

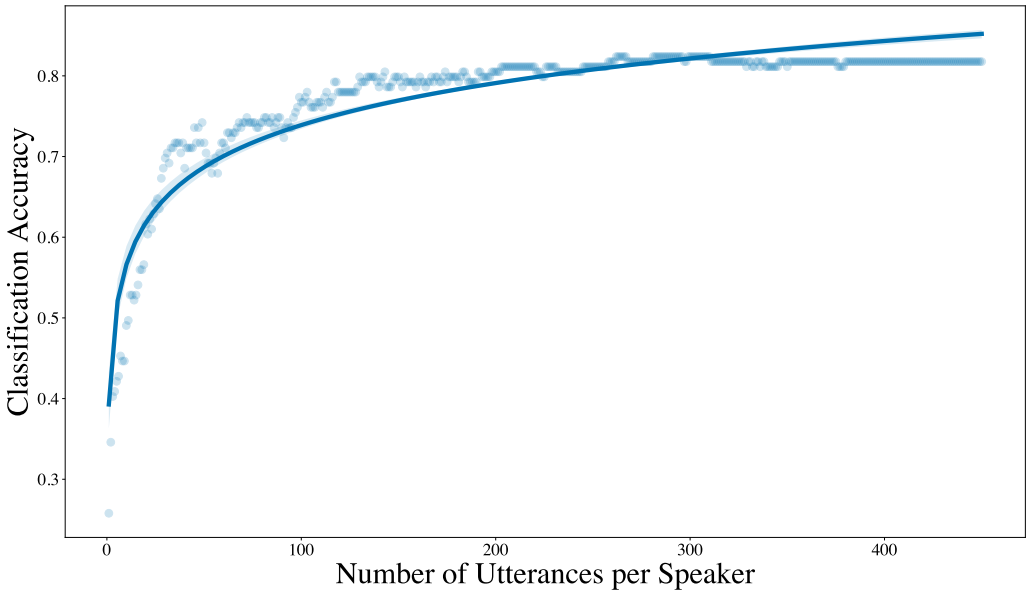


Figure 6.8: The overall classification accuracy of the Region Profile classifier with increasing number of utterances per speaker.

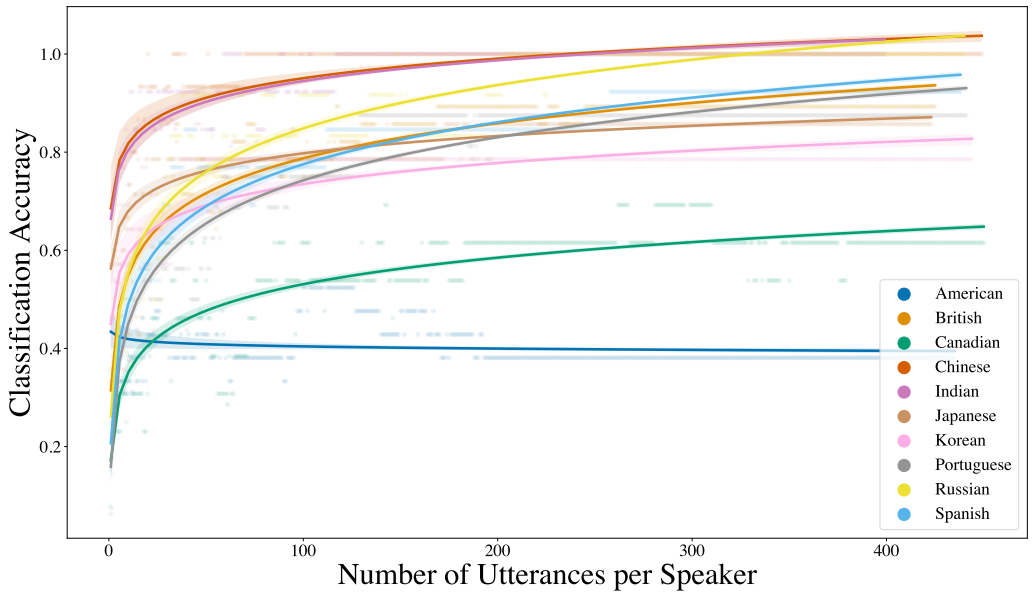


Figure 6.9: The classification accuracy of the Region Profile classifier for each region in the Accented English dataset with increasing number of utterances per speaker.

constructed for individual test speakers each using an increasing number of utterances. Classification using the region profile method was then carried out on these models to determine how the number of utterances impacts accuracy. If the region of a speaker can be classified with reasonable accuracy then the number of utterances used to build that speaker model can be considered adequate in capturing the features of phonetic variation common to that region.

The overall classification accuracy of the system compared with the number of utterances used per speaker can be seen in Figure 6.8 with Figure 6.9 displaying a breakdown per region. Note that not all speakers recorded the same number of utterances and so, where the number of utterances being tested was greater than the total number of utterances available for an individual speaker, the classification was obtained from the maximum number of utterances. As such, higher gains in accuracy could be expected between 200-400 utterances in reality. Nevertheless, it was observed that performance improved dramatically in the early stages with 75% accuracy being achieved at just 115 utterances per speaker. This is, on average, approximately 8 minutes of recorded speech per speaker which is promising for research on underrepresented and low resource varieties. With limited annotated data a speaker's variety can be modelled in enough detail to predict their region with relative confidence and the phonetic and phonological features of their variety can begin to be analysed.

6.3 Investigating Sources of Misclassification

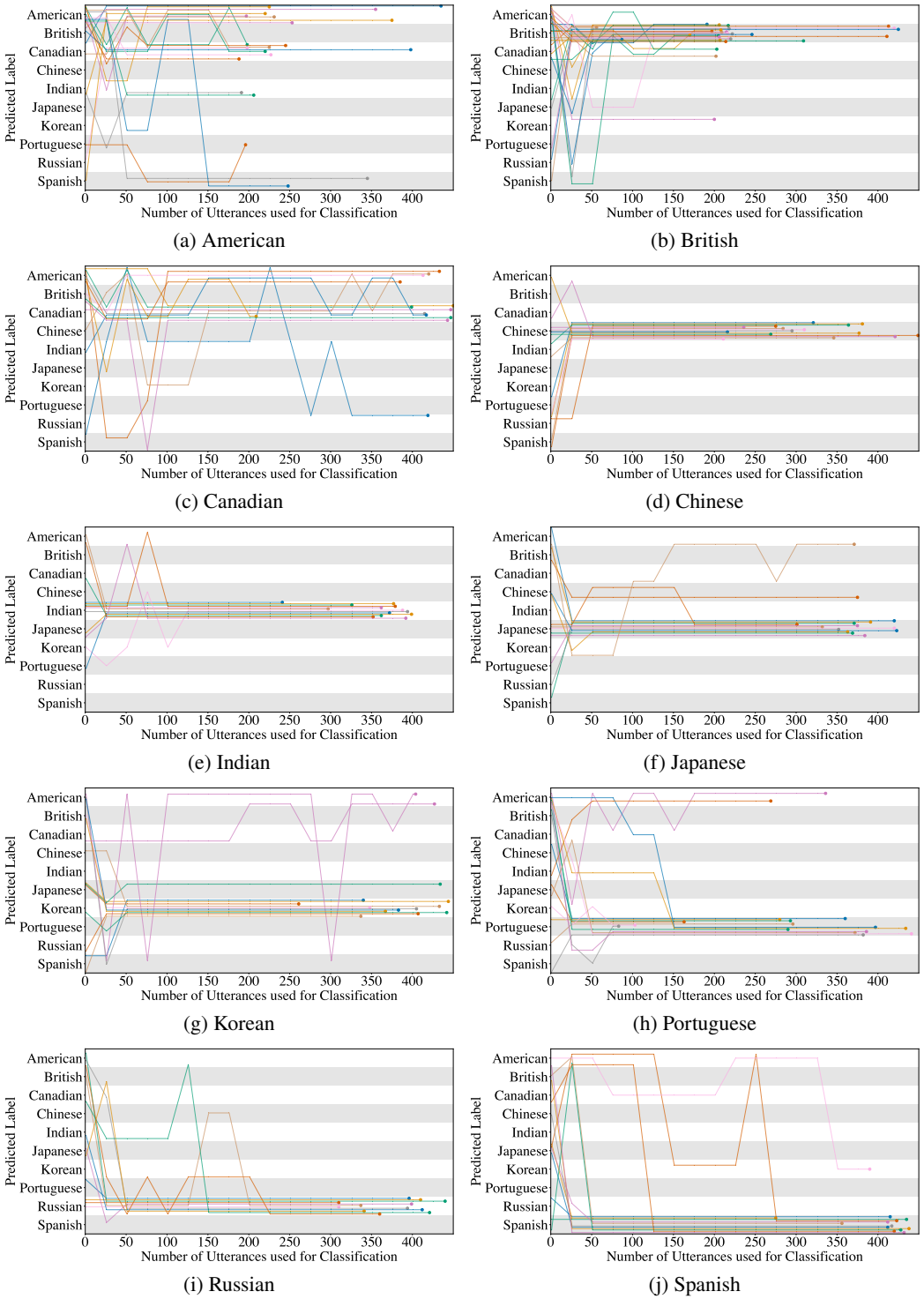
The relatively high classification accuracy achieved through the use of these speaker models based on phoneme similarity supports the idea that the representations themselves capture pronunciation features common to speakers with the same variety. However, individuals do not necessarily always conform to the prototypical English used in their region and their productions could be influenced by other factors as was seen in the visualisation presented in Figure 6.3. It is possible, then, that the misclassification of a specific speaker results not from a poor model of their spoken variety, but from their speech exhibiting features more strongly associated with another region. This hypothesis is explored further in this section.

Figure 6.10 depicts how the predicted region changed with increasing numbers of utterances (in increments of 25) for individual test speakers in each of the regions. These depictions indicate how early a speaker was consistently classified as being from a particular region or how often the predicted region of a speaker changed as an increasing number of utterances were used for classification. This reveals a lot of information regarding speaker variation within a region and commonalities across regions.

Firstly, the American and Canadian speakers were often confused and as the number of utterances increased, speakers tended to flip between the two predicted regions. This suggests that both varieties are similar or that there are no overtly distinguishing features that would indicate a speaker was definitively from either region. This aligns with the previous discussion of how the American and Canadian varieties share a number of features in common but are also the varieties for which there are the fewest mismatches between the prompts and ASR outputs since the system is trained to perform well on these varieties.

The Chinese, Indian, and Russian groups are those for which all speakers were eventually correctly classified. Note that the Chinese speakers converged earliest, followed by the Indian and then Russian speakers. These results correlate with the PER of the ASR on the corresponding varieties. Of the three groups, the system produced the most phoneme misrecognitions with the Chinese speakers thus providing more information in fewer utterances on which to base the classification. Similarly, the PER for the Russian speakers was the lowest of the three meaning more utterances were required to reach sufficient numbers of observed misrecognitions with which to accurately predict the region of a speaker.

Indeed, it would appear that PER plays a significant role in classification, with speakers who attain very low PER being most often classified as being of one of the regions for which the ASR best performs. This can be seen to occur with some British speakers being consistently classed as Canadian, a Portuguese speaker consistently predicted as American, and a number of speakers, specifically one Japanese, one Korean, and one Portuguese, that swap between American and Canadian classifications. Unfortunately, since this modelling and classification approach relies on erroneous ASR output, insufficient information is obtained for speakers who elicit high recognition accuracy.



(a) American (b) British

(c) Canadian (d) Chinese

(e) Indian (f) Japanese

(g) Korean (h) Portuguese

(i) Russian (j) Spanish

Figure 6.10: A depiction of the changing classifications predicted by the Region Profile classifier with increasing number of utterances for each speaker in the 10 regions of the Accented English dataset.

There are some speakers who are misclassified within a relatively small number of utterances and remain consistently predicted as being of this other region. This likely stems from these individual speakers exhibiting particular features which are more closely associated with another region. For example, one British speaker was regularly predicted as being Korean from an early stage. Examination of this individual's matrix representation revealed that they often produced /DH/ phonemes which were recognised as instances of /D/ - a feature which is not indicative of the British region profile but which is prominent in the Korean profile. Similarly one Japanese speaker was continuously classified as Chinese likely due to their /S/ realisations being recognised as /Z/ and one Korean speaker's production of /L/ being misrecognised as realisations of /R/ is likely the reason they were consistently predicted to be in the Japanese group.

6.4 Limitations

Several limitations to the speaker modelling approach presented in this chapter should be noted. Firstly, the wav2vec 2.0 ASR system outputs orthographic forms which are then converted to phoneme sequences using a grapheme-to-phoneme tool. This process does not guarantee an accurate portrayal of the sound sequences recognised by the ASR and has the potential to both lose information and introduce artefacts to the data. For example, often some speakers would produce a word final /Z/ without voicing such that it sounded more like an instance of /S/. The ASR output denoted this with the letter 's' but when the grapheme-to-phoneme tool was applied it mapped this word-final 's' back to the /Z/ phoneme thus losing that information about the speaker's particular pronunciation. Since this work was carried out, other transformer models have emerged which output phonetic and phonemic sequences directly and so would remove the need for the grapheme-to-phoneme tool and would provide an interesting area for future work.

Furthermore, since the models are constructed entirely on the mismatches between read prompts and ASR output, less informative speaker representations are obtained when the ASR performs well for that speaker. Speakers with North American varieties or similar, on which the system was originally trained, elicit much lower PERs and so produce less populated confusion matrices. This in turn leads to poorer classification potential since there is much more limited information on which to make these predictions or investigate the spoken variety.

