# Towards Phonetically-Informed Automatic Speaker Recognition

Elliot Holmes

Doctor of Philosophy

University of York

Language and Linguistic Science

June 2023

**Abstract**

This thesis explores novel applications of phonetic theory to enhance our understanding of Automatic Speaker Recognition (ASR). Previous studies typically only explore the performance of one phonetic feature in isolation; instead, this thesis explores bespoke, systematically-validated combinations of many different phonetic features. Sociophonetic-tailoring is also uncommon in previous literature, so this thesis also explores how these features can be fused together in optimised ways for different accents and speech styles. This thesis finds that all of the tested phonetic features can be effective for ASR, but tailoring approaches to different accents and speech styles is the most important consideration in terms of overall performance. That said, higher formants were generally found to be most effective for ASR whilst features relating to non-modal voicing were found to be least effective. As the tested (socio)phonetic features are all explainable, a potential future application of these findings is to improve the explainability of ASR systems. ASR systems are increasingly present in modern society and they are undeniably powerful, but their inner workings are not fully explainable; they are considered 'black boxes' by researchers like Rudin (2018) and they are becoming increasingly distrusted by triers-of-fact (van der Veer et al., 2021). When tested on their own, the bespoke combinations of explainable phonetic approaches performed worse than state-of-the-art ASR systems, but this reflects the known trade-off between explainability and performance (Moez et al., 2016). However, this thesis also finds that its best-performing phonetic approaches to ASR do not have a detrimental impact to the performance of off-the-shelf ASR systems when they are fused together; as a result, these explainable, bespoke, combinatory phonetic approaches could be fused with ASR systems to add an extra element of explainability to them without concern for performance.

**Contents**

**List of Tables**

**List of Figures**

**Acknowledgements**

I am sincerely and truly grateful to my supervisors at the University of York *Vincent Hughes, Philip Harrison,* and *Dominic Watt* and my supervisors at Aculab PLC *Steve Beet* and *Ladan Ravary.* No matter the technical or theoretical problem, you have all been there to support me every step of the way and this thesis would not exist without any of you.

I would additionally like to thank *George Bailey* and *Volker Dellwo* for examining this thesis; the improvements to this work would not be possible without your time and input.

I must also thank my friends, my family, and above all my wife for keeping me grounded and focused.

This work was supported by the Arts and Humanities Research Council (grant number AH/R012733/1) and Aculab PLC through the White Rose College of the Arts & Humanities.

**Author's Declaration**

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as references.

## 1. Introduction

This thesis explores how phonetics can be used to enhance approaches to ASR. It does this using a novel, bottom-up approach that tests multiple phonetic features, measured from different phonemes, in different combinations. From this, it identifies the best combinations of features for characterising speakers varying across different sociophonetic dimensions, such as their accent. Unlike other research in this area, only a limited number of prior decisions are made with feature selection; instead, every possible combination of the selected features is tested to identify the best combination without bias. This combinatorial approach, plus the observation of multiple sociophonetic dimensions, is novel to the field. This thesis successfully identifies that all of the tested phonetic features can be effective for ASR, but tailoring approaches to different accents and speech styles is the most important consideration in terms of overall performance.

In terms of real-world applications, this information could inform the explainability of Automatic Speaker Recognition (ASR) in future studies. ASR refers to the task of comparing or verifying speakers with minimal human input based solely on information that can be captured from a recording of a speaker's voice. The presence of ASR in modern society is rapidly expanding: judicially, French (2017) writes that ASR is employed in courtrooms around the world as a form of forensic evidence that compares and evaluates audio recordings containing known and unknown voices, such as incriminating phone calls. Commercially, such systems have also been employed by banks like HSBC (2023) since 2016 to verify the identity of their customers over the phone. The ASR systems used for these tasks are undeniably powerful: Morrison and Enzinger (2019) found that current DNN-based, state-of-the-art systems are capable of extremely accurate and reliable recognition, especially when they are trained on case-specific data. However, it is not easy to explain how the output of these systems is consistent with the input they draw upon and process. This is because

modern ASR systems are trained on huge and diverse data sets, modelling spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs) in ways that cannot be easily explained, even by their developers. System output can still be presented as validated evidence, but researchers like Rudin (2018) argue that ASR systems are now 'black boxes' because the information and features used to generate their conclusions cannot be explained clearly. This problem has gotten worse over recent years due to the massive rise in the use of machine learning within these systems, as the following chapter discusses.

As a consequence of this, van der Veer et al. (2021) argue that triers-of-fact now want to understand machine learning systems better so that they can better trust their output. However, 'understanding' can take many forms: Pruksachatkun et al. (2023) write that machine learning models should be *interpretable*, meaning the output can be understood and trusted as valid, and *explainable*, meaning the processes generating the output can also be understood. Explorations of interpretability are frequently prioritised over explainability in related, forensic settings: for example, Garrett and Mitchell (2013) found that, in regards to fingerprint evidence, information about how to interpret fingerprint evidence is more impactful to triers-of-fact than a scientific explanation of how a fingerprint is analysed.

While it is widely recognised that interpretability and explainability are also relevant aspects of forensic voice evidence, there appears to be less consensus on how they are defined or implemented in practice. For example, Morrison et al. (2021) write that triers-of-fact do not need to be able to understand how ASR systems work, they only need to understand that they have worked before under similar conditions. Morrison et al. (2021) label this reading of interpretability as *validation*.

The scientist's goal here, overall, is to minimise the probability of a miscarriage of justice. Interpretability and explainability can both guide approaches towards this goal. That said, the

reasons behind the above prioritisations of interpretability are logical: explainability is significantly more complex and potentially unnecessary if the evidence generated by such systems can already be trusted based on prior validation. However, there is rising interest surrounding the potential benefits of explainability: for example, Eldridge (2019) provides an overview of literature from cognitive science which suggests that triers-of-fact can struggle to see the importance of validation statistics alone when it comes to decision-making. Eldrige (2019) hypothesises that some triers-of-fact may trust validation statistics more if they can understand more about how they were generated, specifically by drawing on prior understandings and perceptions of the evidence that is being validated.

ASR applications aside, the present thesis primarily explores novel methods of characterising speakers using (socio)phonetic theory. Specifically, this thesis explores how measurements of selected phonetic features (which have all been found to be capable of characterising speakers individually, as the following chapter shows) can be combined to characterise speakers from different sociophonetically distinct groups. Only after establishing this does it then suggest that these phonetic approaches are all somewhat explainable (given that they can all be correlated to features of the voice that triers-of-fact can perceive, as the following chapter argues), and as such they could potentially aid in explaining ASR.

Many studies already exemplify how phonetically-informed approaches to ASR can be tested and how such results can help improve explainability in ASR: for example, Hughes et al. (2019a) combined phonetic features that characterise laryngeal voice qualities, such as mean harmonics-to-noise ratio which perceptibly correlates to breathy and creaky voicing, with an ASR system. The inclusion of this feature did not change ASR system performance significantly for most of the replications, but the processes behind the output of the tested ASR system can now be considered more explainable: the validated ASR system output is

now known to be based on information related to breathy and creaky voicing features that triers-of-fact can perceive and understand.

One could argue that, alternatively, these results show that fusing phonetic approaches to ASR with current ASR approaches is pointless: if ASR systems already perform so well, then including these additional phonetic elements that have no impact on performance is unnecessary effort. This is a valid perspective, but it disregards van der Veer et al.'s (2021) concerns about triers-of-fact trust in validation statistics and Eldridge's (2019) recommendation for more explainable information to be included to support validation statistics. Having the ability to include more explainable elements, such as phonetic features, without any cost to performance could therefore be a low-risk way of accommodating these concerns and recommendations that, ultimately, seek to minimise the probability of a miscarriage of justice. Additionally, Hughes et al. (2019a) found that the performance of ASR approaches declined substantially with poorer-quality recordings whilst the performance of phonetic approaches remained more consistent; thus, from an ASR performance perspective, phonetic approaches may aid in maintaining consistent ASR performance across different recordings, particularly when some recordings are of poorer quality. They also found that phonetic approaches were generally more consistent when recordings were of a short duration, like the isolated vowels that this thesis explores.

In some of their replications, Hughes et al.'s (2019a) study even showed that mean harmonics-to-noise ratio could improve the performance of these ASR systems. While this is only an additional potential benefit of consulting phonetic approaches, a performance boost inadvertently also helps explain the ASR system better: it shows that the underlying processes employed by this ASR system may not have originally accounted for breathy or creaky elements of the voice, but with the inclusion of mean harmonics-to-noise ratio now

boosting performance, this means that these perceivable elements of the voice are now accounted for and must be important to the improved output.

This study provides evidence that phonetic approaches can be useful for ASR, that they can already be fused into ASR systems, and that they could improve the explainability of an ASR system's output. However, studies like Hughes et al.'s (2019a) could be taken further: their study, alongside many others, only focuses on a single phonetic feature or segment at a time. Additionally, they rarely ever consider sociophonetic variation: they only look at one accent or one style and the specified accent is typically over-generalised with tags like 'British English'. This thesis aims to address this research gap through a more in-depth, phonetically-informed exploration of ASR. It does this in a novel and original way: it investigates fused combinations of multiple phonetic features and segments and identifies how these combinations can be tailored to different sociophonetic variables. All of this is explored by asking the following research question:

**What explainable phonetic approaches to ASR, ranging from features to segments, are best for recognising different speakers and speech styles?**

The original contributions of this thesis are twofold. Firstly, a replicable, modifiable, semi-automatic methodology for testing novel, combinatory phonetic approaches to ASR has been created. This methodology, as detailed in chapter 3, allows users to identify the best combinations of phonetic features for characterising speakers from different segments of speech. The second contribution of this thesis is three examples of how this methodology can be used to identify tailored combinations that characterise sociophonetically distinct speakers and speech. The three tested groups are:

- Southern Standard British English (SSBE)-accented speakers producing read speech.
- SSBE-accented speakers producing free speech.
- West Yorkshire-accented speakers producing free speech.

There are multiple applications of these findings. Most importantly, this thesis provides novel and fundamental evidence for understanding the phonetic bases of how voices differ. Secondly, it proposes a bespoke, bottom-up way of selecting features on a group-by-group basis for ASR. Thirdly, the tailored combinations of phonetic approaches identified here could now, theoretically, be measured and fused with ASR systems in scenarios where the target speaker and speech matches the sociophonetic profile of one of the tested datasets. Taking and including these tailored phonetic approaches could provide the option for increased explainability, should triers-of-fact require such additional information, with minimised cost to performance. That said, this thesis employs controlled data; thus, the specific combinations generated here serve more as proof-of-concept that this approach can work rather than practical, usable approaches. Future work should test more forensically-realistic data, and this leads to the final major application of these findings: with these three examples as templates, future research could now test different datasets to identify more tailored phonetic approaches for different speakers and speech.

This thesis is now structured as follows. The following chapter reviews pertinent literature to this thesis, mostly focusing on previous ASR studies which explore the different phonetic features, segments, and sociophonetic variables that are tested in this thesis. Chapter 3 then details the developed methodology for testing and validating novel combinations of phonetic approaches for ASR. Chapters 4-6 then detail the results of this study in light of data extractability, how these (socio)phonetic approaches perform individually, and how these approaches can be combined to optimise ASR performance with the three datasets. This will then lead to a full discussion in chapter 7 of how phonetic approaches can be combined in

optimised ways, why certain trends emerged across the data regarding the performance of the selected features and phonemes, how these optimised phonetic approaches can now help explain ASR output, and how these phonetic approaches can be practically fused with current ASR systems without any detrimental impact to performance.

## 2. Literature Review

Given that this thesis is exploring phonetic features and their capacity for ASR from an angle of eventually supporting ASR, literature concerning modern ASR approaches, and the currently minimal use of phonetic approaches, is reviewed first in (2.1). This is done by providing:

- A general explanation of ASR today.
- An overview of the history and evolution of ASR approaches from their phonetically-informed origins to the fast and efficient approaches used today.
- A look at current state-of-the-art approaches to ASR with particular focus on the features used and how these perform with speed and accuracy at the cost of a fully-explainable approach.
- A look at the findings of other researchers and developers who are already incorporating phonetics into ASR approaches.

This section helps identify where phonetic approaches to speaker recognition began and how they once informed ASR, how these approaches could still assist in explaining ASR output today, how other research projects and authors are already improving ASR explainability using phonetic approaches to speaker recognition, and how present speaker recognition research builds on both the history of ASR and on the work of current researchers applying phonetic theory to ASR.

Following this short section, the rest of the literature review (2.2-2.5) is dedicated to establishing the full range of phonetic approaches that are explored by this thesis for ASR. This is done by exploring literature on how explainable and accurate a number of phonetic features have already proven to be for ASR. Additional attention is also paid to how

responsive these phonetic features are to capturing sociophonetic differences and how these features can be measured from different phonemes in speech. This review therefore enables this thesis to explore novel, combinatory phonetic approaches to ASR from a feature, segment, and sociophonetic perspective. Finally, the final section of this chapter (2.6) summarises the directions of this thesis based on all of the reviewed literature.

## 2.1. ASR

ASR refers to the technological process of comparing and verifying speakers based on their unique vocal characteristics. Using many features and methods that are detailed extensively below, modern ASR systems follow the following broad processes detailed by Kamiński and Dobrowolski (2022):

1. Speech signal acquisition, where the recordings of speakers who will be identified or verified are collected along with the training database.
2. Signal pre-processing, where silence is removed and the frames of speech for analysis are selected.
3. Extraction and selection of distinctiveness features, wherein optimised features and measurements for recognising the speakers are created automatically for use in ASR. The nature of these features is discussed throughout this section.
4. Decision-making, wherein the probability that a signal came from one speaker or another is calculated using the generated ASR features.

More detail surrounding how ASR systems utilise, vary along, and developed towards the use of these stages is covered throughout the following account of the history of ASR, building up explanations and detail about systems that are relevant to contemporary forensic ASR and what techniques they employ. Shaver and Acken's (2016) succinct account of ASR's development over the past century is consulted to track ASR developments in this section.

The importance of this section is to highlight how ASR development has been motivated to employ advances in computer science and engineering in order to improve validity, but moving away from phonetics has led to a decline in explainability. This section ultimately draws attention to the strengths and weaknesses of both modern ASR approaches and traditional phonetic analyses: the ease of validating (but difficulty explaining) ASR methods and the ease of explaining (but difficulty validating) traditional phonetic analysis.

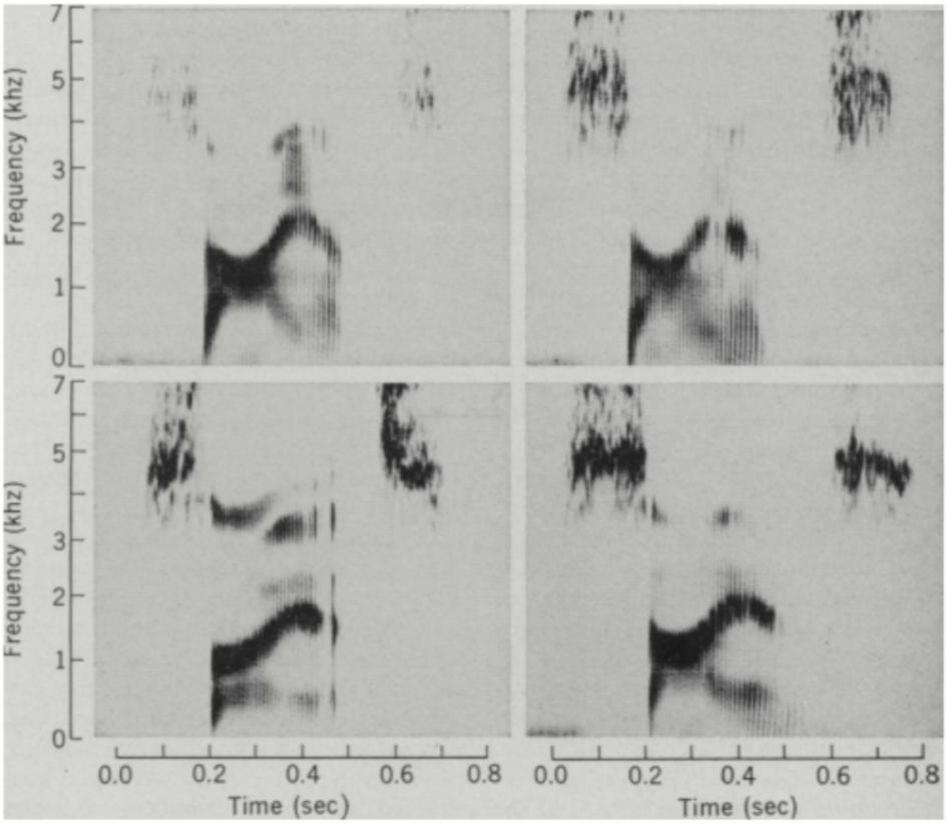
### 2.1.1. The History of ASR

Starting with the origins of ASR, the field was originally born out of phonetic theory and from work that was not, initially, automated. Unfortunately, this early work is plagued by serious methodological failure and its developments have been widely denounced by the forensic phonetics community as a result. The most prolific example of this early work is Kersta's (1962) study on 'voiceprints'. Working for Bell Laboratories, Kersta (1962) developed the first approach to manual speaker recognition. His method rudimentarily observed basic phonetic attributes of a speaker's voice (frequency and intensity) as a means of profiling them. Measurements of these phonetic features can be valid when measured effectively; as is discussed, they are still seen in modern ASR research today. The problem here, however, arose from how ineffectively Kersta (1962) measured these features: he vaguely eyeballed spectrograms, which he labelled as 'voiceprints', as a means for recognising speakers from speech. Spectrograms are visual representations of speech laid out in a specific way: the x-axis charts time, the y-axis charts frequency, and the z-axis charts intensity. These axes can be measured in milliseconds, kilohertz (kHz), and decibels (dB) respectively. Examples of voiceprints are included and criticised below.

Kersta (1962) claimed that the voiceprint of a new utterance recorded from a given speaker could be visually compared to older voiceprints of the same utterance produced by the same

or different speakers as a means of recognising whether the same speaker or a different speaker produced the new utterance. Unfortunately, this claim is empirically and fundamentally flawed. Not only is eyeballing, or even detailed visual inspection, too unspecific to be used as forensic evidence, it does not work even when attempting to employ it as a methodology for speaker recognition. Bolt et al. (1969) show this through a visual comparison of the four voiceprints found in Figure 1 below. These voiceprints are of three different speakers producing the word "science": the top two are produced by the same speaker and the bottom two are produced by two different speakers. According to Kersta (1962), the top two 'voiceprints' should appear the most visually similar because they were produced by the same speaker. However, here they actually appear more visually distinct than similar; especially in amplitude. This highlights the first problem: speakers never produce the same utterances in exactly the same way every time they produce them. Any differences in frequency or amplitude will therefore result in a visually-distinct second 'voiceprint'. In further critique of Kersta's (1962) 'voiceprints', the bottom two 'voiceprints' are not visually distinct enough from the first 'voiceprint' to conclude confidently whether they were produced by different speakers or not; the bottom two 'voiceprints' could easily have been produced by the same speaker as the first 'voiceprint' but louder, given the visual change in intensity. The overarching problem, therefore, lies in the inaccuracy of the eyeballing method: without a quantifiable and verifiable methodology for the actual comparison of 'voiceprints', Kersta's (1962) method cannot reliably capture the similarities between samples of the same speaker's speech or the differences between different speakers. At best, the only reliable conclusion that can be drawn from these 'voiceprints' is that the same utterance was produced each time.

**Figure 1: Spectrograms Produced by Three Different Speakers Saying "Science" (Bolt et al., 1969)**




Kersta's (1962) method was therefore denounced by researchers like Bolt et al. (1969) for being too unreliable and methodologically problematic. The initial implementation of traditional phonetic analyses is therefore immediately poor. All that was taken from this study moving forward was the simple method of comparing speakers, the use of the spectrum, the measurements of time, frequency, and intensity, and the clear need for a quantifiable and automatable approach; everything else is too problematic.

It would take a further 15 years for the first arguably successful approach to ASR to arise. As reported by Doddington (1985) concerning their earlier work from 1977, Doddington (1985) also chose to analyse the spectrum but did so with a much more successful methodology for recognition. Moving away from vague and undefined visual comparisons, Doddington (1985) used Digital Filter Banks (DFBs) that awarded much deeper analyses. DFBs separate a

recording of speech into separate frequency sub-bands that can be analysed separately and independently using frequency and intensity. Note that these are the same phonetic features seen in Kersta's (1962) method; however, Kersta (1962) was effectively analysing the entire spectrum all at once. Using DFBs, Doddington (1985) instead broke the spectrum down into smaller domains for separate, quantified analyses. This approach awarded much more specificity in speaker recognition. It also retained explainability; frequency and intensity are phonetic features that are already well-understood.

Doddington's (1985) measurements taken from the DFBs of one recording are then compared to those taken from another recording to judge whether the same speaker produced both utterances or if the utterances were produced by different speakers. For the comparison task, Doddington (1985) employed a more defined metric: Euclidean Distances (EDs). These are measurements of (dis)similarity between two groups, here the data from the DFBs taken from the two recordings, calculated as the shortest possible distance between the groups. Shorter distances indicate more similarity between the groups, which here would show that the same speaker produced both utterances. By contrast, larger distances indicate more dissimilarity between groups, which here would show that different speakers produced the utterances.

With the measures of (dis)similarity collected as EDs, the success rate of these EDs for speaker recognition is then calculated using Equal Error Rate (EER). EER is the combinatory score of False Acceptance Rate (FAR) and False Rejection Rate (FRR). FAR is when a different speaker is incorrectly recognised as the same speaker. Conversely, FRR is when the same speaker is not recognised as the same speaker when they should have been. Fewer FARs and FRRs are therefore better: an EER of 0% represents perfect performance, but 50% represents failure. This is because 100% would represent the direct opposite of perfect performance; one could therefore easily achieve perfect performance by doing the inverse of this.

The use of traditional phonetic analysis features here may make the recognition process more explainable, but it is difficult to validate. Whilst EDs are an improvement over Kersta's (1962) eyeballing methods of validation, they are outdated today; speaker recognition tasks validated using these metrics do not adhere to modern standards. This is because EDs only take into account an overall degree of (dis)similarity; many, like Morrison and Enzinger (2019), write that modern metrics should take into account the differences between speakers with more specificity. Similarly, EER has also been deemed inappropriate for assessing ASR performance for forensic purposes by Morrison et al. (2010) on the grounds that it does not weigh the different strengths of these FAs and FRs; for EERs, a severe FA or FR is valued as equal to a weaker FA or FR. Log-Likelihood Ratio Costs ($C_{llr}$), alternatively, attributes more weight to more severe FAs and FRs. This modern metric, deemed the modern standard by Morrison et al. (2010), is used in this thesis and is discussed in greater detail later.

Keeping these critiques in mind, Doddington (1985) still used these feature measurements and performance metrics to generate a speaker recognition approach that yielded an EER of <1% in 1977. Though problematic, Doddington (1985) still showed that speaker recognition approaches could be validated in more verifiable ways by 1977. That said, it still cannot be concluded that the DFB-based approach designed by Doddington (1985) performs impressively by modern standards because of the above problems surrounding its outdated feature measurements and performance metrics. Furthermore, any conclusion about the performance of Doddington's (1985) approach is made more problematic when one considers the data that Doddington (1985) used. Doddington (1985) exclusively used short recordings of text-dependent speech, wherein the speaker is told exactly what to say in a scripted, controlled environment. The analysis of text-dependent speech is therefore simpler than that of text-independent speech, wherein the speaker is producing speech freely. This raises concerns about the applicability of Doddington's (1985) approach: as a result of this

increased data control, better performance is to be expected. Doddington's (1985) approach may therefore not perform as strongly outside of this controlled data environment, especially when more realistic, forensically-diverse data is used. This critique holds true for any study, even today: as Morrison and Enzinger (2019) write, ASR systems always perform best on the data they were trained on and performance declines outside of this. Furthermore, it is well-documented that systems trained on text-dependent speech perform much worse when they are tested on text-independent speech (Dufour et al., 2014).

This is not to say that there are no benefits to analysing text-dependent speech. In commercial settings, for example, a fast, simple, and well-performing ASR system trained on text-dependent data would be beneficial for scenarios where a customer is uttering a text-dependent password that can be compared to a previous recording of that password. Judicially, however, text-dependent speech has less use. Suspect and offender speech is unlikely to be so controlled: a phone call, for example, will likely only contain natural, uncontrolled, unscripted, text-independent speech. Greater focus on text-independent speech can also benefit commercial audiences anyway: being able to verify the identity of a customer from only their natural speech would offer a much more streamlined user experience that is less intrusive than the use of a password that the customer may not wish to utter in a public setting. As a result, whilst text-independent speech may present greater challenges in the form of more variability, training ASR on such data offers greater rewards for users in the form of wider applicability.

This tangential discussion of text-dependent and -independent speech is important because the recognition of this text-dependence issue in Doddington's (1985) study, plus the desirability of recognising speakers from text-independent speech for judicial and commercial ASR purposes, drove ASR research forward. It also shows that sociophonetic variables, here speech style, have always been important to consider in ASR. Problematically,

however, ASR researchers also recognised that using text-independent speech meant processing more variables, and current phonetic approaches limited the opportunities to do this.

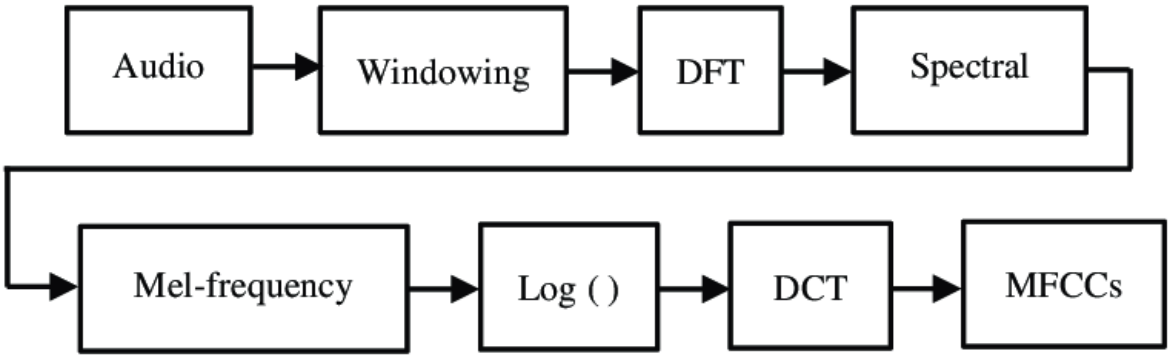
This processing research gap was one that computational and engineering-related researchers would now come to address. With this, an era of ASR research more concerned with ASR modelling and processing began. It was not an immediate change of goals, however; there were still some more developments related to ASR features that were established in this era that will first be discussed. Most notably, Luck's (1969) investigation into how analyses of the cepstrum could be employed in ASR research was overlooked at the time, but Luck's (1969) study continues to inform ASR today because it predicted the development of Mel-Frequency Cepstral Coefficients (MFCCs) which are discussed momentarily. Cepstral analyses separate spectral analysis, such as those seen above in Doddington's (1985) work on DFBs, into further component sources and filters. Theoretically, the source here is the underlying periodic waveform of voiced speech produced by vocal fold vibration in the larynx whilst the filter is what modifies this waveform; this can be different manners of articulation, for example. These filtering elements complicate the source by modifying it; if one wants to capture information about the source and the unique make-up of one's vocal tract, the effects of the filtering elements must be accounted for and subtracted, and vice versa for the information about the filters. Whilst this provides more avenues for measuring more elements of a speaker's voice more specifically, Titze (2008) writes that the proposed de-coupling of the source and the filter is not borne out of empirical research; it is a theoretical assumption. They point out that this theory is challenged by the fact that when measuring information about cepstral coefficients, information about the source and filter are captured together; they cannot be as cleanly divided as the theory suggests.

The practical potential of cepstral analyses, however, would go unnoticed for 5 years until Atal (1974) found that cepstral analyses outperformed all other approaches to ASR at the time: based on 50 milliseconds of text-dependent data, the cepstral analysis method accurately recognised the speaker 70.3% of the time. Atal (1974) also found that the cepstral analysis method's success rate increased to 98% when 500 milliseconds of data was used. At the time, this was the best-performing approach; however, this study's design unfortunately does not hold up against modern standards because it used minimal data from only 10 speakers, all of which was text-dependent.

The turn to ASR modelling and processing, and the influx of computational and engineering-based research, began in the 1970s as a direct result of the desire to approach more complex, text-independent speech. The cepstral analysis development was used as a springboard into studying more abstract representations of the voice using computational and engineering-based approaches that simplified the necessary processing. More specifically, cepstral analyses led to the first major advancement in this period of ASR's development: Mermelstein's (1976) Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs are crucial features for ASR generated via the following processes: firstly, firstly, **audio** is segmented into overlapping timeframe **windows**, typically around 20 to 25 milliseconds with a 10-millisecond shift.. Each of these frames is then multiplied by a window function to reduce edge effects and is then transformed into the frequency domain using a **Discrete Fourier Transform (DFT)**.,. This breaks each frame down into its component frequency **spectrum**, and from this, the power spectrum is computed. Next, a set of overlapping triangular filters called **Mel** filter banks is applied to the power spectrum. These filters are spaced according to the **Mel** scale, which approximates how human hearing perceives pitch: it is more sensitive to changes at lower frequencies and less so at higher frequencies. To reflect this, the filters are more densely packed in the lower frequencies. Each filter captures the total energy within its

frequency band by summing the power spectrum values it overlaps. These resulting energy values are then **logarithmically** scaled to mimic the human perception of loudness. As such, a general focus on interpretability remains with MFCCs. The final step is to apply the **Discrete Cosine Transform (DCT)** to the log-scaled filter bank energies. This transformation compacts the information by decorrelating the features and emphasizing the overall spectral shape. The output is a set of values called **Mel-Frequency Cepstral Coefficients (MFCCs)**, which represent the short-term power spectrum of speech in a perceptually meaningful way.The following visualisation in Figure 2 by Son et al. (2019) visualises these stages, and each stage in the above description has been flagged in **bold**.

**Figure 2: Stages of MFCC Extraction (Son et al., 2019)**




Technical explanations aside, MFCCs essentially represent a way of combining multiple important features into one measurement for ASR. Specifically, they represent a combination of frequency and amplitude measurements taken from the voice through the multiple transformations detailed above. These transformations and combinations still represent the information from the original signal, and are therefore still explainable to an extent. This is an important breakthrough for ASR: Kinnunen (2003) writes that, with traditional phonetic methods prior to MFCC-based approaches, it was computationally inefficient to analyse the spectrum with such specificity. Breaking it down into smaller, more computationally-efficient components, such as MFCCs which capture the entirety of the intensity spectrum's most

salient frequencies, allows for faster ASR that can tackle the needs of text-independent speech in ways that could not be achieved before.

All that said, MFCCs are not without controversy: recently, Gourisaia et al. (2024) found that MFCCs are too sensitive to background noise and this can affect the reliability of the feature extraction process. They recommend alternative measurements, such as Short-Time Fourier Transforms (STFTs), which they found to be more effective with noisy data. Other solutions involve improving MFCCs by normalising cepstral means, as Hautamaki et al. (2008) recommend, or compensating for the included noise by subtracting the average of the noise in each band (Nasersharif and Akbari, 2007).

The next big developments came in the 1980s and 1990s when ASR research began to focus more on modelling than on features. The first of these modelling techniques, developed in the 1980s, was the Hidden Markov Model (HMM). These were popular amongst ASR developers because HMMs created abstract probabilistic models of linear sequences of data, such as MFCCs. They did this in a reduced, conceptual, mathematical space that is much more computationally efficient than the original features taken from the actual recording; in short, this development sped ASR up even further.

At this point, it is worth acknowledging emerging debates at the time. Levinson (1994) overviews how the emergence of these more abstract approaches led some researchers at the time to declare this era as revolutionary based on the speed and accuracy of these newfound approaches. However, Levinson (1994) criticises these researchers for mistaking developments in computer hardware, which are responsible for the achieved speeds and accuracy, as developments in speech recognition. Levinson (1994) also draws specific attention to the use of unrealistic data, such as text-dependent speech, as further infringing on these conclusions about the strength of these technological advancements. However, whilst

Levinson's (1994) claim that further advancements were needed in ASR is strong, MFCCs were still born out of this era and they are still used in ASR today; this was still, therefore, a critical period for the advancement of ASR.

After HMMs came Gaussian Mixture Models (GMMs), a related major development that occurred in the 1990s. Furui (1997) succinctly summarises GMMs as weighted density sums computed from Gaussian components measured from speech. More specifically, these sums are the results of parametric statistical models that measure the probability distribution of normally-distributed, continuous dependent variables found in the original recording; the Gaussian components. These components are composed of the cepstral features previously discussed, such as MFCCs.

Alongside GMMs, another vital development in the 1990s for ASR modelling came in the form of Universal Background Models (UBMs). UBMs are representations of any data using any number of features that are pooled from any number of sources; they effectively offer a generic, widespread snapshot of what data looks like across a large corpus. For the purposes of ASR, UBMs represent features of the voice pooled from multiple speakers and therefore offer an image of what speech generally looks like. UBMs are regularly combined with GMM approaches to create GMM-UBM ASR systems. First, the UBM is a large GMM trained on speech from a wide range of speakers. This captures a general picture of speaker-independent characteristics. A GMM for a target speaker is then also created by adjusting this UBM using recordings from the target speaker. When a new recording from the target speaker is evaluated, the system compares how well the acoustic features of the new recording fit that target speaker's GMM versus the generalised UBM. If the new recording is more likely to fit the speaker model than under the UBM, the speaker is more likely to be accepted; if it's more likely to fit the UBM, the speaker is more likely to be rejected. These calculations are made using $C_{llr}$, which will be discussed soon. After GMMs, the next notable

development to discuss is i-vectors. These were popularised in the early 2010s as a way of simplifying ASR processing further: instead of modelling elements of speech like previous approaches have, this approach takes the whole recording, irrespective of its content, and represents the speaker-dependent and channel-dependent features of the recording in the lowest dimensional space possible. This is a computationally-efficient approach that abstracts measurements of a recording without consideration for what is and is not the voice; it is fast and efficient as a result.

Evolving from this, the most recent major development in ASR has been the use of Deep Neural Networks (DNNs) and classifiers. These can identify the most important features of a speech signal to measure for ASR purposes all on their own; minimal input from researchers and developers is needed. Such approaches have proven extremely powerful performance-wise, as seen in the results of Morrison and Enzinger's (2019) paper comparing the performance of modern ASR systems, but they are also less explainable. The measurements created by these DNNs and classifiers cannot be fully explained by their human developers: how these measurements are created by the system and the data that they use can be understood, but it is unclear what information from the data has actually been considered important or unimportant enough to be included or rejected in the created measurements.

Exploring these DNNs and classifiers further, they convert measurements like MFCCs into even more abstracted representations of data called x-vectors (Shaver and Acken, 2016). These x-vectors are singular measurements that combine multiple measurements, such as MFCCs, into a further unitary measurement. By reducing all of these measurements of speech into one singular vector, the computational cost of comparing a new recording to a prior recording and a UBM is reduced significantly, allowing the system to perform much faster. By containing all of this information, they are also extremely accurate, as seen from Morrison and Enzinger's (2019) tests of multiple ASR systems.

The emergence of modern ASR approaches like x-vectors has made ASR better than ever before in terms of accuracy and speed. However, through abstraction, a decline in explainability has occurred. The previous chapter highlighted literature from Morrison et al. (2021) suggesting that explainability is less important than validity, but literature was also highlighted from Eldrige (2019) and van der Veer et al. (2021) suggesting that some degree of explainability may improve declining public trust in ASR systems. Thus, the present thesis seeks to explore ways of reintegrating explainability by consulting traditional phonetic analysis methods that are more explainable, as the rest of this chapter establishes. However, from the literature reviewed in this section, this thesis recognises the detrimental effect that phonetic approaches have had on ASR historically and how, from a performance standpoint, they have also been far surpassed by modern ASR approaches. As such, this thesis primarily focuses on the *potential* of phonetic approaches for speaker recognition. From this, it then goes on to explore the practicality of implementing these phonetic approaches to speaker recognition alongside current ASR approaches. Thus, this thesis does not suggest that traditional phonetic analyses should be used as a replacement for any element of modern ASR, nor does it suggest that traditional phonetic analyses can be used as an end-to-end ASR approach. It instead suggests that explainable phonetic approaches found to be successful for speaker recognition could complement ASR output by adding an additional element of explainability to the system. This is how the present thesis will build upon the history of ASR. The developments seen in traditional phonetic analyses for speaker recognition will now be reviewed in the following sections.

## 2.2. Considering Phonetic Approaches in ASR

Phonetic theory can achieve a degree of explainability in ASR, and reviewed research has already begun to show this. As discussed in previous chapters, Hughes et al.'s (2019a) study successfully incorporated mean harmonics-to-noise ratio, a phonetic feature correlated to non-modal voice qualities such as breathy voice and creaky voice, alongside commonplace ASR features like MFCCs. Zarate et al. (2015) found that human listeners find phonation information critical in speaker characterisation tasks, so it is expected that this phonetic feature can perform well. Hughes et al. (2019a) were not looking to replace current ASR approaches; only explore and assist them using phonetic theory. They used EER to evaluate the performance of an ASR system with and without the laryngeal voice quality features in same-speaker scenarios and different-speaker scenarios. They also replicated their procedure 20 times to factor in sampling variability. As they predicted, performance was unchanged for most replications; however, as mean harmonics-to-noise ratio can be correlated to an explainable and perceivable element of a speaker's voice, this means that including this measurement as part of the ASR input added an additional element of explainability to the system with no detrimental impact to ASR performance. Through the lack of change in performance, this also implies that the current ASR approach may already capture information about these laryngeal voice quality features, so these are now better understood too. This explainable, assisting position of phonetic information is what this thesis seeks to explore but on a bigger scale, incorporating more features, analysing speech from different sociophonetically-controlled groups, and considering phoneme-level analyses, as the coming sections establish.

In two of Hughes et al.'s (2019a) replications, the combination of prior methods and the included phonetic approaches yielded an impressive EER of 0%. This could suggest that phonetic approaches can even improve upon performance, given that this was the best EER

reported in the study; that said, it must be noted that this is an exceptional result. Such extreme performance may also be due to the fact that the data was controlled and used repeatedly in all scenarios. In short, the materials were favourable for performance. The ASR system tested was also not current to the time of study and may not have reflected current ASR performance. Similarly, the use of EER here is outdated: as discussed in the previous chapter, Morrison and Enzinger (2019) write that EER cannot capture the different severities of different errors like $C_{llr}$, a more modern metric.

It must also be discussed that, despite over half of the replications being positive for the future use of phonetic approaches, 5 of Hughes et al.'s (2019a) replications saw performance actively decrease with the inclusion of laryngeal voice features. This may indicate that phonetic approaches could be harmful to ASR. Despite this, Hughes et al. (2019a) also found that the rate by which performance worsens when the audio quality is reduced is always slowed by including laryngeal voice quality features. This indicates that phonetic approaches to speaker recognition may still prove useful for improving the overall performance of ASR, particularly when poor-quality audio is used.

Finally, Hughes et al. (2019a) also found that, when considered in isolation as the only features used for ASR with no other modern approaches, laryngeal voice quality features can be used to create an ASR system with an EER of 6%. This is solid performance, but as seen, better performance was observed in combination with ASR approaches. As such fusions are possible, their study therefore overall shows that there is no need to consider phonetic approaches in isolation for practical applications anyway; they can likely be combined with ASR systems, offering an additional element of explainability to the system, without impact to system performance.

Other studies have also investigated how phonetic approaches can be employed to improve ASR explainability. For example, Gonzalez-Rodríguez et al. (2014) analysed the false acceptances and rejections of an ASR system from a phonetic perspective to try and diagnose why these false hits were returned. They found that differences in voice creakiness, a perceivable and explainable phonetic voice quality, could account for whether a speaker was accepted or rejected. This phonetically-informed finding improves the explainability of the tested ASR approach: it is now known that the methods used by this particular ASR system must not measure voice creakiness. Explainability this way can help ASR developers improve the reliability of their ASR systems: it indicates that performance should improve if information relating to voice creakiness is measured. This could be achieved through fine-tuned adjustments to more powerful ASR features, like MFCCs, or by integrating related phonetic features, like mean harmonics-to-noise ratio. Therefore, the issues facing ASR approaches can also be diagnosed and explained using phonetic theory.

Continuing this theme of using explainable phonetic theory to explore these modern ASR systems, research in this area can be generally categorised in two ways:

1. Using explainable features, like phonetic features, either to diagnose how these modern approaches work or to add an additional element of explainability to the system. This was seen in Hughes et al.'s (2019a) above study of mean harmonics-to-noise ratio.
2. Looking at the errors generated by an ASR system and inferring what explainable phonetic aspects of the voice they are failing to observe. This was seen in Gonzalez-Rodríguez et al.'s (2014) above study of voice creakiness.

The following study by Skarnitzl et al. (2019) falls into the first area of research above. They aimed to interpret the features employed by VOCALISE, a modern ASR system. They

measured the EER performance output of this system in two languages under two conditions: one with the data simply as it was recorded, one with the fundamental frequency (*f0*) of the recording raised by 4 semitones and the intensity raised by 8%. The two languages consulted were Persian and Czech and the databases for both contained 100 speakers. The speakers were all male and were recorded under the same conditions to reduce the confounding effects of any variables relating to speaker identity and recording conditions.

Skarnitzl et al. (2019) found that VOCALISE performed better in Persian in the baseline data condition with an EER of 1.12% than it did in the Czech baseline condition where the EER was 4.2%. However, the inverse was true when the *f0* and intensity were altered: the Persian EER raised to 2.44% and the Czech EER lowered to 2.42%. These findings are important to ASR for a number of reasons: firstly, this study shows that two well-studied phonetic features, *f0* and intensity, are integral to the processes employed by this ASR system because altering them in both languages affected ASR output. Secondly, it shows that different languages must have different needs in ASR that VOCALISE cannot holistically accommodate, but these differences can be accounted for using phonetic features: *f0* and intensity proved necessary for ASR in Czech because performance improved, but not ASR in Persian because performance declined.

This study broadly demonstrates that phonetic approaches to speaker recognition can be used to support ASR for two reasons: firstly, it shows that the output of modern approaches can be explained to some degree with phonetic approaches. By altering *f0* and intensity and seeing a change in the output, this provides a direct insight into how the ASR system worked, specifically that the present system must consider some measurement of *f0* and intensity in some way and to some extent when characterising the speaker. These are perceivable elements of the voice that triers-of-fact know of, and it may therefore help boost their trust in ASR output by explaining that this phonetic information is important to the output. Secondly,

it identifies language diversity as a variable that can directly affect the output of an ASR system. Acknowledging this, ASR output could be made more explainable by profiling and tailoring approaches to specific language varieties that triers-of-fact recognise.

### 2.3. Considering Phonemes in ASR

Thus far, the focus of this review has been primarily on phonetic features for ASR; however, it is equally important to consider where these phonetic measurements should be taken from, and little research has approached this from a fully phonetically-informed perspective. Most ASR systems use measurement periods wherein the content of the speech used is typically random, and considerations for the data used are typically limited to whether the speech is text-dependent or text-independent, as has been seen throughout many of the above studies. However, phonemes, in particular vowels, have historically been used in speaker recognition studies, but the literature on their use in ASR is very limited. Vowels are expectedly useful for speaker recognition because they offer insights into the shape of a speaker's vocal tract as well as capturing sociophonetic information about a speaker's accent. For example, speakers from the North and South of England differ in the vowel they use in words like 'trap': Northern speakers use /æ/ whilst Southern speakers use /ɑ/ (Koshy and Tavakoli, 2021). Furthermore, vowels generally represent the different open articulations that a speaker uses and they can also be measured quantifiably with phonetic features like formants which are discussed later in this chapter.

That said, one notable-but-outdated study by Paliwal (1984) showed that vowels are particularly useful for ASR and that certain vowels outperformed others. Paliwal (1984) used EDs to compare formant measurements from eleven vowels based on their discriminatory power. From these EDs, Paliwal (1984) identified the following rank order of effectiveness for vowels in ASR: /ə/, /ʊ/, /ɪ/, /u/, /o/, /ʌ/, /a/, /ɔ/, /æ/, /i/, and /ɛ/.

Problematically, few studies since Paliwal's (1984) have continued research into the use of vowels. Furthermore, Paliwal's (1984) study itself has significant constraints and problems. Most notably, the paper is rudimentary: little is said about the data it was collected from, the number of speakers, and why, exactly, this rank order was found. Due to limited research in this area, little is still known about why different vowels perform differently for ASR tasks. Today, ASR studies focus more on how to take successful measurements from any segment of speech as opposed to how individual segments may behave differently in ASR.