In terms of the classification carried out, the method is disadvantaged by the lack of a full linguistic background of the speakers. It may be the case that some speakers labelled as one region are heavily influenced by another as a result of the languages they speak or the areas they have lived. Thus, it is difficult to accurately evaluate the success of a classifier without reliable ground truths. Additionally, due to the unbalanced number of speakers, the test sets

for some regions were relatively small and so the 100% classification accuracy obtained for certain regions is too generous. However, this is a more minor limitation since the purpose of carrying out a classification task was to evaluate the models themselves rather than the classifier.

6.5 Summary

This chapter addressed Research Question 4: *Can an individual's variety of English be modelled based on phoneme confusability and do speakers with similar varieties produce similar representations?*. By leveraging the erroneous output of an ASR system, it was shown that individual speakers could be modelled based on the confusability of phonemes and their variant realisations. The resultant representations of speakers from the same region were shown to be similar through both clustering and classification approaches. Thus, the errors captured within the ASR output model pronunciation features shared by speakers of a particular variety. Furthermore, it was demonstrated that an accurate representation of a speaker's variety could be modelled on relatively little labelled data which supports the use of this approach in modelling varieties which are low resource or historically underrepresented.

Chapter 7

Exploring ASR Sensitivity to Pronunciation Variation

In Chapter 6, speakers with similar varieties were shown to produce similar representations when a speaker modelling method which leverages erroneous ASR output was applied. The success of a region-based classifier demonstrated that the representations captured commonalities across speakers from the same region who likely exhibited similar patterns of pronunciation variation. Evidently, there were specific phonetic features shared by groups of speakers to which the ASR was sensitive and which resulted in misrecognitions. This prompted further investigation into what can be learned about a variety through the interpretation of these speaker models and how this might be useful from the perspective of testing and improving ASR systems and in the case of incorporating ASR technology into linguistic research. As such, this chapter seeks to answer Research Question 5: *How does the information captured by these models compare to existing knowledge of a spoken variety and what can they tell us about the sensitivity or robustness of ASR systems to pronunciation variation?*

In order to address this question, the previously constructed region profiles from Chapter 6 are analysed through the lens of existing literature on variation in different Englishes and on the non-English languages commonly used within the regions. It is shown that the confusions resulting from erroneous ASR output relate to established and studied phonetic features of the varieties and language transference effects. The modelling technique is then applied to a second corpus of L2 English speech which contains human annotations at the phoneme level. This allows for a comparison between the confusions which occur in the ASR output and those identified by the annotators. Through this comparison it is demonstrated that the specific phoneme substitutions detected through leveraging the ASR output are consistent with human judgements. Having shown how the confusion based models can be interpreted to extract information about the sensitivity of the ASR system to phonetic variation and the pronunciation features of specific language varieties, the applications of this analysis are discussed from two perspectives. Firstly, consideration is given to how this knowledge can be used to elicit and collect additional training data for the fine-tuning and improvement of an ASR system for a specific variety. Then, from a linguistic research standpoint, the use of ASR to generate automatic transcripts for manual correction and the benefits of understanding the performance of a specific ASR system on a particular spoken variety are discussed.

The main contributions presented in this chapter include a demonstration of the interpretability of the speaker models and region profiles. It is shown that the erroneous ASR output is systematic and consistent across speakers with similar spoken varieties, is explainable through existing literature on the pronunciation features of the different Englishes, and that the observed confusions typically align with the observations of human annotators. This has real world significance both for the development and improvement of ASR systems that are capable of handling different language varieties and for the use of ASR technology as a tool in annotation and variation analysis. The work discussed in this chapter is included in part in O'Neill and Carson-Berndsen (2023) and in O'Neill and Carson-Berndsen (in preparation). The outline of the chapter is as follows. Section 7.1 analyses and explains the highest valued confusions in each of the generated region profiles in terms of the pronunciation variation expected to be exhibited by speakers in that region. Section 7.2 takes this a step further by applying the same modelling techniques to the L2-Arctic corpus and comparing the substitutions observed in the ASR output with the phoneme level annotations provided by human experts. The applications of such models and the information gained through interpreting them are then discussed in Section 7.3 both in the context of improving ASR performance for specific spoken varieties and from the perspective of incorporating ASR technology into the annotation process for linguistic research. The limitations of these analyses are presented in Section 7.4 followed by a summary of the work presented in this chapter in Section 7.5.

7.1 Interpreting the Region Profiles

In this section, the Region Profiles generated previously in Chapter 6 are examined in more detail in order to determine how the features captured by the speaker modelling process relate to existing linguistic theory and literature. The phoneme substitutions corresponding to the five highest matrix values, which can be considered as the most characteristic of a region, are analysed in relation to the varieties of English used in the region, and, where appropriate, the phonetic and phonological features of other languages spoken in the region and how these might influence the English productions of speakers. A summary of these characteristic confusions for each region is given in Table 7.1 and they are discussed more thoroughly in the upcoming sections. Note that there are often many languages spoken within a region and many more dialectal varieties within these languages. Since the profiles are intended to be representative of commonalities between speakers in a region, the observed characteristic features are explained in terms of the most commonly used languages and varieties where possible. A more thorough investigation which includes an examination of all of the varieties within each region or a full acoustic-phonetic analysis is outwith the scope of this work.

Region	Target	Substitute	Value	Region	Target	Substitute	Value
	/T/	ε	0.255		/L/	/R/	0.196
	/T/	/D/	0.200		/AH/	/AA/	0.196
American	/AH/	/IH/	0.198	Japanese	/L/	ε	0.194
	ε	/D/	0.198		/AH/	/AE/	0.187
	ε	/T/	0.165		/R/	ε	0.166
	/T/	ε	0.225		/IH/	/IY/	0.223
	/R/	ε	0.221		/T/	ε	0.212
British	ε	/R/	0.180	Korean	/Z/	/S/	0.201
	/AH/	/IH/	0.172		/DH/	/D/	0.195
	/AH/	ε	0.163		/EH/	/AE/	0.189
	/T/	ε	0.270		/IH/	/IY/	0.258
	ε	/D/	0.238		/AH/	/IH/	0.206
Canadian	/AH/	/IH/	0.226	Portuguese	/S/	/Z/	0.186
	ε	/T/	0.206		/Z/	/S/	0.181
	/IH/	/AH/	0.172		/D/	/T/	0.163
	ε	/AH/	0.230		/IH/	/IY/	0.284
	/L/	ε	0.224		/Z/	/S/	0.252
Chinese	/S/	/Z/	0.195	Russian	/D/	/T/	0.222
	/EH/	/AE/	0.195		/S/	/Z/	0.188
	/IH/	/IY/	0.193		/AH/	/IH/	0.181
	/T/	/D/	0.416		/Z/	/S/	0.283
	ε	/R/	0.207		/IH/	/IY/	0.230
Indian	ε	/D/	0.179	Spanish	/D/	ε	0.214
	/Z/	/S/	0.169		/T/	ε	0.207
	/R/	ε	0.165		/AH/	/IH/	0.186

Table 7.1: The top 5 highest valued phoneme confusions from each region profile including the target phoneme and substituted phoneme.

7.1.1 American

According to 2018 census data, English is the majority language used throughout America with over 90% of the United States' population reportedly speaking English at home or, where another language is used at home, speaking English very well (Zeigler and Camarota, 2019). Of the non-English languages spoken at home, Spanish is the most common but there are many others including a multitude of Indic languages such as Hindi and Bengali. This is potentially influential in the American speakers observed to group around the Spanish and Indian cluster centres visualised in Chapter 6. The generated region profile is explored under the framework of North American English literature and research. The deletion of /T/ phonemes was found to have the highest value in the American profile. Alveolar plosive elision is a well-documented feature of American English and English in general

and the literature examining its occurrence and source includes that of Labov (1972), Guy (1980), Raymond et al. (2006) and Raymond et al. (2016), and Priva (2015). The absence of /D/ deletion in the five highest valued edit operations is likely due to the comparative prevalence of /T/ which occurs in the text prompts almost twice as often as /D/ and, as such, is observed to be deleted much more often. Another heavily studied aspect of American English pronunciation is that of flapping (or tapping) where, in specific contexts, the /T/ phoneme can be realised as a short voiced flap (Boberg, 2015; Turk, 1992; De Jong, 2011). This voicing likely causes the ASR to misrecognise these instances of /T/ phonemes as being /D/ thus resulting in the /T/ -> /D/ confusion having the second highest value of the American profile. The third highest valued confusion is that of /AH/ -> /IH/. The CMU Pronouncing Dictionary and the grapheme-to-phoneme tool used in this work to obtain phoneme sequences both lack a unique label for the English schwa. Instead the /AH/ phoneme label is most often used where a reduced schwa vowel is likely to occur in a word. However, work by Flemming (2009), has demonstrated that American English varieties have two distinct reduced vowels; the mid-central schwa typically found word finally and a high schwa which occurs word medially. The height of this word medial schwa vowel is the likely source of the /AH/ -> /IH/ confusions. Finally, the fourth and fifth highest values correspond to the insertion of alveolar plosives /D/ and /T/ respectively. Work by Fourakis and Port (1986) has previously discussed the tendency for American English speakers to insert plosives into sonorant-fricative clusters.

7.1.2 British

Across Britain, English is the most commonly spoken language but it is home to other languages including Welsh and Scottish Gaelic (Mac Sıthigh, 2018). There are also many well documented varieties of English used throughout Britain which vary extensively in pronunciation, grammar, and vocabulary (Hughes et al., 2013). The highest valued edit operations of the region profile are discussed here in terms of general British English phonetics and phonology with features specific to certain varieties discussed where relevant. The most prominent feature of the British profile was /T/ deletion. Much like with American English this is a common feature of British English and has been discussed in works such as that of Tagliamonte and Temple (2005), and Baranowski and Turton (2020). The second and third highest values correspond to /R/ deletion and insertion respectively. The English varieties used across much of England and Wales are non-rhotic where only the /R/ sounds occurring before vowels are realised with a rhotic consonant (Ogden, 2017). This likely accounts for the observed deletions due to the ASR being trained on rhotic English varieties. The insertions could also be a result of the non-rhotic Englishes used in Britain since British English often exhibits linking-r (Broadbent, 1991; Lewis, 1975) where word final /R/ sounds are realised when they occur before word initial vowels where they otherwise would not. This tends to be overgeneralised by non-rhotic speakers leading to a phenomenon known

as intrusive-r (Ogden, 2017). Finally, the operations with the fourth and fifth highest values concern the /AH/ vowel, namely /AH/ -> /IH/ confusions and /AH/ deletions respectively, which are attributed to this label being used to denote the schwa vowels. Whilst it is American English varieties which are reported as having two distinct reduced vowel realisations, work by Kondo, 1994 demonstrated that schwa is targetless in terms of F2 meaning a more fronted reduced vowel could lead to the misrecogntion of /IH/ instead. Pretonic schwa elision is also an observed phenomenon in British English and English in general (Glowacka, 2001; Davidson, 2006b) which likely contributes to the observed /AH/ deletions.

7.1.3 Canadian

English and French are the official languages of Canada, although the majority of French speakers are located in and around Quebec with English being the dominant language outside of this province Dollinger, 2019. The Canadian 2016 census also provided data on almost 70 different Aboriginal languages although these are spoken by only a minority of the population (Anderson, 2018). This section focuses on Canadian English for analysing the characteristic features of the region profile. As with the American and British profiles, the highest value in the Canadian profile corresponds to that of /T/ deletion. Indeed, the Canadian profile shares many of the same most prominent features as the American profile including /D/ and /T/ insertion (second and fourth highest values respectively) and the /AH/ -> /IH/ confusion (third highest). These features likely stem from the same sources as with the American profile detailed previously in Section 7.1.1. See Woods (1993) for further details on the similarities between Canadian English and American English. A final confusion unique to the Canadian profile is the fifth highest value /IH/ -> /AH/. This is likely a result of a pronunciation phenomenon known as the Canadian Shift where the short front vowels are lowered and retracted (Boberg, 2008; Sadlier-Brown and Tamminga, 2008). It is likely that this effect is seen more prominently with the /IH/ vowels than /EH/ or /AE/ (which also undergo this shift) due to its higher frequency within the text prompts where occurrences of /IH/ are greater than /EH/ and /AE/ combined.

7.1.4 Chinese

Approximately 91% of China's population speak Sinitic languages, the largest branch of which being Mandarin dialects spoken by around 67% of the population (Chappell and Lan, 2016). Mandarin, specifically that based on the Beijing variety, is also the official language of China (Weng, 2018) and thus it is used as a basis for the following region profile investigation. /AH/ insertion has the highest value in the Chinese region profile. Yang et al. (2021) notes that vowel epenthesis is a feature common in the English productions of speakers from Central and Northern China and the Yunnan province. Vowel insertions, and specifically schwa insertions, typically occur with consonant patterns that do not exist in

Mandarin which does not have consonant clusters and allows for only vowels or nasals in syllable-final positions (Siqi and Sewell, 2012). For example, Broselow et al. (1998) observes vowel epenthesis following word-final obstruents. Similarly, Nogita and Fan (2012) suggests that the schwa-like insertions in English consonant clusters produced by Mandarin speakers are intrusions stemming from gestural mistiming whereby the physical production of the consonants do not overlap and result in an inserted schwa-like sound (Davidson, 2006a). /L/ deletion, the second highest value, often occurs in the English productions of Mandarin speakers when it occurs in syllable final position following a back vowel (He, 2014). Again this is not a phoneme that can occur in this position according to Mandarin phonology. Unexpectedly, the third highest valued confusion was that of /S/ -> /Z/. /Z/ does not exist in the Mandarin phonological inventory and a number of studies observe the reverse pattern in Mandarin Accented English (/Z/ -> /S/). It is hypothesised that this confusion arises from /S/ sounds in non-words denoted by the letter 's' in the ASR output being converted to /Z/ phonemes by the grapheme-to-phoneme tool. For example, the word 'deepest' was recognised as 'dipis' and the target /S/ sound was produced as expected but the predicted phoneme sequence for 'dipis' was /D IH P IY Z/. Analysis of a sample of the prompts and ASR outputs confirmed this typically happens with word final /S/ non-words. Finally, the fourth and fifth highest values correspond to the confusions /EH/ -> /AE/ and /IH/ -> /IY/. Mandarin does not have a phonemic tense/lax distinction and studies by Huang and Pickering (2014) and Khanal et al. (2021) demonstrate that Mandarin speakers tend to use these vowel pairs interchangeably.

7.1.5 Indian

India is a very linguistically diverse country with 22 officially recognised languages, and hundreds of other languages, dialects, and 'mother tongues' as they are termed in census data (Bhattacharya, 2017; Mohanty, 2006). English is often used in education and government throughout the country and is labelled the *associate official language* (Mukherjee and Bernaisch, 2020). Indian English is considered the third largest English variety by number of speakers although there are technically many differing varieties of Indian English which vary significantly based on regional differences (Mukherjee and Bernaisch, 2020; Gargesh, 2008). Despite this, research into varieties of Indian English has revealed a number of common features which serve to distinguish the 'general' Indian English variety (Sailaja, 2012; Trudgill and Hannah, 2017). In particular, the realisations of /T/ and /D/ are typically the retroflex [ʈ] and [ɖ] respectively for all Indian English speakers. This, coupled with the tendency to produce syllable-initial voiceless plosives without aspiration (Gargesh, 2008; Sailaja, 2012), is likely the reason for the highest value of the Indian profile being the /T/->/D/ confusion. The question of whether or not Indian English is rhotic or not would appear to be the subject of debate. Gargesh (2008) and Nihalani et al. (1979) argue that Indian English is rhotic, whilst Sailaja (2012) and Trudgill and Hannah (2017) describe it as non-

rhotic. Sahgal and Agnihotri (1988) and Bansal (1990) describe rhoticity as variable across speakers and influenced by sociolinguistic factors with a lower rate of rhotic productions correlating with education and formality. A degree of non-rhoticity would, however, explain the /R/ insertions and deletions, the second and fifth highest valued confusions, for the same reasons they appear as characteristic features of the British English profile (see Section 7.1.2). Regardless, it is generally agreed that /R/ realisations in Indian English can be approximants, trills, or flaps (Fuchs, 2016). The flapped realisation would be similar to the American English realisation of /T/ (which the ASR often mistakes for an instance of /D/). As a result, the third highest of Indian English, associated with /D/ insertion, is likely the result of /R/ realisations being mistaken for /D/ instances and the two phonemes being too dissimilar to be considered a substitution during alignment. Wiltshire (2006) examines the productions of word final consonants and consonant clusters in Indian English for L2 English speakers with different L1s. They note that for speakers of an L1 which does not allow for voiced obstruents in coda position, it is common for devoicing to occur. In particular, the devoicing of fricatives is more common than plosives and it is noted that the plural marker in English is often produced as [s] in all contexts by many Indian English speakers. This is the most likely cause of the /Z/->/S/ confusion being the fourth highest valued in the Indian profile.

7.1.6 Japanese

In Japan, Modern Japanese (specifically the Tokyo variant) is used throughout the country as a standardised variety for writing and formal speaking situations (Kubozono, 2015; Gottlieb, 2005). The spoken varieties used by the majority of people in Japan are typically defined as varieties of Japanese although there is a smaller minority who use varieties of other languages including Ryūkyūan, a sister branch to Japanese in the Japonic family tree (Boer and Robbeets, 2020), and Ainu, classified as a language isolate (Tranter, 2012). In this section the analysis is constrained to examining the influence of general Japanese phonetics and phonology and literature describing the English of Japanese speakers. The 5 most characteristic substitutions of the Japanese profile can be separated into two main features. The first involves the liquid sounds /R/ and /L/. The /L/ -> /R/ substitution has the highest value in this profile, with /L/ deletion and /R/ deletion being the third and fifth highest respectively. Japanese phonology does not distinguish between the /R/ and /L/ sounds found in English and, rather, has a single /R/ phoneme that is typically realised as an apico-alveolar tap (Kubozono, 2015). Indeed, many studies, for instance those of Miyawaki et al. (1975), Sheldon and Strange (1982), and Strange and Dittmann (1984), have noted the difficulties faced by Japanese speakers in the perception and production of the English /R/-/L/ distinction. Yoko et al. (2019) describes producing the two liquid sounds to be the most challenging for Japanese speakers of English and observes a tendency for speakers to substitute the Japanese tapped consonant with which they are more familiar. It is likely,

therefore, that this particular realisation is being recognised by the ASR as an /R/ phoneme and thus resulting in the /L/ -> /R/ substitution. Further, the /R/ and /L/ deletions likely stem from the frequency with which Japanese /R/ is the target of assimilation (Labrune, 2014). Indeed, many of the deletions occur in consonant clusters where it would appear the liquid has taken on the characteristics of the neighbouring consonant to the point of apparent deletion. The other feature of this variety which features strongly in the region profile is the tendency to substitute the lax vowel /AH/ with a tense counterpart /AA/ or /AE/ (the second and fourth highest valued substitutes respectively). As noted previously, the /AH/ label is used for both the lax near-low central vowel and the schwa - neither of which exist in the phonological inventory of Japanese. Ohata (2004) suggests that Japanese learners of English are likely to use /AH/, /AE/, and /AA/ interchangeably as a result of Japanese having one low vowel produced centrally. In addition, studies like that of Kondo (2000) and Lee et al. (2006) demonstrated language transfer from Japanese onto productions of English schwa, with late bilinguals having not acquired the reduced vowel and less fluent English speakers producing a schwa with F2 values not significantly different from the distribution of F2 values of Japanese /AA/.

7.1.7 Korean

Korean is used across Korea and, despite there being a number of Korean dialects based on geography, it is considered to be relatively homogeneous in that the dialects are mutually intelligible (Yeon, 2012). As such the Korean region profile is analysed in terms of the phonological inventory of contemporary Korean and studies into native Korean speakers of English. The most characteristic feature of the Korean profile is the /IH/ -> /IY/ confusion. The Korean vowel inventory does not have a tense/lax distinction and the vowel length distinction is being lost (Yu Cho and Iverson, 1997). In their work with Korean students of English, Choi (2007) observes difficulties in distinguishing between the /IH/ and /IY/ sounds of English due to this primary interference coupled with the secondary interference where the transcription model of English into Korean results in the same vowel sound being denoted and produced. Similarly, they also note difficulties in discriminating between /EH/ and /AE/ which Yu Cho and Iverson (1997) describe as a near-merger in Korean. This is likely the source of the fifth highest valued confusion /EH/ -> /AE/. Cho, 2016 notes that one of the most prominent features of Korean phonology is consonantal assimilation and that Korean /T/ has ‘special status’ since it assimilates completely to a sonorant following it, and takes on place features of a following obstruent. It is also deleted in consonant clusters. These processes in Korean might transfer to English productions for Korean speakers thus resulting in deletions of /T/ being the second highest value of the Korean profile. The other characteristic features involve the voiced fricatives /Z/ and /DH/, neither of which occur in the Korean phonological inventory (Brown and Yeon, 2015). The /Z/ -> /S/ confusion has the third highest value despite Choi (2007) stating that students tended to have no real difficulty

in discriminating between the two sounds. However, they do note that poorer discrimination performance was observed when the sounds occurred word finally - a finding supported by Major and Faudree (1996). Analysis of a sample of the Korean ASR outputs and text prompts confirmed this effect with the majority of /Z/ -> /S/ confusions occurring in word final positions. The same studies also describe difficulties with producing /DH/ which might account for the fourth highest value corresponding to the /DH/ -> /D/ confusion. (Choi, 2007) reports considerable difficulty amongst students in discriminating between /DH/ and /D/ and Major and Faudree (1996) notes the difficulty in acquiring /DH/ compared with other phonemes.

7.1.8 Portuguese

Portuguese is the official language of Portugal although the lesser spoken Mirandes is also recognised as a co-official regional and minority language (Marques, 2021; Ferreira, 2013). This section focuses on European Portuguese and specifically the Lisbon variety which is used as a standard variety in mainland Portugal (Cruz-Ferreira, 1995). The highest valued confusion in the Portuguese profile is that of /IH/->/IY/. This phonemic distinction does not exist in European Portuguese and studies into the acquisition of this contrast have shown that, without specific perceptual training, the English /IH/ phoneme is typically assimilated into the Portuguese /IY/ category with overlap in the productions of /IH/ and /IY/ (Rato and Rauber, 2015; Rato et al., 2014). The reduced schwa vowel in European Portuguese is typically realised in a higher position than in other Englishes and is transcribed as [ɨ] (Veloso, 2007; Vale and Perpétua, 2020). This would appear to result in the /AH/->/IH/ confusion having the second highest value since, as discussed previously, the /AH/ category is used for both the low central vowel and the reduced schwa vowel. The fourth and fifth highest values in the Portuguese profile correspond to /Z/->/S/ and /D/->/T/ respectively. This is likely due to consonant devoicing which is common for both plosives and fricatives in European Portuguese (Pape and Jesus, 2015; Cruz-Ferreira, 1995; Jesus and Shadle, 2002). Indeed, the devoicing of plosive consonants is a distinguishing feature of Portuguese over other Romance languages like Spanish or Italian which tend to maintain voicing throughout the duration of the plosive production (Pape and Jesus, 2011). An unexpected confusion with the third highest value in the Portuguese region profile is that of /S/->/Z/. The phonological inventory of European Portuguese features both the voiced and voiceless alveolar fricatives and, as previously discussed, devoicing of fricative consonants is common. As such, it would not be expected to find the voicing of /S/ to /Z/ to be a characteristic feature of Portuguese Accented English. It is thus hypothesised that, similarly to the Chinese profile discussed in Section 7.1.4, this is most likely the result of the grapheme-to-phoneme tool incorrectly converting orthographic word-final ‘s’ to the /Z/ phoneme.

7.1.9 Russian

The official national language of Russia is Russian (Chevalier, 2006) and, whilst 277 languages are spoken in Russia according to the 2010 census data, 90% of Russian citizens speak Russian as their native language (Mustajoki et al., 2021). As such, this analysis is constrained to the pronunciation features seen in Russian English speakers and the potential transference effects of the Russian language onto English productions. The most characteristic confusion of the Russian profile is that of /IH/->/IY/. This is not unexpected since this contrast does not exist in Russian. Shafiro and Kharkhurin (2009) demonstrate that Russian-English bilinguals showed reduced accuracy in identifying vowel contrasts which are absent in Russian - one of which being the /IH/-/IY/ contrast. Furthermore, a lack of differentiation in vowel length where it would normally occur in English between tense and lax vowels is found in the productions of Russian speakers (Bondarenko, 2014; Jureková, 2015). The second and third highest values in the Russian region profile would appear to be linked by Russian word final devoicing. Both /Z/->/S/ and /D/->/T/ are substitutions which were observed in the English productions of Russian English as a Foreign Language (EFL) learners by Bondarenko (2014) stemming from the absence of word-final voiced obstruents in Russian (Yanushevskaya and Bunčić, 2015). Indeed, both Tumshevits (2019) and Jureková (2015) note the characteristic trait of consonant devoicing in Russian. Again, the /S/->/Z/ confusion arises unexpectedly in the highest valued substitutions of a region profile. The Russian phonological system has both /S/ and /Z/ and, as previously mentioned, it is devoicing which is characteristic of Russian and Russian English. Thus, the same grapheme-to-phoneme effect discussed previously in Sections 7.1.4 and 7.1.8 is surmised to be the cause. The final confusion relates to the /AH/->/IH/ substitution. This most likely stems from word stress effects and vowel reduction processes. Since Russian does not have secondary and tertiary word stress in the way English does, Vishnevskaya (2011) notes the tendency for Russian speakers of English to overstress vowels which would typically be reduced. This could potentially result in expected English schwa vowels (labelled as /AH/) to be produced with more stress and recognised as another vowel (in this case /IH/). Furthermore, vowel reduction in Russian typically leads to the raising of low vowels, sometimes to the degree of merging with [i] (Barnes, 2007). This vowel raising might transfer to English productions resulting in a much higher realisation of reduced schwa vowels than is expected by the ASR system.

7.1.10 Spanish

Castilian Spanish (henceforth referred to as Spanish) is the official language of Spain although four other co-official languages are recognised (Aranese, Basque, Catalan, and Galician) (Cassany, 2005). These languages also have numerous varieties that some argue constitute distinct languages of their own (Pou, 2004). Nevertheless the following analysis

focuses primarily on the Spanish language and Spanish Accented English. The highest valued substitution seen in the Spanish region profile is /Z/->/S/. In Spanish, voiced fricatives are typically allophones of their voiceless counterparts and [z] is an allophone of /S/ typically occurring before voiced consonants (Anderson and Centeno, 2007; Gómez González and Sánchez Roura, 2016). This is also in agreement with work by You et al. (2005), who found the /Z/->/S/ substitution to have the highest likelihood in an analysis of Spanish Accented English produced by children. The /IH/->/IY/ confusion, the second highest valued, is also a well established feature of Spanish Accented English due to the lack of /IH/ in Spanish phonology (Anderson and Centeno, 2007). Gómez González and Sánchez Roura (2016) notes that /IH/ is typically assimilated to Spanish /IY/ both because of the phonetic similarity of their realisations and their shared orthography since both sounds in English are typically represented by the letter 'i'. The third and fourth highest values pertain to the deletion of the alveolar plosives /D/ and /T/ respectively. As has been discussed previously, word-final plosive deletion is common in many English varieties and the majority of these deletions, from the sample of utterances analysed, occur in word-final position. Flege and Davidian (1984) also found high rates of word-final plosives in Spanish speakers although participants were primarily Mexican Spanish speakers which has more constraints over word-final consonants. Finally, the investigation by You et al. (2005) also notes the large variation in the production of /AH/ in Spanish Accented English likely as a result of its nonexistence in Spanish phonology. Whilst they found the most likely substitutes to be /EH/ or even /UW/, it is evident from these substitutions of more high vowels that some degree of raising happens which likely prompts the fifth highest valued confusion /AH/->/IH/.