Additionally, Paliwal's (1984) study has also not been validated using more modern metrics than EDs. This is where this thesis updates Paliwal (1984): $C_{llr}$, the current state-of-the-art performance metric for validating ASR systems as recommended by Morrison and Enzinger (2019), is used to assess the performance of vowels for ASR in this thesis. Paliwal (1984) also only considered formants despite the fact that, at the time, other phonetic features like *f0* and intensity were in use. This is another way in which this thesis updates this study: it explores novel, combinatory phonetic approaches to ASR which consist of multiple features, and phonemes. Finally, Paliwal (1984) did not factor in phonological context, namely the different stresses that can be put on different vowel sounds. This is also explored in this thesis.

Beyond vowels, however, there has been some interest in nasals for ASR. For example, Eatock and Mason (1994) found that English nasals outperformed vowels when used as the input data for an ASR system with an average EER of 18.8% compared to the average EER of vowels which was 21.1%. Of these nasals, /ŋ/ performed best with an EER of 19.7%. /n/ had an EER of 23% and /m/ had an EER of 23.2%. All of these nasals out-performed /a/ in the study, which yielded a much higher EER of 29%.

Alsulaiman et al. (2017) also looked at the effectiveness of Arabic nasals in ASR. Using Multi-Directional Local Features with Moving Averages (MDLF-MAs), they found that /n/ scored a high Recognition Rate (RR) of 88% whilst /m/ scored 82%. However, /a/ scored an RR between them of 84%. Comparing this study to the above study by Eatock and Mason (1994), this again shows that different language varieties have different needs but now at the segmental level; not just the feature level, as Skarnitzl et al. (2019) found in their study of VOCALISE. Specifically, nasals appear more useful for ASR in English than in Arabic. Gallardo (2015) provides a theoretical explanation for why nasals perform well: the resonance cavities used to produce nasals differ significantly speaker-to-speaker. However, even though explanations exist for why vowels and nasals perform well, little research exists on why different vowels and nasals outperform each other.

Turning now to important and notable work by Heeren et al. (2022), they found that a speaker recognition system that uses formant measurements, specifically formants 2 and 3, taken from combinations of vowels and nasals, specifically /a:/, /e:/, /n/, and /m/, led to a solid performance: a $C_{llr}$ of 0.22. Such work is critical for justifying the present thesis' work: this study shows that a system consisting of combinatory phonetic features, here formants, measured over phonemes, here vowels and nasals, can generate a well-performing speaker recognition system. The present thesis, however, will incorporate more features and more vowels. Also, as discussed, competitive performance is not the goal of this thesis; instead, the goal here is optimising performance of phonetic approaches for ASR. This will show that explainable and perceivable phonetic approaches to ASR can be integrated without detriment to more powerful ASR approaches. This study was also conducted on Dutch data; the present study will look at English and incorporate sociophonetic variation, as is discussed in the following section.

Heeren (2020) also looked at the context of phonemes, in particular comparing tokens of /a:/ in content words and function words. They found that this phoneme in Dutch is typically shortened in function words, but speaker recognition performance remains consistent. Thus, based on Wilemijn Heeren's work above, it is clear that vowels, and phonetic measurements taken from vowels, are important, but their context is not. This justifies the present thesis' approach to isolate and study every token of every chosen vowel, as discussed in the next chapter.

These studies overall show that explainable phonemes, such as vowels and nasals, can be employed in speaker recognition studies which can then inform ASR. The present study therefore uses phonemes as the phonetically-informed basis for its testing of phonetic features for speaker recognition which can later inform ASR. This allows the thesis to identify which explainable features work best on a given perceivable vowel which is in a recording that triers-of-fact hear.

### 2.4. Considering Sociophonetic Variables in ASR

Turning now to sociophonetic variables, it has already been noted that the style of the speech tested can affect the performance of an ASR system, specifically whether that data is text-dependent or text-independent. It has also already been seen that the language variety can affect the performance of an ASR system with Czech and Persian (Skarnitzl et al., 2019). However, different types of speakers within the same language can also affect the performance of ASR systems: broadly, Misnikov (2019) found in their exploration of automatic approaches to recognising speech that the performance of gold-standard speech recognisers is affected by sociophonetic variables like accent, age, and sex. More specifically, Misnikov (2019) found that the speech of young female speakers with accents originating from the West of the USA was easiest to recognise from a database of US speakers, but they

did not find why. Regardless, this study ultimately indicates that sociophonetic information, particularly that relating to qualities such as age, sex, accent, and ethnicity, are important to consider in speaker characterisation tasks for ASR. Automatic speech recognition technologies generally crossover with ASR technologies, so the findings here expectedly pertain to ASR too; however, they have rarely been directly explored in ASR.

Looking more specifically at the few studies that have investigated the impact of sociophonetic variables in ASR, Safavi et al. (2018) found that the EER of a GMM-UBM ASR system decreased as the age of the participants increased; this means that current approaches may be tailored to older speakers, and different methods may be required to improve the effectiveness of characterising younger speakers, akin to how VOCALISE could to be tailored to Czech and Persian (Skarnitzl et al., 2019). This may be due to a lack of age variety in the UBM data. Sociophonetic explanations of errors such as this are important to ASR development, hence why the present thesis also seeks to explore the importance of sociophonetic variables. Problematically, Safavi et al.'s (2018) findings challenge Misnikov's (2019) above findings that ASR systems perform better for younger speakers. However, this contradiction further justifies the need to explore sociophonetic variables in ASR more thoroughly and how performance can vary as sociophonetic variables are considered.

Kajarekar et al. (2006) found that when the same speaker speaks with a different accent to that which they recorded initial speech for with an ASR system, the EER increases; this suggests that ASR systems are sensitive to accent information. This was a matched-guise study; thus, as it was always the same speaker in each scenario, this meant that accent was the only variable that could account for the changes in performance. Sociophonetics has therefore been used here to diagnose and explain ASR system output.

Finally, Hutiri and Ding (2022) found that ASR demonstrates sociophonetic biases and performs worse with female, non-US speakers as a result. These researchers go on to recommend that ASR could, in the future, be tailored towards sociophonetic variables by using subgroups in training datasets. This principle of sociophonetic tailoring informs the sociophonetic portfolio approach that this thesis adopts for ASR purposes.

These studies indicate that ASR performance is sensitive to, and evidently biased towards, sociophonetic information. As such, ASR systems could be tailored towards sociophonetic variables like age and accent as well as style. Based on these findings, different ASR approaches could be generated and explored for different sociophonetically-tailored groups. One can build optimised portfolios of ASR approaches to accommodate the different needs of different speaker groups. This thesis does this using phonetics.

### 2.5. Phonetic Toolkit

This section will now explore the specific phonetic features that are employed in this thesis' analysis. These features are *f0*, formants, intensity, mean harmonics-to-noise ratio, mean autocorrelation, jitter, and shimmer. The goal here is to explore speaker-specificity across multiple features. Each feature is discussed in light of how it can be measured, what it perceptibly and physiologically correlates to in speech, how it has already been studied for ASR, where one should take measurements of each feature from a segment of speech, and how sociophonetic differences can be captured by each feature. Exploring each feature's perceptible, physiological, and sociophonetic correlates is important to establish their explainability.

An acoustic-based toolkit has been selected based on the long-term success of such approaches in the field: Sambur (1975) writes that, due to how they theoretically capture variations between vocal tracts, acoustic features like these should capture speaker

differences that can characterise speakers. More detailed explorations of each individual feature are the focus of the coming sections.

Importantly, Sambur (1975) also found that vowels are effective segments of speech to take phonetic measurements from because they are voiced, capture source signal energy from the vocal fold vibrations, and are directly influenced by the shape of the speaker's unique vocal tract makeup. Sambur (1975) used linear distances, similar to those seen above with EDs, to determine how well acoustic features would perform when characterising speakers. Using only acoustic measures on vowels, their approach performed well for the time: the error rate was only 0.312%. Sambur's (1975) study justifies the use of vowels in this thesis; however, this thesis expands on studies like Sambur's (1975) by investigating different combinations of features and vowels, sociophonetic variables, and a more modern metric for performance validation ($C_{llr}$).

The toolkit is composed of features that have been selected based on their theoretical promise and empirical evaluations of their performance, all of which are explored in the coming sections. These acoustic features will all be explored with vowels for the present study due to their use in prior acoustic phonetic research and, for technical reasons that are explored in the next chapter, most of these features have to be taken from voiced speech like vowels anyway. They could also be used on nasals, as some studies showed above, but not voiceless sounds like many consonants. This is in some way a limitation but, as Gao et al. (2012) write, consonants may contain less useful information for ASR purposes anyway because speakers will typically produce consonants with minimal variation in a given context and will likely do so with little vocal activity; especially when those consonants are voiceless. Almadeed et al. (2016) write that vowels, on the other hand, use the whole of a speaker's vocal tract in production and thus capture more speaker-specific information. Thus, aiming to measure

these acoustic features on vowels is more fruitful for the given task of characterising speakers with phonetic information.

### 2.5.1. *f0*

Looking first at fundamental frequency (*f0*), this is an acoustic phonetic feature that, physiologically, quantifies a speaker's rate of vocal fold vibration. Perceptibly, it maps on to a speaker's perceived pitch: faster rate of vocal fold vibration has higher perceived pitch whilst a slower rate of vibration has a lower perceived pitch. Quantitatively, this rate of vibration can be measured in Hertz (Hz) and this measurement counts the number of cycles of vibration activity per second; thus, perceptibly higher voices have a higher rate of vibration and a higher average Hz. Perceptibly lower voices have a lower rate of vibration and thus a lower average Hz. This is perceptible information, observable by triers-of-fact, that enables *f0* to be explainable.

*f0* has frequently been employed in ASR: for example, Atal's (1972) study explores an early ASR system that exclusively employed *f0* to attain a 97% success rate. This illustrates that phoneticians have employed *f0* in ASR systems since the beginning of ASR research and have done so with some success, but this study shares many of the problems already seen in other early ASR research: it only used 10 speakers, only used data that was text-dependent, and only used 6 recordings from each speaker. Lacking data is especially problematic because *f0* is very variable within speakers; there is likely not enough data here to capture how variable *f0* can be.

Turning to more modern studies, Zhu et al. (2009) found that *f0* can still be used to improve the performance of ASR systems. Specifically, they found that when ASR systems use MFCCs alongside more phonetically-informed features, such as *f0*, they yield an improvement in recognition rate of 5% compared to the baseline system that exclusively

employed MFCCs. *f0* improving the performance of an ASR system using MFCCs is predictable: returning to Almadeed et al.'s (2016) above study, they specifically found that MFCCs do not capture significant information on speaker *f0*; thus, including this missing information expectedly improves ASR performance.

Other evidence for *f0*'s usefulness comes from Teixeira et al. (2013) who found that *f0* proves useful for finding unique characteristics in a speaker's voice that relate to pathologies, such as vocal fold polyps. This evidence is important because Teixeira et al. (2013) found that phonetic features that prove useful for identifying pathologies also tend to prove useful for ASR.

*f0* will not be used alone in this thesis, however. As shown throughout the above literature, combining multiple features for ASR is beneficial and this is the core exploration of this thesis. *f0* has rarely been considered on its own anyway, as seen above.

Turning now to sociophonetic considerations and the applicability of *f0*, this feature has already proven useful for characterising speakers of different ages, sexes, and accents. Specifically, Lortie et al. (2015) found that *f0* typically lowers with age but rises again in advanced old age. Schmid and Bradley (2019) also found that women tend to have a higher *f0* range than men. Finally, Grabe et al. (2000) found that some accents of English have more *f0* variability, namely Cambridge English, whilst Leeds English has relatively little. Sociophonetics is important to explore in ASR because it can make the process of characterising an individual speaker more specific. The process of reducing a pool of comparable speakers based on sociophonetic information has already been employed in forensic settings: as Atkinson (2015) writes, *f0* is used frequently to establish whether the offender is male or female to remove suspects that do not broadly fit the *f0* range of the offender they are trying to identify. Not only does the present study therefore exemplify how

sociophonetic considerations can be useful for ASR, it also shows that *f0* is a well-established, explainable phonetic feature that is already used as courtroom evidence.

Beyond sociophonetics, *f0* is also important for technical reasons in forensic settings, particularly when phone calls are involved. This is specifically pertinent to Lombard speech: this is where speakers instinctively increase their *f0* and intensity in a phone call due to an expectation of background noise or interference, even if none is present. This is important to consider because phone calls are frequently used as forensic evidence.

### 2.5.2. Intensity

Intensity measures the power of a speaker's speech. It perceptibly correlates to loudness, an explainable quality that triers-of-fact can perceive. Like the above features, intensity is also quantifiable and is measured in decibels (dBs). Also like the above, intensity has a long-standing presence in ASR research: seen in the early days of ASR research, Lummis (1973) details an approach to ASR that incorporated *f0* and intensity and yielded an average error rate below 1%. However, like other early ASR studies detailed above, such as Atal's (1972), the external validity of this approach does not hold up against modern standards: it is based on minimal data consisting of 152 text-dependent utterances and the session variability was high. Regardless, it still highlights the point that intensity has been historically considered and successfully employed in ASR to some degree. Furthermore, Lummis' (1973) study builds on Atal's (1972) above study by incorporating more data, thus giving greater credence to the use of *f0* and intensity both as phonetic features for ASR and as features that can be combined.

Intensity is also used in modern studies: for example, Jia et al. (2021) measured intensity in their approach to ASR alongside formants. Their approach proved accurate, though not as accurate as other approaches; however, the use of phonetic features here did prove useful for

speeding up an ASR system's training and recognition time beyond the speeds seen in less explainable modern ASR approaches. This indicates that phonetic features may, in fact, be processable fast enough to meet the demands of modern ASR users and may yield some performance benefits for ASR.

It is also worth noting that intensity's long-standing place in ASR research can be seen in MFCCs: as discussed above, MFCCs are representations of intensity, just abstracted; ASR researchers have therefore employed some measure of intensity to great success over all of the MFCC-incorporating studies listed above.

Like *f0* above, intensity can also vary between speakers based on sociophonetic variables that can distinguish groups of speakers from each other. For example, Chen (2007) found that men are louder than women; their intensity range was higher. With this in mind, consider the expected contrasts in the two feature measurements thus far: men have a higher intensity range yet a lower *f0* range whilst women have a lower intensity range yet a higher *f0* range. At a basic level, this illustrates how phonetic features could be combined to complement each other to aid in sociophonetically-informed ASR: as both features behave differently for each sex, in combination they therefore prove fruitful for distinguishing different sexes.

Regarding age, Lortie et al. (2015) found that age does not necessarily affect intensity measurements. However, age has been found to affect *f0* measurements above. This comparison between intensity and *f0* justifies the need to study multiple features and to combine them: if only *f0* is sensitive to the age distinction, but both features discussed thus far allow for more specific explorations of sex variation, then both features can be combined to identify as many specific differences as possible between these sociophonetic groups. This is because the combination of these features allows for the shortcomings of intensity analyses

to be mitigated whilst still employing this feature for its strengths, ensuring more detailed analyses for ASR.

In practical analysis, however, it is important to note that previous literature has found that intensity can be affected by confounding variables that must be controlled if one is to measure it reliably. For example, Titze and Winholtz (1993) write that microphone distance can affect measurements of intensity: the closer one is to the microphone recording their speech, the louder one's speech will be measured in dB from this recording. However, given the diversity of forensic data and the range of scenarios in which speech can be recorded, this poses a problem for using intensity for forensic purposes: recordings from different microphones at different distances from different suspects will be in use, and as such these recordings will vary in intensity as a result of these variables instead of anything speaker-specific. These issues related to the microphone and input levels are discussed in more detail in Chapter 3.

In terms of other confounding variables, Pfitzinger and Kaernbach (2008) found that emotional speech, such as aggressive speech, is louder than calm speech; thus, in addition to technical variables like microphone distance, one must also consider content variables such as the topic of speech. These studies therefore suggest that for one to explore intensity reliably, these variables must be taken into account in one's experimental design. However, whilst these could be considered problems, the performance of intensity under different conditions could, and should, be profiled. This is because tailoring the use of intensity towards different conditions has forensic and commercial benefits: phone microphones capture speech differently to studio microphones, incriminating evidence may be captured on a phone, and a customer may access their bank details via their phone. For any of these tasks, the speaker may also be in a different emotional state or at a different distance from the microphone capturing their speech. Whilst this presents a lot of variables to control and tailor

for intensity, the present thesis' proposed methodology could explore the effects of these variables in the future using different databases of speech to build a variety of tailored portfolios for the different scenarios that may impact intensity recordings.

Finally, channel mismatch poses an issue for intensity. Channel mismatch refers to the comparison of speech with different recording qualities, such as a comparison between high-quality studio audio and audio recorded from mobile phones. As Hughes et al. (2019b) found, intensity is expectedly sensitive to channel mismatch and its performance declines when channel is not kept consistent. This is particularly problematic in forensic settings when different recordings have different channels, such as different mobile phone microphones. Thus, intensity may not prove to be universally useful for ASR.

### 2.5.3. Formants

Looking next at formants, these are frequency peaks with high concentrations of acoustic energy that build up around particularly resonant areas of the vocal tract in a given configuration, specifically resulting from the interaction between the filter and the periodic, voiced, source signals. In terms of how formants are measured, they are also measured in Hz, like *f0,* which allows them to be quantified. Formants occur regularly, with one formant in roughly every 1000Hz band, and are labelled as formant 1, 2, 3, 4, 5, and so on. Perceptibly, Abhang et al. (2016) write that formants can characterise different vowels; as different vowels are produced with slightly different configurations of the vocal tract, the changes to where the concentrated acoustic energy is located can mark differences in vowel production. For example, formant 1 changes with vowel height and formant 2 changes with vowel frontness. As our vocal tracts all have unique physiological set-ups, formants offer a way of charting this vocal tract space during the production of a vowel that is unique to the speaker.

As such, this correlation is perceptible to triers-of-fact and can therefore offer an explanation for how vowels differ and why vowels sound different when produced by different speakers.

In terms of the use of formants in ASR, Ali et al. (2006) developed a successful method for ASR that was built around formant measurements. This system was very accurate across multiple languages, showing that formants can be effective for ASR. This study is important because it shows that formants can perform well irrespective of language variety; some features, such as *f0* and intensity, have been shown to work better in some languages more than others, as seen in Skarnitzl et al.'s (2019) comparison of Czech and Persian above. This is important to flag because, whilst sociophonetic tailoring is central to this thesis' exploration of ASR, it shows that some phonetic features, here formants, may prove to be universally useful irrespective of any tailoring.

Challenging this, however, is the effect of channel mismatch on formants. On top of intensity, Hughes et al. (2019b) found, formants are particularly sensitive to channel mismatch and their performance declines when channel is not kept consistent. This is particularly problematic in forensic settings when different recordings have different channels, such as different mobile phone microphones. Thus, formants may not prove to be universally useful for ASR.

Considering other critiques, Harrison (2013) found that there are issues with the measurement validity of formants. They used synthetic vowels, which therefore have already known formant values, and found that different measurement extraction tools vary in the values they output; the available measurement tools are therefore inconsistent with each other. This poses another issue for the use of formants in ASR that is addressed in the following chapter.

In terms of other uses in ASR, Jesse et al.'s (2014) i-vector ASR system achieved an EER of 18.1% on text-independent speech based on measurements of formants 1-3. However, higher

formants like formant 4 and 5 should theoretically perform best in ASR. This is because, as Lammert and Narayanan (2015) write, higher formants capture more speaker-specific information relating to the size and shape of their vocal tract. Problematically, however, there are some measurement issues regarding the higher formants: Derdemezis et al. (2016) found that higher formants are unreliable to measure because their energy can either be too weak or the spectral bandwidth may be insufficient due to a too low sampling rate. Hughes et al. (2018) also write about the telephone effect: they write that phones have a bandpass filtering effect that can affect the measurement validity of higher formant measurements that exceed the 3400Hz ceiling. This may include formant 3 and will affect formants 4 and 5. However, Hughes et al. (2018) flag that a solution to this problem is to consider speaker-specific and vowel-specific formant settings. Regardless, these studies may therefore account for why lower formants were more useful in Jesse et al.'s (2014) study.

In terms of furthering developments into how formants can be used in ASR research exactly, Nolan and Grigoras (2005) developed a method of combining analyses of multiple formants across all voiced segments of speech into a single measurement called long-term formant distribution, as discussed above. This serves to speed up ASR processes by reducing the number of potential elements that need fusing together, thus keeping the speed of such systems in mind for modern ASR users. Concerns about speed are discussed further throughout the following chapter that focuses on the developed methodology and the fusion of further phonetic features.

Turning finally to sociophonetics, formants offer a well-researched route for distinguishing groups of speakers by accent. For example, northern English speakers will produce the BATH vowel as /a/ whilst southern speakers will use /ɑː/, as discussed above. The difference here is in frontness and length, so formant 2 will prove particularly useful for distinguishing these groups, as an example. Furthermore, as formats relate to the vocal tract space,

differences in age and sex will also be measurable with formants: men will have lower formants than women due to men having longer vocal tracts, and adults will have lower formants than children as adult vocal tracts are longer than child vocal tracts. Overall, formants should prove useful for tailoring ASR approaches to different sociophonetic groups.

All that said, the integration of formants in ASR has not always been successful. In Hughes et al.'s (2017) study, they found that the integration of formants, in some replications, did not alter the performance of the given ASR system in any way. They write that this may have resulted from the fact that MFCCs partially capture the same information and, as a result, little change was observed. With this in mind, the present thesis aims to navigate this issue through portfolio building: if a given phonetic feature proves to be not useful ASR for a given data- or speaker-type, it can simply be discarded. However, if no such change is observed, as it is here, it could still be included as a means of showing that this explainable, perceptible information is still being considered in some way by the ASR system, and its inclusion adds an additional element of explainability to the ASR output.

### 2.5.4. Mean Harmonics-To-Noise Ratio

Mean harmonics-to-noise ratio represents the amount of voiced sound produced by vocal fold vibration compared to the amount of voiceless sound. It is measured in dB, like intensity; thus, it specifically compares the intensity of voiced sound to the intensity of voiceless sound in a production. Perceptibly, mean harmonics-to-noise ratio is a way of quantifying speech productions that can be characterised as non-modal speech, in particular voice qualities where speech is broken up by inclusions of voiceless noise. As an example, Yumoto et al. (1984) write that mean harmonics-to-noise ratio can map onto how creaky a speaker's voice is given that creakiness can, perceptibly, be distinguished from modal voicing due to the breaks in voicing that occur. Klug et al. (2019) also found that mean harmonics-to-noise ratio can map

onto breathy voice, given that breathiness can, perceptibly, be distinguished from modal voicing due to the inclusion of more voiceless aspirated noise. These are explainable features that triers-of-fact can perceive and understand. Physiologically, Ferrand (2002) writes that the reason mean harmonics-to-noise ratio captures different voice qualities is because it captures and quantifies the reduction in the amount of voiced activity as a result of the additional voiceless noise added to the speech signal via turbulent airflow passing through the glottis. Voiced speech is periodic and voiceless noise is aperiodic; thus, mean harmonics-to-noise ratio is the ratio of periodic to aperiodic speech. Non-modal voicing has a lower mean harmonics-to-noise ratio whilst modal voicing has higher mean harmonics-to-noise ratio. Yumoto et al. (1984) also report that the standard mean harmonics-to-noise ratio expected of adults when producing vowels is around 7.4dB.

Concerning why vowels are the ideal segments to measure mean harmonics-to-noise ratio from for ASR, vowels are, in principle, fully voiced; but speakers do not always achieve full voicing. Therefore, mean harmonics-to-noise ratio captures the specific fluctuations in a speaker's vocal fold vibration that fail to achieve full modal voicing consistently and result in the non-modal voice qualities discussed above.

Mean harmonics-to-noise ratio may therefore offer an alternative approach to analysing intensity that can potentially capture more individual differences below the sociophonetic level that may prove useful for characterising individual speakers belonging to a specific sociophonetic group. This is because Yumoto et al.'s (1984) above average of 7.4dB is based on modal speech; thus, mean harmonics-to-noise ratio may prove useful for ASR with atypical voice qualities that diverge from this average. Of the features reviewed thus far, variability in modality can only be detected by mean harmonics-to-noise ratio, giving further credence to the necessity of combining phonetic approaches to enhance one's approach to ASR.

Additionally, speakers also generally tend to have breathier and creakier voices earlier in the day (Yumoto et al., 1984). On the basis of this, mean harmonics-to-noise ratio may therefore offer a phonetic approach that can characterise speakers in spite of session variability, specifically how a recording done in the morning may differ to a recording taken in the afternoon. This highlights how phonetic approaches can be used to accommodate further variables that can affect speech and ASR performance.

Like the above features, mean harmonics-to-noise ratio has also seen prior success in ASR studies: as Long et al. (2011) found, incorporating mean harmonics-to-noise ratio features into a system already employing current approaches to ASR, namely MFCCs, reduced the error rate of the system by 2.72%. The success of this study therefore reinforces the notion that one could combine phonetic approaches to ASR together alongside current ASR approaches, such as MFCCs, for the overall goal of increasing explainability without any negative effect on performance. In forensic contexts, Hughes et al. (2019a) fused mean harmonics-to-noise ratio into an ASR system as another laryngeal voice quality feature that, as seen above, also improved performance via the inclusion of phonetic features. Furthermore, Klug et al. (2019) successfully used mean harmonics-to-noise ratio to characterise suspects and offenders by breathy voice, specifically in their examination of intensity measurements taken between the lower two harmonics and between the lowest harmonic and the harmonic closest to the first formant. Building on mean harmonics-to-noise ratio's potential, Teixeira et al. (2013) also found that mean harmonics-to-noise ratio is another quantifiable phonetic feature that can be successfully used to diagnose pathologies from a recording of a speaker's voice, specifically to track divergence from modal voicing over time. As discussed above, they concluded that phonetic features useful for the identification of pathologies are also useful for ASR. Overall, the above findings further demonstrate how mean harmonics-to-noise ratio can be used for ASR and further justify the

goals of this study to test and integrate mean harmonics-to-noise ratio alongside other features for ASR.

Mean harmonics-to-noise ratio will not necessarily be the most accurate measurement, however. Yun et al. (2022) found that this feature is dependent on the reliable identification of pitch boundaries. Thus, for this feature to be valid for ASR, *f0* measurements must also be accurate. The accuracy of *f0* is discussed in greater detail in the later chapters of this thesis.

Turning now to mean harmonics-to-noise ratio's usefulness for exploring and characterising sociophonetic distinctions, some patterns have been identified in the literature. Looking first at sex, mean harmonics-to-noise ratio averages are generally higher in men than in women, as Sheena et al. (2022) found; essentially, this finds that there are more harmonics per kHz for men than women, so there is a higher proportion of periodic sound. Problematically, this shows that averaging irrespective of sociophonetic differences, as Yumoto et al. (1984) did above to generate their average of 7.4dB, may actually ignore important and quantifiable sociophonetic distinctions that can aid ASR.

Turning now to age, Ferrand (2002) found that mean harmonics-to-noise ratio was significantly lower in the speech of older speakers and that mean harmonics-to-noise ratio was particularly powerful at exploring this age dimension. This effectively finds that voicing activity in the speech of older speakers is more affected by voiceless noise activity than the voicing activity in the speech of younger speakers. This is expected: in older speakers, there is greater glottal leakage owing to imperfect vocal fold function. This, perceptibly, correlates to the fact that older speakers typically have creakier and breathier voices. It also shows that, whilst intensity was found not to distinguish age well, another feature measuring the same acoustic element of loudness can: mean harmonics-to-noise ratio. This gives further credence to the combining of multiple approaches together for ASR.

Finally, Bergstrom (2017) found that accents can vary based on the grand, overall mean harmonics-to-noise ratio of speakers' production of vowels. Vowels have proven particularly prolific in mean harmonics-to-noise ratio research: for example, Murphy and Akande (2007) focused predominantly on vowels in their analysis of mean harmonics-to-noise ratio and found that this feature is powerful for characterising speakers from different vowel productions. Overall, variation along three important sociophonetic axes (age, gender, and accent) can therefore be captured in some capacity by mean harmonics-to-noise ratio, justifying its use in the present study in light of its goal to capture and tailor ASR approaches to specific speaker groups to accommodate their different needs.

### 2.5.5. Mean Autocorrelation

Mean autocorrelation is the measure of the periodic consistency of the speaker's *f0* with itself; more specifically, the process of autocorrelation involves comparing a segment of the given sound file with a shifted version of the same sound file. In terms of interpretability, mean autocorrelation is therefore, effectively, a measurement of how consistent voicing activity is across a speech signal. Mean autocorrelation is measured quantitatively on a scale of 0 to 1, where 0 represents fully voiceless noise and 1 represents perfectly periodic signals. Perceptibly, this therfore also correlates to non-modal voice qualities as mean harmonics-to-noise ratio did, but Gonzalez-Rodríguez (2014) found that it is particularly adept at measuring and characterising the creakiness of a speaker's voice. Thus, mean autocorrelation complements mean harmonics-to-noise ratio in measuring non-modal voice qualities; both features may therefore prove useful for characterising speakers with creakier voices and for exploring session variability, specifically when speakers have creakier voices at different times of the day. This is how this feature can be interpreted.

In terms of its use in ASR studies, Gonzalez-Rodríguez (2014) found that mean autocorrelation can be used as a useful phonetic feature for diagnosing errors in current ASR approaches. As discussed, diagnostic studies can be used to make ASR systems more explainable, and by then including this phonetic feature in one's approach to ASR, they can provide a potential way of improving the performance of modern approaches by solving the errors they generate. In their study, Gonzalez-Rodríguez (2014) looked at the False Rejections (FRs) of a modern ASR system. These are the instances where the ASR system failed to recognise the correct speaker. They found that, amongst all of the FRs, a common feature was that the speech was noticeably creakier. As creakiness is best detected with mean autocorrelation, this study shows how phonetic approaches can make ASR more explainable: it indicates that the tested ASR system must not have been analysing mean autocorrelation measurements already. Researchers, as a result, now know what the present ASR system does not measure. With this knowledge, the performance of the present ASR system could, hypothetically, be improved: by incorporating a measurement of creakiness, such as mean autocorrelation, the FRs that resulted from creakiness could be rectified. If the feature used is explainable, like mean autocorrelation, then the system can now be better explained to triers-of-fact, too.

Further evidence for the usefulness of mean autocorrelation in ASR comes from Bidondo et al. (2013) who developed a novel ASR system that used mean autocorrelation measurements. Their motive was more abstract than other studies, however: they tested whether mean autocorrelation could be used as a quantifiable representation of how the human brain naturally characterises speakers by comparing new speech to memorised speech. The rationale for focusing on how the human brain characterises speakers is strong, to an extent: it is an already functional system for speaker recognition that triers-of-fact naturally also have, so finding ways of imitating its functionality in ASR provides researchers with a premade

framework for how an effective system should work. Furthermore, the link to human recognition can be explained to a trier-of-fact.

Bidondo et al. (2013) write that their approach is based on the notion that the human brain recognises speakers via natural calculations of distances. They claim that this neurological phenomenon can be captured as a comparison between sound vectors, taking one vector from the current utterance and comparing it to the stored vector of a previous utterance. The brain then hypothetically compares these two sound vectors for similarity to determine whether or not the sound came from the previous source. They found that the phonetic feature measurement that can be used to imitate this task and measure the statistical difference between two vectors most effectively was mean autocorrelation; thus, this study indicates that phonetic and psycholinguistic theory could be employed to create ASR methods that replicate naturally-occurring subconscious processes in the human brain relating to speaker recognition, specifically the task of identifying whether a new speech signal came from the same source speaker as a previous signal or from a different source speaker.

This demonstrates how explainable theories from phonetics and psycholinguistics could be used in ASR, thus improving explainability. That said, this study has some noticeable flaws: firstly, it is based on the assumption that the human brain is a perfect recognition system; problematically, however, the human brain is also capable of false recognition. Furthermore, the study treats the brain as a mathematical computer: maths and computer science are two completely different fields of science that were never made originally to imitate the brain, so using these fields as the rationale for how the brain works is fundamentally flawed. This also exemplifies the tension between machine learning and neuroscience. Furthermore, the idea of comparing sound vectors and finding the shortest distance between them is actually an outdated method of ASR already seen in this thesis: this approach is effectively the same as EDs discussed above. Finally, given that this study uses the human brain as a model for an

effective ASR system, it would have benefited from a comparison to results taken from human participants judging the same data to test just how well-replicated the human brain's speaker recognition ability is; as it stands, this study simply identifies mean autocorrelation as an effective method of speaker recognition based on statistical modelling with no real link to any real cognitive processes.

In ASR studies, mean autocorrelation has been found to be a strong measurement of emotional speech. As Mouawad (2017) found, mean autocorrelation can be used to distinguish speech deemed 'pleasant' and 'unpleasant' quantifiably, with 'pleasant' speech having a higher average mean autocorrelation and 'unpleasant' speech having a lower mean autocorrelation due to the proportion of modal phonation present. This means that mean autocorrelation will be effective at handling between-session, within-speaker variability, particularly in light of a speaker's mood or in light of different topics of conversation with different emotional weighting. This may therefore prove useful in building more comprehensive speaker portfolios that capture how a speaker speaks in different moods. Mean autocorrelation's unique sensitivity to these variables therefore complements other phonetic features discussed here, thus allowing for more aspects of a speaker's speech and more session variability to be accounted for through combinatorial approaches.

Further sociophonetic research on mean autocorrelation, however, is minimal; this is therefore another research gap that this thesis addresses. However, it is likely that mean autocorrelation will capture sociophonetic differences because such differences are already detectable with intensity and mean harmonics-to-noise ratio, the two features most similar to mean autocorrelation.

In terms of where to take mean autocorrelation measurements in speech, vowels once again prove powerful. This is illustrated by Kaur and Jain's (2015) study which employed mean

autocorrelation in an ASR system that focused on vowels and saw comparable performance to modern ASR systems.

### 2.5.6. Jitter

Jitter is a measure of the variations and fluctuations in *f0* between cycles that are caused by a speaker's lack of precise control over their vocal fold vibrations. Perceptibly, these capture one-off events like voice breaks as well as more information about non-modal voice quality features like breathiness and creakiness. Non-modal voice qualities, as discussed above, can be useful for characterising inter-speaker variation because speakers within the same sociophonetic group differ in voice quality. They can also be useful for characterising intra-speaker variation because session variability can affect these voice qualities, such as when a speaker's voice is creakier earlier in the day.

In terms of measuring jitter, there are a multitude of avenues to explore because there are a variety of measurements and algorithms to use. These will each be discussed here in light of their functionality, variability, and ultimately which proves best for measuring jitter overall. Starting with the measurements first, there is the local measurement of jitter. This is a percentage measurement calculated by dividing the average absolute difference between the *f0* peaks of two consecutive periods by the length of an average period in the speech file. A period is the inverse of the sampling frequency: as discussed in the next chapter, the typical sampling frequency is 48kHz for high-quality recordings. Thus, a period is typically 1 divided by 48,000, which is 0.0000208333333 seconds. An *f0* peak of a period is hereon defined as the maximum positive amplitude in the glottal cycle.

There is also the local, absolute measurement of jitter. This is where the average absolute difference between the two *f0* peaks of two consecutive periods is counted in milliseconds.

Next, there is the Relative Average Perturbation (RAP), another percentage measurement. This divides the average absolute difference between the *f0* peak of a period and the *f0* peaks of the two neighbouring periods by the average length of a period in the speech file. This effectively captures the same information as the local measurement but with more room for variability by incorporating more comparisons between *f0* peaks.

The next measurement, the 5-Point Perturbation Quotient (PPQ5), is similar: it's another percentage measurement that takes the average absolute difference between the *f0* peak and the *f0* peaks of its four closest periods and divides this by the average length of a period in the speech file. This builds further on the RAP and local measurements by introducing more variability through data from more *f0* peaks.

The final measurement introduces the greatest amount of variability: this is the Difference of Differences of Periods (DDP). This is a percentage measurement that incorporates data from all *f0* peaks in the speech signal: it calculates the average difference between all consecutive *f0* peaks and divides this by the average length of a period in the speech file. Essentially, all percentage measurements here capture variability in *f0* as a percentage with different amounts of data included, whilst the local, absolute measurement measures variability in *f0* in terms of cycle duration. From an interpretability perspective, these different measurements therefore all quantify information about non-modal voicing, but over different lengths of data.

In terms of previous uses in ASR, Jones et al.'s (2001) study incorporated jitter and found that it was a feature capable of distinguishing speakers based on their breathiness and creakiness, as expected based on the above literature. Further evidence for the efficacy of jitter can be drawn indirectly from studies into voice pathologies. Teixeira et al. (2013) again found that jitter is a particularly notable feature for the detection of pathologies; thus, as features that can identify pathologies also tend to prove useful for ASR, it can be assumed

that jitter may prove useful for ASR. Beyond this, however, jitter's efficacy for ASR has proven problematic: Hughes et al. (2022), for example, found jitter to be detrimental to ASR performance. Also, there is a key problem with the above studies in support of the use of jitter: they did not use real productions of words. Instead, they used held vowels which are long, consistent productions that make taking these measurements easier in laboratory settings with no view to real-world usages of jitter, particularly in forensically-diverse settings.

Turning now to sociophonetic variation, jitter has proven particularly useful for characterising speakers of different sexes: returning to Teixeira et al.'s (2013) study, the values found for male and female speakers without pathologies differed significantly. These results can be found in Table 1 below, and as seen the male jitter values were consistently lower with the exception of jitter (local, absolute). Teixeira et al. (2013) took these measurements from monophthong vowels under the justification that monophthongs are consistent, open-approximation productions that carry speaker-specific information about a speaker's unimpeded vocal tract. Despite the already mentioned problems with sustained vowels, this study further justifies the choice to study vowels as the optimised area for taking phonetic measurements for ASR. It also shows that jitter can capture sociophonetic distinctions, giving further credence to tailoring ASR approaches to different speaker groups, and supports the use of jitter for ASR in general.

**Table 1: Teixeira et al.'s (2013) Jitter Measurements from Female and Male Monophthong Productions Taken Using Boersma and Weenink's (2023) Praat Algorithm**

| Jitter Measurement | Female Speech | Male Speech |
|---|---|---|
| **Jitter (Local)** | 0.39% | 0.26% |
| **Jitter (Local, Absolute)** | 17ms | 18ms |
| **Jitter (RAP)** | 0.23% | 0.15% |
| **Jitter (PPQ5)** | 0.25% | 0.15% |

Problematically, jitter does not only have a variety of different measurements; each of these measurements also has different algorithms and approaches that can be used to take them. These have emerged from debates about the reliability of jitter as a feature and how its measurements are taken. For example, the above results taken from Teixeira et al.'s (2013) study are the results taken from Boersma and Weenink's (2023) Praat, a program for acoustic phonetic analyses that has its own unique algorithms for taking these jitter measurements. Teixeira et al. (2013), however, also took the same jitter measurements from the same data using their own algorithm, and the measurements extracted from this algorithm have been compared to Praat's in Table 2 below. As there are differences between these approaches, this indicates that one approach may not be as accurate as the other, or perhaps that they are both equally inaccurate.

**Table 2: Teixeira et al.'s (2013) Jitter Measurements from Female and Male Monophthong Productions Taken Using Teixeira et al.'s (2013) Own Algorithm**

| Jitter Measurement | Female Speech | Difference from Praat's Algorithms | Male Speech | Difference from Praat's Algorithms |
|---|---|---|---|---|
| **Jitter (Local)** | 0.66% | (+0.27%) | 0.43% | (+0.17%) |
| **Jitter (Local, Absolute)** | 29ms | (+12ms) | 30ms | (+12ms) |
| **Jitter (RAP)** | 0.43% | (+0.20%) | 0.28% | (+0.13%) |
| **Jitter (PPQ5)** | 0.46% | (+0.21%) | 0.28% | (+0.13%) |

This difference in measurements led Teixeira et al. (2013) to explore which algorithm is most reliable for jitter extractions. For this experiment, Teixeira et al. (2013) synthesised a vowel which was designed to generate jitter measurements of 0 across all features. Their algorithm achieved 0 for all measurements, as seen in Table 3 below; however, this study was conducted with the authors' own algorithm in mind. It is therefore potentially subject to author bias. Furthermore, the difference in performance between their algorithm and Praat's algorithms on this synthesised vowel was negligible: as Table 3 below shows, the largest difference was 0.003% for jitter (local). However, as the local measurement here produced a more divergent measurement than expected compared to the other measurements, this complicates issues further: it indicates that the different approaches to taking different jitter measurements also perform differently. This study therefore highlights the potential methodological concerns regarding the best ways of taking these different measurements of jitter, but problematically it does not offer a notably different or more reliable alternative. Interestingly, however, it indicates that a rank order of jitter measurements based on their

reliability can emerge, and as a result some jitter measurements may prove more viable for ASR than others.

**Table 3: Comparative Performance of Teixeira et al.'s (2013) Algorithm and Boersma and Weenink's (2023) Praat Algorithms for Taking Jitter Measurements from a Synthesised Vowel**

| Jitter Measurement | Teixeira et al.'s (2013) Algorithm | Praat Algorithm |
|---|---|---|
| **Jitter (Local)** | 0 | 0.003 |
| **Jitter (Local, Absolute)** | 0 | 0.00003 |
| **Jitter (RAP)** | 0 | 0.00002 |
| **Jitter (PPQ5)** | 0 | 0.00002 |

Such methodological problems will be heightened in forensic contexts where data is more variable, and reports of jitter's methodological unreliability do not stop here, either. Researchers such as Leong et al. (2013) found that measurements of jitter were only actually reliable for male speakers. Reliability in this study is measured by the metric of intra-class coefficients: these are statistical values that measure how data from a group is similar to other data from the same group. Here, the data are the measurements of jitter and the group is the tokens from a single speaker. A measurement of >0.8 is deemed reliable, and the full results of this study can be seen below in Table 4. Note that this table also includes shimmer results; shimmer, and its results here, are discussed in detail in the following section.

**Table 4: Results of Leong et al.'s (2013) Jitter and Shimmer Reliability Investigation Using Intraclass Correlation Coefficients (Perturbation Measures)**

| | Female | | | | Male | | | |
|---|---|---|---|---|---|---|---|---|
| **Measure** | **Week 1** | **Week 2** | **Week 3** | **Over All 10 Sessions** | **Week 1** | **Week 2** | **Week 3** | **Over All 10 Sessions** |
| **Jitter** | 0.7 | 0.78 | 0.81 | 0.5 | 0.7 | 0.81 | 0.89 | 0.91 |
| **Shimmer** | 0.67 | 0.64 | 0.59 | 0.56 | 0.57 | 0.54 | 0.73 | 0.53 |

Looking at these results optimistically, they first show that jitter can capture session variability: as seen here, the jitter results differ each week. Jitter may therefore prove useful for capturing session variability and intra-speaker recognition. The frequency of unreliable coefficients (<0.8), however, validates concerns about jitter, especially for female speakers. However, this may give further credence to this thesis' goal to create tailorable profiles based on sociophonetic considerations. If jitter is more reliable for male speakers, then a phonetically-informed ASR approach should only employ jitter when attempting to characterise male speakers. This illustrates how tailoring ASR approaches can make them more applicable to given sociophonetically-controlled groups. That said, Leong et al. (2013) acknowledge that their results were still variable beyond this sex distinction: some successful recognition was still seen with the female participants and the male participants' measurements were not consistently reliable. They write that their results could therefore still be affected by measurement errors in line with the other studies of jitter reviewed here, and they therefore conclude that a much more standardised testing protocol and metric for reliability should be established and used in place of their intra-class coefficients. As a result, the conclusions drawn here about the efficacy of jitter for ASR are still up for scrutiny.

Furthermore, they again do not reflect diverse, forensic conditions: jitter's successes have therefore only been observed in stable, held vowels; not real, forensically-diverse productions. More positively, though, this study abstractly demonstrates that approaches to ASR must accommodate session variability and sociophonetic differences. It also employed the use of vowels for testing ASR, giving further credence to the use of these phonemes for ASR investigations.

Though studies such as Leong et al.'s (2013) raise some concerns about the use of jitter in ASR, some studies have seen outright success. Turning to Farrús et al.'s (2007) study, they implemented jitter as a method of ASR alongside MFCCs and found that the inclusion of jitter improved ASR performance; the information it captures about non-modal voice qualities must benefit current approaches, and thus its inclusion makes current approaches more explainable whilst also potentially improving performance.