7.2 ASR Sensitivity to Phonetic Variation in L2 Englishes

In this section, the speaker modelling approach detailed in Chapter 6 is applied to a second corpus. Whilst the L2-Arctic corpus has significantly fewer speakers, a subset of the audio recordings include manual annotations at the phoneme level for insertions, deletions, and substitutions. This makes it well-suited for comparison with the observed confusions in the ASR output. Speakers with the same L1 are shown to produce similar representations which, again, appear to cluster in the multi-dimensional space. Furthermore, by examining the most common substitutes of poorly recognised phonemes, it is demonstrated that the behaviour of the ASR typically aligns with the judgements of human annotators. However, it is noted that the level of agreement and the rates of recognition can vary across the different spoken varieties.

Speaker	L1	No. of utterances	Minutes of speech
ABA	Arabic	150	10.32
SKA	Arabic	150	8.21
YBAA	Arabic	149	9.10
ZHAA	Arabic	150	8.38
BWC	Mandarin	150	10.62
LXC	Mandarin	150	9.64
NCC	Mandarin	150	9.35
TXHC	Mandarin	150	8.73
ASI	Hindi	150	7.51
RRBI	Hindi	150	8.69
SVBI	Hindi	150	7.06
TNI	Hindi	150	8.32
HJK	Korean	150	7.54
HKK	Korean	150	8.91
YDCK	Korean	150	10.17
YKWK	Korean	150	8.81
EBVS	Spanish	150	9.91
ERMS	Spanish	150	10.44
MBMPS	Spanish	150	11.52
NJS	Spanish	150	8.09
HQTV	Vietnamese	150	9.20
PNV	Vietnamese	150	9.53
THV	Vietnamese	150	9.44
TLV	Vietnamese	150	10.32
Total		3599	219.81

Table 7.2: A summary of the speakers included in the annotated subset of the L2-Arctic Corpus including the L1 of the speaker, number of utterances annotated and total number of minutes of speech.

7.2.1 Speaker Modelling

The audio data used is taken from the L2-Arctic Corpus (Zhao et al., 2018). This corpus contains recordings of 24 non-native speakers of English across six L1s, namely Arabic, Hindi, Korean, Mandarin, Spanish, and Vietnamese. For each L1 there are four speakers. Whilst each speaker recorded approximately one hour of read speech using the Arctic prompts (Kominek and Black, 2004), this work focuses on the subset of recordings which are manually annotated at the phonemic level for insertions, deletions, and substitutions. The number of utterances and total time recorded for each speaker is given in Table 7.2.

Again, the transformer-based wav2vec 2.0 model (Baevski et al., 2020) was used as the ASR component and, from a cursory examination, the outputs appeared capture the phonetic properties of the speech. For example, the prompt text;

"It was simple in its way and no virtue of his"

was read by an L1 Mandarin speaker and produced the ASR transcript;

"It was simbol in its way and a no virtue of ease"

This would appear to capture common phonetic features of Mandarin Accented English including vowel epenthesis and difficulties with the tense/lax distinction (Siqi and Sewell, 2012; Broselow et al., 1998; Nogita and Fan, 2012; Huang and Pickering, 2014; Khanal et al., 2021).

As before, the prompt texts and ASR transcripts were converted to their corresponding phoneme sequences in ARPAbet notation using the CMU Pronouncing Dictionary (Weide, 1998) and, for non-words, the grapheme-to-phoneme tool trained on this dictionary (CMUSphinx, 2016). This method is beneficial for two reasons. Firstly, the annotated portion of the L2-Arctic corpus also uses phoneme sequences from the CMU Pronunciation Dictionary and so direct comparisons are possible in the forthcoming analysis. Furthermore, since the pronunciations in this dictionary are typically suited to North American varieties of English, they capture the likely pronunciation of the speech which the ASR model was trained on. Thus, whilst these phoneme sequences for the words in the prompt text are treated as a 'ground truth' when investigating phoneme insertions, deletions, and substitutions, it is not suggested that these are the 'correct' pronunciations but, rather, the pronunciations which are 'expected' by the ASR. Phoneme sequences are, again, aligned using a weighted edit distance algorithm whereby more likely operations carried a lower cost during alignment and were thus preferred. The weights are taken from the phoneme similarity matrix constructed in Chapter 4. After alignment, counts of the observed edit operations were stored in a confusion matrix for each speaker where, as with previously discussed matrices, insertions and deletions are treated as substitutions with the empty string (ε). A repository storing the resultant confusion matrix for each speaker in the L2-Arctic Corpus can be found in Appendix A.1.3.

The confusion matrices for each of the 24 speakers were then reshaped into 1600 dimensional vectors and, as in the previous chapter, a k-means clustering approach (MacQueen, 1967) was used to determine whether speakers of the same L1 were typically located in the same area in the multi-dimensional space. In this case, 6 clusters are used to match the number of distinct L1s in the corpus. Whilst the clustering is carried out in 1600 dimensions, t-SNE (Van der Maaten and Hinton, 2008) is, again, used to visualise the speaker representations and cluster centers in 2 dimensions. This can be seen in Figure 7.1 which demonstrates how speakers of the same L1 tend to be located around each other in the space and so presumably have similar vector representations resulting from similar confusion matrices.

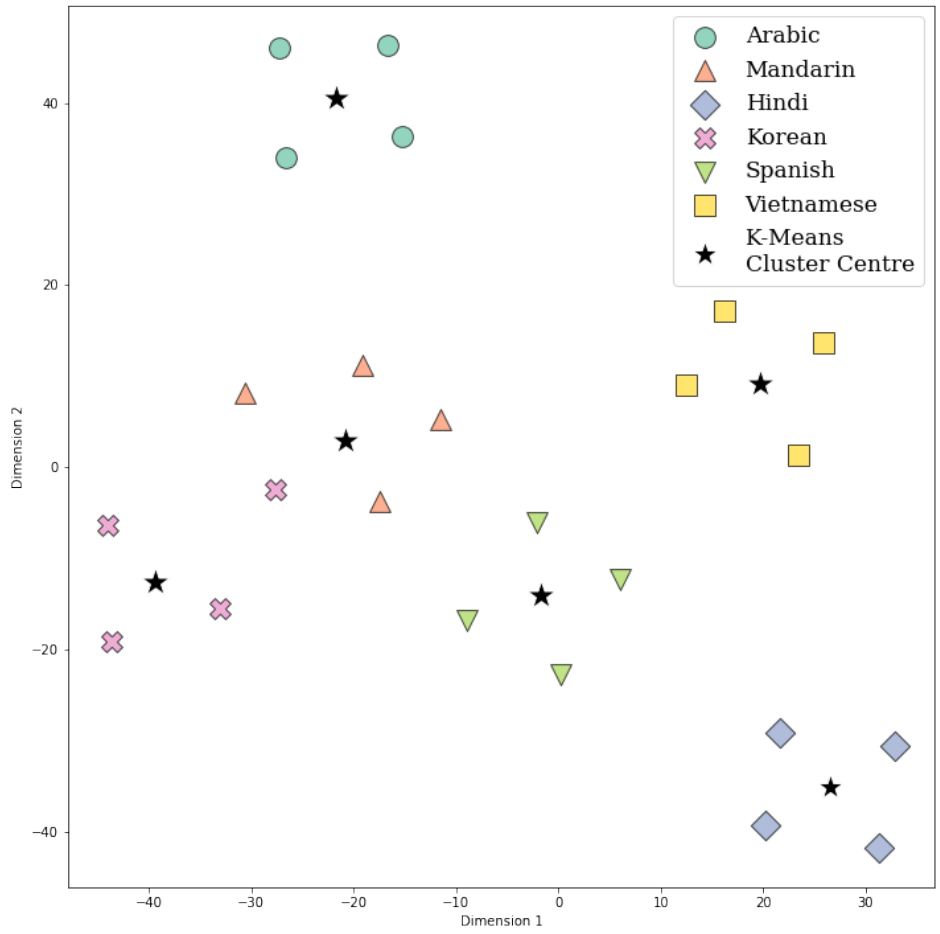


Figure 7.1: A visualisation of the L2-Arctic speaker vectors and the predicted cluster centres from K-means clustering after tSNE dimensionality reduction.

7.2.2 Comparison with Human Annotator Judgements

Having established that the ASR output produces patterns of errors in relation to systematic pronunciation variation across speakers with similar varieties, a more fine-grained analysis of its performance with specific phonetic variation is carried out through comparison with human annotator judgements. As discussed previously, the audio data used to construct the speaker level confusion matrices was the subset of the L2-Arctic corpus that was manually annotated at the phoneme level for substitutions, insertions, and deletions. Thus, the frequency with which a phoneme is realised in such a way that the ASR produces erroneous output can be compared with the frequency with which a human annotator believes the phoneme to have been substituted for another. Furthermore, the specific phoneme substitutions, as determined by the ASR output, can be compared with the phoneme substitutions identified by the annotator to determine whether the ASR output captures the same judgements as made by a human listener. A summary of this comparison is given in Table 7.3.

L1	Target Phoneme	Recognition Rate (ASR)	Recognition Rate (HA)	Most Common Substitute (ASR)	Most Common Substitute (HA)	MCS Substitution Rate (ASR)	MCS Substitution Rate (HA)
Arabic	/EH/	74.8%	89.0%	/IH/	/IH/	13.9%	4.1%
	/P/	76.0%	71.4%	/B/	/B/	21.6%	25.8%
	/TH/	79.2%	81.0%	/S/	/S/	7.5%	13.1%
Mandarin	/TH/	70.2%	50.9%	/S/	/S/	12.9%	37.9%
	/ZH/	70.8%	62.5%	/SH/	/SH/	8.3%	29.2%
	/UH/	71.7%	89.3%	/UW/	/UW/	5.0%	6.6%
Hindi	/OY/	57.5%	95.0%	/AY/	/AY/	25.0%	5.0%
	/AW/	69.7%	79.3%	/OW/	/AO/	15.9%	8.6%
	/TH/	71.8%	41.4%	/T/	/T/	21.1%	45.9%
Korean	/ZH/	60.0%	47.4%	/Z/	/JH/	15.0%	26.3%
	/TH/	75.0%	75.4%	/S/	/S/	12.1%	15.3%
	/JH/	78.6%	72.3%	/Z/	/CH/	7.6%	10.0%
Spanish	/UH/	72.0%	85.0%	/UW/	/UW/	9.0%	9.3%
	/TH/	74.1%	39.7%	/T/	/T/	16.7%	44.0%
	/ZH/	75.0%	20.0%	/SH/	/SH/	20.0%	75.0%
Vietnamese	/ZH/	54.2%	37.5%	/S/	/S/	25.0%	33.3%
	/JH/	55.6%	32.3%	/CH/	/CH/	10.5%	17.7%
	/AA/	58.3%	70.5%	/ER/	/AO/	11.9%	11.4%

Table 7.3: A summary of the top 3 most commonly misrecognised phonemes by the ASR system for each region. The recognition rate of both the ASR and the Human Annotators (HAs) is given, alongside the most commonly identified substitute for each target phone and the rate of which this substitution was identified by both the ASR and the HAs.

The three most frequently misrecognised phonemes for each L1, based on the ASR output, are given in Column 2. Columns 3 and 4 give the recognition rate of each phoneme according to the ASR output (ASR) and the human annotator judgements (HA) respectively. The recognition rate is given as the percentage of occurrences of each phoneme in the prompt texts which were matched in the ASR output or were marked as correct by the human annotator. It can be seen here how the behaviour of the ASR compares to that of the annotators. There are some instances of very similar recognition rates, for example Arabic /TH/ or Korean /TH/. However, more importantly, there are cases where the ASR is either more robust or more sensitive to pronunciation variation compared with human judgements. For instance, Spanish /TH/ is recognised by the ASR 74.1% of the time whilst the human annotator labels it as correctly produced in only 39.7% of occurrences. This would suggest that the ASR is relatively robust to the production of /TH/ by Spanish L1 speakers and, whilst its realisation may be considered more /S/ like, it does not have as significant an impact on the recognition accuracy as might be expected. Conversely, consider the case of Hindi /OY/ which the ASR correctly recognises only 57.5% of the time whilst the human annotator marked 95% of occurrences as correct. Evidently, the realisation of the /OY/ phoneme by Hindi L1 speakers was not considered different enough from the “canonical" production as to be labelled a substitution by the human annotator but yet it caused problems with the ASR and contributed to a higher WER for speakers with this L1.

In the majority of cases, the most commonly substituted phoneme as determined by the ASR output (given in Column 5) matches that identified by the human annotator (given in Column 6). This supports the idea that the behaviour of the ASR is in line with the judgements of a human listener and lends credence to the possibility of leveraging erroneous ASR output to investigate patterns of variation in specific spoken varieties. However, the substitution rates of these Most Common Substitutes (MCS) can vary across L1s and across target phonemes. The substitution rate is given as the percentage of target phoneme occurrences in the prompt text which were recognised as the specific substitute by the ASR or were noted as a substitution with the specific substitute by the human annotator. As would be expected, the substitution rates between the ASR and human annotators have a high degree of variability when the target phoneme recognition rate varies between the two. However, there are also cases where the target phoneme recognition rate is relatively similar from both the ASR and human annotator, but where the substitution rate of the most common substitute differs to a larger extent. For example, the Arabic /TH/ has a recognition rate of 79.2% by the ASR and 81.0% by the human annotator and both find /S/ to be the most common substitute. However, the ASR detects this substitution in 7.5% of /TH/ instances whilst the human annotator notes it in 13.1% of occurrences. Thus, whilst evaluating an ASR component in this way can reveal a lot of information about its weaknesses and highlight common patterns of phonetic variation across groups of speakers, it does not supply the same phonemic annotation as a human annotator would.