Overall, this section has reviewed different studies that found jitter to be useful for ASR, but ultimately unreliable to some degree: different measurements measure different aspects of speech, and the measurements and the algorithms used to calculate them have variable performance. In light of this, and in light of the small amount of recent research that fully considers all aspects of jitter, the present thesis employs all measurements of jitter to test which is most reliable. As the goals of the present thesis are to identify which phonetic approaches are best for given data and speakers, including all of these measurements presents no cost to reliability; they may simply be eliminated if they prove to be non-viable.

### 2.5.7. Shimmer

Turning finally to shimmer, this measures the variations and fluctuations in amplitude between cycles that are caused by a speaker's lack of precise control over their vocal fold vibrations. As a result, shimmer is very similar to jitter, but shimmer captures variations in intensity instead of *f0*. Perceptibly, shimmer also captures non-modal voice quality features such as breathiness; this, as Teixeira et al. (2013) write, is because shimmer is capturing information about the amount of resistance in the glottis.

Also like jitter, shimmer can be measured in different ways. Continuing shimmer's parity with jitter, many of these measurements are similar; they just observe intensity variation instead of *f0* variation. The two measures tend to be correlated, even if they are not necessarily correlated in principle. Starting with the local measurement, this is again a percentage measurement calculated by dividing the average absolute difference between the amplitude deviations of two consecutive periods by the magnitude of the average intensity of the speech file.

Next, there is the local, dB measurement of shimmer. This takes the average base-10 logarithm of the difference between the amplitude deviations of consecutive periods and multiplies this by 20. This differs from jitter's local, absolute measurement in that it does not measure time, but it still continues to consider the variability between periods in a speaker's speech to capture fluctuations.

The next three measurements are similar to jitter's RAP and PPQ5 measurements in that they consider increasing amounts of between-period variability with each measurement, but these measure intensity variation instead. These three measurements all measure the Amplitude Perturbation Quotients (APQs) and are, like many jitter measurements, percentage measurements. The first of these, APQ3, is a three-point APQ measurement that divides the

average intensity deviation of a period and its two flanking periods by the average intensity of the speech signal to identify how variable this section of speech is.

Next, APQ5 is a 5-point APQ measurement that divides the average intensity deviation of a period and the two flanking periods either side of it by the average intensity of the speech signal. It is the same as the APQ3 measurement, but incorporates more data from one more neighbouring period on each side of a period.

The final of these three APQ measurements is APQ11. This is an 11-point APQ measurement that escalates the amount of data drawn upon further: it divides the average intensity deviation of a period and the five flanking periods either side of it by the average intensity of the speech signal.

The final shimmer measurement to discuss is shimmer's own DDP measurement, as jitter had above. However, instead of *f0*, this calculates the average difference between all consecutive intensity deviations and divides this by the average length of a period in the speech file. Like jitter, too, from an interpretability perspective these different measurements also all quantify information about non-modal voicing, but over different lengths of data.

In terms of the use of shimmer in ASR, Teixeira et al.'s (2013) study must be returned to one last time. They found that shimmer can also be used effectively to identify pathologies in the voice that manifest themselves through breathiness, specifically as a result of more resistance in the glottis that can be quantified through below-average shimmer measurements. The importance of such findings for ASR, to reiterate, is that features proving useful for detecting pathologies also typically prove useful for ASR; thus, this is indirect evidence for the usefulness of shimmer in ASR. Also looking at shimmer's efficacy for ASR, Farrús et al. (2007) implemented shimmer into a GMM-UBM ASR alongside MFCCs. They found that the inclusion of shimmer improved the performance of the approach. This infers that the ASR

system did not consider any non-modality aspects of the voice relating to shimmer. With the inclusion of shimmer, it now does, and this adds an additional element of explainability to the ASR system. Not only does this justify the testing of shimmer as a feature for ASR, it gives further credence to the use of phonetic features to improve the explainability and, potentially, performance of current approaches to ASR, and shows that fusing phonetic features with ASR systems is possible.

Problematically, however, shimmer shares many of the measurement validity issues that plague jitter. Starting with Teixeira et al.'s (2013) already discussed study, they also created an algorithm for shimmer measurement extraction and they once again found that their algorithm performed slightly differently to Boersma and Weenink's (2023) Praat algorithm, as seen above with jitter. These results can be found in Table 5 below.

**Table 5: Teixeira et al.'s (2013) Shimmer Measurements from Female and Male Monophthong Productions Taken Using Teixeira et al.'s (2013) Own Algorithm and Boersma and Weenink's (2023) Praat Algorithms**

| Shimmer Measurement | Female Speech; Teixeira et al.'s (2013) Algorithm | Female Speech; Praat Algorithm | Male Speech; Teixeira et al.'s (2013) Algorithm | Male Speech; Praat Algorithm |
|---|---|---|---|---|
| **Shimmer (Local)** | 2.43% | 2.28% | 2.01% | 1.72% |
| **Shimmer (Local, dB)** | 0.45dB | 0.2dB | 0.1% | 0.15dB |
| **Shimmer (APQ3)** | 2.7% | 1.3% | 1.37% | 1% |
| **Shimmer (APQ5)** | 0.7% | 1.37% | 0.79% | 1.07% |

Importantly, they once again employed their synthetic vowel test and this once again found that their algorithm was marginally more accurate than Boersma and Weenink's (2023). However, as Table 6 below shows, neither approach performed perfectly for shimmer: neither ever achieved 0, but both generated very similar results close to 0. This therefore highlights that there is more unreliability for shimmer than there was for jitter; no method proved perfect even when the authors specifically designed their system to be perfect on the given data. There is also an indication of another sex distinction here with male speakers typically having lower shimmer measurements than women, as seen in Table 5 above; however, considering the variability demonstrated in Table 6 below that challenges the measurement validity of both approaches, these conclusions are difficult to support. Furthermore, the use of synthetic speech is again concerning: these findings may have been generated as a result of laboratory conditions and not real speech. These findings may therefore not be generalisable to real-world speech.

**Table 6: Comparative Performance of Teixeira et al.'s (2013) Algorithm and Boersma and Weenink's (2023) Praat Algorithms for Taking Shimmer Measurements from A Synthesised Vowel**

| Shimmer Measurement | Teixeira et al.'s (2013) Algorithm | Praat Algorithms |
|---|---|---|
| **Shimmer (Local)** | 0.0003 | 0.0008 |
| **Shimmer (Local, dB)** | 0.00002 | 0.00007 |
| **Shimmer (APQ3)** | 0 | 0.0003 |
| **Shimmer (APQ5)** | 0 | 0.0001 |

This study indicates that shimmer may have some potential for ASR, but it simultaneously indicates that shimmer as a measurement for ASR may be even more unreliable than jitter. Teixeira et al. (2013) still could not create an idealistically perfect approach for analysing shimmer. This indicates that shimmer, despite its potential, should be used cautiously; however, given that it does show some potential, it will still be included in this thesis as a tested phonetic approach hypothetically capable of contributing to effective ASR methods and all shimmer measurements will be included. The risk of including this is low given that, in testing for efficacy, shimmer measurements may simply be excluded from the combinations if they perform poorly.

This study also gives credence to Leong et al.'s (2013) above conclusion that standardised testing protocols need to be established and implemented to investigate the reliability of such measurements. Looking back at Leong et al.'s (2013) results in Table 4 above, it is also seen that shimmer performed even worse than jitter for them too as no shimmer result was ever deemed reliable in their experimental design; no shimmer measurement for ASR scored >0.8 for male or female data, further backing the lack of reliability of shimmer. This shows that this feature may perform universally badly; unlike jitter, it performs badly even when sociophonetically-tailored to sex.

In terms of other studies, Brockmann et al. (2011) contrastingly found strong support for measurements of jitter and shimmer distinguishing speakers by sex; men, for all measurements considered, had lower average values. Moreover, their study supports the notion that jitter and shimmer values are best taken from vowels because their jitter and shimmer measurements differed between vowels: thus, they may prove useful for distinguishing between speakers of different accents that utilise different vowels in different words. They also found that session variability between vowels can be detected well by jitter and shimmer, reiterating the above point that these features may capture the intra-speaker

differences between speaker productions of these vowels well. This indicates that vowels should be targeted for ASR analyses, that jitter and shimmer may prove useful for detecting sociophonetic differences by accent that can aid in the tailoring of phonetic approaches, and that jitter and shimmer may help account for intra-speaker variability between sessions.

Whilst this study gives some hope for the use of jitter and shimmer in ASR, the issue of reliability remains. Thus far, it has mainly been considered whether the algorithms used to calculate all jitter and shimmer measurements are reliable; now, it must be considered which of these individual measurements is most reliable. Briefly returning to Teixeira et al.'s (2013) above results, they created an algorithm that was promising for jitter, but held author bias in the experimental design and made no such progress on the analysis of shimmer compared to Boerma and Weenink's (2022) algorithms used for Praat. It is worth noting that, a year later, Teixeira and Gonçalves (2014) produced an algorithm for taking jitter and shimmer measurements that did outperform Boersma and Weenink's (2023) Praat by a wider margin than Teixeira et al.'s (2013) approach. This does indicate that instrument reliability may be the problem, but this is challenged by the fact that, seven years prior, Farrús et al. (2007) reliably employed Boersma and Weenink's (2023) Praat algorithms to achieve success in their investigation of the efficacy of jitter and shimmer in ASR. In their study, they contrastingly found that the problems with jitter and shimmer were not a result of the algorithms used to calculate the measurements, but of the measurements themselves. They specifically found that the local measurements they explored, specifically the jitter (local, absolute) and the shimmer (local, dB) measurements, had the lowest EERs when analysed independently for ASR of 26.9%. The highest EERs recorded occurred when jitter (RAP) was analysed, which had an EER of 34.2%, and shimmer (APQ11), which had an EER of 33.8%. This may be expected: these measurements factor in the most amount of data and variability, meaning there is a wider margin for errors. The local measurements are more

specific, and potentially more accurate as a result. On the basis of these results, which can be found in full in Table 7 below, Farrús et al. (2007) only brought forward jitter (local, absolute), shimmer (local, dB), and shimmer (APQ3) to their system as they were the only moderately successful features for ASR with EERs below 28.1%.

**Table 7: Results Taken from Farrús et al.'s (2007) Study of Jitter and Shimmer in ASR Scenarios**

| Jitter/Shimmer Measurement | EER |
|---|---|
| **Jitter (Local, Absolute)** | **26.9%** |
| Jitter (RAP) | 34.2% |
| Jitter (PPQ5) | 33.8% |
| **Shimmer (Local, dB)** | **26.9%** |
| **Shimmer (APQ3)** | **28.1%** |
| Shimmer (APQ5) | 32.9% |
| Shimmer (APQ11) | 33.8% |
| **Optimised Jitter and Shimmer Approach (Threshold: 28.1%): Jitter (Local, Absolute), Shimmer (Local, dB), and Shimmer (APQ3)** | **22.5%** |

This study therefore demonstrates two vital considerations for the present thesis: first, it shows that building portfolios of phonetic features for ASR is a practice that can work: the combination of the selected measurements of jitter and shimmer generated the lowest EER recorded at 22.5%. Though this performance is not wholly impressive, this still shows that the goal of this thesis (to identify, combine, and tailor phonetic approaches to ASR) is

possible and has already been successful in a study that only considered jitter and shimmer measurements, not more reliable phonetic features which jitter and shimmer is combined with in the present thesis. Secondly, whilst caution is still necessary, this study shows that the fundamental problems with jitter and shimmer may not relate to how the measurements are extracted, but the measurements themselves; thus, given the moderate success seen here through the use of Boersma and Weenink's (2023) Praat algorithms, these are used in the present thesis and all measurements are included as a partial recreation of Farrús et al.'s (2007) study to test which jitter and shimmer measurements, irrespective of the algorithm used, could be brought forward for use in sociophonetically-tailored ASR. This is also partially motivated by the need for the present thesis' methodology to be efficient: if all phonetic measurements are extracted from the same program, the creation of these tailored profiles will be faster. Thus, if everything is extracted in Boersma and Weenink's (2023) Praat, as the following chapter explores, the phonetic approaches and research output of the present thesis will be more commercially and forensically viable.

### 2.6. Research Questions

In summary, this chapter has provided the basis for the present thesis and now uses this literature to propose an exploratory research question: "What explainable phonetic approaches to ASR, ranging from features to segments, are best for recognising different speakers and speech styles?". Exploring this allows for a novel exploration of the proposed bespoke, bottom-up, combinatory phonetic approaches to ASR.

More specifically, this thesis explores the usefulness of combining different phonetic features correlated to perceivable elements of the voice such as pitch, loudness, vowel differences, and non-modal voicing. These features are *f0*, intensity, formants 1-5, mean harmonics-to-noise ratio, mean autocorrelation, jitter, and shimmer. It must be noted here that practical

issues for forensic materials, whilst flagged above, are not considered in this thesis. These practical issues include the noted difficulties surrounding measurements of higher formants, intensity variation due to microphones, and the different measurements of jitter and shimmer. Instead, this thesis serves as an exploration of the potential of combining these phonetic approaches for ASR, and practical, forensic applications can be addressed in future work. That said, these issues do arise and are discussed: for example, higher formants were expected to perform better, but were harder to measure. Furthermore, local measurements of jitter and shimmer performed better than other measurements of jitter and shimmer, as predicted above.

In this exploration, this thesis also explores the usefulness of different vowels for ASR. It ties together and expands upon the underpinning conclusions of much of the above research that vowels prove effective for ASR whilst also allowing for the only study to directly assess the performance of different vowels, Paliwal's (1984), to be modernised using the validation metric $C_{llr}$ which has been seen in more up-to-date studies, such as those by by Wilemijn Heeren (2020; 2022).

Finally, this thesis also explores sociophonetic variables discussed throughout this chapter. Specifically, it uses data from different accents to explore variation in the performance of these feature combinations for different accents, as above studies have for different languages with some of the features, and how to tailor different combinations to different accents. It also uses text-dependent and -independent data to explore differences in the performance of these features related to style, given the historical focus on text (in)dependence and how phonetic features originally presented issues for text-independent ASR.

In order to facilitate answering this research question, as well as future-proofing further research regarding the testing of phonetic approaches to ASR, this thesis' primary

contribution to the field is a replicable methodology for testing novel, combinatory, phonetic approaches to ASR. This is detailed in the following chapter. Then, the chapters after that will focus on demonstrating the potential of this methodology for creating optimised, sociophonetically-tailored combinations of phonetic features that can be used for ASR. The portfolios of phonetic approaches created here, for the specified accents and styles, could then (in principle) be used by ASR developers; as seen in this chapter, researchers can already fuse phonetic approaches to ASR with current ASR approaches to add an additional element of explainability to current ASR systems without replacing them. The practicality of fusing the tested combinations found in this thesis is explored in the final chapter. However, it must be noted that controlled data is employed here; therefore, these portfolios are better served as proofs of concept: future research can now follow the methodology to produce optimised portfolios like these for different sociophonetic groups with more forensically-realistic data. These portfolios could then aid in explaining ASR output better to triers-of-fact, specifically by mapping perceptible, explainable elements of the voice onto phonetic measurements that have been integrated into ASR.

## 3. Methodology

The output of this thesis includes a replicable and adaptable methodology for testing novel phonetic approaches to ASR. More specifically, this methodology identifies the best combinations of phonetic features to employ when recognising speakers from a specific sociophonetically-controlled group. The results of this thesis are all generated by this methodology. This chapter details how this methodology generates these optimised portfolios of phonetic approaches for ASR by breaking down the methodology into its main component stages: corpus selection (3.1), forced alignment (3.2), feature extraction (3.3), and portfolio creation (3.4). Each of these stages is discussed in detail here to show how the present thesis applies them and how future researchers can replicate and change elements for other experimental designs. It is important to note that, given that this methodology seeks to test phonetic approaches for ASR that can later inform ASR researchers, automation has been considered at all stages. It is also important to reiterate that this thesis does not suggest that these portfolios can stand alone as independent ASR approaches; the goal here is only to identify what phonetic feature combinations, measured on selected vowels, are best to analyse for characterising specific speakers and data types. The optimised portfolios could then be fused with current ASR systems to complement them with more explainable and perceivable elements, as discussed in the previous chapter.

Looking at this methodology, one may argue that using sociophonetically-controlled data directly impacts the external validity and generalisability of the findings; the results will only be applicable to the selected sociophonetic group. However, this is actually the purpose of the present study: the aim is to investigate sociophonetic specificity in ASR methods.

As a final note, the specific data used in this thesis for generating the example combinations is controlled, as the coming section discusses; it is therefore not forensically-realistic. As

such, these generated combinations are better served as proofs-of-concept. Future work can go on to explore more diverse and forensically-realistic data, but the goal of this thesis' examples is only to establish that more detailed, bespoke, bottom-up, phonetically-informed approaches to ASR are possible.

### 3.1. Corpus Selection

The first stage of this methodology involves the researcher selecting a database of sociophonetically-comparable speech produced by sociophonetically-comparable speakers. The choices made here must be informed in light of two considerations: firstly, to ensure the methodology can function correctly, there are technical requirements of the data that must be met and confounding variables that must be controlled. These are discussed first in this chapter. Secondly, and more importantly, the selected speakers and the speech must be sociophonetically-controlled because the choices made here sociophonetically-informs one's investigation, which can be as diverse as one chooses for forensic purposes. As discussed throughout this thesis, different speakers and different data types may have different characterisation needs that must be explored. Thus, the chosen data informs the portfolios that are generated: the results will therefore only be applicable to the selected data group. Re-running this methodology on different data in future studies will therefore create different portfolios. For example, a database of text-independent speech recorded using mobile phones that was produced by young female speakers with Southern Standard British English (SSBE) accents will generate a portfolio specific to this style, data, age, sex, and accent profile. Effectively, all future investigations should vary along these axes to build further portfolios. These can then be compared to each other to identify any generalisable findings across sociophonetic boundaries, as this thesis explores. As such, this is the first point at which future researchers can tailor the methodology: they may wish to use data from different

speakers producing different data to find out the unique needs of a given group of speakers for ASR.

For this thesis, the sociophonetic variables of accent and style are tested. The present thesis employs three different corpora, each varying along these axes. Two of these datasets are from Nolan et al.'s (2009) Dynamic Variability in Speech (DyViS) Corpus, which contains speech from males aged 18-25 with SSBE accents. However, the two selected datasets vary in speech style type: one is text-dependent, one is text-independent. The text-dependent task involves reading out a news article about a crime; the text-independent task involves a staged phone call admitting guilt to a crime. Comparing data along this speech style axis allows for variation in speech style to be tested. Then, one dataset from Gold et al.'s (2018) West Yorkshire Regional English Database (WYRED) Corpus will be analysed, in particular the dataset that collected text-independent speech from male speakers aged 18-25 speaking with West Yorkshire accents. A comparison between this and the text-independent dataset from Nolan et al.'s (2009) DyViS Corpus therefore allows for the accent axis to be explored and tested. Age and sex were kept consistent because, as seen throughout the previous chapter, these variables can affect speech production and thus the likely within- and between-speaker variation for given features. Thus, this experimental design allows for only the variables of style and accent to be tested in order to explore whether different sociophonetic speakers and speech have different needs.

Before reviewing these databases in detail, however, the baseline technical requirements for any investigation using this methodology must be explored. Firstly, all of the data must be in the WAV format. This is essential for two reasons: the first reason is that the WAV format is compatible with all of the software used in the following stages of this methodology. Secondly, the WAV format is considered the most desirable way to store speech data for forensic testing anyway. This is because, as van Son (2002) writes, it is a completely lossless

and uncompressed format. The lossless format is desirable because it means the speech is recorded and stored in the highest quality available; thus, it has been preserved and stored as closely and as accurately to the original production as possible. The uncompressed format is important because compressing audio and reducing the file size can result in jump errors that can deteriorate the quality of the recording. These errors will affect the validity of the phonetic feature measurements extracted later in this methodology. It is possible to salvage compressed audio from these jump errors, should one need to use data stored in a compressed format, but one would need to calculate and decrease the Root-Mean-Square Error (RMSE) in their analysis of vowel *f0* and formant measurements. RMSE is a measurement that represents the standard deviation of data points from the predicted regression line of the data; simply put, it calculates how far the data diverges from expectations. These jump errors are such divergences, hence why calculating RMSE offers a solution to accounting for and correcting errors created by the use of compressed audio. That said, however, one can avoid even encountering this issue and needing to add this unnecessary step by opting to use data in the uncompressed WAV format, hence why it is the optimum format for testing. That said, this is only the case at the time of publication; future audio formats may exist that are more appropriate for testing, and as long as they are compatible with the rest of the methodology, future research may wish to employ them for whatever benefits they hold.

Moving on to a different technical requirement, the sampling rate of the audio used must be higher than 7.8 kilohertz (kHz). This is a constraint imposed by the other software employed in this methodology: McAuliffe et al.'s (2019) Montreal Forced Aligner (MFA). The MFA, when it aligns speech to its component phonemes, requires the extraction of MFCCs between 0kHz and 7.8kHz. This is where the lower frequency 'bins' are, as discussed above. Thus, the sampling rate of the selected data must be higher than 7.8kHz so that all of the required MFCCs for the MFA to work can be extracted. It is important to note that this is unlikely to

affect any study today: most high-quality data is sampled at 44.1kHz and most mobile phone recordings, whilst lower in quality, are still sampled above this threshold at 8kHz. This is typical of phone transmission. One would also want the sampling rate to be above 7.8kHz for successful formant extraction anyway; any lower, as discussed above, and the frequency range will only extend to 3.9kHz and an unwanted ceiling to the data will be created which can affect higher formant measurements.

Next, confounding variables in data selection that may affect the quality of one's investigation are discussed here. Firstly, one must ensure that the microphone and the microphone set-up is consistent across all speech recordings analysed in the dataset. This is because, as van Son (2002) writes, the microphone used and the microphone set-up have a direct impact on the recording quality. These methodological discussions are particularly important to intensity, as mentioned above. Using RMSE as a metric for identifying and measuring the effect that changing the microphone can have on recording quality, van Son (2002) found that the RMSE of the pitch and formant measurements of the recorded speech surpassed one semitone when different microphones were used to record the same speaker. When the same microphone is used, however, the RMSE of the recorded speech is always below one semitone. Semitone is a measure of *f0*; thus, a change in microphone will affect any phonetic features measuring *f0* in some capacity due to this higher variability. As a result, the microphone used can be a confounding variable in a study. Consistent microphone use is therefore a way of ensuring the speech used for a given speaker and data group is consistent, thus making the portfolios more reliable for use.

This impact of the microphone used, as van Son (2002) writes, is to be expected: all microphones differ in production quality, size, and position within the recording device itself. As explained further by Tiete et al. (2017), the space available to different microphones that is used for receiving and converting acoustic pressure into electrical energy affects the quality

of a recording directly. Therefore, microphone variability is a confounding variable that must be controlled to ensure one takes phonetic feature measurements that are consistent in quality across all of the speakers in the selected group for the portfolio being created. Confounding variables such as this must be controlled during the data selection stage of the methodology because many databases intentionally offer data recorded from different microphones, such as studio microphones and mobile phone microphones, to allow researchers to study the effects of different microphones on recorded speech. Such explorations would be possible with this thesis' proposed methodology in the future: one could compare the results from different microphones to identify which phonetic approaches can always be reliably used irrespective of microphone quality and which only work for a given microphone quality. Matching and mismatching the microphones used in all included data, selected or created, is therefore an opportunity to shape another investigation along the data type axis. This may be particularly important in forensic cases wherein evidence will have expectedly been recorded by a variety of different microphones.

One must also account for the set-up of the microphone as another confounding variable. This is because the position and distance of the microphone to the speaker and the sensitivity of the device itself can affect measurements of phonetic features like intensity, as Titze and Winholtz (1993) write. More specifically, the closer a speaker is to the microphone, or the more sensitive the microphone is, the louder they will appear in the recording. With this in mind, session variability is another confounding variable one must control because this can affect microphone set-up: the researcher may slightly alter the position, distance, or sensitivity of the microphone between recording sessions or the speaker may position themselves differently between recording sessions, as Rouvier et al. (2011) write. Alternatively, however, this is yet another way in which one can shape one's investigation along the data axis. One may wish to identify what phonetic approaches are best to measure

in light of session variability so that phonetic approaches to ASR can be more adaptable to different recording settings, here changes to the microphone set-up. Again, this will be particularly pertinent to forensic cases where data is much more diverse.

Another confounding variable to consider is background noise to the recorded speech, as Rouvier et al. (2011) also flag. Background noise can be other speakers, as would be expected when recording in a public space, or noisy objects in a room, such as a fan. Background noise is important to consider in an ASRinvestigation using this thesis' methodology because it can impact the quality of the phonetic feature measurements taken; a louder noise in the background may skew an intensity reading, for example. It should therefore be controlled because it can be difficult to remove background noise from a recording; selecting or creating data without background noise is easier in an experimental design. However, this is again another avenue for one to shape future investigations: one may wish to identify what phonetic approaches are best to accommodate session variability between recordings caused by different or increased background noise. This has particular commercial benefits, as customers may wish to access their accounts in busier spaces than the quiet home, and forensic benefits, due to the increased diversity of forensic data. One could therefore use the present thesis' methodology to explore how, exactly, phonetic approaches fare when the data has no background noise and when the same data has background noise. Alternatively, many studies have already researched and developed methods of de-noising speech signals using modern approaches. Such technologies could also be employed in the present methodology to accommodate the confounding variable of background noise.

The sociophonetic variables one selects for testing are now discussed further. Concentrating on some of the technical impacts of these sociophonetic variables, Eide and Geish (1996) note that the vocal tract differs significantly in length between men and women on average. This, in turn, results in a higher frequency range for women and a lower frequency range for

men. This will directly affect *f0* and formant measurements, as discussed above. More specifically, this difference between male and female vocal tracts is so widely accepted that it manifests itself in the tools and approaches already being applied in this methodology: Boersma and Weenink's (2023) Praat, which is used for feature extraction later in this methodology, requires one to set different *f0* ranges based on the sex of the speaker analysed to ensure the feature extraction process is as reliable as possible. Specifically, one sets the *f0* range lower for men than women. As such, within a singular investigation using this methodology, one should keep sociophonetic variables like sex consistent as a technical requirement: these distinctions are so well-documented that different configurations already exist within this methodology to accommodate them and make feature extraction more reliable. This, itself, is evidence that sociophonetic tailoring is already crucial for phonetically-informed ASR.

Exploring the decisions here allows future researchers to tailor to their investigations using the present thesis' methodology. In order to exemplify this, the following sections will discuss three specific datasets selected for the present thesis' investigations. Exploring these will allow this methodology to be rigorously tested and explore variation along the sociophonetic axes of accent and style, all whilst meeting the controls and considerations specified here. It should be noted that many of the considerations explored in this section are derived from Drygajlo et al.'s (2015) literature.

Being so transparent about the data and approaches used are efforts to reduce uncertainty by allowing the methodology to be more understandable, as Wang and Hughes (2022) recommend. It also adds to the explainability of the process, which benefits this thesis' core goals. However, Wang (2021) also flags that even variability in the speakers selected as test speakers from a database can cause random variability in the scores for performance, especially when the sample size is small. This means that, despite the amount of data used

from each dataset, the external validity and applicability of any portfolios may still be affected by which speakers were selected as test speakers for each experiment. For more realistic applications of this thesis' methodology, future replications with different setups of speakers are therefore recommended. Here, the goal is simply to identify whether these bespoke, bottom-up, phonetically-informed approaches to ASR are viable.

### 3.1.1. Nolan et al.'s (2009) DyViS Corpus (Text-Dependent and -Independent Tasks)

Looking at the selected datasets more closely, the first corpus selected for the present study is Nolan et al.'s (2009) DyViS Corpus. All of the speakers are aged 18-25, male, and have SSBE accents. Two datasets were selected from this corpus: one containing text-dependent speech and one containing text-independent speech. As the speakers were kept the same, this thesis' explorations therefore identify the optimum, phonetically-informed characterisation needs of young, male, SSBE speakers producing text-dependent or -independent speech.

Looking first at why specifically this data can be selected, it meets all specified technical requirements of this stage and all confounding variables have already been controlled by the corpus creators. Firstly, all of the files are saved in the WAV format. They are therefore compatible with the methodology at present. Furthermore, the files are all 18-bit and 48kHz; they are therefore compatible with the later stages of this methodology, namely McAuliffe et al.'s (2017) MFA.

Focusing first on the text-dependent sub-corpus, all of this data has been selected from Task 3 which is a reading of a news report, as discussed. All data was taken from the same session and from recordings using the same high-quality microphones in the same room without any background noise to control the microphone-related and session variability-related confounding variables above. This ensures that all of the read text-dependent speech is comparable between all speakers. By having each speaker read the same script it also ensures

that, when this methodology turns to examining specific vowels, the same number of tokens of each vowel have been collected from each speaker. This ensures that all confounding variables and independent variables are already controlled.

Turning now to the text-independent sub-corpus, efforts have been made to keep this data as comparable as possible to the text-dependent speech with the only difference being the type of speech produced. This data is from Task 2. The same speakers were used, recordings from the same session were used, the same microphone was used, and all technical requirements were again met. Thus, the core difference was that the speakers were now producing spontaneous speech; however, this raises the issue that there is much more variability between recordings now. One potential issue for the present methodology is that text-independent speech cannot control the number of tokens produced of each vowel; thus, there may be an imbalance between speakers here. That said, efforts have been made by the corpus creators to control this: firstly, much more data is collected from this task, meaning there are more opportunities for each vowel to be used a significant amount. Secondly, they kept the speakers on a specific topic, namely an imaginary discussion with an accomplice about a crime the speaker committed. This ensures that speakers will be producing lexemes from the specific semantic field related to crime, thus meaning the data may contain many of the same lexemes and thus vowels. Whilst this is as controlled as it can be without using pre-determined speech, it is also worth considering that any more control than this would risk the data not reflecting real speech; this data is already potentially hindered by the effects of Labov's (1972) Observer's Paradox as the speakers are aware they are being recorded in a controlled environment.

Issues aside, this thesis can still use these datasets to explore the viability of bespoke, bottom-up, phonetically-informed approaches to ASR. By comparing the portfolios generated for each dataset, this investigation can then highlight what phonetic approaches to ASR will

always work for this speaker group regardless of the data produced. It can also identify what phonetic approaches are style-specific, in particular which approaches only work for this speaker group when they produce text-dependent or -independent speech.

Sociophonetic research into the SSBE accent is of particular interest to this investigation. Relevantly, de Jong et al. (2007) found evidence of phonetic changes in progress in SSBE: many phonologically back vowels, such as /uː/, showed increasing degrees of variability in vowel frontness between speakers. This change is diffusing through this sociophonetic group, and as a result formant 2 may prove particularly useful for distinguishing speakers. However, the results of this study could also be interpreted as unclear: the variability identified may instead be due to different segmental contexts causing different co-articulation effects on formant 2 measurements. This may result in increased within-speaker variability which can trigger further between-speaker variability. Such variability in formant 2 will be explored in this thesis as a result.

### 3.1.2. Gold et al.'s (2018) WYRED Corpus (Text-Independent Task)

Moving now to Gold et al.'s (2018) WYRED Corpus, this was specifically designed to be comparable to Nolan et al.'s (2010) DyViS with the only difference being the accent of the speakers; age and sex are kept consistent. With this in mind, all of the above technical and confounding considerations apply here: session variability was controlled, microphone quality and set-up were controlled as above with high-quality microphones and no background noise, the data meets all technical requirements of being WAV, 16-bit, and 48kHz, and the speakers were all kept the same across tasks. However, only the text-independent dataset was selected here. In the speech produced, the speakers also pretend to be offenders on the phone admitting to the crime, so all of the text-independent concerns and controls above apply here too. The only difference of note is the speakers themselves: they

are still aged 18-25 and male, but here they have West Yorkshire English accents. Thus, the portfolio that will be generated by this data will identify the optimised phonetic approaches for characterising young, male, West Yorkshire English speakers producing text-independent speech.

This investigation will create more portfolios, here concerning how best to characterise this speaker group producing text-independent speech. This data has been selected because it invites a comparison to the above SSBE speakers also producing text-independent speech to see how accents vary the performance of the bespoke, bottom-up, phonetically-informed approaches to ASR tested in this thesis. Comparing these portfolios allows one to identify the different needs of different accents for ASR as well as what phonetic approaches may always prove useful for ASR regardless of accent.

Overall, through strategic data selection, the present thesis allows for comparisons to be made along the speech style axis (by comparing the text-dependent and -independent speech from Nolan et al.'s (2010) DyViS) and along the accent sociophonetic axis (by comparing the text-independent speech from Nolan et al.'s (2010) SSBE speakers and Gold et al.'s (2018) West Yorkshire speakers). More focus is paid to text-independence here given its increased real-world applicability, as highlighted throughout previous chapters, but again it must be reiterated that the selected data only serves to show whether bespoke, bottom-up, phonetically-informed approaches to ASR could work; more real-world, forensically-diverse explorations should be conducted with this methodology once its potential has been explored in the present thesis.

### 3.2. Forced Alignment

Now that the datasets have been selected, the speech signal must be segmented into its component phonemes so that the data can be analysed at the phonemic level. This allows different vowels to be tested, specifically to identify which vowels will prove particularly useful for ASR and what features are best to measure across the different vowels.

The goal of this stage is to generate TextGrid files that can be read into Boersma and Weenink's (2023) Praat, the software that will be used in the following stage for extracting feature measurements. These TextGrids will delimitate where each phoneme is in the original speech recording. Generating these therefore ensures that the phonetic feature measurements can be taken from the specific vowels being tested.

For the creation of these TextGrids, McAuliffe et al.'s (2019) MFA has been selected as the software of choice. This is based on the success of the approaches it employs in prior studies that are discussed throughout this section. The specific task of demarcating these speech boundaries is called forced alignment. A forced aligner, as defined by McAuliffe et al. (2019) themselves, is a piece of software capable of taking a speech recording alongside an orthographic transcription of its speech content to generate time-aligned TextGrid objects that demarcate the exact timestamps of all words and, most importantly, phonemes, in the speech. Many forced aligners exist, so the justifications for choosing the MFA specifically must be discussed below; however, following the underlying themes of futureproofing this thesis' methodology, it should be noted here that if more advanced or more reliable forced aligners come to exist in the future, it is possible to use that forced aligner instead as long as it can generate TextGrid files that Boersma and Weenink's (2023) Praat can read for extracting feature measurements later. As will be seen, swapping out this forced aligner when a more

reliable aligner comes to exist may actually be recommended; the MFA has potential flaws that this thesis explores in detail.

In order to conduct forced alignment, there are multiple components involved. Firstly, forced aligners require three elements to be input: the speech files from the corpus selection stage (which have already been discussed), a pronunciation dictionary of the target language of the speech, and a transcript of this speech. Looking first at this dictionary, it must contain information on the phonemes present in every word of that target language so that the forced aligner can use this dictionary to structure the TextGrid output for each speech file, specifically to organise the order of phonemes found in each word in each speech file.

Turning next to the transcript of the speech, this can complicate the methodology. It must be a plain TXT document only containing the spoken words in the speech file for it to work. When the selected corpora can and has included a transcript, no extra step is required; the provided transcript can be used. This is typically the case for text-dependent speech as the speakers are already using a script of some kind. The problem, however, arises for text-independent speech where, by virtue of being spontaneous, no transcript exists. Transcribing this speech automatically may generate errors that human transcribers would not, but transcribing this speech manually can be an arduous task that may still invoke human error due to unclear speech and errors. As an exploration of solutions to this issue, methods for automatic speech recognition that convert speech to text will be used to test their efficacy on the text-independent speech databases used in this thesis. Google Cloud's (2023) speech-to-text software was used due to its reliability and accuracy, and upon manually checking a sample of five transcripts, none had errors. Full transcripts of all text-independent speech were created with this tool and could now fill this gap in the forced alignment process.

Whilst successful, this software does not align with the explainability ethos of this thesis: it is completely reliant on difficult-to-explain machine learning approaches that cannot be understood. This, unfortunately, represents a cost of this thesis at the time of writing: no reliable-yet-fully explainable automatic speech recognition methodology exists, so Google Cloud's (2023) approach had to be used. This is important to flag because, as it stands, this methodology now cannot be wholly explainable. When a more explainable automatic speech recognition system does exist, however, it should be implemented in this methodology.

Issues aside, the process of forced alignment with this speech, dictionary, and transcript will now be discussed. For the process of matching the dictionary-ordered phonemes to the speech content as labelled by the transcript, further speech recognition tools must be employed. The MFA employs Povey et al.'s (2011) Kaldi speech recognition toolkit to recognise the boundaries of the ordered phonemes in the selected data. This is a well-justified choice: Povey et al.'s (2011) toolkit is a reputable speech recogniser, as Matarneh et al. (2017) write, that has been used for the acoustic analysis of data from multiple different languages. Exemplifying this efficacy on multiple languages, Gauthier et al. (2016) used it for successful speech recognition in Hausa whilst Peddinti et al. (2016) used it for successful speech recognition in English. This last study specifically indicates that a system using Povey et al.'s (2011) Kaldi toolkit, such as the MFA, can be successfully and reliably employed for the analysis of English data, such as that selected for this thesis. More broadly, these studies justify the use of the MFA because together they indicate that as long as a dictionary exists for data in a given language, the MFA is reliably compatible with any language variety due to its use of Povey et al.'s (2011) Kaldi toolkit. This means that the present methodology, despite being tested on English data, could therefore be used to identify different phonetic approaches to ASR in different languages so long as a language-specific acoustic model and dictionary exists for that language. This is another potential application of the present

methodology: beyond data and speaker types, different language varieties could also be explored in light of identifying which phonetic approaches are best for a given variety. Whilst not different languages, the specific choice to change the accent variable in the present thesis exemplifies how one could explore differences between language varieties.

Turning now to the underlying technical processes employed by the MFA, these can be broken down into four primary stages. This is important to discuss in light of explainability and compatibility with the methodology. For the first stage, phonemes are very roughly mapped to the speech files based on the dictionary and transcripts using monophone models. These models treat every occurrence of every phoneme the same way. They do not consider the context they occur in beyond the phonological stress information already included in the dictionary, as will be discussed later; they just use a generalised representation of the phonemes in each word of the transcript to create a basic representation of the structure of the speech's phonemes.

Speech, however, is subject to variation based on surrounding sounds, so the second stage of the MFA considers this specifically. This stage takes each individual phoneme's monophone model and considers the monophones of the two flanking sounds. These three monophone models considered together are called triphone models, and they allow the MFA to consider co-articulatory effects caused by certain sequences of sounds. The co-articulatory effects can include changes in the production of a sound based on the preceding or following sound: for example, word-initial /t/ is aspirated if no sound precedes it, as in "top" which is produced as [$t^h$ɒp], but is unaspirated when a /s/ precedes it, as in "stop" which is produced as [stɒp]. Co-articulatory effects are critical to the accuracy of the MFA: as this example shows, plosive length is a variable that is directly affected by aspiration. If the extended or reduced length of a plosive that results from co-articulatory effects is not considered, the MFA could incorrectly assume it has been aspirated or unaspirated and inaccurately capture the length of

the plosive. As a result, the proceeding sound may include the aspirated section of the production. Given that many of the selected acoustic features required the selected segment of speech to be voiced for a measurement to be extracted, such accuracy is therefore critical to avoid the inclusion of any unvoiced segments of speech such as aspiration.

Thus far, the phonemes themselves and their phonological contexts have been considered. However, this assumes that all speakers produce phonemes the same way in each given phonological context. This is not how speakers produce language; they produce phonemes in unique and individual ways, and so the third stage of the MFA takes into account this speaker individuality. As Chodroff (2018) explains, the MFA employs specific elements of Povey et al.'s (2011) Kaldi software for this task, in particular its Linear Discriminant Analysis Maximum Likelihood Linear Transform (LDA-MLLT) tools. Firstly, this takes the MFCCs of the individual phoneme productions for the unique speaker and builds HMM states for each speaker's phoneme productions. In effect, this captures the feature space of the phoneme productions, but reduces the size for efficiency given how much data this stage accrues by looking at individual productions. From these HMMs, a unique transformation of every occurrence of every phoneme is generated for each speaker, thus accounting for intra- and inter-speaker variability entirely.

The final stage of the MFA combines the phonological context information from the triphone models with the speaker context information from the LDA-MLLT stage. To do this, this stage calculates a transformation of the MFCCs for each speaker's individual phoneme productions from the third stage to personalise the monophone model information from the first stage; these are then combined into the triphone models from the second stage. These enhanced triphone models are then used to align the transcript to the speech with the highest degree of accuracy possible. The reliability of this process can be attributed to its clear focus

on phoneme and speaker individuality; this aligns with the views of the present thesis on using phonetic theory and justifies its use in the present study.

However, in practice, there are concerns that must be raised. The MFA utilises a number of the practices that lack full explainability, namely the HMMs discussed in the previous chapter. Due to present limitations, nothing can be done about this; however, this highlights how this stage could benefit future attention from researchers. More explainable approaches to forced alignment could be tested and developed for the benefits of explainability discussed in the earlier chapters and to ensure all stages of this methodology are more explainable. This is for two reasons: firstly, triers-of-fact may seek explanations for this stage of the methodology. Secondly, errors in this methodology may arise from errors in forced alignment, and these cannot currently be diagnosed or fixed as a result.

Turning now to how the MFA is practically used, regardless of the explainability issue, this is relatively simple. One must first select a dictionary. For the present thesis, given that English is the language spoken in all three selected databases, Panayatov et al.'s (2015) LibriSpeech lexicon was employed based on McAuliffe et al.'s (2019) own recommendation. The reasons for this recommendation are twofold: firstly, this library is amongst the largest created and it encompasses most words in English. The second reason is that it encodes all phonemes in a format compatible with most programming languages: the ARPAbet. The ARPAbet is a format for representing phonemes in compatible fonts for coding given that many IPA symbols are incompatible. Though its usefulness in automating phonetic analyses such as this cannot be understated, a critical flaw of the ARPAbet is that it does not account for all phonemes. However, for the present study, the use of the ARPAbet is mostly compatible with the selected phonemes based on literature reviewed above: /ə/, /ʊ/, /ɪ/, /u/, /ɑ/, /ʌ/, /a/, /ɔ/, /æ/, /i/, /ɛ/, then /ɝ/. Most of these can be represented by the ARPAbet: /ə/ can be represented as AH0 (and ER0 when rhotic /ɚ/), /ʊ/ can be represented as UH, /ɪ/ can be represented as IH,

/u/ can be represented as UW, /ɑ/ can be represented as AA, /ʌ/ can be represented as AH1, /ɔ/ can be represented as AO, /a/ can be represented as AE, /i/ can be represented as IY, /ɛ/ can be represented as EH, and, /ɝ/ can be represented as ER1. The numbers at the end of these is discussed momentarily. It should also be noted at this point that the ARPAbet is not a necessity; it is simply a representation of phonemes devised for dictionaries. Future studies may wish to use phonemes beyond the limits of the ARPAbet or consider supra- or sub-segmental features beyond that which the ARPAbet can represent. As such, one may wish to use alternative dictionaries that do not use the ARPAbet; all that matters is that the representations are compatible with the coding languages used in this methodology.

Looking at an example from Panayatov et al.'s (2015) dictionary in closer detail, Figure 3 below takes the English word "abilities" as it is entered in their dictionary. Every entry is formatted as follows: the word comes first, which is what the MFA indexes using the transcript, then the order of the phonemes in the word follows. This allows the MFA to search for every word in the transcript, access the order of the phonemes in the words to build the necessary updated monophone and triphone models, and print them and their durations in the TextGrid file that aligns the transcript to the sound file.

**Figure 3: Entry for "ABILITIES" in Panayatov et al.'s (2015) LibriSpeech Dictionary**

| ABILITIES AH0 B IH1 L AH0 T IY0 Z |
|---|

It is important to note that, in the case of the vowels, this dictionary does not assume all vowels behave the same in every word; this dictionary factors in phonological stress and encodes whether a vowel has no stress, primary stress, or secondary stress by including a "0", "1", or "2" at the end of the ARPAbet code respectively, as seen in the vowels above already.

The inclusion of this information adds greater depth to the present methodology and further justifies the use of this dictionary: by including this stress information, the MFA can now also consider additional information about the phonological stress that each token carries. This opens the door to exploratory analyses of stress. Such specificity has not been considered in previous literature, thus allowing for the present research to explore another research gap concerning further phonological variation in phoneme productions and how this can affect ASR performance.