7.3 Applications of Model Interpretations

The speaker level confusion matrices reveal features of variation in the pronunciation of a single speaker, highlighting weaknesses of the ASR and gaps in its training data. For example, Figure 7.2 depicts one such confusion matrix (again visualised as a heatmap for clarity) for an L1 Hindi speaker where darker cells indicate higher counts of a particular substitution. Here, the highest value and most common substitution across all of the read prompts from this speaker is in the case of an expected /T/ phoneme being represented in the ASR transcript as a /D/ phoneme. This observed substitution stems from the speaker's realisation of /T/ phonemes as either the retroflex [ʈ] or unaspirated [t] in word initial position. Since the ASR model is unfamiliar with such productions, it typically mistakes them for instances of the /D/ phoneme instead thus resulting in high substitution counts and a deterioration of the word error rate.

It can be surmised, from this example, that more instances of retroflex [ʈ] in the training and fine-tuning data would improve the performance of the ASR for that speaker. More significantly, since it has been established that speakers of similar varieties exhibit similar patterns of pronunciation variation, and that the ASR performs similarly across these speakers, this addition to the training data would improve the recognition performance for

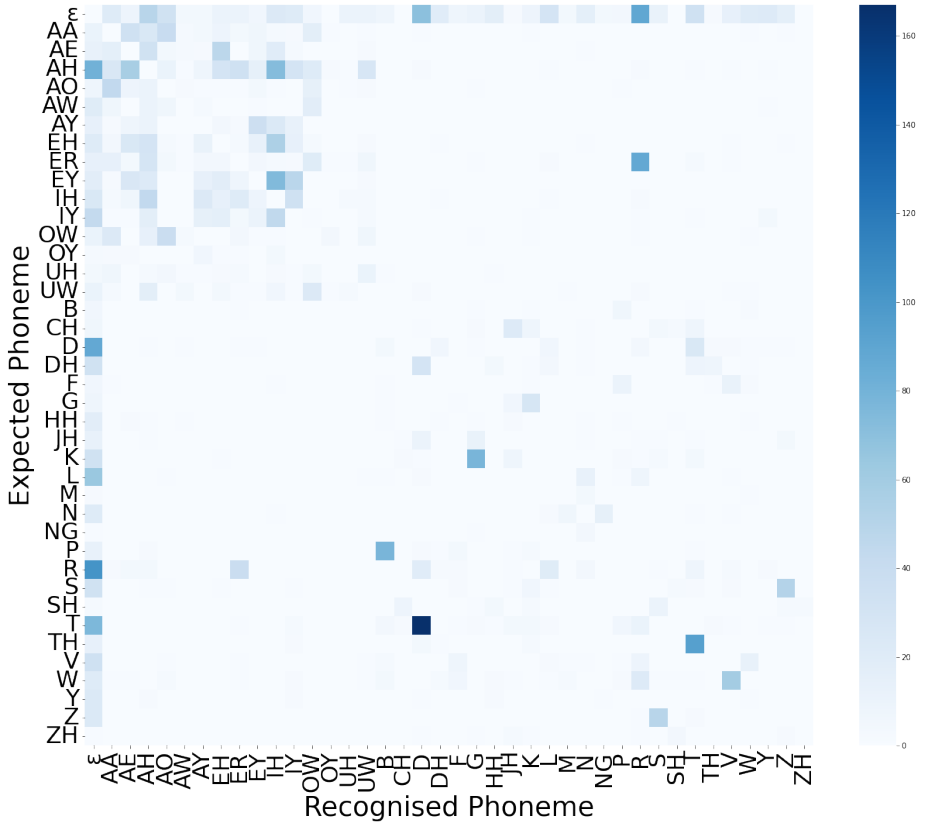


Figure 7.2: A phoneme confusion matrix generated from the ASR transcripts from a single L1 Hindi speaker visualised as a heatmap.

a much larger group of speakers. Thus, this error-based analysis is useful for developing ASR systems and improving their performance with specific spoken varieties. It enables specific phonetic realisations which cause issues during recognition to be pinpointed. With this information, a targeted approach can be applied to the collection of additional training material, perhaps using prompts specifically designed to elicit those productions which the system is currently unable to handle correctly. Since data collection and the required manual annotation is often time, money and resource consuming, the ability to carry out a smaller scale and more focused collection whilst still improving recognition accuracy is a huge benefit. As previously established, similar error patterns are produced for speakers with similar varieties. Thus, improving the ASR performance for a small subset of test speakers will likely result in improvements for a much greater number of speakers who exhibit similar pronunciation features. Knowledge of the sensitivity or robustness of an ASR system to specific variant pronunciations is vital in developing systems capable of accurately recognising underrepresented and minority varieties.

Since this sensitivity investigation into the performance of a specific ASR system reveals a lot about how pronunciation variation is handled for different spoken varieties, it is useful in the context of incorporating ASR technology into the annotation process for linguistic

study. As the use of ASR technology becomes more popular for generating automatic word-level transcriptions of audio data, it is important to understand the limitations of such an approach. Different ASR systems will be trained on different data and will perform differently for different varieties. By testing a specific ASR component on a specific spoken variety or varieties, a better understanding of the expected recognition accuracy can be obtained. ASR systems can be chosen based on their appropriateness for the task and their robustness to style of speech being transcribed. Furthermore, human annotators can be informed of the expected weaknesses of the ASR and the types of errors which are likely to appear within the transcripts thus aiding in the manual correction of automatically generated transcripts.

From a phonetic variationist approach, the use of a highly sensitive ASR component can reveal systematic pronunciation variation through the patterns of errors which occur in the automatic transcripts. It has been demonstrated that the erroneous output is not random but, in fact, captures commonalities between speakers with similar spoken varieties. Moreover, they are typically aligned with the judgements of human listeners, featuring the same phoneme-level substitutions as identified by the annotator. As such, the investigation presented in this chapter can aid in highlighting and discovering variation at the phonetic level across speakers from larger and more general accent groups to more specific fine-grained ones depending upon the spoken varieties of the test speakers. Furthermore, the comparison between phonetic variants captured by the ASR output and those detected by human annotators can inform researchers about the capabilities of a particular ASR component in accurately detecting such variants. This is significant information to obtain before incorporating ASR as a tool in variation analysis and annotation.

7.4 Limitations

Whilst the interpretation of the confusion based models reveal valuable information regarding the pronunciation variation of speakers and the sensitivity of ASR systems to such variation, the approach is not without its limitations. Firstly, in analysing the region profiles, the highest valued confusions were explained in terms of notable phonetic features of the variety as described in existing literature. However, to undoubtedly confirm that such confusions were the direct result of these variant productions, a more detailed analysis of the audio recordings at the acoustic level would be required. This was, unfortunately, outwith the scope of this work and only a sample of audio files were examined by ear in order to confirm the variant productions discussed.

In addition, the one-to-one mapping of expected versus recognised phonemes is not able to capture contextually dependent variation. Whilst the simplicity and interpretability of these confusion models are seen as advantageous for analysis purposes, some phonetic features of a variety occur only in specific environments. For instance, the vowel epenthesis which is

characteristic of Mandarin Accented English is prompted by consonant patterns in English which do not exist in Mandarin. The region profile of Mandarin alone is not sufficient to convey this constraint and an examination of the ASR outputs was required.

Finally, it is proposed that through analysis of these models and the identification of pronunciation features which inhibit ASR recognition, a targeted approach could be applied to the collection and fine-tuning of an ASR system so as to improve the performance for a specific variety. Whilst this would seem to be a natural conclusion, experimentation using this method would be required to confirm its benefit and gauge the level of recognition accuracy improvement.

7.5 Summary

This chapter sought to answer Research Question 5: *How does the information captured by these models compare to existing knowledge of a spoken variety and what can they tell us about the sensitivity or robustness of ASR systems to pronunciation variation?* The interpretability of the region profiles of Chapter 6 allowed for an inspection of the characteristic features of each region in relation to existing literature on phonetic variation. Through the qualitative analysis of the highest valued phoneme confusions of each region profile, it was demonstrated that these confusions typically aligned with documented features of the English variety of the region and the language transference effects from other languages spoken in the region. The ASR outputs used in the construction of these models were also shown to match well with the judgements of human annotators. As such the models present a useful method of investigating the sensitivity of an ASR system to specific variant pronunciations. This has potential applications both in terms of improving ASR through the targeted collection of additional data for fine-tuning, and for exploring the use of ASR as a tool in the annotation process of linguistic research.

Chapter 8

Conclusions and Future Work

This chapter reflects upon how the thesis addresses the five central research questions as detailed in Chapter 1. The main contributions of each chapter are discussed and the particular novelties of the research are highlighted. Whilst the limitations of the various approaches were detailed in the relevant chapters, they are summarised here as the basis for what the future directions of the research may be.

8.1 Research Question 1

What factors influence a speaker's intuitions regarding phoneme similarity?

In Chapter 3, having established that elements of speaker perceptions of phoneme similarity and confusability cannot be fully captured by phonological features alone, it was hypothesised that the distribution of phonemes in everyday use might influence an individual's intuitions. This was prompted by the fact that a speaker's mental representation of a phoneme category is malleable and specific internal phoneme representations are adapted based on exposure and context. Using a phoneme embedding approach, vector representations of English phonemes were constructed based on the syllabic environments in which they occur. By measuring the distance between these vectors, the distributional similarity between phonemes could be analysed. It was then demonstrated that particular aspects of perceptual similarity are likely influenced by the distributional properties of phonemes captured by the generated phoneme embeddings. Thus, initial evidence of the influence of phoneme context and exposure was presented by modelling distributional similarity in an entirely data-driven manner. Continuing with this bottom up approach to modelling phoneme similarity, the phoneme embedding information was combined with ASR confusability data to generate a phoneme similarity matrix based on both the distribution of phonemes and the acoustic properties of their realisations. This similarity matrix is a foundational component of the thesis from which the following chapters stem.

The comparison between the different models of similarity was limited in that perceptual similarity was modelled based on a single confusability experiment. For a more thorough investigation into speaker intuitions regarding similarity, a larger number of empirical studies could be analysed for common elements. Indeed, confusability may not be the best method for measuring an individual's judgement of similarity and other approaches might result in different similarity hierarchies. Such experimentation could also be used to test the predictive power of the generated phoneme similarity matrix. Whilst its incorporation in later chapters

suggests that the similarity matrix captures useful insights into phoneme similarity, it is likely not the optimal representation and further testing against large perceptual studies could reveal its strengths and weaknesses. Finally, the word2vec algorithm was adapted for the purpose of generating phoneme embeddings primarily because of its popularity in the field at the time. However, as has been discussed, the inability for this approach to fully utilise the phoneme sequences and, in particular, the specific ordering of phonemes, is a notable limitation. Since this work was carried out, transformer based model architectures which make use of self attention have become widespread in human language technologies. It is likely that the encoder component of such a model could generate phoneme representations whilst making full use of the environmental constraints which stem from English phonotactics. The investigation of such an approach presents an interesting possibility for future research.

8.2 Research Question 2

Can a model of phoneme similarity be applied to the development of language technologies; specifically in a spelling correction tool for children?

In chapter 4, a potential use case of the phoneme similarity matrix was explored to discover whether such a model held any benefits for the development of language technologies. In order to automatically correct a child's phonetic misspelling, potential corrections must be ranked in terms of how similar they are to the error, not in orthographic terms, but based on phonological similarity. As such, this was an excellent framework within which to test the phoneme similarity matrix. The similarity values in the matrix were incorporated as phoneme substitution costs and values for insertions and deletions were also added as substitutions with the empty string. This allowed for the comparison of the predicted phoneme sequence of a misspelling with that of a potential correction. Through an analysis of a number of spelling correction tools and misspelling corpora containing differing proportions of phonetic type errors, it was shown that this phoneme similarity based approach was capable of correcting misspellings that other methods could not. Additionally, the proportion of spelling errors corrected by this method alone correlated with the expected proportion of phonetic misspellings.

However, it was noted that S-capade could be improved by using misspelling data to inform steps in the process. For example, the grapheme-to-phoneme tool was trained on the CMU dictionary and thus well-formed English words. If it were instead trained on labelled misspellings and their intended phoneme sequences the accuracy of the translation of misspellings into phonemic forms would likely be improved. Additionally, the cost values for the various edit operations came from a similarity model which endeavoured to capture speaker intuitions but was not specific to the sorts of phoneme substitutions which occur in phonetic misspellings. Whilst the original purpose of this project was to test an application of the matrix, the performance of the automatic spellchecker could be improved through a

more task specific model. Lastly, since the S-capade method was designed specifically to target difficult-to-correct misspellings, it is less suited to common typographic errors. The development of this approach was always with a view to incorporating it into a larger spelling correction tool which combines orthographic similarity and phonemic similarity, perhaps with the inclusion of context informed candidate ranking. Whilst undergraduate interns have since made progress on such a tool it still remains an interesting future direction of this work.

8.3 Research Question 3

How might the pronunciation variation present in a child's English impact the misspellings they produce and can a spelling correction tool be adapted to better perform with a specific variety?