A notable issue with the use of dictionaries, however, is what to do when the transcript includes a word not contained within the dictionary, such as a proper noun. A solution provided by CMUdict (2014), the providers of Panayatov et al.'s (2015) LibriSpeech dictionary, is the LOGIOS Lexicon Tool. This will try to generate transcriptions of any missing words using the ARPAbet automatically. However, upon testing this tool for this thesis, it proved unreliable; particularly with the proper nouns causing the issues in the first place. Due to this issue, the researcher manually transcribed missing words using the ARPAbet. This detracts from the automated focus of this thesis, but it demonstrates the need for a more fully comprehensive dictionary or for a tool which can more reliably transcribe words automatically. Regardless, this was exceedingly rare; very few words posed an issue to the MFA. The overarching point to make here is again that more automatic-yet-explainable methods could be researched and developed to address the issues facing this forced alignment stage.

A potential specific issue with the Panayatov et al.'s (2015) LibriSpeech dictionary is that it is specifically designed to represent General American English; it may therefore not be applicable to further varieties of English. This is an issue because these dictionaries seek to represent how speakers produce language, so sociophonetic variation should be considered in the dictionary. Problematically, however, none of the selected databases for this thesis

contain this accent; they contain SSBE and West Yorkshire accents. This could have a variety of hypothetical effects: for example, it may result in an increase in F1 variation for FOOT/STRUT vowels given that General American English has these vowels split whilst West Yorkshire accents does not. However, as Bailey (2016) reports in their uses of the LibriSpeech dictionary for ASR research purposes, its use does not appear to hinder the analysis of British English varieties in any way; thus, the reliability of this methodology should be unaffected overall. Bailey (2016) acknowledges that British English dictionaries do exist, but they are less detailed than Panayatov et al.'s (2015) LibriSpeech dictionary; they also do not consider stress types, meaning an exploratory analysis of stress would not be possible. They also contain fewer words, meaning the above issue regarding missing lexis will be accentuated by using these dictionaries. At the time of writing, no comprehensive dictionary exists for any British English variety. However, should one be made, or if future research is analysing a variety of speech which has a sociophonetically-tailored dictionary available, one should insert it here during the forced alignment stage.

Once the MFA has run successfully, it automatically outputs TextGrid files for every speech file. These can now be used in the next step of the methodology where phonetic feature measurements will be extracted from the selected vowels. It is important to reiterate that this detailed summary reflects how the forced alignment process was conducted at the time of writing. As mentioned throughout this chapter, concern for the future-proofing of this research includes adaptability to new advancements: not only could more explainable solutions be implemented, especially for converting speech-to-text, but new dictionaries that are more comprehensive and sociophonetically-representative should be implemented wherever possible. Furthermore, if formats more robust than the ARPAbet exist for expressing the IPA symbols in a more code-friendly manner, these could also be used. Finally, should more reliable and explainable methods of forced alignment exist than the

MFA, these could be used. As long as the chosen forced aligner can read in WAV files of speech and TXT file transcripts of speech so that TextGrid files demarcating the phoneme boundaries can be generated, any forced aligner can be easily integrated here instead of the MFA without changing any of the preceding or proceeding stages.

### 3.3. Feature Extraction

In this step, all measurements of every phonetic feature discussed in the previous chapter are collected from every speaker's produced tokens of the selected phonemes. At this point, the speech has already been selected (3.1) and the vowels in these files have already been force-aligned (3.2). As a reminder, the selected phonetic features are *f0*, intensity, formants, mean autocorrelation, mean harmonics-to-noise ratio, jitter, and shimmer. For this section, Boersma and Weenink's (2023) Praat software is used to take the measurements. This software has been selected as the optimum tool for this stage of the methodology for two primary reasons: firstly, it is compatible with the other stages of this methodology thus far and the overall goals of this thesis. This is because it has been specifically designed for acoustic analyses, it can analyse all of the selected phonetic features and phonemes, WAV files can be input as speech files, and TextGrids can be input to demarcate phoneme boundaries in the WAV files. The other reason for selecting this software is its noted reliability for phonetic experiments. Boersma and Weenink's (2023) Praat has been a popular tool for acoustic analysis since its creation and that popularity remains steadfast today with much research continuing to employ this software for acoustic analysis, such as Suess' (2023) recent study of how visual cues, such as lip movement, correlate with formant measurements.

This section will now be split into two further sections for ease of discussion: the first discusses the specific, automated approach created for this thesis for the task of feature extraction in Boersma and Weenink's (2023) Praat. The second section will discuss the

reliability of the Boersma and Weenink's (2023) processes for analysing the selected phonetic features to justify this choice of software further and highlight its potential issues.

### 3.3.1. Processes of Feature Extraction

For this methodology, a Praat script has been developed which can automatically take measurements of the selected phonetic features from every token of the selected phonemes produced by the speakers in each corpus. This ensures that the underlying processes behind this methodology all continue to be automated, aiding any future integrations of these approaches alongside real-world ASR systems. Broadly, this script first takes the speech files in WAV format from (3.1) and the TextGrid files from (3.2) to identify the boundaries of the selected vowels. Then, measurements of each of the selected phonetic features is taken from these vowels. It then outputs the results per token, per speaker in CSV files. This code is available upon request from the researcher.

Important specificities of how this code works and what it does are now discussed transparently so that future researchers know how to follow and modify the methodology based on their future investigations into different speech segments and features. This is in keeping with the overarching themes of longevity and futureproofing present throughout this thesis thus far. This stage could, technically, also be revamped with other software that may prove more suitable or reliable to future investigations testing phonetic approaches to ASR.

With the speech selected and the phoneme boundaries demarcated, extracting the desired phonetic features is focal to this stage, and each must be extracted in the most appropriate way. For each phoneme token, one measurement of each phonetic feature is taken by the script, but this measurement is not necessarily taken the same way for every feature. Some are analysed as mean values, namely mean autocorrelation and mean harmonics-to-noise ratio. These are analysed as means because the features themselves are calculated as mean

values across the selected speech window already. Thus, one measurement of these features already represents the data across the whole phoneme token.

For jitter and shimmer measurements, these are singular measurements representative of data over multiple frames in the speech window; they are not mean values, but they are not values of singular frames either. Thus, one measurement once again represents the data across the phoneme token, but this time in relation to its midpoint. Shimmer (APQ11), for example, is based on the 5 frames either side of the midpoint frame, as discussed above.

The remaining features, however, are measured from individual frames. These are *f0*, intensity, and the formants. It is important not to take mean values of these measurements because these features can fluctuate within a vowel production; thus, mean values could warp the analysis. Therefore, for these features, the midpoint frame will be where the measurement is taken from to ensure the values are not warped by taking means. This works for monophthongs, as the present study is testing, because these are roughly stable productions; this will not, however, work for diphthongs in future studies as the midpoint will only capture the midpoint of the diphthong production. This problem could be addressed by taking multiple frame measurements throughout the diphthongs in future research. Finally, these feature measurement extractions will only work if the demarcated area is accurate and the midpoint is voiced; otherwise, there will be no speech to extract measurements from. Overall, however, this means that each feature is analysed in the most accurate and appropriate way.

### 3.3.2. Internal Reliability of Boersma and Weenink's (2023) Praat

Having now discussed how one practically applies Boersma and Weenink's (2023) Praat for feature extraction, its underlying processes for measuring these phonetic features will now be discussed to highlight the reliability of this software choice for the present methodology as well as where opportunities for evolution arise for future researchers. Firstly, in terms of

input compatibility, Boersma and Weenink's (2023) Praat can read in any WAV file for analysis; thus, it is software completely compatible with the reliability choices discussed above in the data selection stage (3.1) so long as the file is saved in the WAV format. Furthermore, its ability to take measurements from a given phoneme automatically are entirely dependent on the reliability of the input TextGrid files provided; as a result, the reliability of Boersma and Weenink's (2023) Praat for accurately measuring aspects of a phoneme token are dependent on the reliability of the forced alignment stage (3.2). This is why so much work into the reliability of the prior stages is important; as long as they are compatible with Praat, their reliability directly affects the reliability of this stage. The selected data has been established as reliable, but it must be noted that the limitations of forced alignment, despite the reliability of the MFA as it stands, may therefore have some impact on the reliability of the present methodology for feature extraction.

Turning now to the output, Boersma and Weenink's (2023) Praat offers accessible formats for exporting its measurements. It exports the results as CSV files, a commonly used file format which can be opened in R Core Team's (2023) R, where the portfolio creation stage (3.4) will take place. Thus, Boersma and Weenink's (2023) Praat demonstrates full compatibility with the stages of this methodology either side of feature extraction; it does not alter or warp the input and it allows the output to be tailored to one's investigation, thus raising no confounding variables that could affect the present methodology.

Having established its broad reliability, the actual processes that Boersma and Weenink's (2023) Praat uses in its default settings to measure the selected phonetic features must now be analysed. Looking first at how mean autocorrelation is measured, Boersma and Weenink's (2023) processes are based on Boersma's (1993) own developed approach. Despite the potential creator bias here, it is evidently the most reliable method for measuring this feature. They argue that their algorithm is more reliable and accurate for measuring periodicity, the

underlying regularity of a speech signal that mean autocorrelation measurements are taken from, than other methods based on cepstra and combs. They write that this is because other approaches cannot correctly estimate the autocorrelation of a selected periodic speech signal, such as a phoneme. They write that the failure of these other approaches is due to them not dividing the autocorrelation function of the selected periods of the phoneme by the autocorrelation of the entire window of the phoneme. As their method does this, they deem it most reliable. This has been independently verified by another researcher: Jouvet and Laprie (2017) compared a number of algorithms for calculating mean autocorrelation and found that Boersma's (1993) produced the fewest *f0* frame errors in its analysis, thus justifying its reliability. Given the evident measurement validity of this approach, it has been confidently employed as the chosen approach for measuring this phonetic feature in this thesis.

Additionally, the success of Boersma's (1993) algorithm here also indicates that the software's approach to measuring *f0* is the most reliable approach available for measuring this feature too. This is because it employs the exact same algorithm to measure *f0*. Thus, the above discussion justifies that *f0* is also analysed in the most reliable way here. However, in keeping with the themes of adaptability and futureproofing present throughout this thesis, if future methods of measuring *f0* and mean autocorrelation should arise that are more reliable and can be implemented through Boersma and Weenink's (2023) Praat, they should be implemented. Furthermore, should they not be available in Boersma and Weenink's (2023) Praat, any software capable of taking these measurements whilst also retaining compatibility with the input and output specifications above could be employed here too. The same goes for all of the rest of the phonetic feature measurement approaches that are now discussed.

Moving on to formants, Boersma and Weenink's (2023) Praat offers multiple methods for measuring these phonetic features. Thus, ascertaining which is most appropriate and reliable is crucial to the present methodology. The selected method for the present methodology is

Burg's (1967) method, and this decision has been made on the basis of Bhore and Shah's (2015) study which compared three methods of taking measurements of formants. These methods are the cepstral analysis approach, the Linear Prediction-Based Cepstral (LPBC) technique, and Burg's (1967) method. They assessed each method by calculating the RMSE generated in analyses of the same data under each approach. They found that Burg's (1967) method generated the lowest RMSE. Similarly, Harrison (2013) found that the LPC technique generates the most errors, especially when the settings are altered. Burg's (1967) method has been selected based on this research; as a result, this further demonstrates the instrument reliability of the present methodology.

Looking at why Burg's (1967) method is more reliable in more detail, it analyses time series data formulated as iterations, which here are the frames of the individual phoneme tokens. One compares formant measurements from one frame to measurements taken from a reflection coefficient of the same frame in order to reduce errors that can be caused by predicting vectors; as a result, all data used to calculate these measurements are therefore present in the data and no assumptive or predictive data modelling is employed. Linear Predictive Coding (LPC) is then used to calculate the formants from this to identify them with greater accuracy.

Moving on to mean harmonics-to-noise ratio, Fernandes et al. (2018) write that Boersma and Weenink's (2023) Praat employs a method of calculating this phonetic feature that is generally considered the most accurate method of calculating it by the research community, thus further demonstrating the measurement validity of the present study. In essence, the approach used involves calculating, in decibels, the percentage of the periodic voiced energy of the voiced segments of the signal against the remainder of the signal that is simply noise.

Turning now to intensity, evidence for the reliability of Boersma and Weenink's (2023) Praat's approaches come from their long-term usage in phonetic research without any contest. De Jong and Wempe (2009) used an earlier version of Boersma and Weenink's (2023) Praat to identify syllables automatically and reliably using intensity, for example. Through widespread acceptance through the research community, the approaches adopted by Boersma and Weenink's (2023) Praat can be deemed reliable. It must be acknowledged, however, that the performance of this feature can still vary based on the materials used, as discussed above.

Turning finally to jitter and shimmer, the measurement methodologies and reliability issues were discussed in detail in the previous chapter. As a brief reminder, jitter can be measured five different ways: as a local measurement, a local, absolute measurement, a RAP measurement, a PPQ5 measurement, and a DDP measurement, Shimmer, similarly, can be measured as a local measurement, a local, dB measurement, an APQ3 measurement, an APQ5 measurement, an APQ11 measurement, and a DDP measurement. The mathematical processes detailed in the previous chapter are those that Boersma and Weenink's (2023) Praat uses to calculate these measurements. The concerns about their lack of reliability were also discussed above, and the potential measurement errors will be discussed in greater detail in the later chapters of this thesis.

### 3.4. Portfolio Creation

This final section now provides an overview of how the portfolios of combined phonetic approaches to ASR were created. This is all done in R Core Team's (2023) R. The process involves taking the raw phonetic measurements from (3.3) and using $C_{llr}$, a current gold-standard performance validation metric that will be the focus of this section, to calculate how the features perform for ASR on each vowel for each database. $C_{llr}$ is also used to calculate

what the best combinations of features are for each vowel for each database. These portfolios then serve as the tested best phonetically-informed approaches to ASR.

The first stage of portfolio creation involves converting the feature measurements into scores. This is done using Likelihood Ratios (LRs). In essence, this is where every same- (SS) and different-speaker (DS) pair within a test group is compared using the measurements of each feature taken from a given vowel from a given database. Each of these comparisons is represented as a score which is calculated as the similarity and typicality between two speakers based on the given measurements; from these scored individual comparisons, ratios are calculated which represent the probability of the evidence being valid under the SS and DS hypotheses. These are the LRs.

More specifically, the LRs are calculated using a GMM-UBM approach. The consequences of this for the explainability of the methodology will be discussed in the later chapters of this thesis. Of the 100 speakers in each given database, 66 have been compared thus far in the comparisons. 33 of these speakers were used as a test group, and they have been compared to each other and to another 33 speakers who represent the background group. It should be reiterated here that the selection for the test group, as Wang (2021) states, can have an impact on the variability of the results.

Calibration is then done: this involves using the final 34 speakers, which are used as a development group, to develop weights to apply to the LRs thus far to produce Log-Likelihood Ratios (LLRs). This, Morrison (2018) writes, is a process of making the LRs even more reliable by using a further group of data to train the ASR approach further, thus minimising the margin for error further. This is fusion, and this allows one to combine scores while accounting for correlation.

The final stage, and the evaluation technique for overall performance, involves calculating Log-Likelihood Ratio Costs ($C_{llr}$). This cost, as explained by Morrison and Enzinger (2019), is the average of individual costs, specifically the costs of the false same-speaker and different-speaker LLRs made from all of the above pair comparisons. In general, the closer the cost is to 0, the better the performance of that feature on that vowel for that database is. This is because $C_{llr}$ captures the magnitude of false results using higher scores. This magnitude is one of the main benefits of $C_{llr}$ over other performance validation metrics discussed in the previous chapter that make it the current gold standard performance metric for validating ASR performance.

This provides an overview of how the performance of these phonetic approaches is assessed. However, this only generates a cost for a singular combination of phonetic features for a single vowel for a single database. Thus, to build the portfolios, different combinations of features need to be tested for every vowel for every database. As a result, the above process is followed multiple times to generate multiple $C_{llr}$ measurements: to start, the $C_{llr}$ of all phonetic features and vowels combined is calculated. This provides a baseline performance. Following this, the same conditions are tested on each individual vowel to provide a baseline performance for each vowel. Then, $C_{llr}$ is calculated again, but with one phonetic feature removed. From this, the contributions of that feature for performance can be calculated as the change in $C_{llr}$: if the cost decreases and moves towards 0, that means that feature is detrimental to performance as removal improves performance. If the cost increases and moves away from 0, this means that this feature is integral to performance as removal decreases performance. Once these individual removals are complete, one final combination is tested wherein only the features that prove integral to performance are included. This is therefore a ‘top-down’ approach.

From this testing, the best possible combination of features for a given vowel from a given database can be selected based on what generates the lowest cost. This process can then be repeated for every vowel from every database, and from this a library of the best combinations of features to measure for each vowel produced by a given group can be generated. These portfolios consist exclusively of explainable and perceivable phonetic approaches to ASR and could, theoretically, now be used as supplementary ASR materials. As they are the best-performing phonetic approaches to ASR, they will have minimal detrimental impact to better-performing ASR methods, if fused. As discussed, however, the specific portfolios that are generated for this thesis are done so with forensically-unrealistic data because they serve as proofs-of-concept; more diverse and realistic data should be used beyond this thesis when practical applications are being considered.

## 4. "Undefined" Results

The following three chapters discuss the results. These chapters each analyse the raw data from three perspectives: the first of these chapters (4) explores how successfully extractable the results for the phonetic approaches are. The second of these chapters (5) explores how the phonetic approaches perform for same-speaker and different-speaker recognition tasks. The final of these chapters (6) explores how the phonetic approaches can combine to generate optimised approaches to ASR; this is where the novel, combinatory, phonetic portfolios are explored. The raw data for these chapters can be retrieved from the author's website[1].

The previous chapter detailed the methodology under the assumption of a successful data extraction. However, Boersma and Weenink's (2023) Praat can fail in its analyses; certain attempts at feature extraction (3.3) can yield "undefined" results when the measurement cannot be taken. This chapter focuses on this phenomena in order to illustrate the loss of data that automation can bring. This analysis therefore provides an insight into the viability of these phonetic approaches. These results are also focal to data trimming anyway: for portfolio creation (3.4) to work, R Core Team's (2023) R requires the data frames to only contain numeric data; character strings will cause the entire analysis to fail. Data trimming is therefore solely motivated by functionality; successful measurements that manifest themselves as outliers are still retained given that the purpose of this study is to investigate ASR, and the potential for speakers to produce outliers is a real-world scenario that could face ASR tasks. They are therefore retained to avoid misrepresenting the reality of human speech. This, plus this chapter's exploration of feature extraction success rates, is important for forensic audiences and validation.

[1] https://elliotjholmes.wordpress.com/

All of these “undefined” results are immediately trimmed to ensure the portfolio creation stage can proceed; however, it has been discovered that if a phonetic feature returns results wherein over 25% of the total measurements are “undefined”, for whatever reason that must be investigated further, that phonetic feature must be trimmed entirely. This is because including them yet taking 25% of the results away consistently leads to data insufficiencies during the portfolio creation stage. This meant that the portfolios could not be fully created, and thus the given phonetic feature should be trimmed entirely prior to the creation of any portfolio. For the same reasons, it was also found that any phonemes that provide less than 15,000 tokens of successfully extracted data must also be trimmed. Due to this token limit, it was discovered that secondary stress vowels consistently fail due to data insufficiency; thus, all of these phonemes must be trimmed too. This trimming contributes to the goal of creating best phonetic practices for ASR anyway; if a phonetic feature or phoneme generates too many failed measurements (or not enough successful measurements) this indicates that it may be too unreliable for real-world uses as it may not generate enough data for ASR tasks to take place and work reliably. The behaviour of these “undefined” results can therefore offer insights into the inner workings of this phonetically-informed methodology, and thus this chapter provides an analysis into these results to identify and diagnose any problems in this explainable methodology, as many of the above researchers do with their explainable methodologies to repair and improve them.

Turning first to why any phonetic feature, phoneme, or speaker results would generate these “undefined” results, they can arise for many predictable reasons relating to the first three stages of the developed methodology (3.1-3.3). Firstly, they may appear as a result of lacking speech data in data selection (3.1) wherein the speaker did not produce a vowel where they should have. This could be due to a variety of co-articulation effects. This is particularly pertinent in text-independent speech which will be more natural and free-flowing, but may

also occur in certain accents which elide certain sounds. A relevant example here would be the use of glottal stops in place of vowels in West Yorkshire accents, as in how “to” is commonly realised as [tʔ]. This may affect the selected data from WYRED.

Secondly, they may also occur due to misaligned TextGrids created during forced alignment (3.2) or by the Montreal Forced Aligner (MFA) attempting to find the vowels that were not actually produced in the selected data (3.1). The MFA is not capable of skipping unproduced sounds, so will insert them in spaces where they do not exist and may therefore capture voiceless speech. The transcripts for text-independent speech created using automatic speech recognition tools may also have errors wherein they propose a wrong word containing the wrong sounds.

Finally, they may also arise due to inputting mathematical impossibilities during feature extraction (3.3): for example, *f0* needs voiced speech to take a measurement. The inclusion of any unvoiced speech, potentially due to the above errors with data selection (3.1) and forced alignment (3.2), will therefore make the calculation impossible. Similarly, shimmer (APQ11) requires eleven periods to be included, and if the phoneme production was shorter than eleven periods, there will not be enough data to calculate this.

As seen, problems with the earlier stages can build up and affect the later stages. Though these are all possible problems, the previous chapter detailed the efforts that were undertaken to control reliability as best as possible to ensure that the fewest number of these “undefined” results arose: the data was all checked to ensure the advertised speech was included in each file; the output of the speech-to-text software was checked to ensure it generated correct transcripts; a reliable forced aligner was selected to mitigate the risk of misaligned TextGrids; reliable software for feature extraction was selected; and the Praat script has been checked

and tested in multiple preliminary studies by the author to ensure there are no accidental mathematical impossibilities (Holmes, 2021a; 2021b; 2021c; 2022a; 2022b).

Whilst trimming is important, it is equally important to rectify any issues resulting in "undefined" results simply because more data is ideal; it gives the portfolios greater external validity. This is why diagnostic studies are so important, and why they were reviewed in detail above. Some of the "undefined" results are inevitable, however, due to the data produced: if the target phoneme simply was not produced due to a coarticulation effect, no current tool for forced alignment can account for this. However, if this coarticulation effect can be labelled as a variant of the sociophonetic group of the speaker, this gives further credence to the need to investigate sociophonetics in ASR as it grants an insight into what phonemes should be omitted immediately due to the sociophonetic profile of a given data or speaker type. It also gives credence to the notion that dictionaries specific to a sociophonetic group may be important because they can capture variation that can enable greater performance from phonetically-informed approaches.

Thus, for each given corpus, an investigation into these "undefined" results now begins. These results will be grouped in three separate ways to investigate their distribution: by phonetic feature, by phoneme, and by speaker. This identifies which phonetic features, phonemes, and speakers (if any) must be trimmed in each corpus to avoid exceeding the above limits. In doing this, this chapter is also investigating how sociophonetic variables affect the frequency of "undefined" results. For the phonemes, this is also done for every stress condition in line with the proposed exploratory analysis into the effects of phonological stress on phonetic approaches to ASR.

Some expectations have been made for this analysis. Firstly, given the number of variables already controlled above (particularly for text-dependent speech where the same phonemes

have been produced in the same order), minimal variation is expected between speakers. Similarly, it is expected that there will be minimal variation between phonemes per dataset given that they are all monophthong vowels and are all roughly the same length of 0.03s; they are therefore consistent in data length and quality, to an extent. That said, the ratio of "undefined" results may be somewhat higher for text-independent speech based on the increased likelihood of co-articulation effects that may result in a vowel not being produced fully, as discussed above.

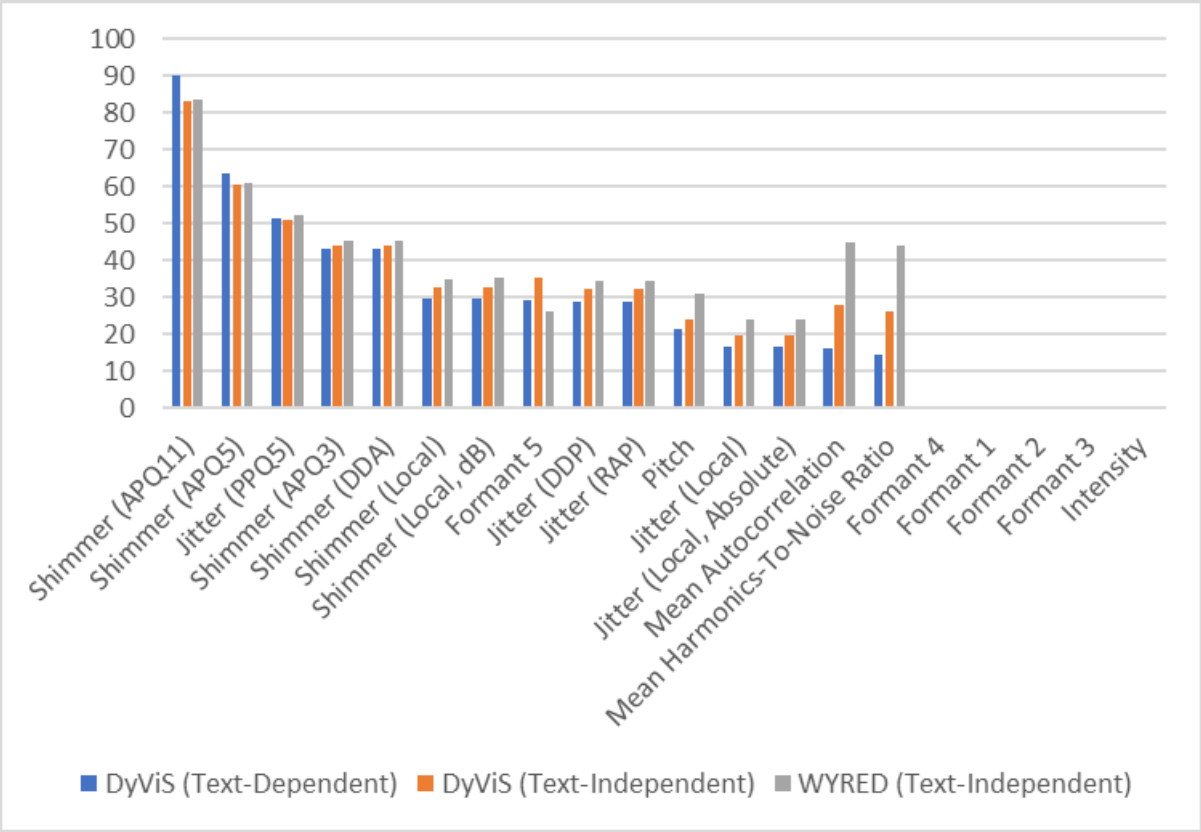
Phonetic features, on the other hand, should prove most variable in the ratio of "undefined" results. Analysing these will therefore offer the most insights into the reliability of the methodology and how effective each phonetic feature is based on its ability to be extracted successfully for each speaker and data type. For example, *f0* should always successfully extract a result because each of the selected vowels are, theoretically, voiced. Boersma and Weenink's (2023) Praat requires this; if parts of the selected segment are unvoiced, the extraction will fail. Thus, the frequency of "undefined" results can tell two things: whether the methodology has issues regarding the placement of phoneme boundaries (thus bringing into question the reliability of the forced aligner which may have misaligned the speech and included surrounding unvoiced segments) or whether the speaker does not produce certain vowels due to co-articulation effects or their accent. This does not bring into question the reliability of *f0* itself for ASR; just the methodology. Contrastingly, however, the performance of phonetic features involving multiple data points, such as shimmer (APQ11), can indicate that the phonetic feature is unreliable; not the methodology. This is because this phonetic feature requires ten neighbouring periods to collect data from and vowels may be too short to do this (especially in text-independent speech). This phonetic feature will likely, therefore, be deemed too unreliable for real-world use.

In summation, this data trimming section will prove useful for ensuring the proposed portfolios for optimised phonetic approaches to ASR can be created whilst also offering insights into the reliability of the methodology: it will identify which phonetic features will be most reliable for producing data as well as how different accents and styles can impact the data extraction processes. It should also be reiterated that, in light of the automation goals of this thesis for its potential commercial and forensic audiences employing it for ASR investigations, that all of this trimming has been automated. This ensures fluid, automatic phonetic analyses for ASR. This also ensures that, in the future, any issues can be diagnosed and rectified with the present methodology.

### 4.1. "Undefined" Results by Phonetic Feature

Figure 4 below shows the percentage of "undefined" results, grouped by phonetic feature, for each corpus. These percentages are calculated from 68,872 tokens of the selected phonemes for the text-dependent data from Nolan et al.'s (2009) DyViS corpus, 125,998 tokens for the text-independent data from Nolan et al.'s (2009) DyViS corpus, and 106,937 tokens for the text-independent data from Gold et al.'s (2018) WYRED corpus.

**Figure 4: Percentage of "Undefined" Results per Phonetic Feature**




Firstly, certain phonetic features are consistently successful irrespective of the sociophonetic variation between the datasets. Some features yielded no "undefined" results for any dataset, and these are formants 1-3 and intensity. Formant 4 generated the least "undefined" results beyond this in all datasets: 0.48% in Nolan et al.'s (2009) DyViS text-dependent data, 0.54% in their text-independent data, and 0.21% in Gold et al.'s (2018) WYRED text-independent data. All of these features are therefore clear of the 25% threshold; however, beyond this, only jitter (local) and jitter (local, absolute) clear the threshold in every dataset, with both measurements generating 16.74% in Nolan et al.'s (2009) DyViS text-dependent data, 19.64% in their text-independent data, and 23.87% and 23.94% respectively in Gold et al.'s (2018) WYRED data.

Around the 25% threshold is where the datasets begin to differ. *f0* is only clear of the 25% threshold for the DyViS data, generating 21.34% "undefined" results in the text-dependent

data and 24.14% in the text-independent data. In Gold et al.'s (2018) WYRED data, it generated 30.75%. The success of this feature is therefore dependent on the sociophonetic variable of accent.

Additionally, mean harmonics-to-noise ratio and mean autocorrelation only clear the 25% threshold in Nolan et al.'s (2009) text-dependent data with 14.61% and 16.28% "undefined" results respectively. In their text-independent data they generate 26.1% and 27.78% respectively, and in Gold et al.'s (2018) text-independent data they generate 43.96% and 44.83% respectively. The success of these features is therefore dependent on the sociophonetic variable of style.

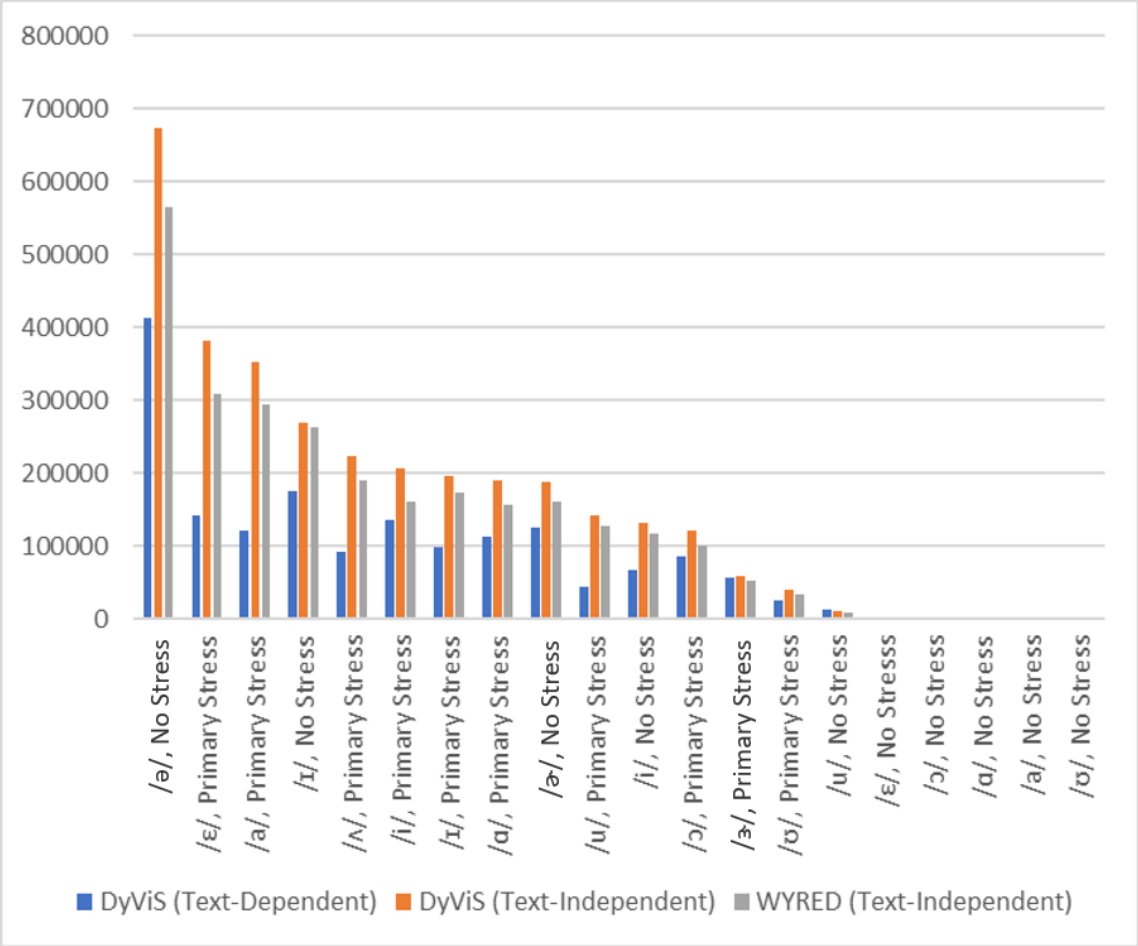
The remaining features consistently fail to clear the 25% threshold in any dataset: formant 5, jitter (RAP), jitter (DDP), jitter (PPQ5), and all shimmer measurements. Thus, from these results, an overall rank order of extractability is visible: intensity, formants 1-4, and local measurements of jitter are always extractable, formant 5, larger jitter measurements, and all shimmer measurements are always unextractable, the success of *f0* extraction is subject to accent, and the success of mean harmonics-to-noise ratio and mean autocorrelation is subject to style. Breaking it down into phonetic feature groups, however, there are other rank orders of note here too. The lower formants (formants 1-3) are most effective, followed by formant 4 which is still usable, then formant 5 which is unusable. Local measurements of jitter performed best and were the only usable jitter measurements. Finally, whilst none of the shimmer measurements were usable, a rank order still emerged with the local measurements performing best and the larger measurements getting progressively worse, as seen by shimmer (APQ11) performing worse than shimmer (APQ5) and shimmer (APQ3) .

### 4.2. “Undefined” Results by Phoneme

Now the results are sorted by phoneme. Remembering that all secondary stress conditions have already been trimmed, it must be stated that a number of further phonemes have been trimmed because no tokens of these phonemes were produced at all in the selected data. These are /ʊ/ (no stress), /ɔ/ (no stress), and /a/ (no stress).

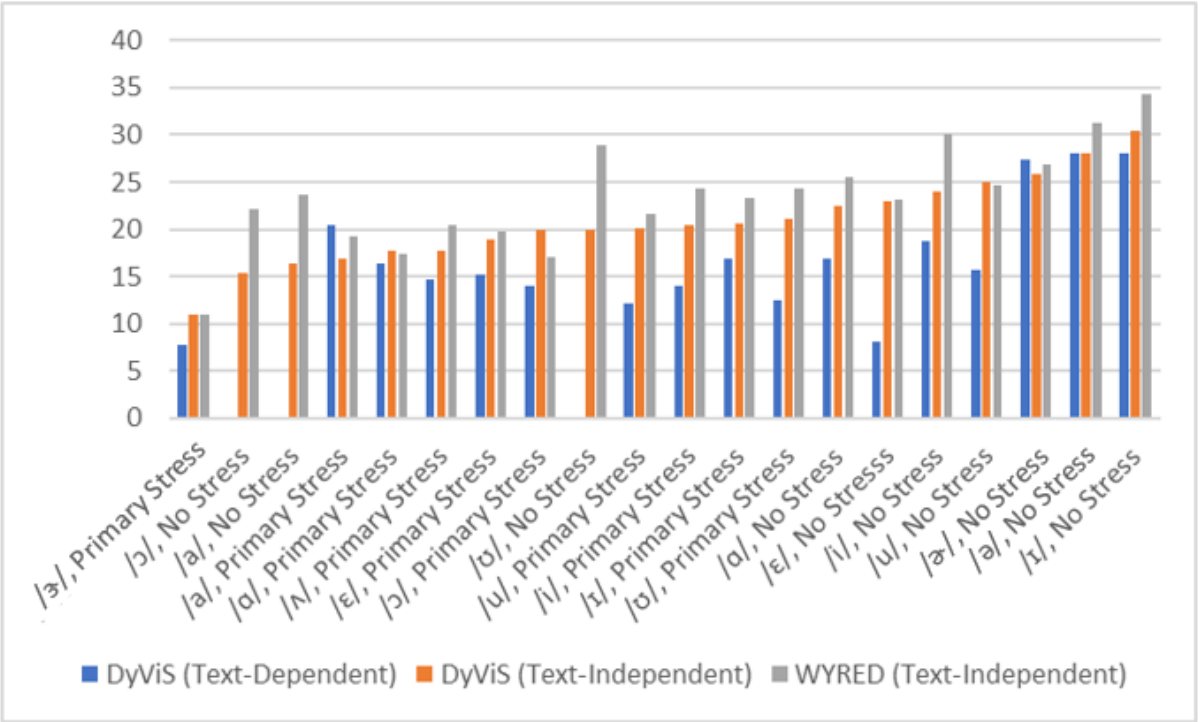
This offers a segue to an important fact to highlight: the rest of the phonemes all, expectedly, vary in frequency of production. This is simply because certain phonemes are more common than others. To illustrate this, the frequency of successful measurements for each phoneme can be found in Figure 5 below, and this shows how significant the disparity can be: here, /ə/ (no stress) consistently has over 200,000 more associated measurements than any other vowel, for example. More crucially, these frequency results show that some of these phonemes generate <15,000 tokens of data which, as discussed above, means they must also be trimmed to ensure analyses can occur. In all datasets, these are /ɑ/ (no stress), /ɛ/ (no stress), and /u/ (no stress). There are no dataset-specific trimmings here, and this is expected: the datasets are consistent in language and in topic, so the lexical items available and the phonemes used will be similar, as discussed in the previous chapter.

**Figure 5: Frequency of Successful Measurements per Phoneme**




Turning now to the percentage of "undefined" results with the remaining phonemes, Figure 6 below shows that sociophonetic variables do appear to impact successful data extraction from vowels; there appears to be more "undefined" results in text-independent data and in West Yorkshire-accented data particularly. That said, there are trends irrespective of sociophonetic variables: the best-performing vowel is always /ɝ/ (primary stress) and the worst-performing vowel is always /ɪ/ (no stress). As a final note, it should be said that no vowel quality measures (height; frontness) or stresses predict frequency or success rate.

**Figure 6: Percentage of "Undefined" Results per Phoneme**



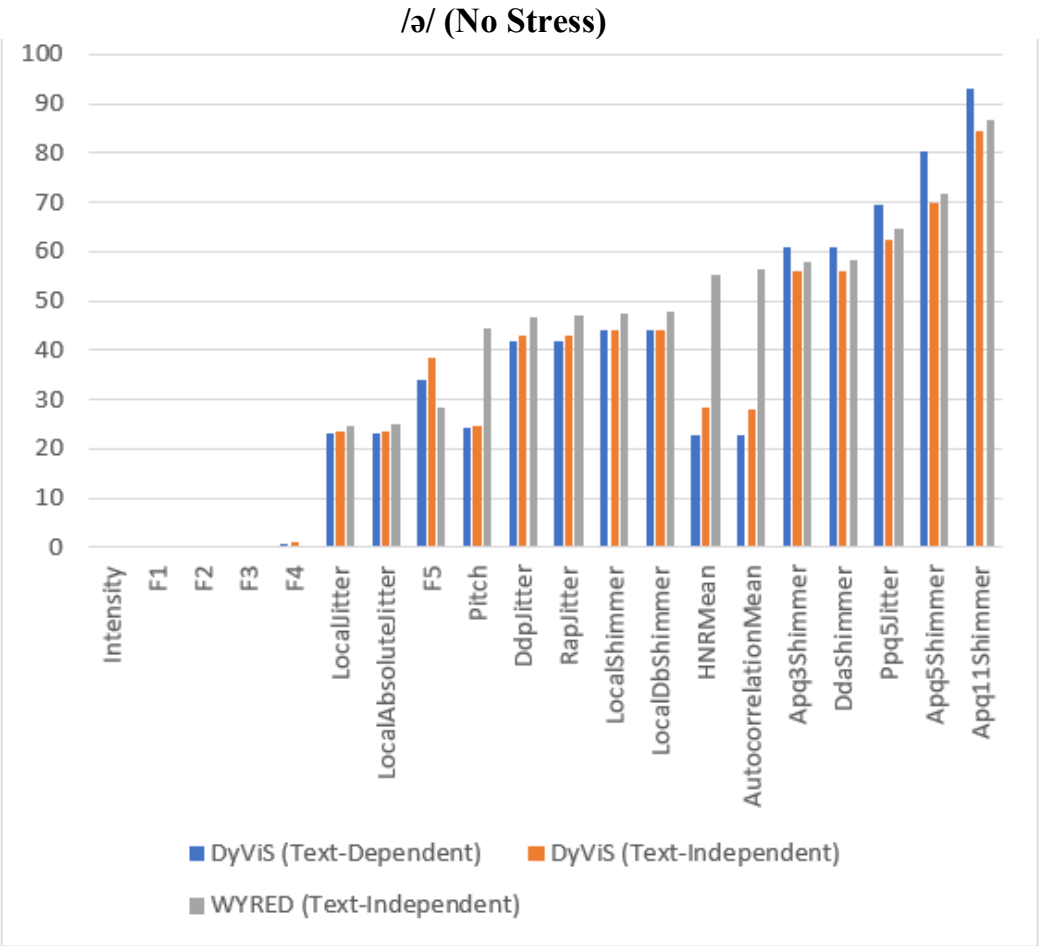

### 4.3. Interactions Between Feature and Vowel

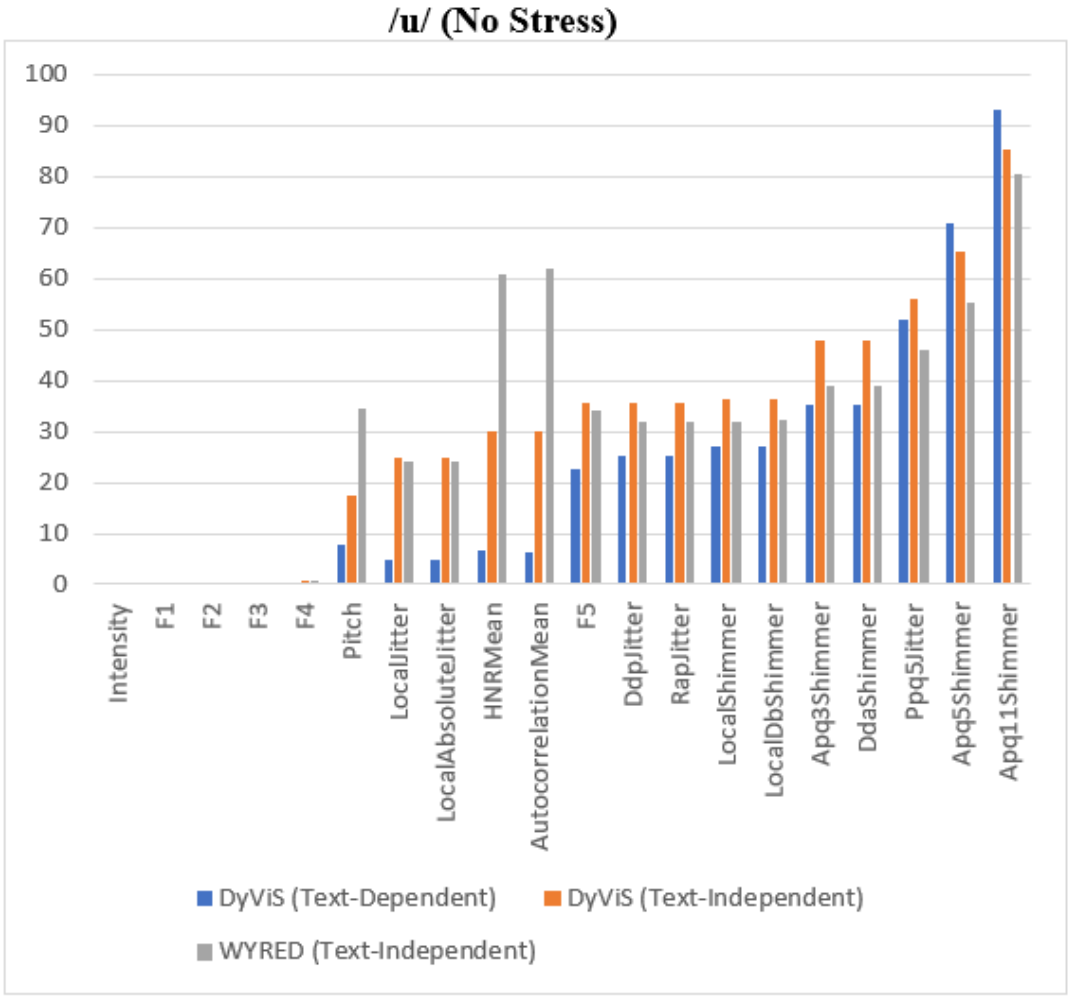
Having now isolated the results by feature and by vowel, a short investigation into the interactions between feature and vowel can now be conducted. Specifically, the success rate of each measurement's extraction will now be looked at on individual vowels; not across all vowels, as (4.1) did. This will show whether certain features work on specific vowels and not others, or whether the global trends identified in (4.1) apply at the individual vowel level as well.