The hypothesis that pronunciation variation impacts both the misspellings produced by a child and potentially their literacy acquisition was investigated in Chapter 5. This idea stemmed from the observation that a number of spellings in the Irish Accented English corpus, used to test S-capade, appeared to encode the *local Dublin English* variety. However, the existing research regarding the interplay of spoken variation and written literacy was considered inadequate. Thus, preliminary evidence for such an effect was demonstrated through the adaptation of a spelling correction tool to a specific variety of English and subsequent qualitative analysis of the spelling errors. Since adaptation was shown to improve the accuracy of the system, this would suggest that pronunciation plays a role in the types of misspellings produced by an individual. Furthermore, this demonstrated the benefits of designing language technologies for specific variants in order to improve user experience.

Whilst these initial results support the hypothesis, there are some limitations which should be considered and which present areas for future work. Firstly, many of the pronunciation features which are characteristic to a particular spoken variety are context specific. Certain phonetic realisations and phonological processes only occur in specific environments. However, the similarity matrix used to determine operation costs and the alignment algorithm for comparing two sequences of phonemes both work in a context independent manner. As such, there is no way to distinguish between a phoneme substitution which is very likely in one environment but less so in another. A more complex method of comparing similarity between the phoneme sequences would be an interesting extension of this work. Furthermore, to provide more concrete evidence of the influence of pronunciation on spelling, a much larger corpus is required. Steps towards this data collection to date have included an undergraduate honors project which developed a mobile game for children. The underlying concept was that additional lives or points could be earned, not by watching ads, but by spelling words. Maximum rewards would be given for a correct spelling whilst partial rewards could be earned by making an attempt that was close to correct. In this way, children's spelling attempts can be collected at a larger scale with the target word annotated in advance.

8.4 Research Question 4

Can an individual's variety of English be modelled based on phoneme confusability and do speakers with similar varieties produce similar representations?

After demonstrating that a phoneme confusion matrix could be adapted to a specific language variety, it was considered whether or not confusability could be used to model an individual's spoken variant in Chapter 6. Specifically, this chapter presented a process of modelling speakers by leveraging the phoneme confusions observed in erroneous ASR output as a result of their sensitivity to pronunciation variation. It was shown that speakers from the same region, who most likely have similar language varieties, generate similar representations, cluster together and can be reliably classified by region. Significantly, accurate speaker models which capture pronunciation features common to a region can be sufficiently captured with a relatively small amount of labelled data. Thus, this method lends itself to the research of underrepresented or low resource languages and varieties.

The research presented in Chapter 6 (and in Chapter 7) makes use of the wav2vec 2.0 ASR model which outputs character level representations of the speech. These outputs had to be translated to phonemic sequences, often using a grapheme-to-phoneme tool in the case of non-words. As has been discussed already, this tool is not necessarily trained on this specific task and is prone to errors. However, since the research of this chapter was carried out, ASR models with similar architecture have emerged which are trained to output phonetic or phonemic sequences. It would be an interesting area of future work to investigate one such model and its use in this speaker modelling process since it would remove the need for the grapheme-to-phoneme tool. Additionally, a major limitation noted in the chapter is the lack of information on the linguistic background of the speakers present in the corpus. Some interesting outliers were identified during clustering and it was proposed that some speakers in a particular region might exhibit features characteristic of another due to the effects of language transference. This would explain their clustering with another region (for example, the American speakers found to cluster with the Spanish speakers). The acquisition of a speech corpus with this speaker background information would allow for investigation of this language transference hypothesis. The following chapter uses the L2-Arctic corpus in an effort to examine L1 influence on L2 productions and how they manifest in the speaker models. However this corpus is limited in the number of speakers. A large-scale, speech corpus collection, with detailed linguistic histories of the speakers would be an excellent resource to add to the field. Finally, expanding on the observation that some speakers produced representations which were more characteristic of other regions than their own, a possible area for future research lies in testing human classification abilities. It would be interesting to see how a human listener, linguistically trained or otherwise, would classify each speaker in terms of their region and whether or not this matches the behaviour of the region classifier.

8.5 Research Question 5

How does the information captured by these models compare to existing knowledge of a spoken variety and what can they tell us about the sensitivity or robustness of ASR systems to pronunciation variation?

Chapter 7 primarily focused on demonstrating the interpretability of the region profiles by exhibiting the linguistic information they appeared to capture. This was done, firstly, through a detailed analysis of the region profiles in the context of existing literature regarding the English variant and observed pronunciation features typical of the region. Then, in order to demonstrate how this modelling approach can be used to test the sensitivity of an ASR to particular phonetic variants, the phoneme substitutions, insertions, and deletions identified from the ASR output were compared with those detected by human annotators. In this way the ASR's performance on variant pronunciation could be assessed with respect to specific pronunciation features. Finally, this chapter discussed how the information extracted through the interpretation of the models could be applied to both the development of ASR systems for specific spoken variants and to the incorporation of ASR technology as a tool for linguistic annotation.

The region profiles captured specific traits of the English varieties in line with existing research on pronunciation variation. However, as discussed in the chapter, only a small sample of the audio recordings were used to confirm the existence of the phonetic features and this was done through manual listening and the author's perception. The Accented English corpus, however, presents an excellent resource for a more fine-grained acoustic phonetic analysis to be carried out in order to analyse the characteristics of each speaker and the different regions. Moreover, it was suggested that the information regarding variant pronunciations common to particular regions could be used in order to fine tune an ASR system to a specific spoken variety by adopting a targeted approach to the collection of additional training data and exposing the system to the phonetic variants it is sensitive to. The testing of this hypothesis would make for a valuable extension of this thesis work. Additionally it could be explored whether or not an ASR adapted to one specific variety would produce improved performance on speakers who exhibit similar characteristics of that variety. For example, would a Spanish tuned model perform better with the speech of the American speakers who were observed to cluster around the Spanish speakers? If this were to be the case, it would have a significant impact on ASR for low resource languages which typically do not have enough available annotated data on which to train an ASR system. If a similar language or variety can be identified that shares some pronunciation features in common, an ASR system trained on this other variant could act as a basis for constructing an ASR system for the low resource variety.

8.6 Final Comments

This thesis has presented a breadth of work contributing to a number of research fields. From modelling human perceptions of the abstract phoneme categories, to tackling the automatic correction of children's misspellings. From constructing explainable models of individual speakers to testing an ASR system on its sensitivity to variant pronunciations and its potential use for linguistic annotation. It is hoped that the work of this thesis casts a spotlight on the benefits of multidisciplinary research and promotes collaboration across fields of study.

Appendix A

Links to Supplemental Resources

A.1 Similarity and Confusion Matrices

A.1.1 Baseline English Phoneme Similarity Matrix

https://github.com/emmaon/VariationModelling/blob/main/original_distance_matrix.csv

A.1.2 Accented English Region Profiles

https://github.com/emmaon/VariationModelling/tree/main/RegionProfiles

A.1.3 L2 Arctic Speaker Matrices

https://github.com/emmaon/VariationModelling/tree/main/L2Arctic_SpeakerMatrices

A.2 Code

A.2.1 English Syllabifier

https://github.com/emmaon/syllabifier

A.2.2 S-Capade

https://github.com/ucd-csl/Scapade

Bibliography

Anderson, Raquel T and José G Centeno (2007). "Contrastive analysis between Spanish and English". In: *Communication disorders in Spanish speakers: Theoretical, research and clinical aspects*, pp. 11–33.

Anderson, Thomas (2018). *Results from the 2016 Census: Aboriginal languages and the role of second-language acquisition*. Statistics Canada.

Atkinson, Kevin (2002). *Aspell Spell Checker Test Data*. Last accessed May 2020. URL: http://aspell.net/test/cur-all/batch0.tab.

Atkinson, Kevin (2004). *GNU Aspell - How Aspell Works*. Last accessed November 2022. URL: http://aspell.net/0.50-doc/man-html/8_How.html.

Aw, AiTi, Min Zhang, Juan Xiao, and Jian Su (2006). "A phrase-based statistical model for SMS text normalization". In: *COLING/ACL*, pp. 33–40.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in neural information processing systems* 33, pp. 12449–12460.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Bailey, Todd M and Ulrike Hahn (2005). "Phoneme similarity and confusability". In: *Journal of Memory and Language* 52.3, pp. 339–362.

Bansal, Ram Krishna (1990). "The pronunciation of English in India". In: *Studies in the pronunciation of English: A commemorative volume in honour of AC Gimson*, pp. 219–230.

Baranowski, Maciej and Danielle Turton (2020). "TD-deletion in British English: New evidence for the long-lost morphological effect". In: *Language Variation and Change* 32.1, pp. 1–23.

Barnes, Jonathan (2007). "Phonetics and phonology in Russian unstressed vowel reduction: A study in hyperarticulation". In: *Ms., Boston University*.

Barrus, Tyler (2018). *PySpellChecker*. Last accessed May 2020. URL: https://pypi.org/project/pyspellchecker/.

Best, Catherine T et al. (1994). "The emergence of native-language phonological influences in infants: A perceptual assimilation model". In: *The development of speech perception: The transition from speech sounds to spoken words* 167.224, pp. 233–277.

Best, Catherine T, M Tyler, O Bohn, and M Munro (2007). "Nonnative and second-language speech perception". In: *Language experience in second language speech learning*, pp. 13–34.

Bhattacharya, Usree (2017). "Colonization and English ideologies in India: A language policy perspective". In: *Language policy* 16, pp. 1–21.

Boberg, Charles (2008). “Regional phonetic differentiation in standard Canadian English”. In: *Journal of English Linguistics* 36.2, pp. 129–154.

Boberg, Charles (2015). “North American English”. In: *The handbook of English pronunciation*, pp. 227–250.

Boer, Elisabeth de and M Robbeets (2020). “The classification of the Japonic languages”. In: *The Oxford guide to the Transeurasian languages*, pp. 40–58.

Bondarenko, Olga (2014). “Does Russian Rnglish Rxist”. In: *American Journal of Educational Research* 2.9, pp. 832–839.

Bowers, Jeffrey S and Peter N Bowers (2017). “Beyond phonics: The case for teaching children the logic of the English spelling system”. In: *Educational Psychologist* 52.2, pp. 124–141.

Brill, Eric and Robert C Moore (2000). “An improved error model for noisy channel spelling correction”. In: *ACL*, pp. 286–293. DOI: 10.3115/1075218.1075255.

Broadbent, Judith (1991). *Linking and intrusive r in English*. Tech. rep. UCL working papers in linguistics.

Broselow, Ellen, Su-I Chen, and Chilin Wang (1998). “The emergence of the unmarked in second language phonology”. In: *Studies in second language acquisition* 20.2, pp. 261–280.

Brown, Lucien and Jaehoon Yeon (2015). *The handbook of Korean linguistics*. John Wiley & Sons.

Cassany, Daniel (2005). “Plain language in Spain”. In: *Clarity: Journal of the international association promoting plain legal language. 2005;(53): 41–4.*

Chappell, Hilary and Li Lan (2016). “Mandarin and other Sinitic languages”. In: *The Routledge encyclopedia of the Chinese language*. Routledge, pp. 643–666.

Charity, Anne H, Hollis S Scarborough, and Darion M Griffin (2004). “Familiarity with school English in African American children and its relation to early reading achievement”. In: *Child development* 75.5, pp. 1340–1356.

Chevalier, Joan F (2006). “Russian as the national language: an overview of language planning in the Russian Federation”. In: *Russian Language Journal/Русский язык* 56, pp. 25–36.

Cho, Young-mee Yu (2016). “Korean phonetics and phonology”. In: *Oxford research encyclopedia of linguistics*.

Choi, Jae-Oh (2007). “Teaching English pronunciation and listening skills”. In: *English Language & Literature Teaching* 13.2, pp. 1–23.

Chomsky, Noam and Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.

Church, Kenneth W and William A Gale (1991). “Probability scoring for spelling correction”. In: *Statistics and Computing* 1.2, pp. 93–103.

CMUSphinx (2016). *Grapheme-to-phoneme tool based on sequence-to-sequence learning*. URL: https://github.com/cmusphinx/g2p-seq2seq.

Collins, Beverley and Inger M Mees (2013). *Practical phonetics and phonology: A resource book for students*. Routledge.

Connor, Carol McDonald and Holly K Craig (2006). "African American preschoolers' language, emergent literacy skills, and use of African American English: A complex relation". In: *Journal of Speech, Language, and Hearing Research* 49, pp. 771–792.

Corrigan, Karen P, Richard Edge, John Lonergan, Bettina Migge, and Máire Nı Chıosáin (2012). "Is Dublin English 'alive alive oh'?" In: *New Perspectives on Irish English.*, pp. 1–28.

Cruz-Ferreira, Madalena (1995). "European Portuguese". In: *Journal of the International Phonetic Association* 25.2, pp. 90–94.

Cutler, Anne, Andrea Weber, Roel Smits, and Nicole Cooper (2004). "Patterns of English phoneme confusions by native and non-native listeners". In: *The Journal of the Acoustical Society of America* 116.6, pp. 3668–3678.

Daffern, Tessa and Sarah Critten (2019). "Student and teacher perspectives on spelling". In: *Australian Journal of Language and Literacy* 42.1, pp. 40–57.

Damerau, Fred J (1964). "A technique for computer detection and correction of spelling errors". In: *Communications of the ACM* 7.3, pp. 171–176.

Davidson, Lisa (2006a). "Phonotactics and articulatory coordination interact in phonology: Evidence from nonnative production". In: *Cognitive Science* 30.5, pp. 837–862.

Davidson, Lisa (2006b). "Schwa elision in fast speech: Segmental deletion or gestural overlap?" In: *Phonetica* 63.2-3, pp. 79–112.

De Jong, Kenneth J (2011). "Flapping in American English". In: *The Blackwell companion to phonology*, pp. 1–19.

Do, Youngah and Ryan Ka Yau Lai (2019). "Measuring Phonological Distance in a Tonal Language: An Experimental and Computational Study with Cantonese". In: *Proceedings of the Society for Computation in Linguistics* 2.1, pp. 371–372.