In order to conduct this short investigation, the data from the most and least frequently produced vowels from (4.2), /ə/ (no stress) and /u/ (no stress), has been isolated. As seen below in Figure 7, the overall trends from (4.1) can indeed be seen in the results for individual vowels too: for both, the only features that are consistently used are intensity, formants 1-4, and the local jitter measurements. The only features that are consistently not used are formant 5, the remaining jitter measurements, and all shimmer measurements. *F0* is

sociophonetically-conditioned by accent and mean harmonics-to-noise ratio and mean autocorrelation are sociophonetically-conditioned by style.

**Figure 7: Percentage of "Undefined" Results per Phonetic Feature for /ə/ (No Stress) and /u/ (No Stress)**






### 4.4. "Undefined" Results by Speaker

Looking finally at speakers, the "undefined" results expectedly vary by the sociophonetic variable of style. Looking at the data from Nolan et al.'s (2009) DyViS corpus to illustrate this, in the text-dependent data the frequency of successfully extracted measurements per speaker is very stable: there is an average of 18,088 tokens per speaker with a very low standard deviation of 308. This stability is expected because the speakers all produced the same text-dependent transcript and, as per the design of this experiment to control speaker variables too, they are all similar in profile. The small variability must still be flagged, however, because the range is higher than 0 despite all speakers producing the same script.

This could be due to individual variation, or it could be a result of the potential inaccuracy of the MFA. How the MFA performed for the different databases is explored in later chapters.

Conversely, in Nolan et al.'s (2009) text-independent data, the frequency of extracted measurements per speaker is unsurprisingly much more variable than the text-dependent DyViS dataset: for example, Speaker(90) produced 66,040 tokens whilst Speaker(6) produced 14,014. This is expectedly variable; the speakers did not all produce the same data because this data is text-independent. Trends from Gold et al.'s (2018) text-independent data also support this: Speaker(54) produced 51,792 whilst Speaker(19) produced 15,496. Speakers simply produce different amounts of tokens of the target phonemes to each other in text-independent speech.

In terms of successful data extractions, certain speakers produced more tokens that invoked more "undefined" measurements than others. In the text-dependent data from Nolan et al.'s (2009) DyViS corpus, Speaker(88) produced 32.06% "undefined" results compared to Speaker(76) who produced 7.47% "undefined" results, for example. Whilst this range is smaller compared to the features and phonemes, this speaker variation shows that different speakers still produce tokens of the selected phonemes in more successfully measurable ways than others; it is evidence of individual variation beyond the sociophonetic level. It cannot be due to any other shared sociophonetic qualities between certain speakers because the speakers cannot be grouped in any further way that could predict the success rate of data extraction; this grouping was already controlled during data selection (3.1). This is therefore individual variation and could be beneficial to monitor for individual ASR as a result.

That said, the success rate of data extraction with these individual speakers is not predictable when style varies: different speakers produced the most and least "undefined" results in the text-dependent and -independent data from Nolan et al.'s (2009) DyViS despite the exact

same speakers being used. In the text-independent data, Speaker(85) produced 33.97% "undefined" results compared to Speaker(60) who produced 13.89% "undefined" results. As discussed above, in the text-dependent data Speaker(88) produced 32.06% "undefined" results compared to Speaker(76) who produced 7.47% "undefined" results. This indicates that there are individual factors beyond style that can affect successful data extraction, and this again may be a window into speaker individuality.

In summary of this chapter, the main differences between the databases lay in the features that can be successfully extracted: intensity, formants 1-4, jitter (local), and jitter (local, absolute) were always extractable, but mean harmonics-to-noise ratio and mean autocorrelation were only extractable in the text-dependent data and *f0* was only extractable in the SSBE data. Beyond this, similar degrees of phoneme and speaker variation were seen in the databases, but text-dependent and SSBE-accented speech tended to have more successful data extraction in general.

## 5. Variation

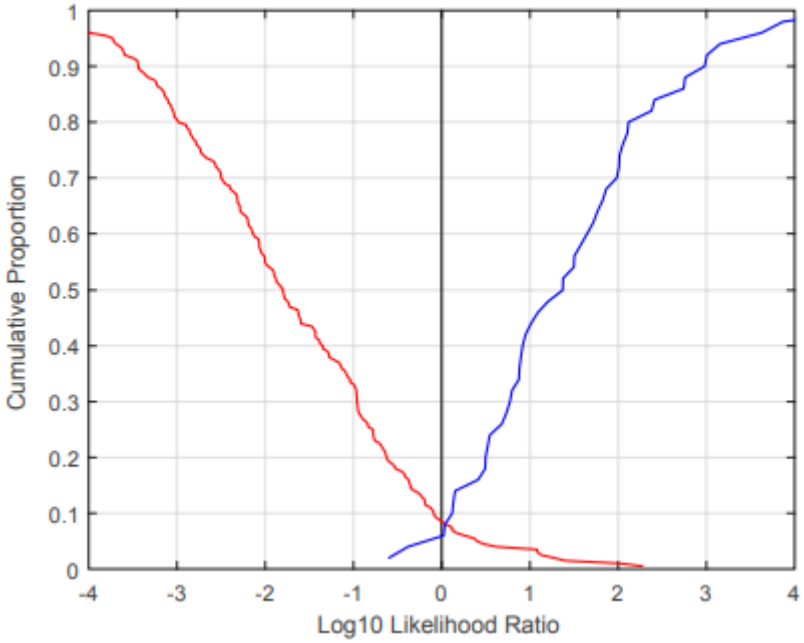
Having now filtered out the "undefined" results and investigated the information they offer regarding the extractability of the phonetic features, phonemes, and speakers, this chapter now looks at the ASR results and the performance of these phonetic approaches. This investigation will be split into four main sections: in the first section, the Tippett plots used to explore these results will be described in detail (5.1). In the second section, the overall performance of every feature and vowel combined together will be explored to show what these combinatory phonetic approaches are capable of without any (socio)phonetic tailoring (5.2). Then, sociophonetic tailoring related to accent and style will be considered to gain deeper insights into how these combinations work and what the effects of sociophonetic tailoring are on performance (5.2). The third section will then explore the performance of individual features (5.3) and the fourth section will explore the performance of individual phonemes (5.4). The later sections (5.3-5.4) therefore dive deeper into the combinations identified in (5.2) to gain better insights into what individual phonetic elements are contributing to performance. The purpose of this chapter is to explore the raw performance of phonetic approaches. This will set up the final results chapter wherein the optimised combinations of features and segments for different sociophonetic groups will be explored.

### 5.1. Tippett Plots

In order to visualise the performance of all of these phonetic features and segments together, Tippett plots have been used (Meuwly, 2001). These, as Morrison et al. (2021) write, allow for the probability distributions of same-speaker and different-speaker $\log_{10}$ LRs (Likelihood Ratios) to be visualised together. These should always supplement $C_{llr}$ values, which the following chapter explores. Tippett plots provide rich information about the output of any recognition system, including the magnitude of the strength of evidence the system is capable

of producing, the magnitude of the contrary-to-fact results, the extent to which the system is well calibrated, and the overall validity of the system. In order to explain Tippett plots in greater detail, Figure 8 will be used which illustrates a well-performing speaker recognition system from Morrison et al. (2021).

**Figure 8: An Example of a Well-Performing Speaker Recognition System from Morrison et al. (2021)**




In Figure 8, the y-axis represents the cumulative proportion, or the percentage of comparisons made between speakers. The x-axis represents the comparisons as $\log_{10}$ Likelihood Ratios (LLRs). LLRs were explained in (3.4), but in essence these are calibrated scores representing every same-speaker (SS) and different-speaker (DS) comparison possible using the feature measurements as input (3.3). Before explaining the slopes, Table 8 below from Champod and Evett (2000) will be reviewed which translates these LLRs into verbal descriptions which ease the interpretation of Tippett plots. Note that their terminology refers to SS evidence as 'prosecution' evidence and DS evidence as 'defence' evidence. This terminology will be used hereon to avoid confusion when referring back to Table 8.

**Table 8: Champod and Evett's (2000) Descriptions of $Log_{10}$ LRs**

| Evidence | $Log_{10}$ LR | Description |
|---|---|---|
| **Support for Prosecution (SS)** | 4 to 5 | Very Strong Evidence |
| | 3 to 4 | Strong Evidence |
| | 2 to 3 | Moderately Strong Evidence |
| | 1 to 2 | Moderate Evidence |
| | 0 to 1 | Limited Evidence |
| **Neutral** | 0 | Neutral Evidence |
| **Support for Defence (DS)** | 0 to -1 | Limited Evidence |
| | -1 to -2 | Moderate Evidence |
| | -2 to -3 | Moderately Strong Evidence |
| | -3 to -4 | Strong Evidence |
| | -4 to -5 | Very Strong Evidence |

Turning back to Figure 8 and the slopes, it will now be explained why this Tippet plot shows that this system is well-performing. Here, the slope extending to the right (blue) visualises the cumulative distribution of the SS LLRs whilst the slope extending to the left (red) visualises the inverse cumulative distribution of DS LLRs. As seen, over 90% of the SS pairs present support for prosecution by extending to the right of 0 (positive integers). 10% of that is strong evidence (3 to 4), 20% of that is moderately strong (2 to 3), 30% is moderate, and the remaining 30% is limited (0 to 1). Conversely, less than 10% of the SS pairs present support for the defence, extending to the left of 0 (negative integers) and this evidence is only limited (0 to -1).

Looking at the DS pairs, over 90% of these pairs present support for the defence by extending to the left of 0 (negative integers). 20% of these present strong evidence (-3 to -4), 20% present moderately strong evidence (-2 to -3), 30% present moderate evidence (-1 to -2), and 20% present limited evidence (0 to -1). Conversely, less than 10% of the DS pairs present support for the prosecution, extending to the right (positive integers). The majority of this evidence is limited (0 to 1) at 6%, but 3% is moderate (1 to 2) and 1% is moderately strong (2 to 3).

Overall, this Tippett plot shows that the system performs well for SS and DS comparisons, with over 90% of the SS and DS pairs supporting the prosecution and defence respectively, as intended. This supporting evidence also includes strong evidence, and the SS and DS pairs which do not support the prosecution and defence respectively are minimal and do not include any strong evidence. Furthermore, as the lines intersect close to 0 on the x-axis, this shows that the system is well-calibrated between the SS and DS LLRs with no bias towards SS comparisons, wherein the intersect would be higher than 0 (to the right), or DS comparisons, wherein the intersect would be lower than 0 (to the left).

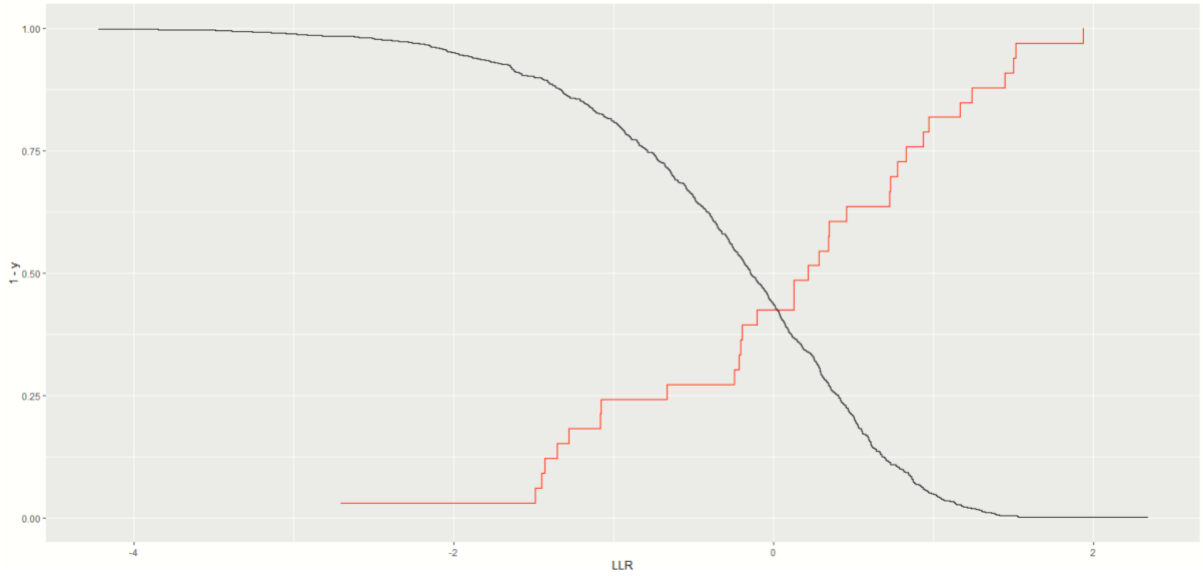
### 5.2. Overall Performance of Combinatory Phonetic Approaches

Having now explained Tippett plots, the data from this thesis will be analysed using them. In the following graphs, the line extending to the right that visualises the cumulative distribution of the SS LLRs is now red (this was blue above) whilst the line extending to the left that visualises the inverse cumulative distribution of DS LLRs is black (this was red above).

Figure 9 below presents the results from combining all possible data irrespective of feature, vowel, or sociophonetic group. This was done using the test, background, and calibration speakers from all 3 datasets, following the methodology in (3.4). This graph will now be interpreted just like the example above. Looking first at the SS LLRs (the red line), only

around 56% of the comparisons present support for the prosecution (extending to the right). Of this, only roughly 20% can even be considered moderate evidence (1 to 2); the rest is all considered limited evidence (0 to 1). A high 44% of comparisons presents problematic support for the defence (extending to the left), of which some extends to moderately strong evidence (-2 to -3). Turning now to the DS LLRs (the black line), only 56% of the comparisons present support for the defence here (extending to the left), though some of this is very strong evidence (-4 to -5). That said, 44% of the comparisons therefore presents problematic support for the prosecution (extending to the right), and some of this extends as far as moderately strong (-2 to -3). Therefore, this Tippett plot shows an overall poor performance due to the high frequency of contradictory evidence for both the prosecution and defence. However, it at least appears well-calibrated as the lines intersect at 0; there is no evident bias towards either the SS or DS comparisons.

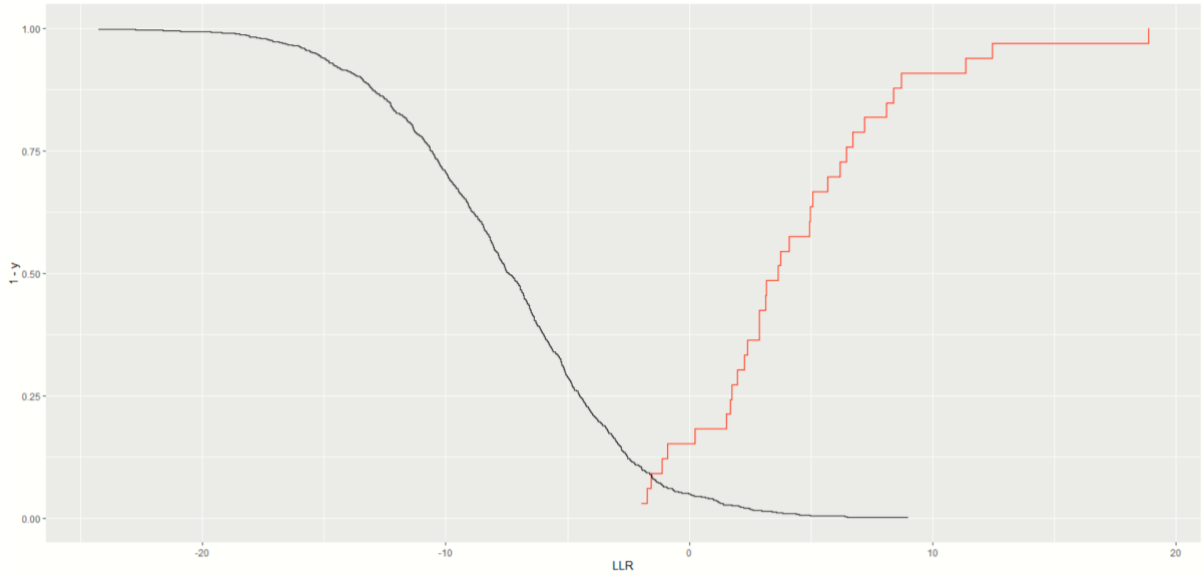
**Figure 9: Performance of All Phonetic Features and Segments Combined from All Databases**




Turning now to Figure 10, this calibrated system includes all features and phonemes but only for one sociophonetically-controlled database, here the text-dependent database from DyViS.

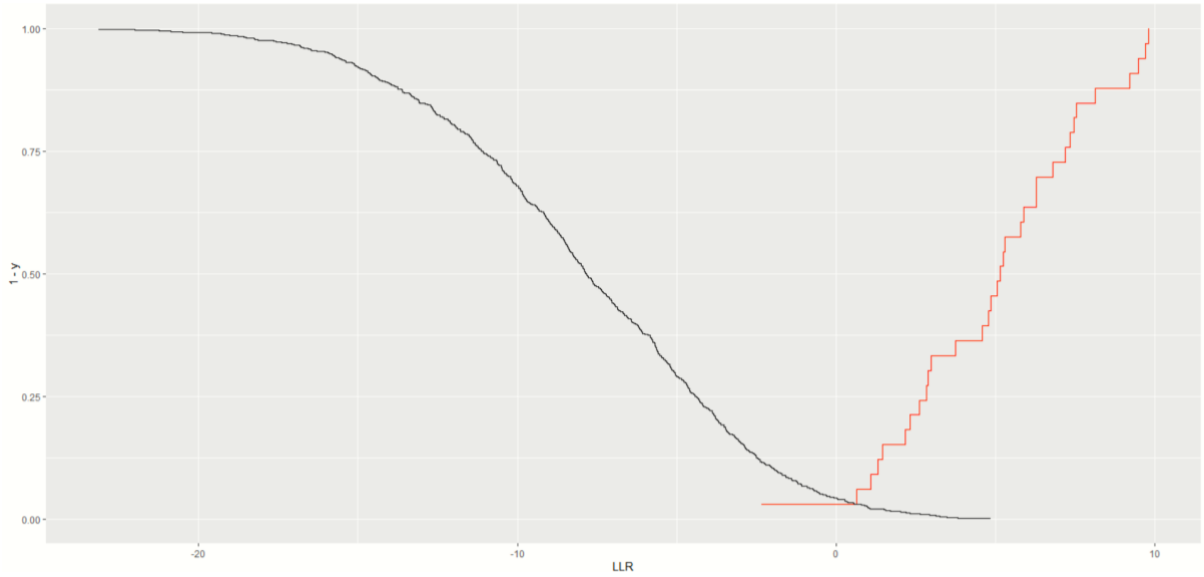
Firstly, some bias towards the DS comparisons may be present: the curves intersect at -2 and not 0. Turning to the SS LLRs, this means that around 80% of comparisons show support for the prosecution. This is much better than the previous system without any sociophonetic tailoring. Of this evidence, some was extremely strong, going as high as +18 (far exceeding Champod and Evett's (2000) scale). The 20% showing problematic support for the defence only ever extended as far as being limited evidence (0 to -1). Due to the bias towards the defence, around 95% of comparisons for the DS LLRs showed support for the defence, and again this evidence was extremely strong with the highest going to -24. This, again, is much better than the prior system. Of the 5% showing problematic support for the prosecution, some of this was extremely strong and extended to +9. This overall shows that the combinatory performance of all phonetic features in the text-dependent database from DyViS is strong, but with minor bias towards the defence that hampers balanced performance. What is most important here, however, is how much stronger this is compared to the previous approach: more SS comparisons support the prosecution, more DS comparisons support the defence, and the strength of evidence from these comparisons is overall much stronger. Sociophonetic tailoring appears to improve performance.

**Figure 10: Performance of All Phonetic Features and Segments Combined for DyViS (TD)**



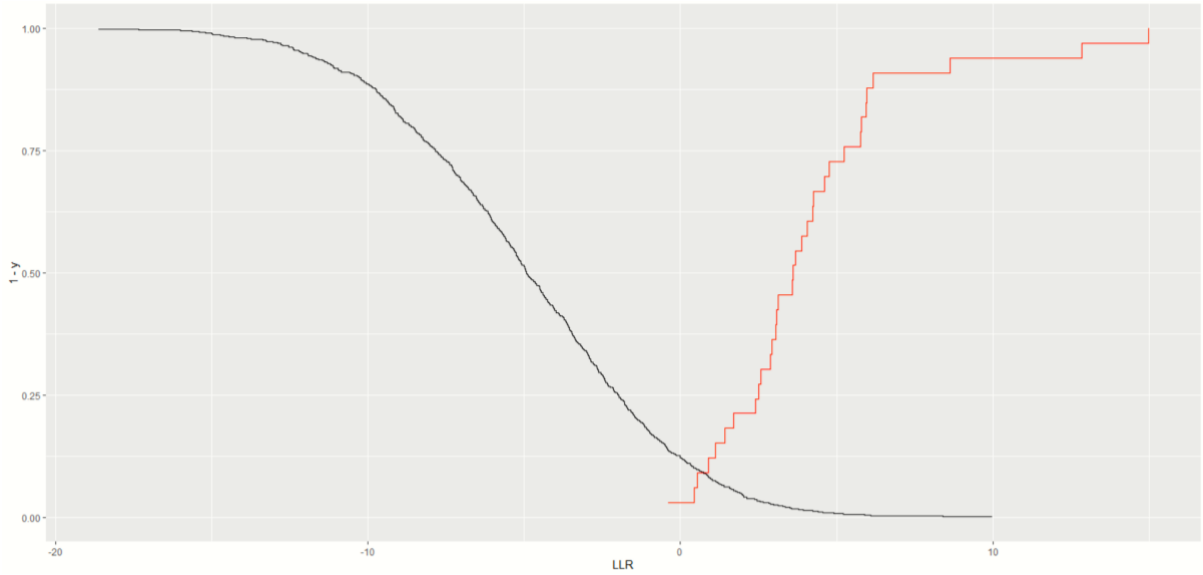

Compare this, however, to the performance of the system using the text-independent database from DyViS in Figure 11 below. There is a minor bias towards the prosecution with the lines intersecting around 1, but looking at the SS LLRs, 97% shows support for the prosecution with some evidence extending as far as +10, presenting extremely strong evidence. The limited 3% of problematic support towards the defence only extends as far as moderate evidence. Turning to the DS LLRs, 95% showed support towards the defence and some of this extended as far as -24, again showing extremely strong support. However, the limited evidence problematically supporting the prosecution extended as far as +5, showing very strong evidence. Despite this small bias towards the prosecution, this is a much more balanced and better-performing approach; phonetic approaches thus far appear better-suited to text-independent data, contrary to the historic movement away from phonetic approaches seen in earlier chapters. This shows that style is a sociophonetic variable that can affect performance.

**Figure 11: Performance of All Phonetic Features and Segments Combined for DyViS (TI)**




Turning finally to the text-independent database results from WYRED in Figure 12 below, these are mostly similar to those seen in the DyViS text-independent database. There is again a minor bias towards the prosecution as the intersection of the lines is at around 1. Looking at the SS LLRs, this is again very strong with around 97% showing support for the prosecution and some of this evidence is extremely strong, extending as far as +15; better than that seen for the SSBE text-independent speech. That said, however, for the DS LLRs performance is slightly worse with around 88% showing support for the defence. Some of this evidence is extremely strong, extending as far as -18, but of the 12% showing problematic support for the prosecution, some of this is very strong at +10. Thus, this is still a well-performing system, but with notable biases towards the prosecution. As the SSBE text-independent dataset performed better than the West Yorkshire text-independent dataset, accent therefore appears to affect performance, again showing that sociophonetic information is important.

**Figure 12: Performance of All Phonetic Features and Segments Combined for WYRED (TI)**




In summary: of the two text-independent databases, performance was better for the SSBE accents than the West Yorkshire accents, though both showed minor bias towards the prosecution. Text-independent speech performed noticeably better than the text-dependent speech, and all of the sociophonetically-tailored approaches greatly outperformed the approach where sociophonetic variables (accent and style) were not considered; this system performed particularly badly.

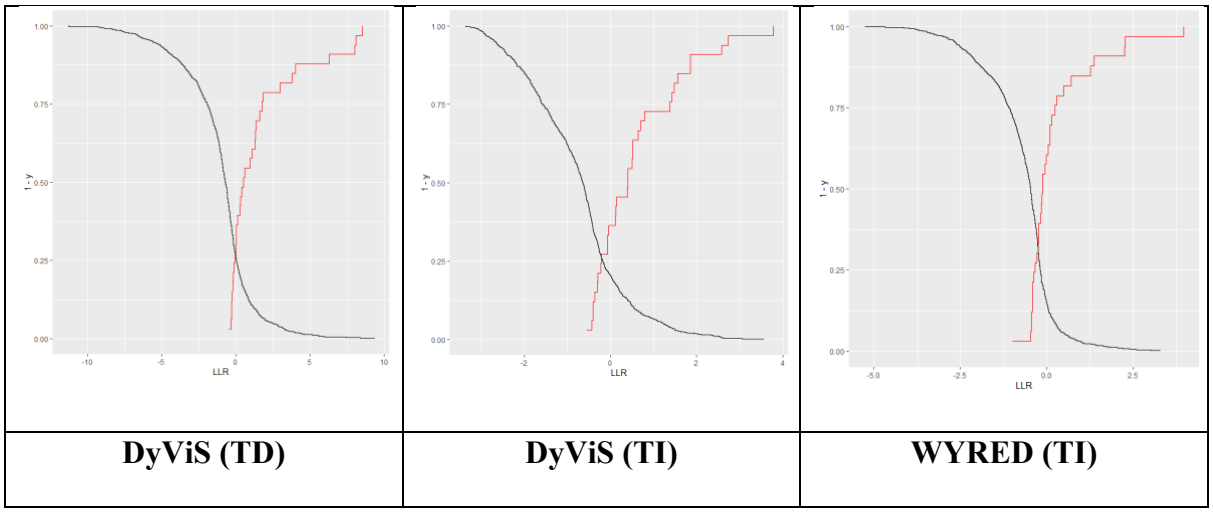
### 5.3. Individual Feature Performance

The previous section indicated that phonetic approaches can be powerful when sociophonetic variables are tailored towards. Now, the individual performance of some of the selected phonetic features that made up those combinations will be isolated and explored in each database. This allows for a deeper dive into the above results and how the features contributing to them behave.

Turning first to intensity, Figure 13 below shows no biases towards the prosecution or defence; they intersect at 0 for all databases and therefore produced well-calibrated results,

unlike the above combinations of all features. That said, these individual feature analyses do not outperform the combinations above: for each, only 70-75% of the SS and DS comparisons support the prosecution and defence respectively. There is notably stronger evidence both ways for the text-dependent speech from DyViS, however, extending as far as +/-10. This feature may therefore be affected by style and is more effective with text-dependent speech.

**Figure 13: Intensity Performance**

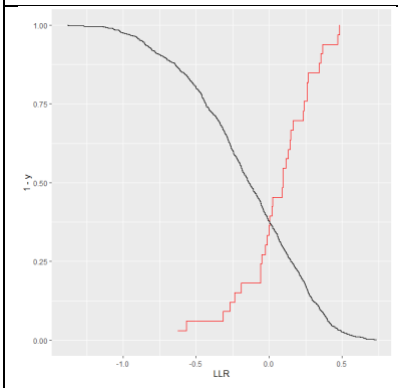
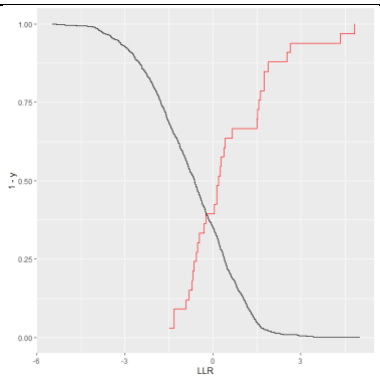
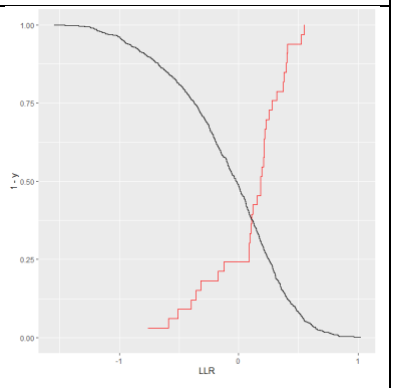
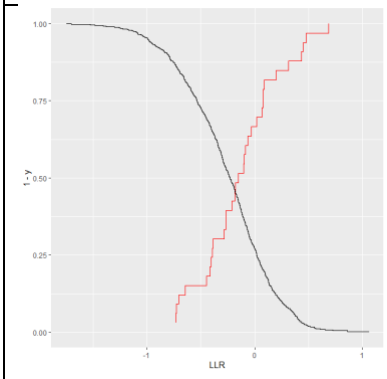
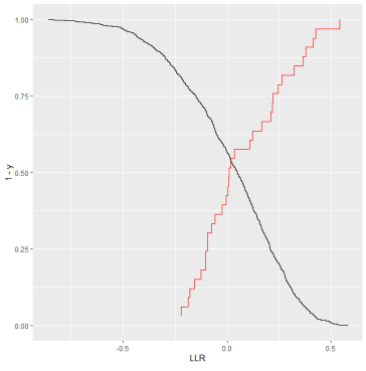
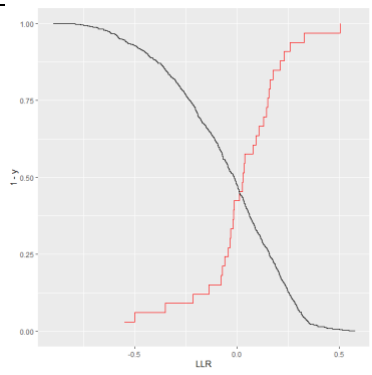



| DyViS (TD) | DyViS (TI) | WYRED (TI) |
|---|---|---|

Next, F1 and F2 will be reviewed together as the lower formants in Figure 14 below. Firstly, F1 appears to perform worst in the text-independent database from DyViS. The LLRs for all databases are well-calibrated with intersects at 0 and performance is always mediocre, shown through only 63% of SS and DS comparisons supporting the prosecution and defence respectively. However, the DS LLRs from the text-independent database from DyViS have stronger contrary evidence problematically supporting the prosecution (-5) and the SS LLRs have stronger contrary evidence problematically supporting the defence (+4.5). By contrast, the strength of contrary problematic evidence only extends as far as +/-1 in the other databases. Of other interest, F2 is typically performing worse than F1 in all databases, with

only 50% of SS and DS comparisons ever supporting the prosecution and defence respectively.

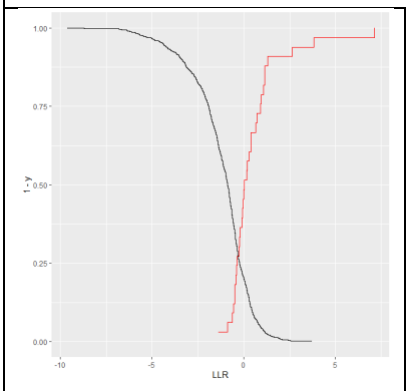
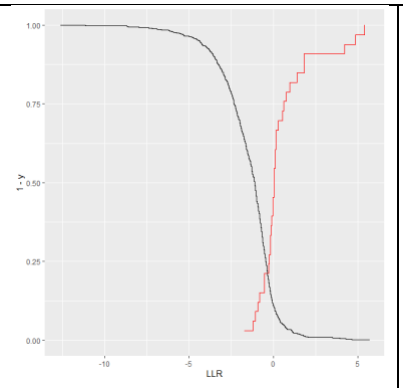
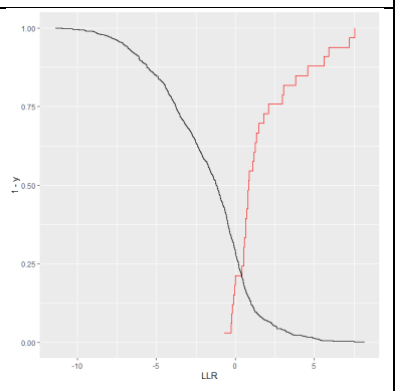
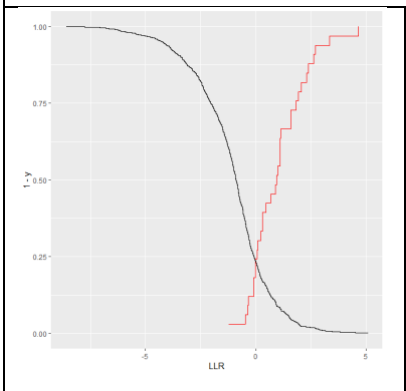
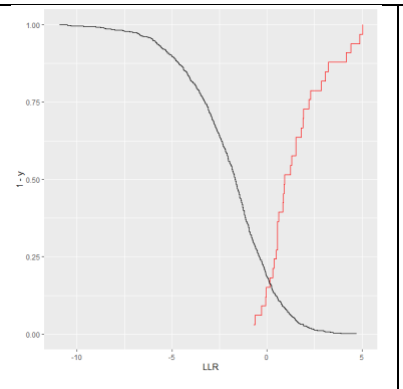
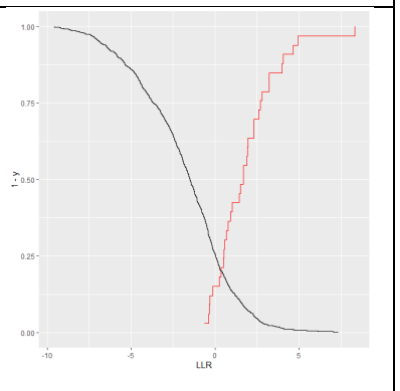
**Figure 14: Lower Formant Performance**

| **Formant 1** | | |
|---|---|---|
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |
| **Formant 2** | | |
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |

Next, F3 and F4 will be reviewed together as the higher formants in Figure 15 below. For all databases, F3 performed much better than the lower formants. This is because, for all databases, over 75% of the SS and DS comparisons support the prosecution and defence respectively. F4, however, is the best performing formant overall: roughly 77-80% of the SS and DS comparisons support the prosecution and defence respectively. For both higher

formants, SS and DS evidence supporting the prosecution and defence respectively can extend as far +/-10, exhibiting extremely strong evidence, but some DS evidence problematically supports the prosecution very strongly (+5).

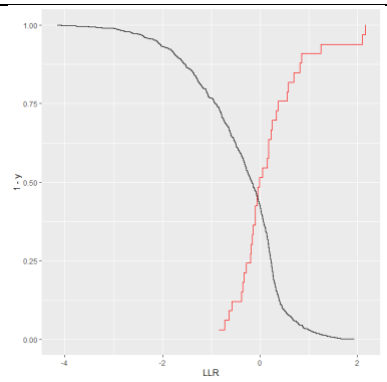
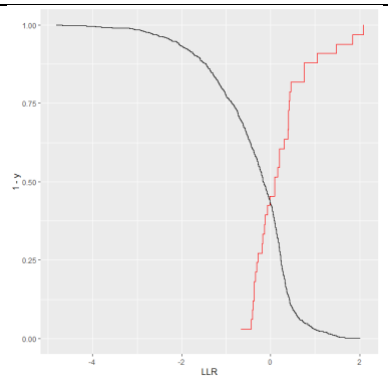
**Figure 15: Higher Formant Performance**

| **Formant 3** | | |
|---|---|---|
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |
| **Formant 4** | | |
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |

The final features of interest to discuss are the mean value measurements in Figure 16 below: mean harmonics-to-noise ratio and mean autocorrelation. These only appeared in the text-dependent database for DyViS and performance can be seen in Figure 16 below. The main point to flag here is that performance is poor: only 50% of the SS and DS comparisons

support the prosecution and defence respectively, and none of this evidence exceeds moderate strength (2).

**Figure 16: Mean Measurement Performances from DyViS (TD)**

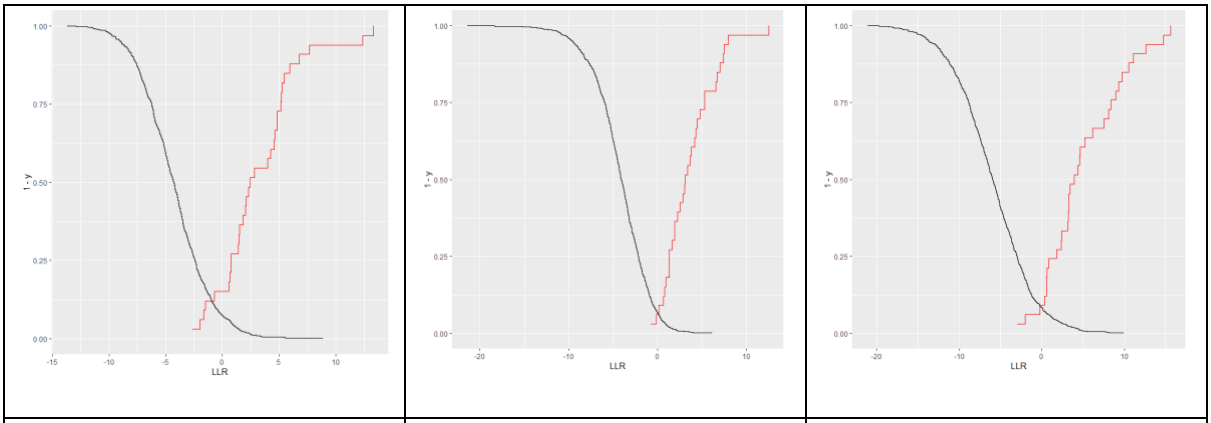
|  |  |
|---|---|
| **Mean Harmonics-To-Noise Ratio** | **Mean Autocorrelation** |

### 5.4. Individual Segment Performance

Now, the individual performances of each phoneme per database, including all features, will be explored. Generally, performance for the vowels is very consistent across the databases, but some specific vowels are worthy of more detailed discussion here.

The first of these to discuss is /ə/. As seen in Figure 17 below, 88-85% of SS and DS comparisons support the prosecution and defence respectively; this is the best performing vowel in that respect. However, of note is that of the DS evidence problematically supporting the prosecution, some goes as far as being extremely strong (+10). This is true for this vowel in every database, as seen. This shows that sociophonetic variables do not affect the performance of some vowels.

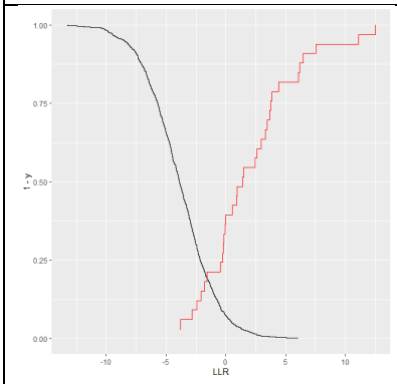
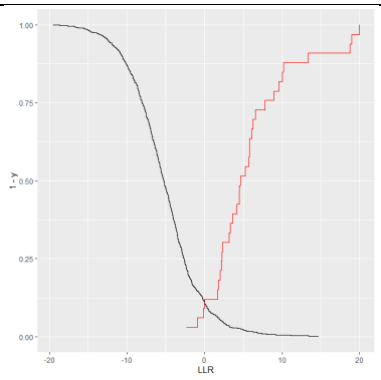
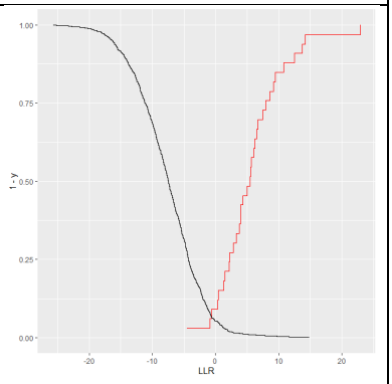
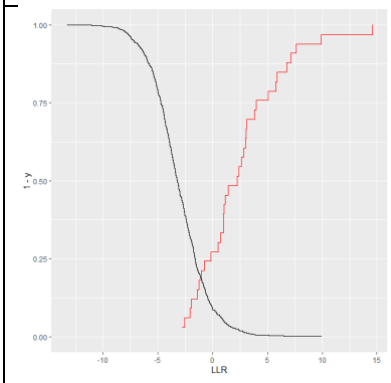
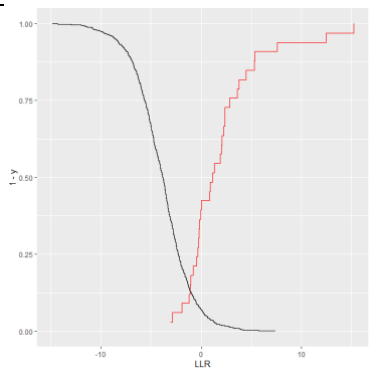
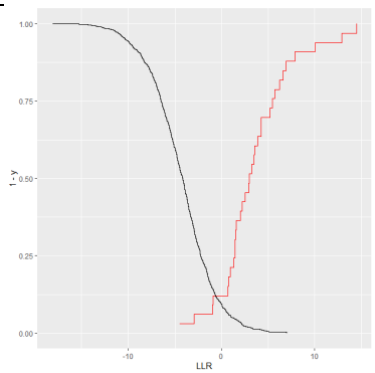
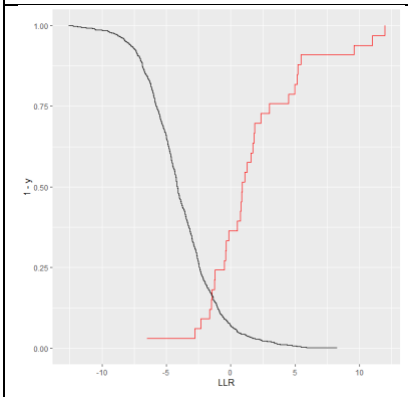
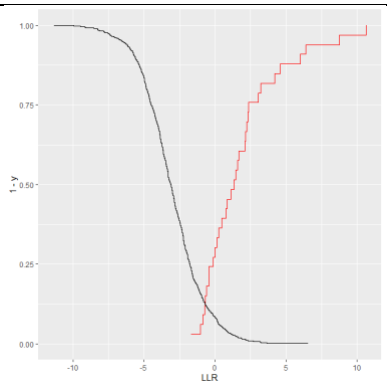
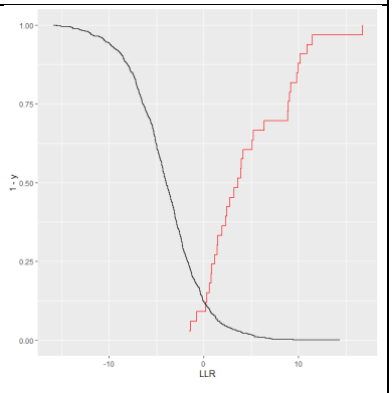
**Figure 17: /ə/ Performance**



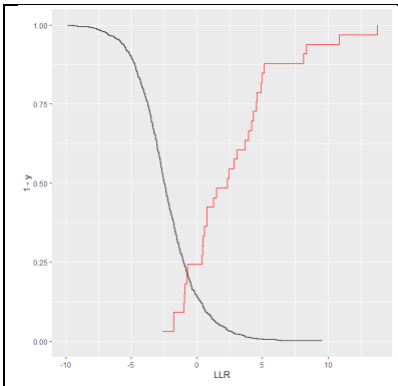
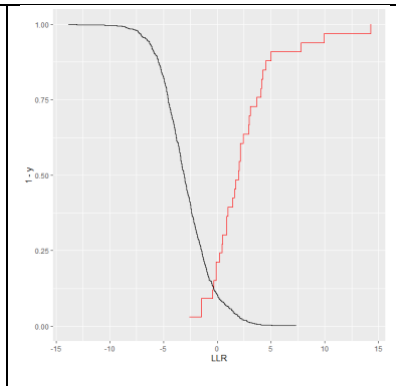
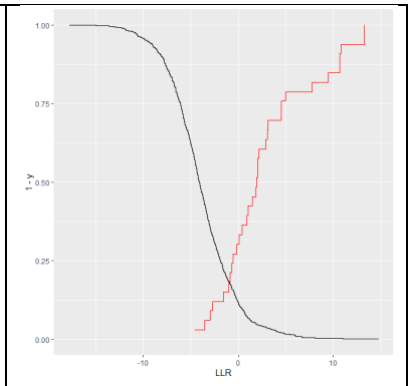
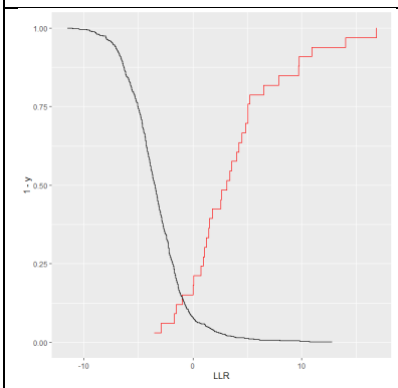
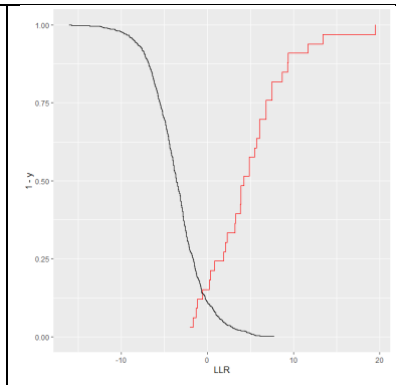
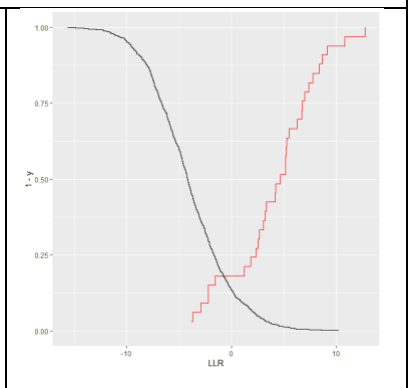

| DyViS (TD) | DyViS (TI) | WYRED (TI) |
|---|---|---|

Building off of this, three other vowels show a similar trend wherein they mostly perform well for SS and DS comparisons supporting the prosecution and defence respectively, averaging around 88%, but also show evidence of strong contrasting problematic evidence. These are /ɛ/, /ɪ/, and /i/. These have been shown below in Figure 18. Also of note is that, generally, the text-independent datasets are outperforming the text-dependent dataset here, and of these text-independent databases DyViS is outperforming WYRED as the DS comparisons problematically supporting the prosecution in WYRED extend as far as extremely strong (+15). This shows, contrastingly, that some vowels are affected by sociophonetic variables.