Docherty, Gerard J (2011). "The timing of voicing in British English obstruents". In: *The Timing of Voicing in British English Obstruents*. De Gruyter Mouton.

Dollinger, Stefan (2019). "English in Canada". In: *The Handbook of World Englishes*, pp. 52–69.

Downs, Brody, Oghenemaro Anuyah, Aprajita Shukla, Jerry Alan Fails, Sole Pera, Katherine Wright, and Casey Kennington (2020). "Kidspell: A child-oriented, rule-based, phonetic spellchecker". In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 6937–6946.

Driessen, Geert and Virgie Withagen (1999). "Language varieties and educational achievement of indigenous primary school pupils". In: *Language Culture and Curriculum* 12.1, pp. 1–22.

Everri, Marina and Kirsty Park (2018). *Children's online behaviours in Irish primary and secondary schools*. Tech. rep. Zeeko, Nova UCD.

Ferreira, Marta Cunha (2013). "Language death". In: *Advanced English and Language Analysis*.

Fisher, William M (1999). “A statistical text-to-phone function using ngrams and rules”. In: *ICASSP*. Vol. 2, pp. 649–652.

Flege, James Emil and Richard D Davidian (1984). “Transfer and developmental processes in adult foreign language speech production”. In: *Applied psycholinguistics* 5.4, pp. 323–347.

Flemming, Edward (2009). “The phonetics of schwa vowels”. In: *Phonological weakness in English*, pp. 78–98.

Fourakis, Marios and Robert Port (1986). “Stop epenthesis in English”. In: *Journal of Phonetics* 14.2, pp. 197–221.

Francis, W. Nelson and Henry Kučer (1964). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Providence, Rhode Island: Brown University.

Frisch, Stefan (1996). “Similarity and frequency in phonology”. PhD thesis. Northwestern University.

Fuchs, Robert (2016). *Speech rhythm in varieties of english*. Springer.

Gallagher, Gillian and Peter Graff (2012). “The role of similarity in phonology”. In: *Lingua* 2.122, pp. 107–111.

Garbe, Wolf (2012). *SymSpell*. Last accessed November 2022. URL: `https://github.com/wolfgarbe/SymSpell`.

Gargesh, Ravinder (2008). “Indian English: Phonology”. In: *Varieties of English* 4.2, pp. 231–243.

Giegerich, Heinz J et al. (1992). *English phonology: An introduction*. Cambridge University Press.

Gimson, Alfred Charles and Susan Ramsaran (1970). *An introduction to the pronunciation of English*. Vol. 4. Edward Arnold London.

Glass, Yu-An Chung James (2018). “Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech”. In: *Proceedings of Interspeech, Hyderabad, India*.

Glowacka, Donora (2001). “Unstressed vowel deletion and new consonant clusters in English”. In: *Poznan Studies in Contemporary Linguistics* 37, pp. 71–94.

Gómez González, Marıa de los Ángeles and Teresa Sánchez Roura (2016). *English Pronunciation for Speakers of Spanish: From Theory to Practice*. De Gruyter Mouton.

Gottlieb, Nanette (2005). *Language and society in Japan*. Cambridge University Press.

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber (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.

Guy, Gregory (1980). “Variation in the group and the individual: The case of final stop deletion”. In: *Locating language in time and space*. Academic Press, pp. 1–36.

He, Yunjuan (2014). “Production of English Syllable Final /l/ by Mandarin Chinese Speakers.” In: *Journal of Language Teaching & Research* 5.4.

Hickey, Raymond (2004). “The phonology of Irish English”. In: *Handbook of varieties of English* 1, pp. 68–97.

Hickey, Raymond (2005). *Dublin English: evolution and change*. John Benjamins Publishing.

Hickey, Raymond (2008). “Irish English: phonology”. In: *Varieties of English* 1, pp. 71–104.

Hickey, Raymond and Carolina P Amador-Moreno (2020). *Irish Identities: Sociolinguistic Perspectives*. Language and Social Life. De Gruyter.

Hinsvark, Arthur et al. (2021). “Accented speech recognition: A survey”. In: *arXiv preprint arXiv:2104.10747*.

Hinton, Geoffrey et al. (2012). “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”. In: *IEEE Signal processing magazine* 29.6, pp. 82–97.

Hodge, Victoria Jane and Jim Austin (2001). *An evaluation of phonetic spell checkers*.

Holbrook, David (1964). “English for the Rejected: Training Literacy in the Lower Streams of the Secondary School.” In.

Houston, Ann Celeste (1985). *Continuity and change in English morphology: The variable (ING)*. University of Pennsylvania.

Huang, Meichan and Lucy Pickering (2014). “Revisiting the pronunciation of English by speakers from Mainland China”. In: *Pronunciation in Second Language Learning and Teaching Conference (ISSN 2380-9566)*, p. 206.

Hughes, Arthur, Peter Trudgill, and Dominic Watt (2013). *English accents and dialects: An introduction to social and regional varieties of English in the British Isles*. Routledge.

International Phonetic Association and others (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Itô, Junko (1989). “A prosodic theory of epenthesis”. In: *Natural Language & Linguistic Theory* 7.2, pp. 217–259.

Jakobson, Roman, C Gunnar Fant, and Morris Halle (1951). “Preliminaries to speech analysis: The distinctive features and their correlates”. In.

Jarosz, Gaja (2019). “Computational modeling of phonological learning”. In: *Annual Review of Linguistics* 5, pp. 67–90.

Jesus, Luis MT and Christine H Shadle (2002). “A parametric study of the spectral characteristics of European Portuguese fricatives”. In: *Journal of Phonetics* 30.3, pp. 437–464.

Johnsen, Sverre Stausland (2011). “Rhyme acceptability determined by perceived similarity”. In: *Paper presented at the 29th West Coast Conference on Formal Linguistics*. University of Arizona.

Johnsen, Sverre Stausland (2012). “From perception to phonology: The emergence of perceptually motivated constraint rankings”. In: *Lingua* 122.2, pp. 125–143.

Jureková, Petra (2015). “The Pronunciation of English in Czech, Slovak and Russian Speakers”. In: *Unpublished bachelor thesis). Brno: Masaryk University*.

Kahn, Daniel (1980). *Syllable-based generalizations in English phonology*. Routledge.

Kane, Mark and Julie Carson-Berndsen (2016). “Enhancing Data-Driven Phone Confusions Using Restricted Recognition”. In: *INTERSPEECH*, pp. 3693–3697.

Kaplan, Aaron and John Woodmansee (2018). “Imperfect Rhymes as a Measure of Phonological Similarity”. In: *UNC Spring Colloquium*.

Kawahara, Shigeto (2007). “Half rhymes in Japanese rap lyrics and knowledge of similarity”. In: *Journal of East Asian Linguistics* 16.2, pp. 113–144.

Kenstowicz, Michael and Charles Kisseberth (1979). *Generative phonology: Description and theory*. Academic Press.

Khanal, Subash, Michael T Johnson, Mohammad Soleymanpour, and Narjes Bozorg (2021). “Mispronunciation Detection and Diagnosis for Mandarin Accented English Speech”. In: *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, pp. 62–67.

Khoury, Richard (2015). “Microtext normalization using probably-phonetically-similar word discovery”. In: *WiMob*, pp. 384–391.

Koenecke, Allison et al. (2020). “Racial disparities in automated speech recognition”. In: *Proceedings of the National Academy of Sciences* 117.14, pp. 7684–7689.

Kökeritz, Helge (1945). “The reduction of initial kn and gn in English”. In: *Language*, pp. 77–86.

Kolachina, Sudheer and Lilla Magyar (2019). “What do phone embeddings learn about Phonology?” In: *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 160–169.

Kominek, John and Alan W Black (2004). “The CMU Arctic speech databases”. In: *Fifth ISCA workshop on speech synthesis*.

Kondo, Yuko (1994). “Targetless schwa: is that how we get the impression of stress timing in English”. In: *Proceedings of the Edinburgh Linguistics Department Conference*. Vol. 94. Citeseer, pp. 63–76.

Kondo, Yuko (2000). “Production of schwa by Japanese speakers of English: An acoustic study of shifts in coarticulatory strategies”. In: *Papers in laboratory phonology V: Acquisition and the lexicon* 5.5, p. 29.

Kubozono, Haruo (2015). *Handbook of Japanese phonetics and phonology*. Vol. 2. Walter de Gruyter GmbH & Co KG.

Kukich, Karen (1992). “Techniques for automatically correcting words in text”. In: *Acm Computing Surveys (CSUR)* 24.4, pp. 377–439.

Labov, William (1972). *Sociolinguistic patterns*. 4. University of Pennsylvania press.

Labov, William (1986). “The social stratification of (r) in New York City department stores”. In: *Dialect and language variation*. Elsevier, pp. 304–329.

Labov, William, Sharon Ash, and Charles Boberg (2008). *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Labrune, Laurence (2014). “The phonology of Japanese/r: A panchronic account”. In: *Journal of East Asian Linguistics* 23.1, pp. 1–25.

Lee, Borim, Susan G Guion, and Tetsuo Harada (2006). “Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals”. In: *Studies in Second Language Acquisition* 28.3, pp. 487–513.

Levenshtein, Vladimir I (1966). “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10.8. Soviet Union, pp. 707–710.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). “Improving distributional similarity with lessons learned from word embeddings”. In: *Transactions of the Association for Computational Linguistics* 3, pp. 211–225.

Lewis, J Windsor (1975). “Linking /r/ in the General British pronunciation of English”. In: *Journal of the International Phonetic Association* 5.1, pp. 37–42.

Li, Jinyu et al. (2022). “Recent advances in end-to-end automatic speech recognition”. In: *APSIPA Transactions on Signal and Information Processing* 11.1.

Liberman, Mark Y (2019). “Corpus Phonetics”. In: *Annual Review of Linguistics* 5, pp. 91–107.

Ma, Jianqiang, Çağrı Çöltekin, and Erhard Hinrichs (2016). “Learning phone embeddings for word segmentation of child-directed speech”. In: *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pp. 53–63.

Mac Sıthigh, Daithı (2018). “Official status of languages in the United Kingdom and Ireland”. In: *Common Law World Review* 47.1, pp. 77–102.

MacQueen, J (1967). “Classification and analysis of multivariate observations”. In: *5th Berkeley Symp. Math. Statist. Probability*. University of California Los Angeles LA USA, pp. 281–297.

Major, Roy C and Michael C Faudree (1996). “Markedness universals and the acquisition of voicing contrasts by Korean speakers of English”. In: *Studies in Second Language Acquisition* 18.1, pp. 69–90.

Markl, Nina (2022). “(Commercial) Automatic speech recognition as a tool in sociolinguistic research”. In: *University of Pennsylvania Working Papers in Linguistics* 28.2, p. 11.

Marques, Sandra (2021). “Easy Language in Portugal”. In: *Handbook of Easy Languages in Europe* 8, p. 413.

McCulloch, Gretchen (2013). “Phonological natural classes and set theory”. In: *All Things Linguistic*.

McQueen, James M, Anne Cutler, and Dennis Norris (2006). “Phonological abstraction in the mental lexicon”. In: *Cognitive science* 30.6, pp. 1113–1126.

Mendonça Almeida, Gustavo Augusto de, Lucas Avanço, Magali Sanches Duran, Erick Rocha Fonseca, Maria das Graças Volpe Nunes, and Sandra Maria Aluísio (2016). "Evaluating phonetic spellers for user-generated content in Brazilian Portuguese". In: *International conference on computational processing of the Portuguese language*, pp. 361–373.

Mielke, Jeff (2012). "A phonetically based metric of sound similarity". In: *Lingua* 122.2, pp. 145–163.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)". In: *arXiv preprint arXiv:1301.3781*.

Miller, George A and Patricia E Nicely (1955). "An analysis of perceptual confusions among some English consonants". In: *The Journal of the Acoustical Society of America* 27.2, pp. 338–352.

Mitton, Roger (1985). "A collection of computer-readable corpora of English spelling errors". In: *Cognitive Neuropsychology* 2.3, pp. 275–279.

Miyawaki, Kuniko, James J Jenkins, Winifred Strange, Alvin M Liberman, Robert Verbrugge, and Osamu Fujimura (1975). "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English". In: *Perception & Psychophysics* 18.5, pp. 331–340.

Mohanty, Ajit K (2006). "Multilingualism of the unequals and predicaments of education in India: Mother tongue or other tongue". In: *Imagining multilingual schools*, pp. 262–283.

Moore, Roger K and Lucy Skidmore (2019). "On the use/misuse of the term'phoneme'". In: *arXiv preprint arXiv:1907.11640*.

Mukherjee, Joybrato and Tobias Bernaisch (2020). "The development of the English language in India". In: *The Routledge handbook of world Englishes*. Routledge, pp. 165–177.

Mustajoki, Arto, Zhanna Mihienko, Natalia Nechaeva, Emma Kairova, and Anna Dmitrieva (2021). "Easy language in Russia". In: *Handbook of Easy Languages in Europe*, p. 439.

National Council of Curriculum and Assessment (2019). *Primary Language Curriculum*. Dublin, Ireland: Government of Ireland.

Nihalani, Paroo, Ray K Tongue, and Priya Hosali (1979). *Indian and British English: A handbook of usage and pronunciation*. Oxford University Press.

Nogita, Akitsugu and Yanan Fan (2012). "Not vowel epenthesis: Mandarin and Japanese ESL learners' production of English consonant clusters". In: *Working Papers of the Linguistics Circle* 22.1, pp. 1–26.

Norris, Dennis, James M McQueen, and Anne Cutler (2003). "Perceptual learning in speech". In: *Cognitive psychology* 47.2, pp. 204–238.

Norvig, Peter (2007). *How to write a spelling corrector*. [Online]. URL: `http://norvig.com/spell-correct.html`.

O'Neill, Emma and Julie Carson-Berndsen (2019). "The Effect of Phoneme Distribution on Perceptual Similarity in English". In: *INTERSPEECH 2019*, pp. 1941–1945.

O'Neill, Emma and Julie Carson-Berndsen (2022a). "Investigating the Sensitivity of Automatic Speech Recognition Systems to Phonetic Variation in L2 Englishes". In: *New Ways of Analyzing Variation 50*. [Abstract].

O'Neill, Emma and Julie Carson-Berndsen (2022b). "Modelling Pronunciation Variation in Different Spoken Englishes". In: *UK Speech Conference 2022*. [Abstract].

O'Neill, Emma and Julie Carson-Berndsen (2023). "Investigating the Sensitivity of Automatic Speech Recognition Systems to Phonetic Variation in L2 Englishes". In: *University of Pennsylvania Working Papers in Linguistics*. Vol. 29.2.

O'Neill, Emma and Julie Carson-Berndsen (in preparation). *Leveraging Erroneous ASR Output to Build Interpretable Speaker Models of Pronunciation Variation in English*. Journal paper in preparation.

O'Neill, Emma, Mark Kane, and Julie Carson-Berndsen (2018). "Two Data-Driven Perspectives on Phonetic Similarity". In: *UK Speech Conference 2018*. [Abstract].

O'Neill, Emma, Joe Kenny, Anthony Ventresque, and Julie Carson-Berndsen (2021). "The Influence of Regional Pronunciation Variation on Children's Spelling and the Potential Benefits of Accent Adapted Spellcheckers". In: *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 674–683.

O'Neill, Emma, Robert Young, Elsa Thiaville, Muireann MacCarthy, Julie Carson-Berndsen, and Anthony Ventresque (2020). "S-capade: Spelling correction aimed at particularly deviant errors". In: *International Conference on Statistical Language and Speech Processing*. Springer, pp. 85–96.

Ogden, Richard (2017). *Introduction to English phonetics*. Edinburgh university press.

Ohala, John J et al. (1990). "The phonetics and phonology of aspects of assimilation". In: *Papers in laboratory phonology* 1, pp. 258–275.

Ohata, Kota (2004). "Phonological differences between Japanese and English: Several potentially problematic Areas of Pronunciation for Japanese ESL/EFL Learners". In: *Language learning* 22, pp. 29–41.

Oostendorp, Marc van (2004). *The Theory of Faithfulness*. Unpublished manuscript.

Pape, Daniel and Luis MT Jesus (2011). "Devoicing of phonologically voiced obstruents: Is European Portuguese different from other Romance languages?" In: *ICPhS*, pp. 1566–1569.

Pape, Daniel and Luis MT Jesus (2015). "Stop and fricative devoicing in European Portuguese, Italian and German". In: *Language and speech* 58.2, pp. 224–246.

Parrish, Allison (2017). "Poetic Sound Similarity Vectors Using Phonetic Features". In: *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. URL: https://aaai.org/ocs/index.php/AIIDE/AIIDE17/paper/view/15879/15227.

Philips, Lawrence (2000). "The double metaphone search algorithm". In: *C/C++ users journal* 18.6, pp. 38–43.

Pou, Jaume Corbera (2004). "The Languages of Spain". In: *JSL: Journal of the School of Language, Literature, and Culture Studies*, p. 127.

Priva, Uriel Cohen (2015). "Informativity affects consonant duration and deletion rates". In: *Laboratory phonology* 6.2, pp. 243–278.

Rato, Anabela and Andréia S Rauber (2015). "The effects of perceptual training on the production of English vowel contrasts by Portuguese learners". In: *ICPhS*.

Rato, Anabela, Andréia S Rauber, Letícia P Soares, and Liane R Lucas (2014). "Challenges in the perception and production of English front vowels by native speakers of European Portuguese". In: *diacrıtica*, p. 141.

Raymond, William D, Esther L Brown, and Alice F Healy (2016). "Cumulative context effects and variant lexical representations: Word use and English final t/d deletion". In: *Language Variation and Change* 28.2, pp. 175–202.

Raymond, William D, Robin Dautricourt, and Elizabeth Hume (2006). "Word-internal/t, d/deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors". In: *Language variation and change* 18.1, pp. 55–97.

Read, Charles (2018). *Children's creative spelling*. Original work published 1986. Routledge.

Řehůřek, Radim and Petr Sojka (May 2010). "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. `http://is.muni.cz/publication/884893/en`. Valletta, Malta: ELRA, pp. 45–50.

Russell, Robert C. and Margaret King Odell (1918). "Soundex". In: *US patent 1,261,167*.

Sadlier-Brown, Emily and Meredith Tamminga (2008). "The Canadian shift: Coast to coast". In: *Proceedings of the 2008 annual conference of the Canadian Linguistic Association*. Vol. 4, pp. 1–14.

Sahgal, Anju and Rama Kant Agnihotri (1988). "Indian English phonology: A sociolinguistic perspective". In: *English World-Wide* 9.1, pp. 51–64.

Saiegh-Haddad, Elinor (2003). "Linguistic distance and initial reading acquisition: The case of Arabic diglossia". In: *Applied Psycholinguistics* 24.3, p. 431.

Sailaja, Pingali (2012). "Indian English: Features and sociolinguistic aspects". In: *Language and Linguistics Compass* 6.6, pp. 359–370.

Scharenborg, Odette and Esther Janse (2013). "Comparing lexically guided perceptual learning in younger and older listeners". In: *Attention, Perception, & Psychophysics* 75, pp. 525–536.

Scharenborg, Odette, Sebastian Tiesmeyer, Mark Hasegawa-Johnson, and Najim Dehak (2018). "Visualizing phoneme category adaptation in deep neural networks". In: *Proceedings of Interspeech, Hyderabad, India*.

Schwartz, Judith I (1982). "Dialect interference in the attainment of literacy". In: *Journal of Reading* 25.5, pp. 440–446.

Selkirk, Elisabeth O (1984). "On the major class features and syllable theory". In: *Language Sound Structure*. Ed. by M. Aronoff and R. Oehrle. Cambridge, Massachusetts: The MIT Press.

Shafiro, Valeriy and Anatoliy V Kharkhurin (2009). "The Role of Native-Language Phonology in the Auditory Word Identification and Visual Word Recognition of Russian–English Bilinguals." In: *Journal of psycholinguistic research* 38.2.

Sheldon, Amy and Winifred Strange (1982). "The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception". In: *Applied psycholinguistics* 3.3, pp. 243–261.

Shi, Xian, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie (2021). "The accented English speech recognition challenge 2020: open datasets, tracks, baselines, results and methods". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6918–6922.

Siegel, Jeff (1999). "Creoles and minority dialects in education: An overview". In: *Journal of Multilingual and Multicultural Development* 20.6, pp. 508–531.

Silfverberg, Miikka, Pekka Kauppinen, and Krister Lindén (2016). "Data-driven spelling correction using weighted finite-state methods". In: *SIGFSM Workshop on Statistical NLP and Weighted Automata*, pp. 51–59.

Silfverberg, Miikka, Lingshuang Jack Mao, and Mans Hulden (2018). "Sound analogies with phoneme embeddings". In: *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pp. 136–144.

Siqi, Li and Andrew Sewell (2012). "Phonological features of China English". In: *Asian Englishes* 15.2, pp. 80–101.

Smith, Caroline L (1997). "The devoicing of/z/in American English: Effects of local and prosodic context". In: *Journal of Phonetics* 25.4, pp. 471–500.

Smolensky, Paul and A Prince (1993). "Optimality Theory: Constraint interaction in generative grammar". In: *Optimality Theory in phonology*, p. 3.

Snell, Julia and Richard Andrews (2017). "To what extent does a regional dialect and accent impact on the development of reading and writing skills?" In: *Cambridge Journal of Education* 47.3, pp. 297–313.

Strange, Winifred and Sibylla Dittmann (1984). "Effects of discrimination training on the perception of /rl/ by Japanese adults learning English". In: *Perception & psychophysics* 36.2, pp. 131–145.

Stüker, Sebastian, Johanna Fay, and Kay Berkling (2011). "Towards Context-Dependent Phonetic Spelling Error Correction in Children's Freely Composed Text for Diagnostic and Pedagogical Purposes". In: *INTERSPEECH*.

Szmrecsanyi, Benedikt (2013). *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge University Press.

Tagliamonte, Sali and Rosalind Temple (2005). "New perspectives on an ol'variable:(t, d) in British English". In: *Language Variation and Change* 17.3, pp. 281–302.

Tatman, Rachael and Conner Kasten (2017). "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions". In: *Interspeech*, pp. 934–938.

Terry, Nicole Patton (2006). “Relations between dialect variation, grammar, and early spelling skills”. In: *Reading and Writing* 19.9, pp. 907–931.

Terry, Nicole Patton (2012). “Examining relationships among dialect variation and emergent literacy skills”. In: *Communication Disorders Quarterly* 33.2, pp. 67–77.

Terry, Nicole Patton and Carol Connor (2010). “African American English and spelling: How do second graders spell dialect-sensitive features of words?” In: *Learning Disability Quarterly* 33.3, pp. 199–210.

Terry, Nicole Patton and Hollis S Scarborough (2011). “The phonological hypothesis as a valuable framework for studying the relation of dialect variation to early reading skills.” In: *Explaining individual differences in reading: Theory and evidence*, pp. 97–117.

Toutanova, Kristina and Robert C Moore (2002). “Pronunciation modeling for improved spelling correction”. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 144–151.

Tranter, Nicolas (2012). *The languages of Japan and Korea*. Routledge.

Trudgill, Peter and Jean Hannah (2017). *International English: A guide to varieties of English around the world*. Taylor & Francis.

Tumshevits, Alina (2019). “Perception of Russian-accented speech by native and non-native speakers of English”. In.

Turk, Alice (1992). “The American English flapping rule and the effect of stress on stop consonant durations”. In: *Working papers of the Cornell phonetics laboratory* 7, pp. 103–133.

Vale, Ana Paula and Rafaela Perpétua (2020). “Early Context-Conditioned Orthographic Knowledge in European Portuguese: The Spelling of the Schwa”. In: *Frontiers in Education*. Vol. 5. Frontiers Media SA, p. 513577.

Van der Maaten, Laurens and Geoffrey Hinton (2008). “Visualizing data using t-SNE”. In: *Journal of machine learning research* 9.11.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.

Veloso, João (2007). “Schwa in European Portuguese: the phonological status of []” In: *Journées d’Études Linguistiques*, pp. 55–60.

Veronis, Jean (1988). “Computerized correction of phonographic errors”. In: *Computers and the Humanities* 22.1, pp. 43–56.

Vishnevskaya, GM (2011). “Pronunciation Error Variables in ESP: The Case of Russian English”. In: *Традиции и новаторство в преподавании родного и иностранного языков в вузе*

Wagner, Robert A and Michael J Fischer (1974). “The string-to-string correction problem”. In: *JACM* 21.1, pp. 168–173.

Ward Jr., Joe H (1963). “Hierarchical grouping to optimize an objective function”. In: *Journal of the American statistical association* 58.301, pp. 236–244.

Weide, Robert L (1998). *The CMU pronouncing dictionary*. URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Weng, Jeffrey (2018). "What is Mandarin? The social project of language standardization in early Republican China". In: *The Journal of Asian Studies* 77.3, pp. 611–633.

Wheeler, Max W. (1972). "Distinctive features and natural classes in phonological theory". In: *Journal of Linguistics* 8.1, pp. 87–102.

Wikipedia (2020). *Wikipedia:Lists of common misspellings*. Last accessed May 2020. URL: https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings.

Wiltshire, Caroline R (2006). "Word-final consonant and cluster acquisition in Indian English (es)". In: *30th Boston University Conference on Language Development*, pp. 1–10.

Woods, Howard B (1993). "A synchronic study of English spoken in Ottawa: Is Canadian English becoming more American". In: *Focus on Canada*, pp. 151–178.

Yang, Rong, Ran Ao, and Ee Ling Low (2021). "Features of Chinese English". In: *English in East and South Asia*, pp. 107–121.

Yanushevskaya, Irena and Daniel Bunčić (2015). "Russian". In: *Journal of the International Phonetic Association* 45.2, pp. 221–228.

Yeon, Jaehoon (2012). "Korean dialects: A general survey". In: *The languages of Japan and Korea*, pp. 168–185.

Yiakoumetti, Androula (2007). "Choice of classroom language in bidialectal communities: to include or to exclude the dialect?" In: *Cambridge journal of education* 37.1, pp. 51–66.

Yip, Moira (1987). "English vowel epenthesis". In: *Natural Language & Linguistic Theory*, pp. 463–484.

Yoko, KITA et al. (2019). "Japanese learners of English and Japanese phonology". In: *Research Bulletin of Naruto University of Education* 34, pp. 209–216.

You, Hong, Abeer Alwan, Abe Kazemzadeh, and Shrikanth Narayanan (2005). "Pronunciation variations of Spanish-accented English spoken by young children". In: *Ninth European Conference on Speech Communication and Technology*.

Yu Cho, Young-mee and Gregory K Iverson (1997). "Korean phonology in the late twentieth century". In: *어학연구*.

Zeigler, Karen and Steven A Camarota (2019). "67.3 million in the United States spoke a foreign language at home in 2018". In: *Center for immigration studies* 29.

Zhao, Guanlong, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna (2018). "L2-ARCTIC: A non-native English speech corpus." In: *Interspeech*, pp. 2783–2787.

Zwicky, Arnold M (1976). "Well, this rock and roll has got to stop. Junior's head is hard as a rock". In: *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill.* 12, pp. 676–697.