**Figure 18: Additional Vowel Performances with Strong False Rejections in Same-Speaker Comparisons**

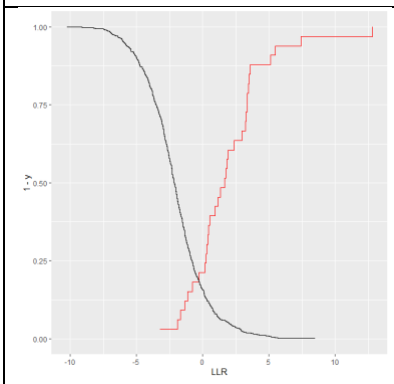
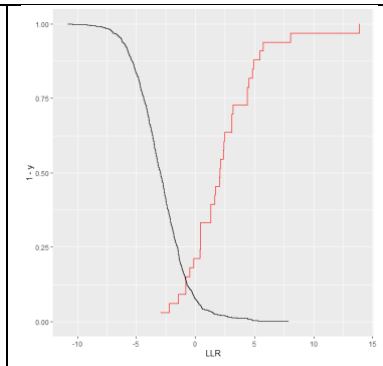
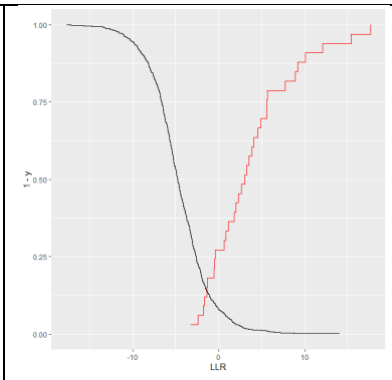
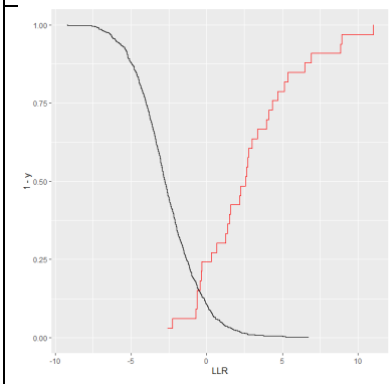
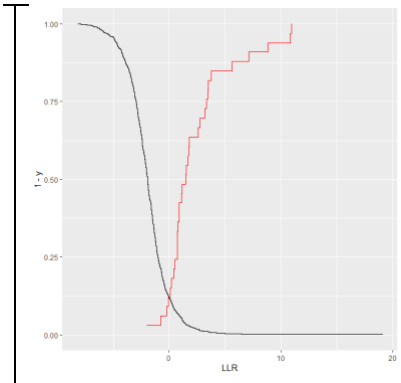
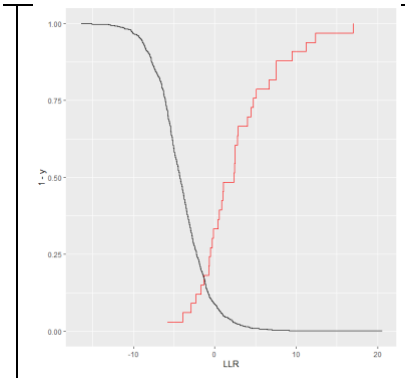
| /ɛ/ (Primary Stress) | | |
|---|---|---|
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |
| **/ɪ/ (No Stress)** | | |
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |
| **/ɪ/ (Primary Stress)** | | |
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |

| /i/ (No Stress) |
|---|

| | | |
|---|---|---|
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |
| **/i/ (Primary Stress)** | | |
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |

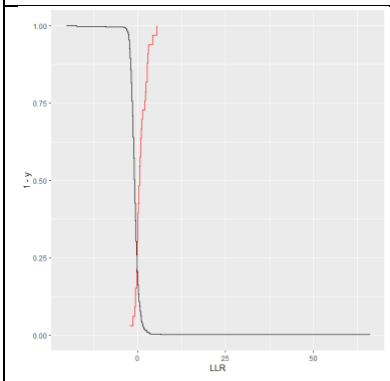
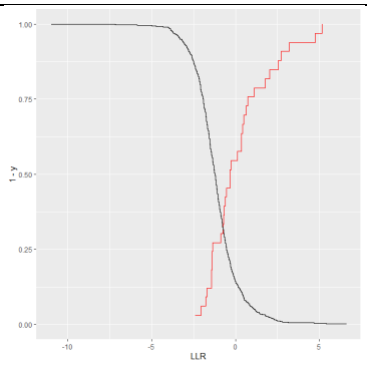
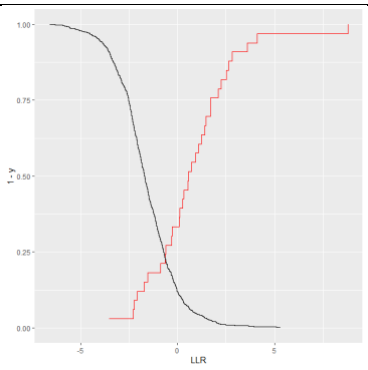
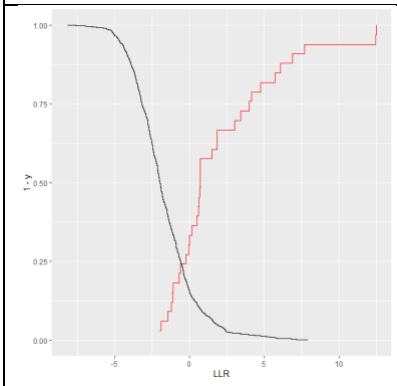
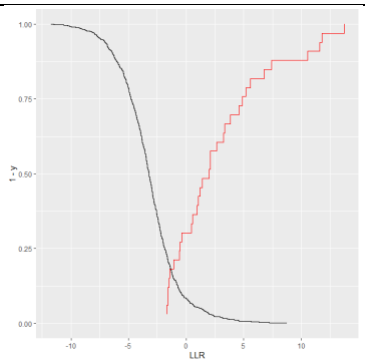
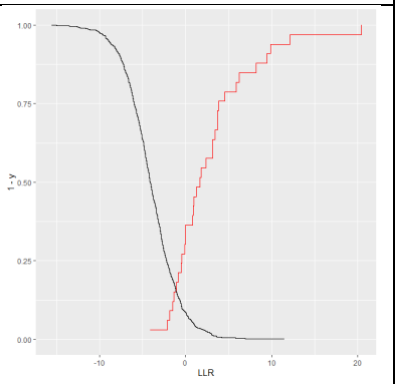
Conversely, /ɚ/ (no stress) and /ɝ/ (primary stress) are other vowels displaying strong performance on average, yielding around 88% of SS and DS comparisons supporting the prosecution and defence respectively, yet extremely strong DS evidence problematically supporting the prosecution (+10-20). This is seen in Figure 19 below. This is particularly true of /ɝ/ (primary stress) where the DS evidence problematically supporting the prosecution is stronger. This vowel exhibits the worst different-speaker comparison performance as a result.

**Figure 19: /ɚ/ (No Stress) and /ɝ/ (Primary Stress) Performance**

| /ɚ/ (No Stress) | | |
|---|---|---|
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |
| **/ɝ/ (Primary Stress)** | | |
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |

The final vowels of particular note are the close back vowels /ʊ/ and /u/. These vowels performed the worst for a variety of reasons, as seen in Figure 20 below: both vowels show roughly only 75% of SS and DS comparisons supporting the prosecution and defence respectively; the worst performance of the vowels. Most notably, for /ʊ/ (primary stress), this performed especially badly for the text-dependent data from DyViS: the SS comparisons problematically supporting the prosecution is moderately strong, but the DS comparisons problematically supporting the prosecution are stronger (+60).

**Figure 20: Close Back Vowel Performances**

| /ʊ/ (Primary Stress) | | |
|---|---|---|
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |
| **/u/ (Primary Stress)** | | |
|  |  |  |
| **DyViS (TD)** | **DyViS (TI)** | **WYRED (TI)** |

## 6. Portfolios

Having now trimmed non-numerical data, explored the performance of the datasets with Tippett plots, and observed individual feature and segment performances, portfolios will now be built which test different combinations of phonetic features and vowels to identify which are outright best for ASR tasks for the different datasets. More specifically, different combinations of features will now be tested for each phoneme for each style and accent, following this thesis' applicable and repeatable methodology for identifying novel, phonetically-informed approaches to ASR. This chapter will overview these portfolios, but detailed descriptions of each of them can be found in Appendix A. Those descriptions serve as a reference guide for the portfolios which future researchers may wish to employ, replicate, or update. This chapter will be structured as follows: in (6.1), the overarching scores and the importance of sociophonetic specificity will be established using $C_{llr}$ values that link to the Tippett plots in (5.2). In (6.2), the importance of each individual feature will be explored using $C_{llr}$ values and the 'top down' approach discussed in (3.4). Finally, the importance of vowel-specificity will also be explored using $C_{llr}$ values and this 'top down' approach (6.3). After this, the portfolios of features per vowel, per sociophonetic group will be summarised from Appendix A (6.3).

### 6.1. The Importance of Sociophonetic Specificity

The overarching $C_{llr}$ values were calculated for the Tippett plots seen in (5.2). These can be found in Table 9 below. As a summation of the above discussions regarding $C_{llr}$, scores closer to 0 are stronger, and those closer to 1 are worse. As seen, the overall combination of all data from all three corpora was 1.05. Whilst this performance is very poor, this still serves as a benchmark which the (socio)phonetic features and phonemes can be compared to as a metric for what effects they have on performance. As seen, all three sociophonetically-tailored

systems outperform this significantly. On top of this, the text-independent speech outperforms the text-dependent speech, showing style as having an impact on performance, and of these text-independent scores the SSBE database performs best, showing that accent has an impact on performance too. This, overall, shows that sociophonetic tailoring is important to ASR performance.

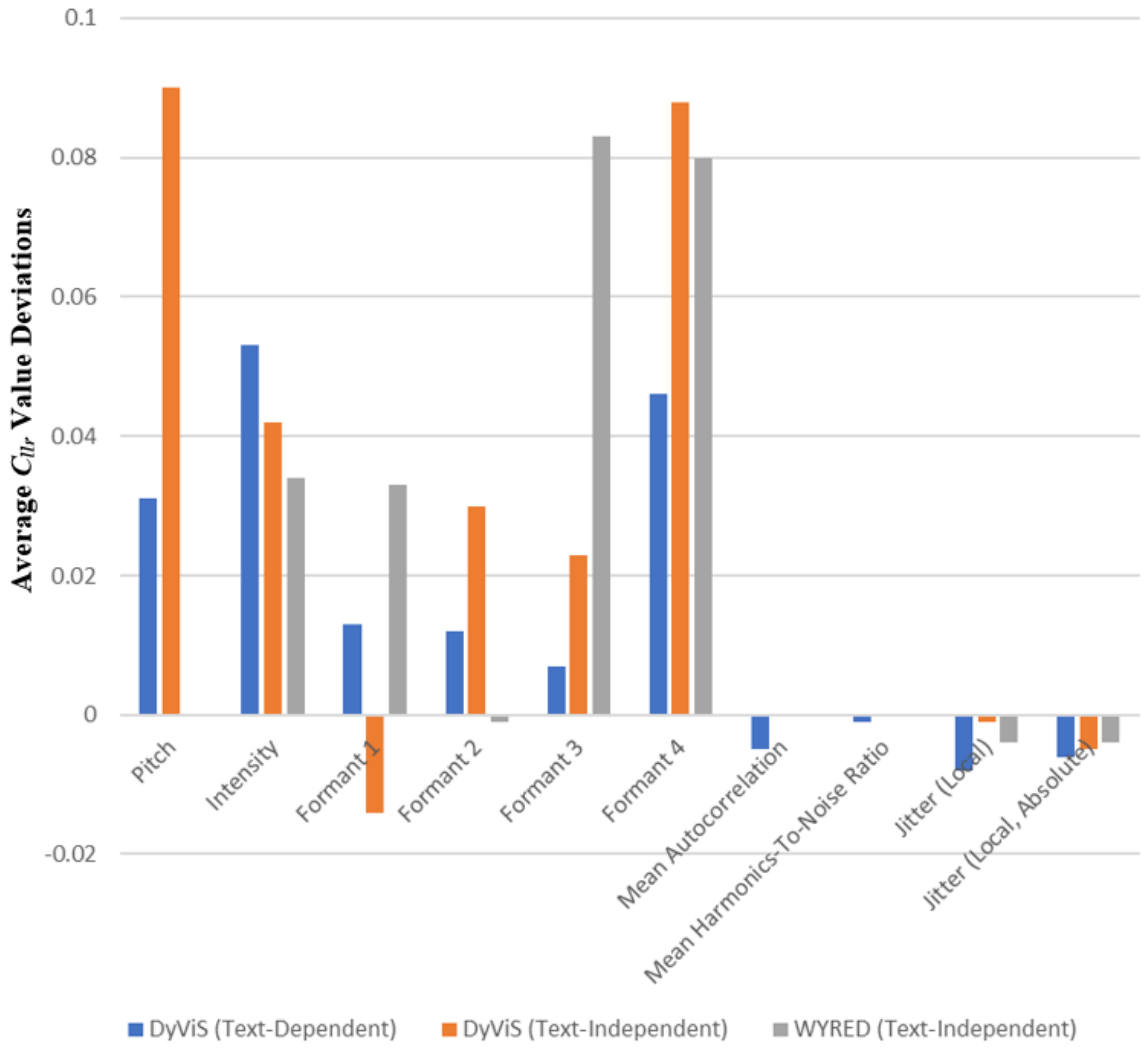
**Table 9: Overall and Baseline $C_{llr}$ Value Deviations**

| | |
|---|---|
| **Full Combination** | 1.05 |
| **DyViS (TD)** | 0.32 |
| **DyViS (TI)** | 0.15 |
| **WYRED (TI)** | 0.23 |

### 6.2. The Importance of Individual Features

From here, the individual feature performances will be assessed using the top-down approach discussed in section 3.4. As a recap, the performance of each feature will be isolated by removing each feature one at a time and charting the effect its removal has on the scores above. This will be done for each individual database. The results in Figure 21 below visualise these changes in performance as deviations from the above baselines: reductions in score mean the approaches perform better without those features, so these features should be removed. Increases in score mean the approaches perform worse without those features. These are the features that need to be retained.

**Figure 21: Average $C_{llr}$ Value Deviations Per Feature, Per Database**




As seen, all features related to non-modal voicing, and those that also proved most difficult to extract in chapter 4, are overall the worst performing on average: mean autocorrelation, mean harmonics-to-noise ratio, jitter (local), and jitter (local, absolute). It could be argued that this is simply a reflection of data insufficiency: of the features that met the thresholds established in Chapter 4, these features were hardest to extract still. However, *f0* also proved difficult to extract in Chapter 4 yet, in the databases it was extractable in, it proved integral to performance. The poor performance of these features could therefore relate to the fact that they all measure the same thing: non-modal voicing features. Beyond this, it is notable that

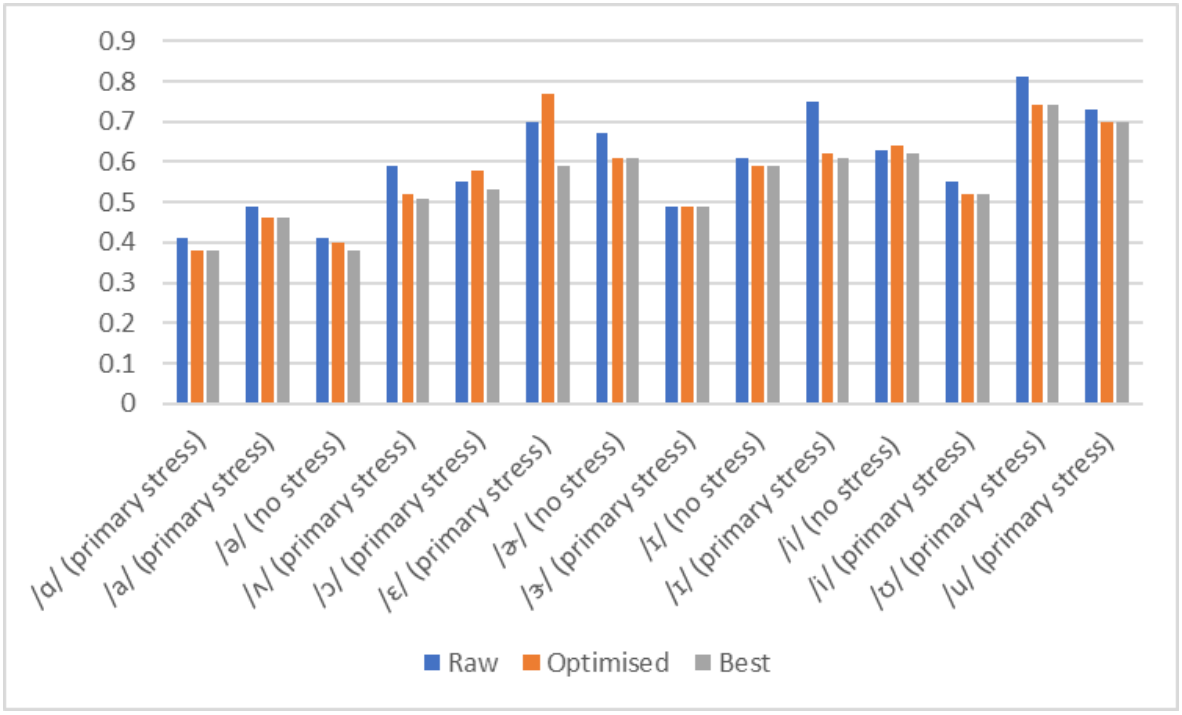
F3, F4, and intensity are the only features that are always useful irrespective of sociophonetic tailoring, but they are useful to different degrees: formant 4 proves best in the text-independent database from DyViS, formant 3 proves best in the text-independent database from WYRED, and intensity proves best in the text-dependent database from DyViS. This reflects the findings from the Tippett plots found in (5.2). Thus, whilst it can be claimed that some features proved universally useful, they still exhibit different degrees of importance depending on the sociophonetic variables of style and accent. Some features, however, demonstrate the need of sociophonetic-tailoring more clearly: F1 does not perform well in the text-independent data from DyViS and F2 does not perform well in the text-independent data from WYRED.

### 6.3. Combining Phonetic Features, Vowel Specificity, and Sociophonetic Tailoring

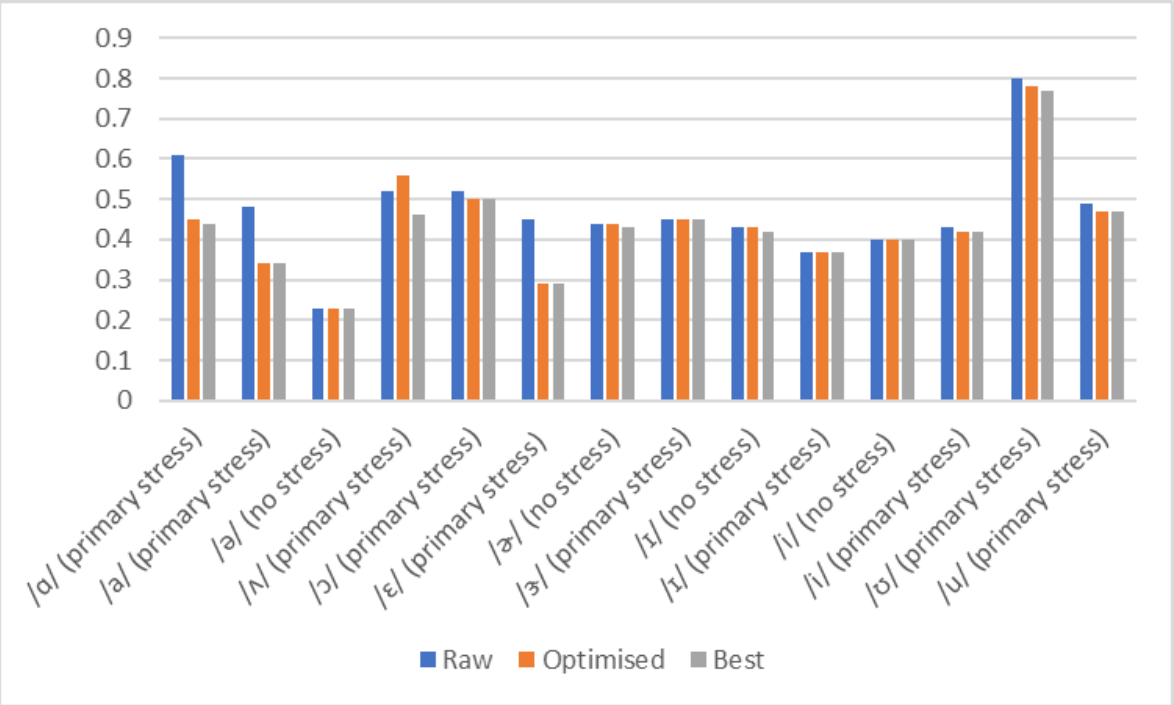
From here, vowel-specific combinations were tested. First, system performance when all of the features are combined together was calculated for each individual vowel level for each database. These 'raw' performances can be found in Appendix A. From these analyses, optimised portfolios of features were calculated using the 'top-down' method discussed above. These 'optimised' combinations are composed solely of the features that, if removed for that given vowel, worsened performance. The features that improved performance when removed, were removed. Figure 22 below summarises the 'optimised' combinations per vowel, per sociophonetically-tailored database. Information about what composes these 'optimised' combinations can also be found in Appendix A. Note, however, that a 'best' combination has also been included. This proved an important inclusion: the 'raw' score is before any features are removed, the 'optimised' score is a score composed of only the features that perform well when isolated, but the 'best' score is simply the best $C_{llr}$ value recorded at all. This has been included because the 'best' score was not always the 'optimised' score; the 'best' score was, occasionally, simply found through the removal of

one specific feature. Sometimes, as seen, optimisation could even worsen performance. Information about what composes these 'best' combinations can also be found in Appendix A.

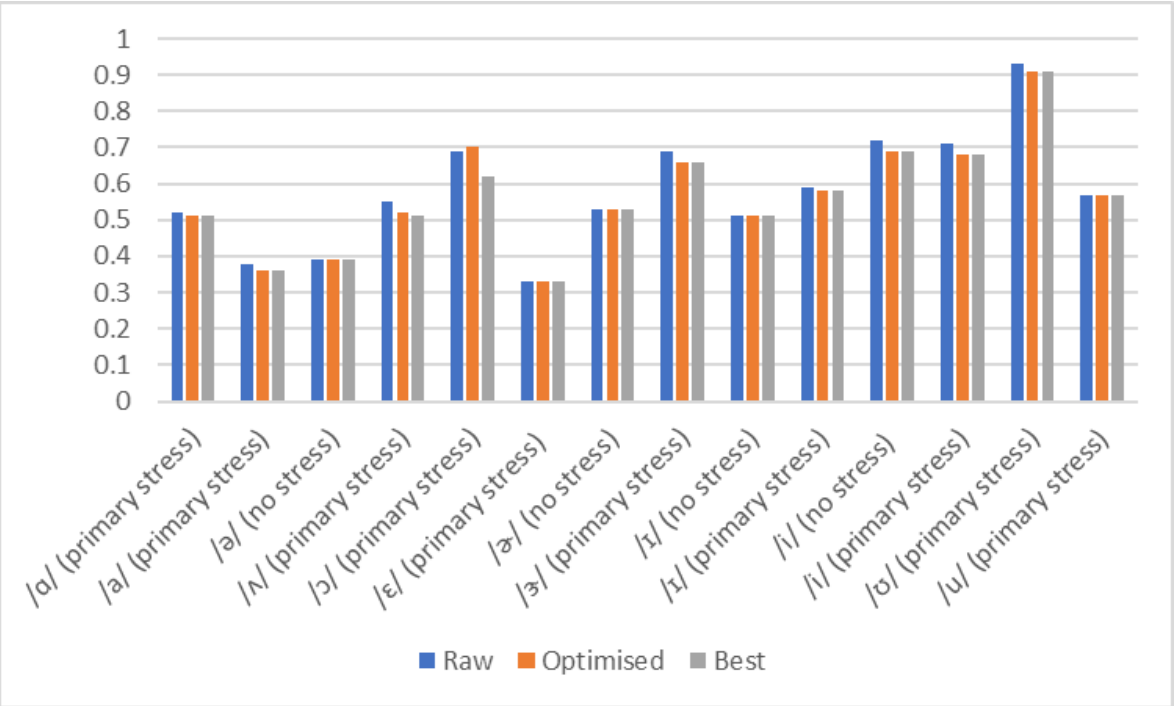
**Figure 22: Vowel-Specific Raw, Optimised, and Best $C_{llr}$ Values per Database**



**DyViS (Text-Independent)**




**WYRED (Text-Independent)**




Though there is no consistent rank order, some notable trends in these graphs include the unanimous poor performance of the back vowels, as also seen in the Tippett plots from (5.3),

and the strong performance of /ə/, also seen in (5.3). These findings also reflect the frequency of token findings from (4.2): the back vowels were amongst the least frequent, and /ə/ was most frequent. This indicates a rather simple finding: more data may therefore correlate to better performance. Such trends will be discussed in much more detail in the coming chapter.

Most crucially, no vowel-specific 'raw', 'optimised', or 'best' combination actually outperformed the all-vowel, all-feature combinations for each database. For reference, these were 0.32 for the text-dependent data from DyViS, 0.15 for the text-independent data from DyViS, and 0.21 for the text-independent data from WYRED. Thus, this chapter shows that vowel-specificity can be tailored towards and feature combinations can be optimised, but it may actually be harmful to do so; tailoring all of the approaches together to different sociophonetic variables (without trimming any extractable features and vowels) appears to be, overall, the most important consideration for best performance. This conclusion is further supported by the fact that the worst score seen was when all databases were considered together without any sociophonetic specificity (1.05).

## 7. Discussion

The previous three chapters explored the "undefined" results, the feature and segment performance variation, and how these features can be combined and optimised for different vowels produced by different sociophonetic groups. Now, these results must all be tied together in light of this thesis' goals: to explore the place of novel, combinatory, phonetically-informed approaches for ASR. This chapter synthesises and breaks these results down more in-depth to isolate what they say about the effectiveness of phonetic approaches to ASR both overall and on a feature-by-feature, vowel-by-vowel, and sociophonetic group-by-group basis. It then uses these breakdowns to comment on current approaches to ASR. This also leads to discussions concerning how the work of this thesis can be carried forward for future ASR tasks and how the approaches identified here can be improved further. Throughout the chapter, some discussions are also included concerning the potential effectiveness of this work for ASR applications. In particular, this chapter ends with an example fusion of the best phonetic approaches found in this thesis with an off-the-shelf ASR system. Using the datasets from this thesis, it shows that such fusions are possible, and that they do not hamper the performance of these ASR systems.

### 7.1. The Performance of (Socio)Phonetic Approaches and Considerations

In the preliminary chapters, it was broadly evidenced that acoustic phonetic features should prove usable for ASR. This chapter is first concerned with the research question established in (2.6): 'To what extent do explainable, acoustic-phonetic approaches provide useful information for ASR?' This research question aligns with the goals of this thesis to explore the use of phonetic approaches for ASR. The phonetic approaches explored in (2.2-2.5), from the features to the vowels, were all established as explainable; thus, the purpose of this section is to establish whether these explainable approaches are, at all, viable (performance-

wise) for use in ASR tasks. A potential application of such findings, as discussed across the earlier chapters, is that they could add an additional element of explainability to ASR systems in the future. In order to explore this broadly, Table 9 has been included again below. This table recaps the full combination $C_{llr}$ value (all databases) and the $C_{llr}$ values for the individual, sociophonetically-controlled databases. The individual features and segments are considered separately later; here, they are all included together.

**Table 9: Overall and Baseline $C_{llr}$ Value Deviations**

| | |
|---|---|
| **Full Combination** | 1.05 |
| **DyViS (TD)** | 0.32 |
| **DyViS (TI)** | 0.15 |
| **WYRED (TI)** | 0.23 |

Firstly, as mentioned in (6.1), the full combination condition generated the worst $C_{llr}$ value across all conditions tested in this thesis. This is the only system tested prior to factoring in any phonetically-informed nuance: all features, segments, and sociophonetically-distinct data were included together. This potentially challenges the stance of this thesis that phonetic approaches can be useful for ASR as this score, consisting of every tested phonetic approach in this thesis, is worse than any other score reported in this thesis. As this score even exceeds 1, this indicates that the system cannot be based on any useful information at all for ASR, as per Morrison et al.'s (2021) overview of the $C_{llr}$ metric.

That said, the best scores out of all of those reported in this thesis were generated when one actively tailors towards the sociophonetic variables of accent and style but still combines every phonetic feature and segment together, as discussed in (6.1). This actively supports the central sentiment of the thesis: approaches devised entirely of phonetic features and

phonemes, yet actively tailored to different sociophonetically-controlled groups, have performed best for ASR. This finding importantly shows that the sociophonetic variables of accent and style are critical to ASR performance: when they were not considered, as seen in the full combination discussed above, performance decreased drastically. The improvement in performance when including sociophonetic information is also reflected in the lower intersections seen in the Tippett plots in (5.2).

It should also be emphasised here, however, that different sociophonetic groups perform differently to each other, with the system based on the text-independent SSBE dataset performing best (0.15). This means that sociophonetic information, whilst the most important consideration for ASR performance, also affects performance. The bias towards text-independence is surprising, given the historical issues that have faced phonetic approaches to ASR covered in (2.1), but the bias towards SSBE over non-standard varieties like West Yorkshire English is less surprising given the biases towards prestige varieties discussed throughout (2-3).

It must also be emphasised that combining all of the phonetic features and vowels together for a given sociophonetic group generated better scores than any vowel-specific, optimised-feature combination generated for any database, as seen in (6.3). Thus, only considering sociophonetic variation, not segmental variation or feature optimisation, generated the best performances seen. The fact that tailoring to segment and feature was found to generate worse scores directly challenges the literature concerning vowels in (2.3) concerned with the differing performances of different vowels. However, this literature is still supported to an extent by the fact that the best-performing combinations were still employing phonetic segments, specifically vowels; they were just employing all of them without any specification. Moreover, the fact that all of the most successful combinations employ every extractable phonetic feature tested supports the literature concerning the use of all of these

features for ASR. This is because, combined together, they have generated the best performances seen in this thesis. From these broad results, it can be summarised that the most important phonetic consideration for ASR is sociophonetic variation: feature- and segment-specificity can have effects, as the rest of this chapter shows, but it is never more important than sociophonetic-specificity. For the features and segments, all that matters is that they are extractable, as per the findings of (4.1-4.4).

The effectiveness of the best phonetically-informed approaches to ASR found in Table 9 above can also be exemplified, to some extent, by comparing these $C_{llr}$ values to those from a modern ASR system. The best portfolio here is from the all-feature, all-vowel system tested on the text-independent DyViS database which generated the best score of 0.15. This score is notably lower than some of those reported in modern ASR performance research, such as Basu et al.'s (2022) recent work that found the best score of a modern ASR system, E3FS3, to be 0.21. However, there are multiple caveats to this conclusion: firstly, most of the reported portfolios in this thesis scored higher than 0.21. In fact, all portfolios related to text-dependent speech and West Yorkshire accents were. Secondly, this may not be a fair comparison to make: Basu et al.'s (2022) data was more forensically diverse. The data used in this thesis was high quality, single session, and channel-matched; it was therefore predisposed to perform better, even though it was text-independent. However, this invites a more thorough examination of how these phonetic approaches could work for ASR in the future. Nonetheless, this score comparison still indicates that phonetic approaches to ASR may prove useful for ASR, and future work could integrate the best portfolios here into a pre-existing ASR system; after all, the 3 best portfolios reported above are close to 0.21. The final sections of this chapter explore this: it fuses these portfolios with an off-the-shelf ASR system to show that they can be successfully added as additional, explainable elements to ASR systems without affecting performance.

This section has offered a generalised image of the effectiveness of phonetic approaches to ASR and shown that sociophonetic information is the most important considered for performance. From a practical perspective only, this is the key finding of this thesis. Now, the effectiveness of the individual phonetic approaches, starting with each phonetic feature, will be explored in more detail to understand more about their behaviours in ASR tasks. This is done despite the fact that feature- and vowel-specificity was found to hamper performance; it is therefore a scientific investigation into why, exactly, feature- and vowel-specificity can hamper performance.

### 7.2. The Performance of Individual Phonetic Features

In this section, it will be shown that some of the individual phonetic features can have a more specific positive effect on $C_{llr}$ values (despite, overarchingly, all features being needed to achieve the best performances seen above). Of the features that were carried forward for extraction after the feature trimming reported in (4.1), only *f0*, intensity, and some of the formants independently improved the vowel-specific scores. Mean harmonics-to-noise ratio, mean autocorrelation, and jitter will all be shown to be detrimental to these vowel-specific scores.

It must be acknowledged, however, that these results only show phonetic approaches improving upon a baseline of solely phonetic approaches; not commonplace ASR features like MFCCs. This can form the basis for future work. Also, the focus will primarily be on the phonetic features but vowels, sociophonetic variables, and methodological concerns will be discussed in this section too as they arise. These will, however, be discussed in much more detail in the later sections of this chapter. Table 10 below shows the average effect that removing each feature has on the $C_{llr}$ values for each database's vowels. This table will guide the following sections.

**Table 10: Average $C_{llr}$ Value Deviations Per Feature, Per Database**

| DyViS (TD) | DyViS (TI) | WYRED (TI) |
|---|---|---|
| Intensity (+0.053) | *f0* (+0.09) | Formant 3 (+0.083) |
| Formant 4 (+0.046) | Formant 4 (+0.088) | Formant 4 (+0.08) |
| *f0* (+0.031) | Intensity (+0.042) | Intensity (+0.034) |
| Formant 1 (+0.013) | Formant 2 (+0.03) | Formant 1 (+0.033) |
| Formant 2 (+0.012) | Formant 3 (+0.023) | Formant 2 (-0.001) |
| Formant 3 (+0.007) | Jitter (Local) (-0.001) | Jitter (Local) (-0.004) |
| Mean Harmonics-To-Noise Ratio (-0.001) | Jitter (Local, Absolute) (-0.005) | Jitter (Local, Absolute) (-0.004) |
| Mean Autocorrelation (-0.005) | Formant 1 (-0.014) | |
| Jitter (Local, Absolute) (-0.006) | | |
| Jitter (Local) (-0.008) | | |

### 7.2.1. The Efficacy of *f0*

Firstly, it was summarised in chapter 2 that *f0* should prove effective for ASR. On the basis of Table 10 above, this claim has been supported to an extent: synthesising all of the portfolio testing from chapter 6, *f0* proves useful for ASR in every database it is included in. This is because removing it, on average, always increases the score; it is therefore integral to

performance. Based on modern ASR systems and speaker recognition studies reviewed in (2.5.1), this is to be expected: Zhu et al. (2009) found that including *f0* measurements in an ASR system improved recognition rate by 5%, for example. Its effectiveness for ASR can be further exemplified by the fact it is always included in every vowel portfolio in Appendix A for the text-independent DyViS database and every vowel portfolio for the text-dependent DyViS database except for /ə/ (no stress), /ɛ/ (primary stress), /ɪ/ (primary stress), and /i/ (no stress).

Turning to these problematised vowels as a first insight into the problems with *f0*, however, their weaker performances may relate to the fact that these vowels generally appear to perform worse for SS comparisons in the Tippett plots from (5.3). The reason this only occurs for text-dependent speech may be due to the higher variability seen between text-dependent tokens of these vowels resulting from suprasegmental pitch effects. These four vowels are amongst the most commonly produced in this dataset, as reported in (4.2), and are therefore more likely to occur in many more different contexts. More specifically, they are more likely to occur in different positions in utterances with different pitch contours. Exploring Speaker(21)'s data as an example, one can compare the two /ə/ (no stress) tokens in "Police announced last night that they have arrested one of two men…". These occur first in the word-initial "Police" and later in the preposition "of". This is a declarative statement and these generally have more pronounced descending pitch contours in text-dependent speech than in text-independent speech, as Malyuga et al. (2017) write. This is why the first token of /ə/ (no stress) here has one of the highest recorded *f0* measurements at 182.81Hz and the latter has the lowest recorded *f0* at 113.92Hz. This shows how more common vowels can occur in more varied contexts which are affected differently by suprasegmental pitch effects. This may account for the worse SS-comparison performance of *f0* on these vowels in text-

dependent speech, and this imbalanced performance between SS- and DS-comparisons accounts for its weaker performance.

Whilst taking this into account, these results still show that *f0* can be effective for ASR; it just has to be used appropriately. Whilst this vowel-specific issue problematises the full effectiveness of this feature for ASR in text-dependent data, it in turn shows that there may actually be rank orders in vowel performance: when it comes to using *f0* for ASR tasks with SSBE speakers producing text-dependent speech, /ə/, /ɛ/, /ɪ/, and /i/ rank at the bottom. This supports claims from (2.3) that some vowels perform better than others to an extent: a rank order of vowels has emerged, but these worst vowels are not the same as those that performed worst in Paliwal's (1984) study. This further justifies the critiques of Paliwal's (1984) study discussed in (2.3). Whilst these problems have been identified, it should be reiterated that (7.1) has already been shown that one does not need to tailor towards different vowels anyway; these differing vowel-specific performances are therefore ultimately unimportant to the practical use of *f0* in ASR.

Moving from vowels to sociophonetics, these *f0* results further show that speech styles should be tailored towards. More specifically, whilst *f0* proved useful to both styles in SSBE speech, *f0* performed much better in text-independent speech; in fact, it was the most integral phonetic feature overall to performance in this database, as Table 10 shows above. This indicates that accent must be important to consider in ASR.

Though seeming effective when usable, the biggest issue with *f0* is that its extraction was not found to be viable for the WYRED database in (4.1). This further supports the importance of tailoring to accent given that *f0* has proven non-viable for ASR tasks with West Yorkshire accents. Even for DyViS, however, this feature was found to have the most "undefined" results in chapter 4 without becoming unusable. This also indicates that feature-specificity is

important to phonetic approaches to ASR, but only for extractability. Simply put: if they are extractable, they should be used, as (7.1) discussed.

The bigger issue here, however, is the implications of this extractability issue for the reliability of this thesis' methodology. Given that *f0* should always be extractable if the speech analysed is voiced, as discussed in (3.2), the high frequency of "undefined" results indicates that unvoiced sections must have been analysed when taking measurements from the theoretically voiced vowels. There must have therefore been an issue with earlier stages of the methodology, specifically forced alignment and the demarcation of the voiced vowel boundaries. Boersma and Weenink's (2023) Praat cannot be responsible: this software will always take a successful *f0* measurement as long as it's from a voiced section, and it did so successfully when the section was voiced. The corpora are somewhat responsible: vowels were sometimes not produced in the original recordings due to co-articulation effects and accent variation which will be discussed in later sections. However, these issues have mainly emerged from unvoiced sections being included in the aligned boundaries of the selected vowels. These issues with forced alignment will be exemplified and elaborated on later in this chapter when the methodological concerns surrounding forced alignment, and how the effectiveness of phonetic approaches has been impeded by these issues, will be the focus of a full section.

### 7.2.2. The Efficacy of Intensity

(2.5.2) also established that intensity should prove effective for ASR. On the basis of these results, this claim has been supported with more confidence than that seen for *f0*. This is because, for all three databases, intensity always increased the score when removed, as seen in Table 10 above. It is therefore integral to performance irrespective of speaker accent or style. It is present in the majority of the best-performing vowel portfolios from every dataset

seen in Appendix A. All of this shows that intensity behaves more reliably across different vowels and sociophonetic groups than *f0* did in (7.2.1).

That said, this inadvertently challenges claims concerning the importance of sociophonetic tailoring: as it always performs well, this instead suggests that sociophonetic variables like accent and style are unimportant to consider when using certain phonetic features, here intensity, because they will always work. That said, intensity was not of equal importance to each database: it was more important to performance in the text-dependent speech than the text-independent speech, and it was more important to performance in SSBE-accented speech than West Yorkshire-accented speech for ASR performance. This is also visualised in the Tippett plots in (5.3). This therefore, still supports claims concerning the importance of sociophonetic tailoring.

The success of intensity in ASR is predictable as it has been seen before in ASR studies: Jia et al. (2021) employed it alongside formants to create a system both faster and more reliable than modern ASR systems. Whilst this thesis' results support Jia et al.'s (2021) claim that intensity can be used reliably, this thesis does not show that intensity can be used quickly for ASR. This is because the processing time behind the results of this thesis was >2 hours. Speed will be discussed later in this chapter as another methodological concern that hinders the hypothetical performance of phonetically-informed ASR approaches for future ASR applications.

Overall, the success and presence of intensity in the portfolios is clear support for the effectiveness of this approach. However, the conditions of this thesis mean its effectiveness cannot be fully verified, especially for forensic purposes. In the present experiments, session variability was not tested. This is a confounding variable already known to affect intensity from chapter 3: microphone distance can change between sessions and this can affect the

intensity of a given recording, as Titze and Winholtz (1993) found, and the emotion in one's speech can change between sessions and can affect intensity, as Pfitzinger and Kaernbach (2008) found. Variation along these axes was not tested here; thus, the evidence for the effectiveness of this phonetic approach here is based on recordings wherein intensity is consistent. This is also likely why it was reported in (5.3) that intensity performed best with the text-dependent data: it is the most controlled. Further testing of session variability with the same speakers and styles of speech is therefore required to understand just how effective intensity is; what this thesis shows, however, is that in controlled environments it has the capacity to be effective. Whether it can be reliably used in forensic casework, where these confounding variables are expected to be present, must be tested further.

### 7.2.3. The Efficacy of Formants

Moving on to formants, (2.5.3) stated that formants should prove effective for ASR, but this has only been supported to an extent. In navigating this discussion of formants, the higher formants will be discussed first for ease of structured discussion.

Starting with the highest formant explored in this thesis, F5, the effectiveness of this formant cannot be supported by this thesis because it generated too many "undefined" results in all corpora tested in (4.1). This reflects an overall trend that emerged concerning the extractability of formants: the higher formants had less successful data extraction than the lower formants. All formants lower than F5 were viable for use, but F4 generated more "undefined" results than F3, which generated more "undefined" results than F2, which in turn generated more "undefined" results than F1. This is predictable: in Derdemezis et al.'s (2016) study cited in earlier chapters, they found that the higher formants tended to be the most unreliable measurements to extract given that they have weak energy or energy outside the used spectrogram range. In forensic recordings, it may also be that the sampling rate is too

low for F5 to be represented in the recording in any case, as discussed in (2.5.3). They suggest that a fix for this may be to increase the dynamic range of the spectrogram in Boersma and Weenink's (2023) Praat to view higher energies, but this may have detrimental effects on other features, as also discussed in (2.5.3). Thus, F5 had to be omitted. This challenges claims concerning the efficacy of higher formants for ASR: the highest formant tested could not even be considered viable for use in ASR tasks.

On the contrary, F4 was always extractable and, as seen in Table 10 above, always the second-most effective feature for ASR irrespective of any sociophonetic variables. It also appears in the majority of vowel portfolios created in chapter 6 and it captures intra- and inter-speaker variation well, as seen in the Tippett plots from (5.3); its performance is therefore strong in both core ASR tasks.

The effectiveness of F4 for ASR is expected: as Lammert and Narayanan (2015) found, the higher formants are less related to the articulations involved in vowel production and more related to vocal tract length which is much more unique to a given speaker. Thus, the individual differences between speakers should be most visible with higher formants than the lower formants which capture more vowel articulation and co-articulation information. The results of this thesis reflect this conclusion, given the overall efficacy of F4.

These consistently good results, however, again challenge claims for the need for sociophonetic specificity. F4 has proven useful irrespective of style or accent, and as a result shows that another phonetic feature may not need to be sociophonetically-tailored in any way because some will always perform well.

Moving down to the lower formants, these were again always extractable. As seen in Table 10 above, F3 always proves effective for ASR irrespective of accent or style. However, it performs noticeably better in West Yorkshire accents: in SSBE text-independent speech, F2

actually outperforms F3, and in SSBE text-dependent speech, F1 does as well. The reason for F3's success with West Yorkshire-accented speech appears to relate to how it performs equally well in the same-speaker and different-speaker comparisons in this database. The results from (5.3) show this: in the WYRED database, F1 and F2 performed as well as F3 did in DyViS when capturing inter-speaker variation. However, in DyViS, F1 and F2 did not perform well for capturing intra-speaker variation reliably. This is likely due to ongoing language change that is affecting F1 and F2 that will be discussed momentarily. F3, on the other hand, performed well in both tasks for WYRED, outperforming the lower formants in intra-speaker variation. This is likely due to greater stability in West Yorkshire accents; no such ongoing language change was occurring at the time of recording. Thus, it is this balanced performance between two ASR tasks that lead it to perform as well as it does in this database. This variation in performance between accents supports claims for sociophonetic tailoring in ASR.

Looking next at F2 in Table 10, this was always extractable but proved most useful for SSBE accents. For West Yorkshire accents, F2 measurements were actually detrimental to performance overall. This challenges claims concerning the efficacy of formants but, in turn, supports claims concerning the importance of sociophonetics: SSBE has proven more responsive than West Yorkshire accents to the use of F2 in ASR.

It was expected that F2 would perform better in SSBE accents because F2 maps onto vowel fronting, as discussed in chapter 2. This is particularly important to SSBE because de Jong et al. (2007) found, using DyViS data as well, that /ʊ/ and /u/ showed high variability in F2 due to a sociophonetic change in progress affecting SSBE: these vowels are becoming more fronted. It is a change not all speakers have adopted yet and are not using consistently, but it is diffusing through this sociophonetic group. As a result of this, F2 will not prove as reliable for same-speaker comparisons in this database due to this higher intra-speaker variation. This

conclusion is supported by the above data: in (5.3), F2 performed worse for same-speaker comparisons in the DyViS databases than the WYRED database where no such ongoing language change is occurring. Simply put, if SSBE speakers are still adopting it, they will therefore not be using it in all applicable contexts yet. This means that their productions may be more variable as a result, affecting the reliability of F2 for ASR.

This language change, expectedly, affects the efficacy of F2 for /ʊ/ and /u/ in particular in the results collected from the DyViS databases. Whilst the inter-speaker variation is captured well in these vowels' results from (5.4), the high intra-speaker variation also seen for these vowels in (5.4) means that they do not perform well for ASR overall. This means that a rank order of vowels has also emerged here regarding the performance of F2, particularly for SSBE accents: /ʊ/ and /u/ perform worst. This supports claims concerning rank ordered vowel performances, but this again does not reflect the rank order from Paliwal's (1984) study. Furthermore, different vowel rank orders are now emerging for different features from the results: whilst /ʊ/ and /u/ emerged as worst in the F2 measurements, they were not the worst in the *f0* measurements, as discussed in (7.2.1). The performance of different vowels is therefore dependent both on the features used when assessing performance and the sociophonetic profile of the database used to test them, given that this rank order only emerged for SSBE. This finding that there are rank orders emerging that do not reflect Paliwal's (1984) original order and appear to be specific to different accents and styles will be expanded upon later in this chapter; however, the sociophonetic-specificity of these rank orders provides further support for the importance of sociophonetic tailoring in ASR.

Moving finally to F1, this proves most useful for West Yorkshire accents. For SSBE accents, it performed best with text-dependent speech for ASR. This further supports claims concerning the importance of sociophonetic tailoring. Turning back to de Jong et al.'s (2007) study reviewed in (2.5.3), the worse performance of F1 in text-independent speech in SSBE

is somewhat expected: taking the vowel /a/ as an example, F1 showed particular variability for this vowel in their study because this was undergoing another change in progress which affected this vowel's height in SSBE speakers. In controlled, text-dependent speech, where speakers are more aware of their speech, variability within the speaker will be more limited, and thus the measurements will be more stable. However, in more natural, text-independent speech, variability within the speaker will be higher, and less stability will be seen across the measurements as a result. This will impact same-speaker recognition negatively in text-independent settings, therefore impeding the efficacy of this feature for ASR in this database. This is directly reflected in the results for this vowel in (5.4): in the text-independent data for SSBE, F1 captures more performance variation for this vowel than the text-dependent results, which were more stable. Thus, F1 is therefore too variable for same-speaker recognition tasks conducted with text-independent data, accounting for the worse performance of this feature in text-independent SSBE speech.

The greater efficacy of F1 in West Yorkshire accents, especially compared to F2, is also predictable. This is because Earnshaw (2021) writes that West Yorkshire accents showed more variation between different speakers through F1 than F2. This manifests itself in the results here too: in (5.3), it was shown that F1 performance was more balanced than F2 because it was better at measuring different-speaker variation. Thus, the greater performance of F1 in this dataset may be due to its greater efficacy for inter-speaker recognition. Overall, these F1 performance variations between different styles and accents are important to recognise because they further support claims that phonetic approaches should be tailored to different accents.

### 7.2.4. The Efficacy of Mean Harmonics-To-Noise Ratio

(2.5.4) also debated whether mean harmonics-to-noise ratio will prove effective for ASR, and this claim has not been supported. Firstly, mean harmonics-to-noise ratio only generated enough data to be viable for use in ASR in the text-dependent dataset (4.1). Its successful extraction is therefore subject to the sociophonetic variable of style. Whilst this challenges this feature's efficacy, this variable extractability still supports the use of sociophonetic information by showing that different sociophonetic variants respond differently to different phonetic approaches. This supports the conclusion that phonetic features should be tailored towards the needs of different sociophonetic groups.

Mean harmonics-to-noise ratio's limited extractability can be explained in relation to how this measurement is taken and the methodological problems that are arising in this discussion concerning forced alignment. Firstly, as this is a mean measurement taken from voiced speech, it requires the demarcated area to be accurate and fully voiced. This is easier with text-dependent speech because it is less subject to co-articulation effects like deletion that can lead these segments to be unvoiced. This may explain why it is only successfully extracted for text-dependent speech. In terms of extractability, mean harmonics-to-noise ratio is less successful than *f0* because *f0* only needed the midpoint to be voiced; as a mean measurement, mean harmonics-to-noise ratio measures across the whole segment and requires the whole segment to be voiced.

Even with extraction issues aside, mean harmonics-to-noise ratio still had a detrimental impact on ASR performance when used in the text-dependent SSBE dataset, as seen in Table 10 above. When removed, the score lowered, therefore improving performance when not included. Based on this, it should not be considered in ASR as a viable phonetic approach even when it can be extracted successfully. Turning to the plots in (5.3), this weak

performance can be explained by the fact that mean harmonics-to-noise ratio is better at different-speaker comparisons; for same-speaker comparisons, it showed higher variability in performance. Thus, mean harmonics-to-noise ratio may only be effective for measuring inter-speaker variation; not intra-speaker variation. The lack of balance between these ASR tasks accounts for its weak performance. These results also challenge the efficacy of phonetic approaches to an extent: only certain phonetic features are proving effective for ASR now, and mean harmonics-to-noise ratio is not one of them. That said, it was still integral to the best-performing system which combined all features and vowels, as reported in (7.1).

Issues aside, Yumoto et al.'s (1984) study was also discussed in (2.5.4) which raised the point that mean harmonics-to-noise ratio may only prove useful in select circumstances anyway, namely for speakers who have atypical, non-modal speech patterns involving greater variability in breathiness or creakiness or for speech recorded in the morning when a speaker's voice is creakier. Thus, there is still some potential for mean harmonics-to-noise ratio to prove more useful for ASR in conditions not covered by this thesis: for example, this feature could be tested on speakers with voice pathologies that are associated with non-modal voice qualities. This is supported by multiple studies from chapter 2, such as Teixeira et al.'s (2013) study which discussed how features useful for recognising pathologies may also prove useful for ASR. Furthermore, it could be tested on a database where session variability correlates to different times of the day when voices are creakier; it may be a feature that only proves useful at certain times.

### 7.2.5. The Efficacy of Mean Autocorrelation

(2.5.5) also debated whether mean autocorrelation will prove useful for ASR, and this has also been challenged. Firstly, mean autocorrelation only generated enough data to be viable for text-dependent speech; its extraction is again determined by the sociophonetic variables of

style, much like mean harmonics-to-noise ratio was. Even more like mean harmonics-to-noise ratio, mean autocorrelation is included in the best-performing systems from (7.1), but in the vowel-specific systems it has a detrimental impact to performance when included, as seen in Table 10 above. Looking at the results from (5.3), it even failed for the same reasons: it captures variability between speakers well, but does not capture stability within the same speaker well. This continues to challenge the claim that phonetic approaches are effective to an extent: only certain phonetic approaches are proving effective for ASR, and these mean value features are now those proving to be consistently ineffective.

Given how these two features capture similar information about the consistency of voicing activity, specifically about breathiness and creakiness, this indicates that the problems could lie in the observance of perceptible features related to non-modal voicing. That said, this only accounts for why these features fail when used for ASR tasks; the fact that they cannot overcome the extraction thresholds for text-independent speech styles, as seen in (4.1), indicates that wider methodological issues may be affecting these phonetic features, in particular surrounding forced alignment. This may be because these features are both measured as mean values and need the entire segment to be voiced, yet the forced aligner may have included unvoiced segments. The high frequency of failed extractions in the text-independent databases may therefore indicate that the forced aligner is not demarcating the voiced activity of the vowels accurately in natural speech. This is potentially due to co-articulation effects like deletion which are more common in text-independent speech, as discussed. Examples of this will be explored in later sections of this chapter when evidence against the reliability of current forced alignment methods will be explored. Overall, it appears that issues may arise due to the observance of features related to non-modality or forced alignment issues.

Whilst mean autocorrelation could still prove effective in scenarios with more session variability or voice pathologies present, as discussed in (7.2.5) regarding mean harmonics-to-noise ratio, the present thesis challenges the conclusions reviewed in (2.5.5) that saw mean autocorrelation be successful. For example, Gonzalez-Rodríguez (2014) found that mean autocorrelation could be used to address research gaps in modern ASR system performance, but only went as far as identifying it as something that could prevent false rejections; they did not test it as a viable approach to ASR. This thesis has tested it as an approach to ASR that could be implemented in ASR and it has found that, under the tested conditions, mean autocorrelation should not be used because it is only extractable for text-dependent data and, when extracted, it can be detrimental to performance.

Furthermore, Bidondo et al.'s (2013) psycholinguistic study of speaker recognition was criticised in (2.5.5) and can be criticised further here. This study problematically suggested that one could use mean autocorrelation in ASR to replicate human cognitive processes. This study was deemed to be flawed on the grounds that the human brain is not a perfect recognition system and is, in fact, capable of false recognition. This criticism has been supported to an extent by these results: mean autocorrelation has been shown as ineffective for ASR, so this study (which assumes that mean autocorrelation can rectify issues with ASR) has been challenged further.

### 7.2.6. The Efficacy of Jitter

It was also debated whether jitter will prove effective for ASR. This has also been challenged overall, but with some minor caveats. Showing some support for the efficacy of jitter, some jitter measurements were consistently capable of successful measurement extraction in all databases: the local and local, absolute measurements. This is predictable: literature reviewed in (2.5.6) established that the jitter measurements that average out the least amount of data

and capture more variability, namely the local measurements, will be the most effective jitter measurements for ASR tasks. This is explainable based on the selected data: as the vowel segments are short, measurements like PPQ5 have insufficient data to collect from, thus showing why the measurements that require more data points failed during data extraction, as shown in (4.1).

These local measurements appear overall important because they are included in the best-performing systems reported in (7.1). Based on the vowel-specific systems, however, this claim has been challenged because they were consistently detrimental to performance. When removed, the system score is always lowered irrespective of style or accent, therefore improving performance via their removal.

The poor performance of jitter challenges previous studies that have seen success with jitter, such as Farrús et al.'s (2007) and Jones et al.'s (2001) studies which found that this feature proves useful for distinguishing speakers based on breathiness and creakiness. In combination with the failures of mean harmonics-to-noise ratio and mean autocorrelation which also perceptibly correlated to these non-modal voice qualities, one of the key developing findings of this thesis appears to be that measurements that capture information relating to these non-modal voice qualities do not appear to be as effective for ASR with the selected data. All measurements relating to non-modal voicing, including shimmer which will be discussed momentarily, are the features that consistently perform badly either in data extraction or in vowel-specific system performance. This may be due to the selected data which includes only modal voices: in datasets containing speakers with more non-modal voicing, or datasets concerning changes in creakiness that are associated with time of day, these features may still prove useful, as discussed above in (7.2.4-7.2.5).

As an alternative explanation, however, flaws in the methodology may account for the failures of jitter as they did for the failures of the mean measurements discussed above in (7.2.4-7.2.5). While the mean measurements may have failed due to forced alignment, jitter may have failed as a result of issues relating to feature extraction. Teixeira and Gonçalves' (2014) study, which was reviewed in (2.5.6), found that common algorithms for extracting jitter measurements, which are employed by Boersma and Weenink's (2023) Praat, were not the most reliable algorithms for taking jitter measurements. This may account for its lack of success here, but the difference between Teixeira and Gonçalves' (2014) better-performing algorithms and the one used by Boersma and Weenink (2023) was negligible anyway; both performed well, so Boersma and Weenink's (2023) Praat is therefore unlikely to account entirely for the failure of this feature in this thesis. It may be more likely that features which measure these non-modal voice qualities are ineffective for ASR with the given data groups.

Returning finally to Leong et al.'s (2013) study from chapter 2, this found that jitter will only be effective for male speakers. The results of this thesis ultimately challenge this conclusion: whilst a future investigation into how sex can be tailored towards in ASR would be needed to test Leong et al.'s (2013) claim fully, jitter was not found to be successful for any male speakers' vowels in this study.

### 7.2.7. The Efficacy of Shimmer

Whilst this section has already covered all of the synthesised data from Table 10, shimmer must also be discussed here. Literature reviewed in (2.5.7) predicted that shimmer will prove useful for ASR, and this has been strongly rejected. This is due to its poor extractability; under no vowels or databases could enough data be collected to test its effectiveness for ASR. It was also established that the shimmer measurements that average out the least amount of data and capture more variability, namely the local measurements, will be most effective for

ASR. This also cannot be supported because, whilst the local measurements did produce more successful extractions, they still did not produce enough data to test shimmer.

This overarchingly challenges a number of assumptions: firstly, it challenges the overarching claim that the tested phonetic approaches work well for ASR because it again shows that not all phonetic approaches will prove useful, here shimmer. More specifically, however, its universally poor performance across all vowels, styles, and accents challenges all assumptions concerning the importance of vowel-specificity and sociophonetic variables: no rank order of vowels emerged and shimmer could not be tailored to different styles and accents; it always performs poorly. It could not even be included in the above all-feature, all-vowel scores for each database that performed best, indicating that at least some degree of feature tailoring, not just sociophonetic tailoring, is also necessary in relation to extractability, showing why the extraction results from chapter 4 are particularly important.

The overall failure of shimmer is a mostly expected outcome based on the trends already identified above: shimmer, again, perceptibly correlates to non-modal voice qualities. So far, the features that perceptibly correlate to these qualities have been repeatedly shown to prove harmful for ASR with the selected datasets. Shimmer may have seen success in previous studies, such as Farrús et al.'s (2007) study which was reviewed in (2.5.7), but the results of the present thesis challenge this conclusion. Furthermore, Brockmann et al. (2011) suggested that shimmer would also prove useful for ASR with male speakers, but the present study also challenges this; it was not effective for any of the male speakers of any of the accents producing any of the speech styles.

Turning to more methodological concerns, this thesis supports researchers who argue that shimmer is unreliable as a measurement. For example, Teixeira and Gonçalves (2014) study showed that Boersma and Weenink's (2023) Praat may not be the best tool for extracting

shimmer measurements, and this may account for the results here given that Boersma and Weenink's (2023) Praat was used in this methodology. Leong et al. (2013) also found that shimmer was the worst performing feature for ASR in their study; even worse than jitter. This has similarly been seen in this thesis.

More positively, however, Farrús et al. (2007) showed that local measurements outperform other shimmer measurements, and this is somewhat visible in the present results: the local measurements were more successfully extractable in (4.1), but they still did not generate enough results to be viable for testing. The better extractability of local measurements can again be explained as a result of using phonemes, however: due to how short phonemes are, larger shimmer measurements are not viable for use because there will not be enough data points to observe. The APQ11 measurement, for example, would need eleven frames that the phonemes could not regularly provide, especially in natural speech styles which have shorter vowel productions. Local measurements, by virtue of needing less data to generate a measurement successfully, expectedly produce more measurements for shimmer just like they did for jitter.

### 7.2.8. Summarising Phonetic Feature Findings

Support for the use of explainable phonetic features in ASR has overall been seen, but with some caveats. Some features appear to not be useful at the vowel-specific level, and these all relate to non-modal voicing: mean harmonics-to-noise ratio, mean autocorrelation, jitter, and shimmer. This modality trend may be due to the tested groups: all data was selected from the same session and from speakers with typical modality. Future studies could revisit these features on the claim that they are effective for non-modal voicing characterisation, which may prove useful for characterising speakers across different sessions, when their voices are creakier, or speakers with atypical voices. Whilst this unifying perceptible quality may

account for their failure here, there are also a mix of methodological issues that will be discussed in the following sections. These relate to forced alignment predominantly, but also Boersma and Weenink's (2023) Praat algorithms and data quantity.

The rest of the features, however, proved useful to some degree with the vowel-specific systems. Some proved effective universally and irrespective of accent or speech style, namely intensity and formant 4. Some, by contrast, only proved useful to select accents and speech styles, like the lower formants and *f0*.

The most important finding, however, is that all features par shimmer were included in the sociophonetically-tailored baselines that generated the best scores in (7.2.1). These best systems show that feature- and vowel-tailoring is, practically, unnecessary. As a rule of thumb, it appears that if a feature or vowel is extractable, it should be included; data quantity appears to prejudice better performance. Overall, however, these results give support to the place of these explainable phonetic approaches in ASR.

### 7.3. The Performance of Vowels

Having discussed the efficacy of different features, the efficacy of different vowels will now be discussed in more detail. It has already been discussed how vowel-tailoring cannot achieve the best scores, but for investigative purposes they will still be explored and compared.

Vowel efficacy specifically relates to claims in (2.3) that certain vowels will prove more effective for ASR than others. The following rank order from Paliwal (1984) was identified in the literature review surrounding vowel performance: /ə/, /ʊ/, /ɪ/, /u/, /o/, /ʌ/, /a/, /ɔ/, /æ/, /i/, then /ɛ/. However, this rank order is problematic: as discussed in (2.3), Paliwal (1984) was vague in detail about why this rank order emerged. Thus, despite finding a rank order amongst the limited literature on vowel performance, it was predicted that this rank order will be challenged. That said, based on other literature reviewed, it was also expected that there

would at least be differences in vowel performance that emerge, likely correlated to the sociophonetic differences explored. Table 11 below synthesises the reported scores from the portfolios generated in Appendix A: specifically, it shows the best-performing portfolios for each vowel from each database and ranks each vowel from best performance to worst performance.

**Table 11: Best-Performing Combinations for Each Vowel in Each Database (Ranked)**

| DyViS (TD) | DyViS (TI) | WYRED (TI) |
|---|---|---|
| /ɑ/ (primary stress) (0.38) | /ə/ (no stress) (0.23) | /ɛ/ (primary stress) (0.33) |
| /ə/ (no stress) (0.39) | /ɛ/ (primary stress) (0.29) | /a/ (primary stress) (0.36) |
| /a/ (primary stress) (0.47) | /a/ (primary stress) (0.34) | /ə/ (no stress) (0.39) |
| /ɝ/ (primary stress) (0.49) | /ɪ/ (primary stress) (0.37) | /ɑ/ (primary stress) (0.45) |
| /ʌ/ (primary stress) (0.51) | /i/ (no stress) (0.4) | /ʌ/ (primary stress) (0.51) |
| /i/ (primary stress) (0.52) | /ɪ/ (no stress) (0.42) | /ɪ/ (no stress) (0.51) |
| /ɔ/ (primary stress) (0.53) | /i/ (primary stress) (0.42) | /ɚ/ (no stress) (0.53) |
| /ɪ/ (no stress) (0.58) | /ɚ/ (no stress) (0.43) | /u/ (primary stress) (0.57) |
| /ɛ/ (primary stress) (0.6) | /ɑ/ (primary stress) (0.44) | /i/ (primary stress) (0.58) |
| /ɚ/ (no stress) (0.61) | /ɝ/ (primary stress) (0.45) | /ɪ/ (primary stress) (0.58) |
| /ɪ/ (primary stress) (0.61) | /ʌ/ (primary stress) (0.46) | /ɔ/ (primary stress) (0.62) |
| /i/ (no stress) (0.62) | /u/ (primary stress) (0.47) | /ɝ/ (primary stress) (0.66) |
| /u/ (primary stress) (0.71) | /ɔ/ (primary stress) (0.5) | /i/ (no stress) (0.69) |
| /ʊ/ (primary stress) (0.75) | /ʊ/ (primary stress) (0.77) | /ʊ/ (primary stress) (0.91) |

Firstly, it is evident from Table 11 that rank orders do emerge. However, Paliwal's (1984) rank order is never present; none of the vowel rank orders, from any of the databases, reflects their rank order. This, to some extent, is expected: Paliwal's (1984) study is dated and has

fundamental issues. Not only does this study use EDs, an outdated metric discussed in (2.1.1), to rank the performances of the ASR approaches per vowel, these approaches were composed only of formant measurements. Paliwal (1984) also did not offer any explanation as to why their rank order emerged.

Beyond Paliwal's (1984) study, there is a lack of any other modern or reliable study seeking to evaluate the performance of different vowels prior to the present thesis; thus, the present study fills a research gap by serving as a needed update on the performance of different vowels. It does this by employing a more modern and reliable metric, $C_{llr}$, to assess the performance of the vowels and by identifying different combinations of phonetic features that are best-suited to the individual vowels. It has also already offered some explanations for why certain vowels perform better, but more explanations behind the performances of the vowels will be explored in more detail momentarily.

Looking at the generated rank orders more closely first, they are not uniform across the different datasets; the results show that the vowels perform differently depending on the sociophonetic variables of style and accent to an extent. This further shows that different accents and speech styles have different needs in ASR. It is therefore further evidence that accents and speech styles need to be tailored towards because different approaches behave differently when these variables are changed.

When considered independently, however, some vowels also perform in universal ways; there are some consistent patterns across these rank orders which will now be discussed. Firstly, /ə/ (no stress) always ranks within the top three best-performing portfolios. The strong performance of /ə/ is somewhat unexpected given the above discussion of /ə/ in (7.2.1): there, it was reported that measurements taken from /ə/ can be highly variable within the speaker due to the high frequency of tokens, as reported in (4.2), and these tokens can

occur in a variety of different environments affected by different pitch contours. As discussed, poorer performance in intra-speaker variation should lead to a worse performance overall for ASR; one could therefore expect /ə/ to perform poorly based on this rationale. However, this high frequency may also be why this vowel performed so well. As it is consistently the most frequently produced vowel, it therefore has more measurements to take from these different segmental contexts and, despite being subject to more variability, therefore also has more tokens to capture this variability across. This may therefore account for why this vowel performed well: enough data is provided to capture the limits of variation in this vowel accurately. High data quantity may therefore account for why this vowel performed so well in the portfolios from every database. The importance of data quantity has also already been highlighted above in (7.2.8).

Another universal vowel behaviour spotted is that /ɚ/ and /ɝ/ always rank amongst the middle. The reason for this builds off of the above discussion of /ə/: these are vowels that similarly occurs in a lot of different contexts, but as seen in chapter 4, they were simply not as frequent in the speech files. Thus, despite being as variable as /ə/, they do not have the quantity of supporting data on these different contexts. As a result, this may be why more inter-speaker variation was captured in the results for this vowel in chapter 5. The lack of intra-speaker stability may therefore be why these vowels performed poorly in the portfolios. This, overall, gives further credence to the importance of high data quantity for performance. Given that the best-performing scores seen in (7.2.1) consider all vowels together, many trends in these results are therefore indicating that data quantity is one of the most important considerations for ASR performance.

Contrastingly, the poor performance of /u/ and /ʊ/ in SSBE accents may have resulted from limitations in the methodology for reasons that differ based on accent. Looking at the DyViS data to exemplify this, these vowels are subject to a change in progress, as discussed in

(7.2.3) when exploring the F2 results. The provided dictionary for the forced alignment stage is fixed; it cannot adapt to this change in production so it will only classify the sound based on what information is in the dictionary. Tokens may therefore be incorrectly tagged for certain speakers who have a more fronted production than others in certain segmental contexts, and this may incur variability that accounts for the poorer performance of these vowels.

Whilst the lack of adaptability in forced alignment could be responsible for this performance issue, this issue may ultimately be an argument against the use of vowels overall: language change is a natural part of language, and when change affects vowels as seen here, it may render comparing tokens of a vowel ineffective due to their increased variability that evidently cannot be captured by modern forced alignment tools. More importantly, however, this thesis has found that vowel-specificity is problematic anyway: the baseline score performances in (7.2.1), which only considered sociophonetic specificity and did not tailor to any vowel, were the best-performing combinations. They therefore show that this vowel issue is entirely avoidable by simply including all vowels together irrespective of variation.

This covers the more universal behaviours in vowel performance. Moving forward to more individual behaviours, this highlights a problem that faced Paliwal (1984) that still faces this study today: due to the overall lack of research into the performance of individual vowels in ASR since Paliwal's (1984) study, it still cannot be determined why, exactly, individually distinct performances arise. More research should therefore be done into what, exactly, makes different vowels more viable for ASR in different groups. Some universal conclusions relating to accent variation and data quantity have been drawn for some vowels above, but the presence of minute variations by style and accent, and the overall existence of completely different rank orders emerging, cannot currently be explained should be investigated further.

That said, there may not be a practical need for this research: this study has found that vowel performance differences do not, ultimately, matter to ASR. The best performances were seen when all vowels were considered together, not apart. This, again, supports one of the important themes in this thesis' findings: data quantity is more important to ASR performance than feature- or vowel-specificity. The best portfolios included the most amount of data possible, caveated only by a need to be sociophonetically-tailored and only to include features that are extractable. As a result, feature- and vowel-specific research may not benefit ASR performance from a practical perspective.

That aside, the next exploration to discuss in relation to vowels is the exploratory analysis into stress; a variable that Paliwal (1984) did not consider. Firstly, it must be noted that secondary stress consistently failed to generate enough extractable tokens to be considered as a viable approach to ASR in (4.2); they were trimmed early in the investigation. Similarly, many "no stress" conditions had to be trimmed; thus, the first important finding of this thesis regarding stress conditions is that data from vowels produced with primary stress are the most successful during extraction. This is, again, due to data quantity: as seen in (4.2), more primary stress vowels were extractable than no stress and secondary stress vowels. From the above discussions, however, it has also been seen that /ə/ in its no stress condition is the most successful vowel for ASR tasks based on score. However, this was also the overall most frequent vowel; this finding may, again, be due to high data quantity. Thus, despite primary vowels being easier to extract, stress does not appear to affect the performance of vowels for ASR; data quantity is evidently more important. Furthermore, the best scores seen combined all no stress and primary stress conditions of each vowel anyway; it can therefore again be concluded that higher data quantity is more important to performance than breaking down the data to individual vowels and their individual stress contexts.

Turning finally to how different sociophonetic groups vary based on vowel performance, one final important discovery here is that the best-performing vowel portfolios were seen in text-independent speech. This was also true of the baseline scores, as discussed in (7.2.1). This is also the speech style that ASR researchers were most concerned about, as discussed in (2.1); thus, the fact that phonetic approaches to ASR are most effective for this speech style means that they may already be more viable for ASR than initially expected. Phonetic approaches may therefore already prove useful for the type of speech data that researchers are most concerned with studying. The success of text-independence may be due to the increased amount of data available for each phoneme, as seen in chapter 4. More text-dependent data may yield similar successes, and this is an avenue for future research, but this finding again suggests that data quantity may ultimately be the most important contributing factor to ASR performance.

The text-independent data from SSBE also generated better scores than the text-independent speech from West Yorkshire. This may reflect biases towards standard varieties that permeate many methodological concerns flagged in chapters 2 and 3; however, it was also the database that most data was collected from. This, again, may therefore support the importance of data quantity.

### 7.4. An Exploration into the Implementation of Phonetic Approaches in ASR

The goal of this thesis is predominantly to explore the potential of phonetic approaches for ASR using empirical methods for validation typically used for testing and evaluating ASR systems. Thus, this chapter has shown the successes and limitations of the tested (socio)phonetic approaches using $C_{llr}$. However, a running theme from the earlier chapters concerning the motivations of this thesis is that these phonetic approaches could be implemented in ASR in the future as an additional element of explainability to the system.

Though this work will take the form of future study, an exploratory analysis into the practical implementation of the phonetic approaches explored in this thesis will be conducted here. This will involve fusing the best approaches identified in this thesis, which to reiterate are the portfolios composed of all features and phonemes tailored to each sociophonetic group (found above in Table 9), alongside Phonexia's (2024) x-vector Voice Inspector ASR system which utilises DNNs (making it representative of the modern forensic ASR systems and the processes discussed in Chapter 2).

In Table 12 below, $C_{llr}$ values from Phonexia's (2024) Voice Inspector have been reported. These were calculated using the exact same datasets used in this thesis, as seen. More specifically, the same pairs of comparisons were tested to produce scores with Phonexia's (2024) Voice Inspector as tested to produce scores with the present thesis' phonetic approaches. However, it should be flagged that, for this exploratory experiment to work, the files had to be split in half for Phonexia's (2024) Voice Inspector as a minimum of 2 separate files per speaker were needed so that same-speaker comparisons could be conducted, as per Phonexia's (2024) Voice Inspector functionality; this likely led to overoptimistic performance, as only 2 same-session files were therefore used per speaker. Beyond this, the scores were then calibrated and fused just like the phonetic analyses above which combined the scores from separate phonetic feature measurements: as the output of Phonexia's (2024) Voice Inspector were scores, these could be processed into $C_{llr}$ values following the same methodology covered in Chapter 3. $C_{llr}$ valuesfrom Phonexia's (2024) Voice Inspector for each database are included below alongside a reminder of the scores for the all-feature, all-vowel phonetic portfolios generated in this thesis that represent the best identified phonetic approaches to ASR. The combined $C_{llr}$ values in the middle will be discussed momentarily.

**Table 12: $C_{llr}$ Values from Phonexia's (2024) Voice Inspector, This Thesis, and the Combination of Both for Each Corpus**

| Corpus | Phonexia's (2024) Voice Inspector $C_{llr}$ | Combined $C_{llr}$ | Phonetic $C_{llr}$ |
|---|---|---|---|
| **DyViS (Text-Dependent)** | 8.46E-34 | 1.01E-15 | 0.32 |
| **DyViS (Text-Independent)** | 5.32E-67 | 6.16E-55 | 0.15 |
| **WYRED (Text-Independent)** | 5.48406476846864E-96 | 4.15823554652027E-97 | 0.23 |

As seen, Phonexia's (2024) Voice Inspector always outranks the phonetic $C_{llr}$ values from this thesis, and as reiterated throughout, this is expected: all of the ASR developments moving away from traditional phonetic analyses reviewed in chapter 2 were motivated by improvements to performance, which manifest here, and speed, which was seen in the much faster collection of these results than those of this thesis, an issue which will be discussed in the final section of this chapter in more detail. Moez et al. (2016) also review the known trade-off between explainability and performance: specifically, they found that more explainable approaches tend to perform worse, and that is seen here. It should also be noted here that Phonexia's (2024) Voice Inspector performance is extremely good, and this is likely due to the unrealistic nature of this data, as also discussed in this chapter.

The most important findings here are the combined $C_{llr}$ values. Specifically, these are the $C_{llr}$ values for the fused combinations of the phonetic and ASR approaches here. As seen, these significantly improve upon the phonetic approaches tested in this thesis. This is expected: this thesis' approach was based solely on phonetic approaches, so when ASR approaches are included which, as discussed and shown above in isolation, are already known to perform better, a performance improvement is expectable. Compared to Phonexia's (2024) Voice

Inspector, though, the difference in performance is essentially zero: the inclusion of traditional phonetic approaches does not affect performance here. The performance of the ASR system is already at ceiling, so there is essentially no room for the phonetic information to improve performance. However, importantly, the addition of the phonetic information does not degrade performance either. This shows that phonetic approaches, which bring with them the explainability benefits reviewed in the earlier chapters, can be integrated alongside ASR approaches without major impact on performance. This means that they can be included as additional elements of explainability, should triers-of-fact need them, without impacting performance. Alternatively, these results could be read as an indication that Phonexia's (2024) Voice Inspector already measures these features of speech: performance change was effectively 0, so they may therefore already be factored in. This diagnostic view also addresses the explainability issue as it can show triers-of-fact that these perceivable elements of speech are potentially already factored into this ASR system's output.

It is important to reiterate that this thesis does not take issue with the stance of Morrison et al. (2021) that triers-of-fact need only know that a system is valid, but it does support the claims linked to Eldridge (2019) and Wang (2022) that a much more holistic approach could be implemented in ASR that also encompasses explainability. This is important because the job of the expert is to help reduce the probability of a miscarriage of justice, and adding an additional element of explainability to the system in the form of explainable phonetic analyses could help here. Moreover, it was suggested that increased explainability may be important to consider because triers-of-fact have growing trust issues with ASR evidence, as reported by van der Veer et al. (2021). This thesis may therefore provide one potential way of addressing this concern.

These results overall indicate that combining phonetic approaches with ASR approaches has no impact on performance. Speed, as will be discussed momentarily, still remains an issue here, but overall these explainable phonetic approaches can add an additional, optional element of explainability to ASR systems. This reflects the conclusions of similar studies reviewed earlier in this thesis: whilst Hughes et al.'s (2019a) study only considered one phonetic feature, one sociophonetically-controlled group, and did not consider phoneme-level analyses, they still found that ASR system performance was unchanged for most replications when mean harmonics-to-noise ratio was considered. However, as mean harmonics-to-noise ratio can be correlated to an explainable and perceivable element of a speaker's voice, this meant that including this measurement added an additional element of explainability to the system at no cost to performance. Thus, phonetic inclusions can aid any potential explainability queries from any triers-of-fact. This thesis effectively replicated this study but explored many more phonetic variables. It showed that tailored combinations of multiple explainable phonetic features, phonemes, and socio-phonetic considerations can be integrated into current ASR approaches to add an element of explainability without impacting performance, and used the current gold-standard validation metric ($C_{llr}$) to show this.

### 7.5. The Future of Phonetically-Informed Approaches to ASR

So far, the results have been discussed in terms of what they show for the effectiveness of phonetic approaches to ASR. Whilst caveated throughout with certain features that needed trimming universally or in specific scenarios, the headline findings thus far have been:

- Certain phonetic features, like formant 4, always perform well in ASR tasks for the given speaker groups.

- Features relating to non-modal perceptible features, such as breathiness and creakiness, perform worst for the given speaker groups.

- Vowels work well for ASR, but not in any predictable vowel-specific or stress-specific ways. More work could therefore be done to explore the efficacy of different vowels.

- Different sociophonetic groups have different needs in ASR that should be tailored towards. These sociophonetic considerations are of the utmost importance; the best performing approaches combined all of the extractable features and vowels for each given sociophonetic group.

- Forced alignment may also be accountable for some of the errors facing the phonetic approaches to ASR tested in this thesis.

- For each sociophonetically-tailored approach, it appears that data quantity may be more important than feature- and vowel-specificity.

- The best phonetic approaches identified in this thesis can be fused with current, off-the-shelf ASR systems without degrading performance. As these phonetic approaches are explainable, this could therefore add an element of explainability without any detrimental impact to the superior performance of these systems.

The remainder of this discussion will now build upon these general conclusions that explainable (socio)phonetic approaches are effective by turning to the future. It will position the developments offered by this thesis in the history of ASR research, in particular how they build upon historical and current methods and performance that were detailed in chapters 1 and 2. It will also explore where developments must occur from here; as discussed throughout, methodological concerns with the present thesis bring into question the viability of phonetic approaches to ASR right now, particularly for future ASR needs. These will be discussed in more detail here.

Though the primary concern of this thesis was always exploring novel, combinatory, and bespoke phonetically-informed approaches to ASR, the motivation for study was partly born out of concern for the explainability of modern ASR systems. Modern ASR systems, such as Phonexia BETA4, VOCALISE 2019A, and Nuance 11.1 were found to perform extremely well by Morrison and Enzinger's (2019) standards. These systems are amongst the modern gold standard for ASR; however, they use features generated by DNNs that cannot be fully explained by human developers and, as Rudin (2018) writes, can fail in ways that cannot be diagnosed and fixed. As a result of these concerns, as van der Veer et al. (2021) write, triers-of-fact are less trusting with ASR evidence due to lack of explainability and perceivability of what the output represents.

This thesis' novel combinations of explainable phonetic features could, as shown, be fused in a supplementary position in ASR to add an additional element of explainability to the system. They are explainable because the selected features have perceptual correlates in the voice, as discussed. They have also already been used successfully in ASR individually, as many studies reviewed in chapter 2 saw. The work of this thesis has shown that they are capable of effective ASR in combination with each other, and the final sections of this thesis showed that these portfolios can be fused successfully with off-the-shelf ASR systems. That said, the specific portfolios tested here are, again, proofs-of-concept; they were not created under forensically-realistic conditions. However, portfolios can be tested that are more forensically-realistic by using this thesis' methodology with more forensically-realistic data.

These portfolios, if integrated alongside current ASR systems, could lead to work that supports Tancock's (2018) stance that ASR needs to be accepted by wider society as a reliable and valid development by showing triers-of-fact that the system is, at least in-part, informed by phonetic measurements correlated to perceptible elements of the voice. It is

evident that ASR systems are not currently trusted as lay authors like Dickson (2020) villainise such systems based on their lack of explainability.

van der Veer et al. (2021) also argued that triers-of-fact need to be able to trust that these ASR systems work. In order to enable this, Morrison et al. (2021) suggested that these systems must be validated. This means that triers-of-fact should not be expected to understand exactly how a system works, only that it has worked and been used before. The work of this thesis does not infringe on or contradict this stance; simply put, adding these additional phonetic elements of explainability to an ASR system could aid validation by allowing triers-of-fact to link the output of ASR systems to elements of the voice that they already know and can perceive, should they require more support. Furthermore, these portfolios are already validated and optimised using the same performance metric ($C_{llr}$) that is commonly used to validate current ASR systems anyway. Concerns for their effects on ASR performance have therefore already been considered: as the final sections of this thesis showed, these portfolios do not have a detrimental effect on ASR system performance when fused.

Rudin's (2018) stance that unexplainable approaches should be nothing more than a placeholder until a more explainable approach exists was also explored in chapter 1. The present thesis strongly disagrees with the strength of this stance: modern ASR systems are already powerful, but their flaw is only explainability; supplementary, explainable information could fill this gap without the need for a full replacement system. Using phonetics as this supplementary explainable information, many studies in chapter 2 have already fused individual phonetic features with ASR systems to improve their explainability. Now, this thesis shows that multiple explainable phonetic features, vowels, and sociophonetic variables can be combined and tailored for performance, thus offering a more bespoke approach to addressing the explainability problems facing ASR.

Rudin's (2018) stance also cannot be fully supported at the time of writing because the tested phonetic approaches are not viable replacements from a practical perspective; they are hampered by modern limitations in methodology that render them too slow to process for real-world applications, especially in commercial settings where speed is more critical. The present thesis has not produced something efficient: it is slow and labour-intensive as a methodology, so utilising already powerful and fast ASR methods should be done to meet efficiency needs. As it stands, the present methodology took >2 hours to run per portfolio tested.

The importance of balancing performance and speed is prioritised by other authors more clearly: for example, Shaver and Acken (2016) concluded their report on the history of ASR with a comment on the future, stating that the primary goals of ASR must always be to recognise speakers with as few errors as possible and as fast as possible. The first of these goals, which concentrates on performance, is already achievable using current ASR methods, as chapter 2 showed. It is also achievable with the phonetic approaches tested here: it was shown that these combinations can perform well in fusion with off-the-shelf ASR systems too. However, the second of Shaver and Acken's (2016) main goals, speed, cannot be currently achieved through phonetics. This is because aligning the vowels, extracting feature measurements, and building and testing each portfolio took >2 hours. ASR developers may wish to recognise new speakers live during a phone call, particularly for commercial purposes; this is not possible under the method followed by this thesis, so prioritising already powerful ASR approaches currently makes more practical sense for ASR.

Fusing phonetic approaches with ASR systems would also enable the portfolio methodology to be used for another purpose: to understand and diagnose already-existing ASR systems. As already exemplified in Skarnitzl et al.'s (2019) and Hughes et al.'s (2019a) study, one can fuse phonetic features into an existing ASR system and measure the change in performance,

just like the 'top-down' method used to create the portfolios in chapter 6, to identify what effects these features have on ASR systems. Should they improve performance, that means that they should now be included because the present system does not incorporate it. If performance is stable, that means the present system already captures this feature and this knowledge makes the ASR system more explainable as a result. If performance declines, however, then the phonetic feature should simply be removed; as with mean autocorrelation, mean harmonics-to-noise ratio, and jitter, this thesis' methodology also enables researchers to identify what does not work for a given speaker or data group. This still falls under diagnosis and explanation, however, as it can be used to show what perceptible elements of the speakers' voices are not relevant here. This further shows how the present thesis' methodology can be employed beyond its original purposes to explain approaches to ASR.

Some final avenues for future research that combine the work of this thesis with modern ASR approaches would be to test current ASR approaches on phonetically-controlled data, such as phonemes, as opposed to a random allocation of speech. Also, sociophonetically-controlled UBMs could be tested instead of randomly-grouped UBMs. Both of these allow for modern ASR systems to be used with more explainability: one can see what the needs of different speech segments, speakers, and speech are with ASR features instead of phonetic features. By doing this, phonetic theory (in the shape of phoneme-specificity and sociophonetic considerations) can still be implemented in ASR in ways that retain the speed and performance of modern ASR approaches but improve explainability.

Speed is not the only issue encountered in this thesis, however. As seen throughout this chapter, there are pressing technical issues with the methodology that must be addressed to make sure these novel phonetic approaches to ASR are more viable. Firstly, in chapter 3, the efforts undertaken to ensure the results of this thesis are reliable were detailed: the selected data contains the expected speech, a reliable forced aligner was selected to mitigate the risk

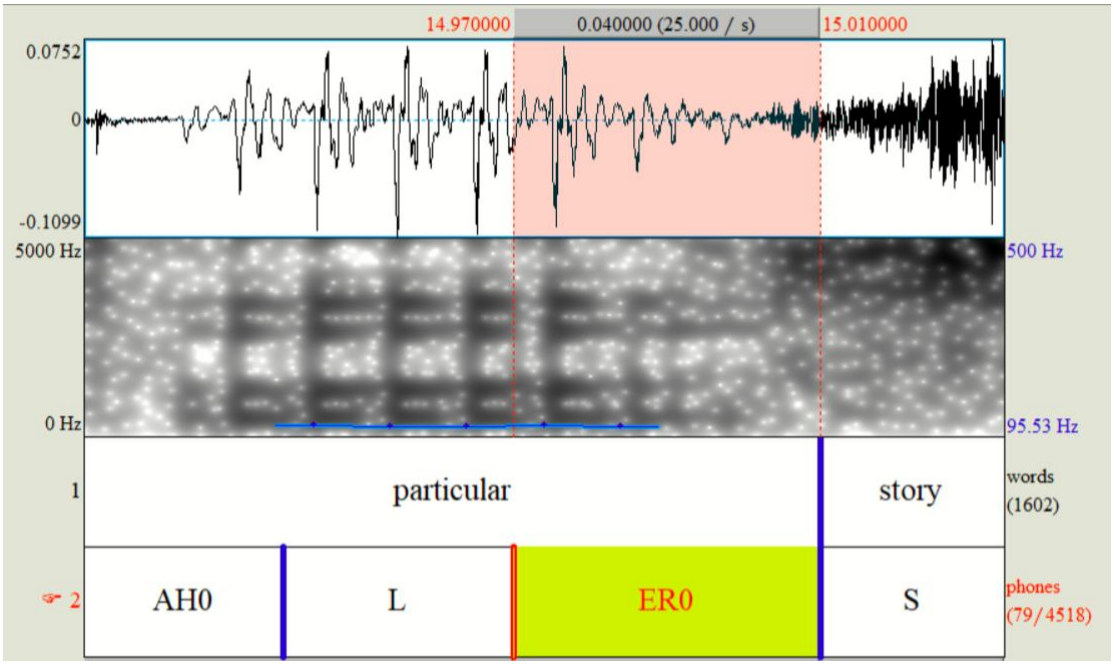
of misaligned TextGrids, reliable software in the form of Boersma and Weenink's (2023) Praat was selected for feature extraction, the Praat script has been checked and tested to ensure there are no accidental mathematical impossibilities, and a reliable performance metric was selected to create these portfolios with ($C_{llr}$). Despite all of this, chapter 4 raised concerns for the performance of the methodology through the unexpectedly high frequency of "undefined" results. Some causes of this issue will now be explored.

Starting with Boersma and Weenink's (2023) Praat itself, its reliability for extracting measurements from the data it is given can be mostly supported by the literature reviewed in chapters 2 and 3 above: Goedemans (2001) writes that Boersma and Weenink's (2023) Praat is generally considered the most reliable tool for phonetic analysis, Bhore and Shah (2015) found that formants are extracted reliably from Boersma and Weenink's (2023) Praat, Fernandes et al. (2018) found that mean harmonics-to-noise ratio is extracted reliably, de Jong and Wempe (2009) showed that intensity can be extracted reliably, and Jouvet and Laprie (2017) verify that Boersma and Weenink's (2023) Praat's approaches to extracting *f0*, and by extension mean autocorrelation, are reliable. Thus, based on prior literature, it is unlikely that the extraction issues originate from Boersma and Weenink's (2023) Praat.

The true cause of this issue can be traced by exploring the *f0* analysis above further. *f0*, based on the above literature, can only be reliably extracted if the segment in question is voiced. All of the selected phonemes are expected to be voiced; they are vowels. Thus, the frequency of "undefined" *f0* results may provide an insight into the performance of the forced aligner because an "undefined" result may indicate that the forced aligner is inaccurately capturing voiceless speech when demarcating the boundaries of the vowels. Upon a qualitative inspection of the collected data, this appears to be true: as seen in Figure 23 below, this is an example of the segmentation of the vowel /ɚ/ being extracted inaccurately from the word "particular" produced by Speaker(13) from the text-independent WYRED data. Over half of

this segment, as indicated by the lack of periodicity in the waveform, is unvoiced; moreover, the segmentation includes the beginning of the following voiceless alveolar fricative /s/. This triggered the return of an "undefined" result for *f0*.

**Figure 23: Failed Vowel Segmentation**

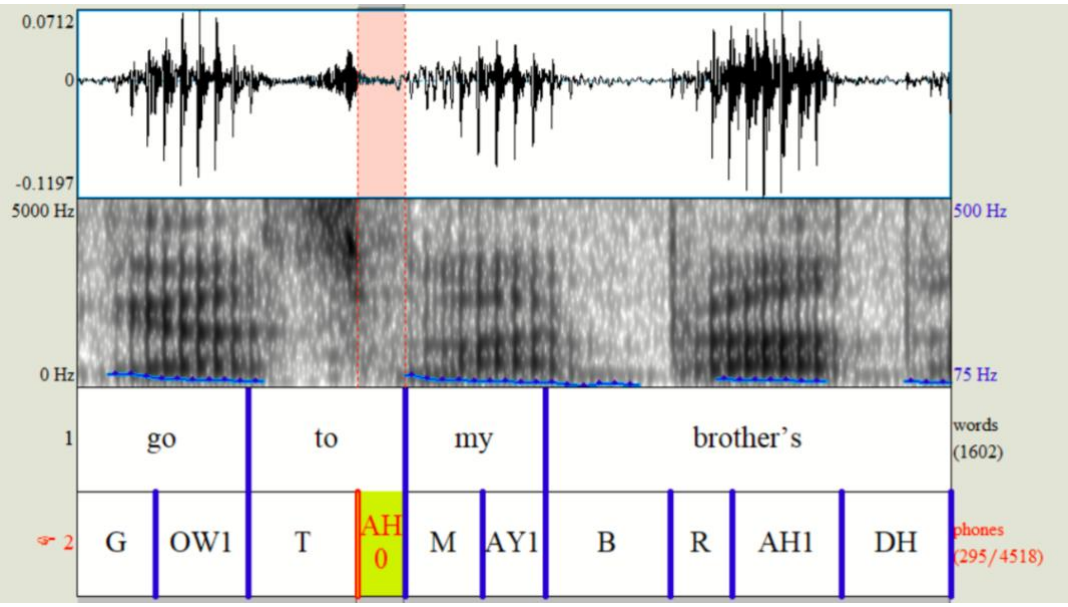



Above is an example of an error with forced alignment that may relate to co-articulation: the following /s/ is unvoiced, so the preceding vowel is cut short in preparation. This may indicate that the MFA's method, detailed in chapter 3, may not capture co-articulation effects as reliably as expected.

Continuing this investigation, there are sociophonetic reasons behind some forced alignment errors. One such reason is that the forced aligner could not account for scenarios where the vowel was not produced due to the style or accent of the speaker. This particularly concerns text-independent speech and select accents, such as West Yorkshire accents, which are subject to co-articulation effects that may omit the vowel. An example of an omitted vowel, but one that the MFA has still tried to align, can be seen in Figure 24 below. Here, Speaker(13) from WYRED produces a glottal stop instead of /ə/ in their production of "to".

This is a feature of West Yorkshire accents and can also be more typical of text-independent, free-flowing speech. This could explain a large amount of issues found in this thesis because there were more *f0* extraction failures in the text-independent databases, specifically in the West Yorkshire accent database too.

**Figure 24: Failed Detection of An Omitted Vowel**




There are further issues with forced alignment which must now be discussed. These problems do not contribute to the inaccuracy of the present methodology as the above issues do; these problems instead show how the intended goal of full explainability via phonetic methods is not, currently, feasible. Firstly, the MFA requires a transcript. This is fine for the text-dependent speech as a script of expected speech is typically provided, but as text-independent speech does not have a script by nature, one may have to turn to speech-to-text technologies, as tested in chapter 3. This is a particularly critical issue for forensic uses of ASR where text-independent speech is more prevalent. These speech-to-text technologies are very reliable and very fast; much like modern ASR methods. However, they are similarly also based on approaches that are not fully explainable. This issue cannot presently be avoided; the technologies available at the time of writing limits this. Thus, future research could turn to

automatic speech recognition in an attempt to make this more explainable. Otherwise, it contradicts the central explainability ethos of the thesis.

As an additional issue with forced alignment, it is reliant on dictionaries that may cause problems when investigating accent differences. For example, the FOOT/STRUT split is present in the General American English dictionary that was used in this thesis. This is only present in SSBE; the West Yorkshire accents do not have this split, so these vowels were technically investigated incorrectly as a result of the dictionary. In order to explore the efficacy of different vowels in different accents more accurately, dictionaries should be created for separate accents to account for sociophonetic variation more accurately. This is, again, a recommendation for future research. More importantly, accent-specific dictionaries could also solve the glottal stop issue observed above.

One may argue that this challenges claims from chapter 3 that the MFA can work with accents other than General American English: for example, due to the lack of FOOT/STRUT split in West Yorkshire English, this has led to tokens of the /ʊ/ vowel being divided across two different ARPAbet tokens given that some tokens classified as /ʌ/ will, in fact, be /ʊ/. Given that data quantity has been found to be so important to these phonetic approaches, this may account for the poor performance of /ʊ/ seen above. That said, however, this thesis also found that vowel specificity may be ultimately unimportant anyway; the best-performing combinations of phonetic approaches did not tailor to specific vowels, so future work may not need to contend with this issue at all.

The final problem with the dictionaries, as discussed above in chapter 3, is the use of proper nouns. If a word is not included in the dictionary, such as many proper nouns encountered in this study, the forced aligner ceases to function. A tool was tested for automatically creating dictionary entries for these, but it proved unreliable, as discussed in chapter 3. Thus, future

forced aligners must be used that can skip these unknown words. Alternatively, a more reliable automatic approach to transcribing these words, perhaps incorporated during the speech-to-text stage which could transcribe these words directly into ARPAbet representations, could be employed. Entering them manually, as was done in this thesis, goes against the automated ethos of this thesis too.

Despite placing accountability predominantly on the forced aligner, it is still possible to criticise Boersma and Weenink's (2023) Praat. Shimmer has been seen to be successful in ASR previously, as Farrús et al. (2007) demonstrated, but Teixeira and Gonçalves (2014) discovered that Boersma and Weenink's (2023) algorithms are not as reliable as theirs for accurate extraction. This may, potentially, account for the overall failure of shimmer in this study and the (mostly) lacking success of jitter. Retesting with their algorithm instead of Boersma and Weenink's (2023) may therefore yield better results that may support their conclusion that shimmer, and more measurements of jitter, can be used in ASR. This can be achieved with the present methodology: as discussed, future revisions may swap in software and approaches other than Boersma and Weenink's (2023) for feature extraction, perhaps even for speed-related concerns, and in doing so can implement more reliable measures.

It is also possible to question the overall explainability of the performance metric. In chapter 3, it was established that the most reliable metric for performance is $C_{llr}$. Prior methods have proven useful, but none capture as much information about same- and different-speaker performance as $C_{llr}$. However, this metric involves heavy abstraction and whilst the measurements used are based on explainable phonetic science, it is possible to also argue that this performance metric denies the system full explainability. Solutions to this issue may also lie in future research.

Finally, there are methodological concerns beyond these technical issues that lie with the data selected. The databases selected do not accurately represent general speech and therefore may not be reliable as the basis for generalisable portfolios. This is because the speech is controlled on topic, namely on the topic of a crime committed. This may work for some forensic scenarios because lexis from the semantic field of crime is being employed that is typical of more forensically-realistic data, but it is still not generalisable. Given the control of topic, these findings may also not be applicable to commercial scenarios wherein one may wish to recognise a speaker using more terminology from the semantic field of commerce. Thus, future research should also test data from the same sociophonetic group producing less crime-focused speech.

These are also simulated tasks recorded in controlled environments; thus, despite being text-independent, the speech still may not represent naturalistic speech and may be subject to Labov's (1972) Observer's Paradox. These research gaps can be addressed through further testing with the methodology under different conditions, specifically with the same speakers talking about different topics or producing more naturalistic data. This future work is critical for forensic implications: as discussed, the work of this thesis demonstrates the viability of phonetic approaches to ASR, but only under desirable conditions. Now, more forensically diverse data should be consulted, especially if real ASR applications are to be considered. This is particularly important for the high frequency of "undefined" results, which may become a more serious issue in less favourable data; here, though some explanations have been identified, they do not account for all of the "undefined" results, and these "undefined" results affect the interpretability of the findings. This offers a segue to a particularly important future study that asks the research question "how do juries react to this interpretable, supplementary, phonetic material for ASR and its limitations?". Now that a system built upon phonetic approaches deemed interpretable has been built and shown to

work without detrimental effects on performance, future work should explore how the system is received.

## 8. Conclusion

This thesis has explored novel, bespoke, and combinatory phonetically-informed methods of ASR. It did so with future practical implementations in mind, namely to aid ASR explainability. To do this, it started with an overview and critique of the history of ASR. From this, a move away from phonetics was observed that resulted in a research gap pertaining to explainability. This research gap is getting increasingly flagged by lay and academic audiences, particularly to ensure triers-of-fact trust in ASR evidence. This thesis, based on the work of prior academics in related speaker recognition fields, then proposed that phonetic approaches to ASR could be taken much further, and in doing so could address this research gap. It showed that phonetic measurements (that are already explainable through their perceptible links to elements of the voice) can be measured and employed for ASR purposes well. It specifically explored multiple combinations of phonetic features to identify and create well-performing, validated portfolios of explainable features for different sociophonetically-controlled groups. It explored how reliably extractable the measurements of the features are, their ability to recognise the same and different speakers, and how best to combine them for a given sociophonetic group or vowel to generate the most effective phonetic approaches to ASR. This combinatorial approach is novel to the field of ASR, and due to a focus on explainability, automation, and methods of validation used in ASR, could now be applied to ASR in future studies to address the explainability research gap directly. It was even shown that fusing the best-performing phonetic approaches with an off-the-shelf ASR system had no detrimental impact on their performance; thus, the explainable benefits of phonetic information can be integrated with well-performing ASR systems without performance concern.

Based on these results, a number of successes and challenges were encountered. Firstly, it was shown that the tested phonetic approaches are effective for ASR: phonetic features

relating to formants and intensity can universally benefit ASR performance whilst those relating to non-modal voice qualities, such as mean harmonics-to-noise ratio, mean autocorrelation, and jitter were detrimental to vowel-specific performance. This was all tested on tokens of monophthong vowels from speech. However, it was found that phonetic approaches ultimately performed best when more data were used anyway, as seen by the fact that considering all vowels together outperformed an analysis of any one vowel. These vowel-specific issues are, at least practically, therefore unimportant.

The biggest finding of this thesis was that the most important phonetic consideration for ASR performance was sociophonetics: for a given sociophonetic group, changing any feature- or vowel-specific parameters did not achieve scores as high as the scores when all tested features and segments were combined. On top of this, when sociophonetic variables were not considered, the worst score overall was observed.

Having established that phonetic approaches can be successful, it was also shown how these phonetic approaches could now be used by future ASR research to improve explainability. It was seen that, when fused alongside Phonexia's (2024) Voice Inspector, that the best combinations of features did not hamper performance whilst adding an additional element of explainability to the analysis.

This thesis does not support Rudin's (2018) argument that explainable approaches should eventually be used instead of less explainable, current ASR approaches; it instead argues that these approaches should be combined to take into account the explainability benefits of these phonetic approaches and the stronger performance of pre-existing ASR methods. Even from a practical perspective it cannot support the use of these phonetic portfolios as viable approaches to ASR on their own. This is because there are technical limitations at the time of writing that slow the processing of the present thesis' methodology. The more pressing issue,

however, is that there are methodological research gaps that hinder the full explainability of the phonetic approach anyway, namely through the use of speech-to-text technologies. This thesis therefore supports the use of fusion to integrate these phonetic approaches alongside current ASR approaches.

Even if researchers do not wish to use the direct output of this thesis, a more abstract use of this thesis is how to employ its principles of explainability in methodologies for ASR. As discussed, this thesis presently recognises that the use of phonetic features may be unattractive to commercial ASR given its lack of speed. However, modern ASR researchers could still take forward the use of combinatory portfolios, in particular by using sociophonetically-informed profiles of *ASR* features instead of *phonetic* features. Furthermore, they could use vowels as controlled and explainable units of voiced speech that have now proven effective for ASR when combined together. Sociophonetically-controlled UBMs could also be considered, given the importance of sociohponetically-controlled datasets here. Using modern approaches, which have larger computational benefits, but with these explainable, phonetic principles in mind, still moves the field into more explainable, phonetically-informed areas that address the explainability problems facing modern ASR.

Whilst this shows how the work of this thesis could be implemented now, attention should also be paid to what needs to be done to make phonetic approaches to ASR, as tested here, more viable in the future. Firstly, the extraction of phonetic feature measurements, as well as the portfolio creation methods, is too computationally taxing to work fast enough for commercial ASR researchers looking for live recognition. Thus, future research must be conducted into speeding up the methods of data extraction and comparison. Some efforts have already been made, such as the discussed long formant measurements that can reduce the amount of formant features to analyse (Nolan and Grigoras, 2005).

Forced alignment presented the most problems for the tested approach. As seen, it appears to be responsible for most failed extractions due to misaligned phoneme boundaries and the inclusion of unvoiced segments. It is inaccurate, and this may be due to its lacking adaptability to co-articulation effects and its dependence on a dictionary that may not accurately reflect the influence of accent on the data. More specifically, the dictionary used was only tailored to General American English, which does not have vowel distinctions that are present in the included accents of this study. Prior literature suggested that the use of this dictionary would be reliable, but this led certain vowels to be misrepresented, namely /ʌ/, due to the dictionary not including sociophonetic variation known to be present in the selected accents, here the lack of FOOT/STRUT split in West Yorkshire accents. This shows that forced alignment may require more varied dictionaries.

Secondly, current forced alignment problems required the use of partially unexplainable tools, here speech-to-text technologies. This is problematic for the ethos of this thesis because an unexplainable approach has had to be employed due to no explainable approach to automatic speech recognition presently existing. This problem is pertinent for text-independent speech which, by nature, does not have a transcript. Solving this problem required turning to speech-to-text technology, which itself slowed the system down. These flagged issues show that the explainability, accuracy, and applicability of forced alignment should be rectified before these phonetic approaches can be employed viably as an effective-yet-explainable approach in ASR. This may take the form of a new forced aligner, a new dictionary, and perhaps an explainable speech-to-text model that transcribes speech using the ARPAbet directly, removing the need for any problematic dictionaries to be used.

In summary of these technical problems, phonetic technologies are currently too limited to compete with modern ASR demands. Boersma and Weenink's (2023) Praat can extract the measurements, but it is too slow. A more dedicated system, perhaps even including better

approaches for measuring jitter and shimmer to test their effectiveness further, should be developed. Modern forced alignment methods offer competitive performance, but they have accuracy problems that cannot be rectified due to their lacking explainability and reliance on dictionaries that cannot capture intra- and inter-speaker variation accurately. Current forced alignment methods also require the use of speech-to-text approaches that generate problems for the explainability of this ASR approach. This issue cannot be rectified until future automatic speech recognition technologies and approaches exist. The problems facing forced alignment are similar to the problems facing ASR that this thesis sought to address: the field is employing less explainable approaches and creating unfixable problems that may benefit from future applications of (socio)phonetic theory. That said, this thesis in no way argues that phonetic approaches to ASR need to be competitive with current ASR methods in terms of processing; some of these technical issues may not be an issue for future work concerning the fusion of these now-tested phonetic approaches.

On top of these technical issues, there is the additional theoretical issue that this thesis has opened up for phoneticians: the individual differences in vowel performance could not be accounted for between the databases. Some universal trends were linked to data quantity issues and wider sociophonetic differences, but the more specific reasons for different rank orders emerging for different accents and styles are unclear. This, therefore, continues to be an under-researched area. At the time Paliwal (1984) discovered their rank order of vowel performance, they could not account for the rank order they discovered due to lacking research; little progress has been made since then. This thesis does successfully update this study's findings using more modern validation metrics, however, by re-affirming that rank orders do emerge in vowel performance and that different vowels behave differently for different accents and styles. However, due to continuingly lacking research, the reasons behind why these different vowels behaved differently could not be discerned beyond data

quantity and wide sociophonetic explanations. This is another avenue for future research; however, the overarching conclusion of this thesis is that these differences between vowels do not matter for the practical reality of ASR; the best performance was observed when all vowels were considered together indiscriminately.

Issues aside, addressing these problems only seeks to improve the foundations that are laid here. Combinatory, bespoke, phonetic approaches to ASR have been shown to work well and sociophonetic considerations have been shown to be more important than initially thought. Future work can now explore fusing these methods into ASR approaches to boost explainability more realistically. The portfolios generated here could technically now be used by ASR researchers, but these are better served as proof-of-concept; they used forensically-unrealistic data. Now, more forensically-realistic portfolios should be made for different accents, sexes, races, styles, recording qualities, and sessions before ASR applications are considered further. If a database of speech is big enough to allow for a singular sociophonetic group to be profiled, a portfolio can be generated for it to produce a clearer picture of the needs of that group for ASR. This work may even involve the creation of more speech databases, which is another avenue for future research. However, at present, it must be reiterated that the portfolios tested are only applicable to the chosen speakers and speech, and they were created under desirable conditions. The portfolios are still usable, but they are not generalisable.

As another future avenue for research, different languages could also be explored; the methodology is, technically, language-independent. This thesis explored variation within the same language, but as long as the databases and dictionaries exist for data in another language, these can also be investigated using the same template methodology found here to generate more portfolios for different speakers and speech.

Any created portfolios can then be continuously used and updated to improve their reliability. It is particularly important to update these portfolios because data quantity has already proven to have a positive effect on these portfolios; more data will therefore improve them. To do this, one could conduct replications that swap out which speakers are used as the test speakers in the portfolio creation stage using the same data. This is because Wang (2021) found that this can capture further variability using the same data. Thus, this opens up more avenues for future research to improve the reliability of the groundwork laid here simply through replications with the same data and the inclusion of further data.

As a final thought, it should also be reiterated that this methodology is modular; each task is done independently, as detailed in chapter 3. This allows the methodology to be futureproof: the component technologies and approaches can all be swapped out with new developments to either explore new ground, such as new data or even features, or to improve the reliability of different components, such as the forced alignment and feature extraction stages.

The position of this research is simply that the individuality of one's voice should be characterised and measured using the best tools for the task. It has been shown that current ASR approaches are already best for speed and performance, but novel, bespoke, combinatorial (socio)phonetic approaches to ASR are also viable and could be integrated in ASR as supplementary tools for explainability purposes. This thesis identified some of these approaches and provided a tool for identifying more of these approaches for different datasets. This thesis recognises that the reality of ASR is constantly negotiated and debated; research in phonetics, computational science, and engineering have all brought revisions to ASR over time, and the present thesis joins other researchers such as Hughes et al. (2019a) in rediscovering phonetics' place in ASR.

**Reference List**

Abhang, P. A., Gawali, B. W., & Mehrotra, S. C. (2016). *Introduction to EEG- and Speech-Based Emotion Recognition* [e-book]. Amsterdam: Elsevier. Retrieved April 27, 2023 from https://www.researchgate.net/.

Ali, A., Bhatti, S., & Mian, M. S. (2006). *Formants Based Analysis for Speech Recognition*. Paper presented at the 2006 IEEE International Conference on Engineering of Intelligent Systems, Islamabad, Pakistan. https://doi.org/10.1109/ICEIS.2006.1703179.

Almaadeed, N., Aggoun, A., & Amira, A. (2016). Text-Independent Speaker Identification Using Vowel Formants. *Journal of Signal Processing Systems*, *82*(3), 345–356. https://doi.org/10.1007/s11265-015-1005-5.

Alsulaiman, M., Mahmood, A., & Muhammad, G. (2017). Speaker recognition based on Arabic phonemes. *Speech Communication*, *86*, 42–51. https://doi.org/10.1016/j.specom.2016.11.004.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research, 18*, 1-78. https://doi.org/10.48550/arxiv.1704.01701.

Atal, B.S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, *52*(6), 1687–1697. https://doi.org/10.1121/1.1913303.

Atal, B.S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America, 55*(6), 1304-1322. https://doi.org/10.1121/1.1914702.

Atkinson, N. (2015). *Variable factors affecting voice identification in forensic contexts*. [Doctoral dissertation, University of York]. White Rose eTheses Online.

Bailey, G. (2016). Automatic detection of sociolinguistic variation using forced alignment. *University of Pennsylvania Working Papers in Linguistics, 22*(2). https://repository.upenn.edu/pwpl/vol22/iss2/3.

Basu, N., Bali, A. S., Weber, P., Rosas-Aguilar, C., Edmond, G., Martire, K. A., & Morrison, G. S. (2022). Speaker identification in courtroom contexts – Part I: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. *Forensic Science International*, *341*, 2–22. https://doi.org/10.1016/j.forsciint.2022.111499.

Bergstrom, B. E. (2017). *Effect of Speaker Age and Dialect on Listener Perceptions of Personality.* [Masters dissertation, Brigham Young University]. BYU ScholarsArchive.

Bhore, S. S. & and Shah, M. S. (2015). Comparison of Formant Estimation Techniques. *International Journal of Electronics, Communication & Soft Computing Science and Engineering, 6,* 127-129. https://ijecscse.org/papers/IETE2015/140.pdf.

Bidondo, A., Sato, S., Kinigsberg, E., Arouxet, M., Andrés, S., Arias, A., Saavedra, A., & Groisman, A. (2013). Speaker recognition analysis using running autocorrelation function parameters. *The Journal of the Acoustical Society of America*, *133*(5), 3293–3293. https://doi.org/10.1121/1.4805422.

Boersma, P. & Weenink, D. (2023). Praat (Version 6.3.09). [Computer Software]. Retrieved April 18, 2023 from https://www.fon.hum.uva.nl/praat/.

Boersma, P. (1993). Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound. *IFA Proceedings*, *17*, 97-110. https://www.fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf.

Bolt, R. H., Cooper, F. S., David, Edward E., Jr, Denes, P. B., Pickett, J. M., & Stevens, K. N. (1969). Identification of a Speaker by Speech Spectrograms: How do Scientists View Its Reliability for Use as Legal Evidence? *Science*, *166*(3903), 338–343. https://doi.org/10.1126/science.166.3903.338.

Brockmann, M., Drinnan, M. J., Storck, C., & Carding, P. N. (2011). Reliable Jitter and Shimmer Measurements in Voice Clinics: The Relevance of Vowel, Sex, Vocal Intensity, and Fundamental Frequency Effects in a Typical Clinical Task. *Journal of Voice*, *25*(1), 44–53. https://doi.org/10.1016/j.jvoice.2009.07.002.

Burg, J. P. (1967). *Maximum entropy spectral analysis*. Paper presented at the 37th Meeting of the Society of Exploration Geophysicists, Oklahoma City: OK.

Champod, C. & Evett, I. W. (2000). Commentary on A. P. A. Broeders (1999) Some observations on the use of probability scales in forensic identification. *Forensic Linguistics*, 7(2), 238-243.

Chen, S. H. (2007). Sex differences in frequency and intensity in reading and voice range profiles for Taiwanese adult speakers. *Folia Phoniatr Logop*, *59*(1), 1-9. https://doi.org/10.1159/000096545.

CMUdict. (2014). LOGIOS Lexicon Generation Tool. [Computer Software]. Retrieved April 18, 2023 from http://www.speech.cs.cmu.edu/tools/lextool.html.

De Jong, G., McDougall, K., & Nolan, F. (2007). Sound Change and Speaker Identity: An Acoustic Study. In Müller, C. (Ed.), *Speaker Classification II,* pp.130–141. Springer. https://doi.org/10.1007/978-3-540-74122-0_12.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385.

Derdemezis, E., Vorperian, H. K., Kent, R. D., Fourakis, M., Reinicke, E. L., & Bolt, D. M. (2016). Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. *American Journal of Speech-Language Pathology*, *25*(3), 335–354. https://doi.org/10.1044/2015_AJSLP-15-0020.

Dickson, B. (2020, July 7). What will happen when we reach the AI singularity? *The Next Web*. https://thenextweb.com/.

Doddington, G. R. (1985). Speaker recognition-Identifying people by their voices. *Proceedings of the IEEE*, *73*(11), 1651–1664. https://doi.org/10.1109/PROC.1985.13345.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2016). Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises. Retrieved April 18, 2023 from http://enfsi.eu/.

Dufour, R., Estève, Y., & Deléglise, P. (2014). Characterizing and detecting spontaneous speech: Application to speaker role recognition. *Speech Communication*, *56*(1), 1–18. https://doi.org/10.1016/j.specom.2013.07.007.

Earnshaw, K. (2021). Examining the implications of speech accommodation for forensic speaker comparison casework: A case study of the West Yorkshire face vowel. *Journal of Phonetics*, *87*, 2-15. https://doi.org/10.1016/j.wocn.2021.101062.

Eatock, J. P., & Mason, J. S. (1994). *A quantitative assessment of the relative speaker discriminating properties of phonemes*. Paper presented at ICASSP '94, Adelaide: Australia. https://doi.org/10.1109/ICASSP.1994.389337.

Eide, E., & Gish, H. (1996). *A parametric approach to vocal tract length normalization*. Paper presented at the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference, Atlanta: GA. https://doi.org/10.1109/ICASSP.1996.541103.

Eldridge, H. (2019). Juror comprehension of forensic expert testimony: A literature review and gap analysis. *Forensic Science International: Synergy.* 1, 24-34. https://doi.org/10.1016/j.fsisyn.2019.03.001.

Farrús, M., Hernando, J., & Ejarque, P. (2007). *Jitter and shimmer measurements for speaker recognition*. Paper presented at the 8th Annual Conference of the International Speech Communication Association, Antwerp: Belgium. https://nlp.lsi.upc.edu/papers/far_jit_07.pdf.

Fernandes, J., Teixeira, F., Guedes, V., Junior, A., & Teixeira, J. P. (2018). Harmonic to Noise Ratio Measurement - Selection of Window and Length. *Procedia Computer Science, 138,* 280-285. https://doi.org/10.1016/j.procs.2018.10.040.

Ferrand, C. T. (2002). Harmonics-to-Noise Ratio: An Index of Vocal Aging. *Journal of Voice*, *16*(4), 480–487. https://doi.org/10.1016/S0892-1997(02)00123-6.

French, J. P. (2017). A developmental history of forensic speaker comparison in the UK. *English Phonetics*, *21*, 271 – 286. https://eprints.whiterose.ac.uk/117763/7/Developmental_History_of_Forensic_Speaker_Comparison_in_the_UK.pdf.

French, J.P. (1990). Forensic applications of phonetics. *Terminologie & Traduction*, *1,* 181-187. https://op.europa.eu/mt/publication-detail/-/publication/57458bb5-82c1-4232-bff3-bdd498230944.

Furui, S. (1997). Recent advances in speaker recognition. *Pattern Recognition Letters*, *18*(9), 859–872. https://doi.org/10.1016/S0167-8655(97)00073-1.

Garrett, B. L. & Mitchell, G. (2013). How Jurors Evaluate Fingerprint Evidence: The Relative Importance of Match Language, Method Information, and Error Acknowledgment. *Journal of Empirical Legal Studies,* 10(3), 484-511. https://scholarship.law.duke.edu/faculty_scholarship/3853/.

Gallardo, L. F., Möller, S., & Wagner, M. (2015). *Importance of Intelligible Phonemes for Human Speaker Recognition in Different Channel Bandwidths*. Paper presented at Interspeech 2015, Dresden: Germany. https://www.isca-speech.org/archive_v0/interspeech_2015/papers/i15_1047.pdf.

Gao, S., Hu, J., Gong, D., Chen, S., Kendrick, K. M., & Yao, D. (2012). Integration of consonant and pitch processing as revealed by the absence of additivity in mismatch negativity. *PloS One*, *7*(5), e38289. https://doi.org/10.1371/journal.pone.0038289.

Gauthier, E., Besacier, L., & Voisin, S. (2016). Automatic Speech Recognition for African Languages with Vowel Length Contrast. *Procedia Computer Science, 81,* 136-143. https://doi.org/10.1016/j.procs.2016.04.041.

Gold, E., Ross, S. & Earnshaw, K. (2018). *The 'West Yorkshire Regional English Database': Investigations into the Generalizability of Reference Populations for Forensic Speaker Comparison Casework.* Paper presented at Interspeech 2018, Hyderabad: India.

Gonzalez-Rodríguez, J. (2014). Evaluating Automatic Speaker Recognition systems: An overview of the NIST Speaker Recognition Evaluations (1996-2014). *Loquens, 1*(1), e007. http://dx.doi.org/10.3989/loquens.2014.007.

Gonzalez-Rodríguez, J., Gil, J., Pérez, R., & Franco-Pedroso, J. (2014). *What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials.* Paper presented at Odyssey 2014, Joensuu: Finland.

Google Cloud. (2023). Cloud Speech-to-Text. [Computer Software]. Retrieved April 19, 2023 from https://cloud.google.com/speech-to-text.

Gourisaria, M. K., Agrawal, R., Sahni, M., & Singh, P. K. (2024). Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. *Discover Internet of Things,* 4(1). https://doi.org/10.1007/s43926-023-00049-y.

Grabe, E., Post, B., Nolan, F., & Farrar, K. (2000). Pitch accent realization in four varieties of British English. *Journal of Phonetics*, *28*(2), 161–185. https://doi.org/10.1006/jpho.2000.0111.

Harrison, P. (2013). *Making accurate formant measurements : an empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements*. [Doctoral dissertation, University of York]. White Rose eTheses Online.

Hautamäki, V., Kinnunen, T., & Fränti, P. (2008). Text-independent speaker recognition using graph matching. *Pattern Recognition Letters,* 29(9), 1427-1432. https://doi.org/10.1016/j.patrec.2008.02.021.

HSBC. (2023). *Voice ID.* HSBC. https://ciiom.hsbc.com/.

Heeren, W. (2020). The effect of word class on speaker-dependent information in the Standard Dutch vowel /a:/. *Journal of the Acoustical Society of America,* 148(4), 2028. https://doi.org/10.1121/10.0002173.

Heeren, W., Smorenburg, L., & Gold, E. (2022). Optimizing the strength of evidence: Combining segmental speech features. *Proceedings of IAFPA 2022.* Prague: Czech Republic.

Holmes, E. J. (2021a). Using Phonetic Theory to Improve Automatic Speaker Recognition. Poster presented at XVII AISV Conference, Virtual conference. 4-5 February 2021.

Holmes, E. J. (2021b). Using Phonetic Theory to Recognise Synthetic Speech. Poster presented at mFiL 2021, Virtual conference. 28-29 April 2021.

Holmes, E. J. (2021c). The Use of Nasals in Automatic Speaker Recognition. Presented at The 29th International Association for Forensic Phonetics and Acoustics (IAFPA), Virtual conference. 22-25 August 2021.

Holmes, E. J. (2022a). Optimising Phonetic Approaches to Automatic Speaker Recognition. Poster presented at The 2022 Colloquium of the British Association of Academic Phoneticians, Virtual conference. 4-8 April 2022.

Holmes, E. J. (2022b). Recognising Socio-Phonetically Comparable Speakers Using Phonetic Approaches to Automatic Speaker Recognition. Presented at The 29th International Association for Forensic Phonetics and Acoustics (IAFPA), Prague, Czech Republic. 10-13 July 2022.

Hughes, V., Cardoso, A., Foulkes, P., French, J. P., Harrison, P. and Gully, A. (2019a) Forensic voice comparison using long-term acoustic measures of voice quality. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS).* Melbourne: Australia.

Hughes, V., Harrison, P., Foulkes, P., French, J. P. and Gully, A. (2019b) Effects of formant settings and channel mismatch on semi-automatic systems in forensic voice comparison. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS).* Melbourne: Australia.

Hughes, V., Harrison, P., Foulkes, P., French, J. P., Kavanagh, C. and San Segundo, E. (2017) Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing. *Proceedings of Interspeech*. Stockholm University: Sweden.

Hughes, V., Harrison, P., Foulkes, P., French, J. P., Kavanagh, C. and San Segundo, E. (2018) The individual and the system: assessing the stability of the output of a semi-automatic forensic voice comparison system. *Proceedings of Interspeech*. Hyderabad: India.

Hughes, V., Llamas, C. and Kettig, T. (2022) Eliciting and evaluating likelihood ratios for speaker recognition by human listeners under forensically realistic channel-mismatched conditions. *Proceedings of Interspeech*. Incheon: Korea.

Jesse, M., Alexander, A., & Forth, O. (2014). *Forensic Voice Comparisons in German with Phonetic and Automatic Features Using Vocalise Software*. Paper presented at the 54th International Conference of the AES, London: UK. https://www.researchgate.net/publication/343254940_Forensic_voice_comparisons_in_German_with_phonetic_and_automatic_features_using_VOCALISE_software.

Jia, Y., Chen, X., Yu, J., Wang, L., Xu, Y., Liu, S., & Wang, Y. (2021). Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network. *Complex & Intelligent Systems*, *7*(4), 1749–1757. https://doi.org/10.1007/s40747-020-00172-1.

Jouvet, D., & Laprie, Y. (2017). Performance Analysis of Several Pitch Detection Algorithms on Simulated and Real Noisy Speech Data. *Proceedings of the 25th European Signal Processing Conference.* Kos: Greece.

Kajarekar, S. S., Bratt, H., Shriberg, E., & de Leon, R. (2006). A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition. *2006 IEEE Odyssey - The Speaker and Language Recognition Workshop*. San Juan, PR: USA.

Kaur, K. & Jain, N. (2015). Feature extraction and classification for automatic speaker recognition system – a review. *International Journal of Advanced Research in Computer Science, 5*(1), 1-6. https://www.semanticscholar.org/paper/Feature-Extraction-and-Classification-for-Automatic-Kaur-Jain/3e75781a18158f6643115ba825034f9f95b4f13b.

Kersta, L. (1962). Voiceprint Identification. *Nature, 196*, 1253–1257. https://doi.org/10.1038/1961253a0.

Kinnunen, T. (2003). *Spectral Features for Automatic Text-Independent Speaker Recognition*. [Licentiate's thesis, University of Joensuu]. The PUMS Project.

Klug, K., Kirchhübel, C., Foulkes, P. & French, J. P. (2019). Analysing breathy voice in forensic speaker comparison: using acoustics to confirm perception. *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne: Australia.

Koshy, A. & Tavakoli, S. (2022). Exploring British Accents: Modelling the Trap-Bath Split with Functional Data Analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics,* 71(4), 773-805. https://doi.org/10.1111/rssc.12555.

Labov, W. (1972). *The Social Stratification of English in New York City.* 2nd edn. Cambridge: Cambridge University Press.

Lammert, A. C., & Narayanan, S. S. (2015). On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. *PloS One*, *10*(7), e0132193. https://doi.org/10.1371/journal.pone.0132193.

Leong K., Hawkshaw M. J., Dentchev D., Gupta R., Lurie D., & Sataloff R. T. (2013). Reliability of objective voice measures of normal speaking voices. *Journal of Voice*, 27(2), 170–176.

Levinson, S. E. (1994). Speech Recognition Technology: A Critique. In D. B. Roe & J. G. Wilpon (Eds.), *Voice Communication Between Humans and Machines* (pp. 159-164), National Academy of Sciences. https://doi.org/10.17226/2308.

Lo, J.J.H. (2021) Cross-Linguistic Speaker Individuality of Long-Term Formant Distributions: Phonetic and Forensic Perspectives. *Proceedings of Interspeech 2021*, Brno: Czech Republic.

Long, Y., Yan, Z., Soong, F. K., Dai, :., & Guo, W. (2011). Speaker characterization using spectral subband energy ratio based on Harmonic plus Noise Model. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague: Czech Republic.

Lortie, C. L., Thibeault, M., Guitton, M. J., & Tremblay, P. (2015). Effects of age on the amplitude, frequency and perceived quality of voice. *AGE*, *37*(6), 117–117. https://doi.org/10.1007/s11357-015-9854-1.

Luck, J.E. (1969). Automatic Speaker Verification Using Cepstral Measurements. *Journal of the Acoustical Society of America*, *46*(4), 1026-1032. https://doi.org/10.1121/1.1911795.

Lummis, R. (1973). Speaker verification by computer using speech intensity for temporal registration. *IEEE Transactions on Audio and Electroacoustics*, *21*(2), 80–89. https://doi.org/10.1109/TAU.1973.1162443.

Malyuga, E., Orlova, S., Krouglov, A., & Ivanova, M. (2017). Methodological Aspects in Training Businesspeople: English Declarative Sentences Intonation Contours in Business Negotiations. *EDULEARN17 Proceedings,* Barcelona: Spain.

Matarneh, R., Maksymova, S., Lyashenko, V., & Belova, N.V. (2017). Speech Recognition Systems: A Comparative Review. *Journal of Computer Engineering, 19*(5), 71-79. https://openarchive.nure.ua/server/api/core/bitstreams/1e23eacd-3bc2-480a-8e59-72596cb28826/content.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. (2017) Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proceedings of Interspeech 2017*, Stockholm: Sweden.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. In: Chen, C.H. (Ed.), *Pattern Recognition and Artificial Intelligence*. Academic Press, New York, 374–388.

Meuwly, D. (2001). *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique.* [Doctoral dissertation, University of Lausanne] N. Brü.

Misnikov, E. (2019). *Sociophonetic factors in automatic speech recognition: a study on American English.* [Master's Thesis, Albert Ludwig's University]. Deutsche National Bibliothek.

Moez, A., Jean-François, B., Waad, B. K., Solange, R., & Juliette, K. (2016). Phonetic content impact on Forensic Voice Comparison. *2016 IEEE Spoken Language Technology Workshop (SLT),* San Diego, USA.

Mokgonyane, T. B., Sefara, T. J., Manamela, M. J., Modipa, M. I., & Masekwameng, M. S. (2020). The Effects of Acoustic Features of Speech for Automatic Speaker Recognition. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Durban: South Africa.

Morrison G.S., Zhang C., Rose P. (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International*, *208*, 59–65. https://10.1016/j.forsciint.2010.11.001.

Morrison, G. S. (2018). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International*, *283*, 1-7. https://doi.org/10.1016/j.forsciint.2017.12.024.

Morrison, G. S., & Enzinger, E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction. *Speech Communication*, *85*, 119–126. https://doi.org/10.1016/j.specom.2016.07.006.

Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W. C., van der Vloed, D., Ypma, R. J. F., Zhang, C., Anonymous, A., & Anonymous, B. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, *61*(3), 299–309. https://doi.org/10.1016/j.scijus.2021.02.002.

Mouawad, P. (2017). *Modeling and predicting affect in audio signals: perspectives from acoustics and chaotic dynamics*. [Thesis, University of Bordeaux]. Archive ouverte HAL.

Murphy, P. J., & Akande, O. O. (2007). Noise estimation in voice signals using short-term cepstral analysis. *The Journal of the Acoustical Society of America*, *121*(3), 1679–1690. https://doi.org/10.1121/1.2427123.

Nasersharif, B. & Akbari, A. (2007). SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features. *Pattern Recognition Letters,* 28(11), 1320-1326. https://doi.org/10.1016/j.patrec.2006.11.019.

Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, *12*(2), 143–173. https://doi.org/10.1558/sll.2005.12.2.143.

Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *The International Journal of Speech, Language and the Law*, *16*(1), 31–57. https://doi.org/10.1558/ijsll.v16i1.31.

Ogden, R. (2009). *An introduction to English phonetics.* Edinburgh: Edinburgh University Press.

Paliwal, K. K. (1984). Effectiveness of different vowel sounds in automatic speaker identification. *Journal of Phonetics*, *12*(1), 17–21. https://doi.org/10.1016/S0095-4470(19)30846-0.

Panayotov, V., Guoguo Chen, Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore: Singapore. https://doi.org/10.1109/ICASSP.2015.7178964.

Peddinti, V., Manohar, V., Wang, Y., Povey, D., & Khudanpur, S. (2016). Far-field ASR without parallel data. *Interspeech 2016*, San Francisco: USA.

Pfitzinger, H. R. & Kaernbach, C. (2008): Amplitude and amplitude variation of emotional speech. *Proceedings of Interspeech 2008*, Brisbane: Australia.

Phonexia. (2024). *Phonexia Voice Inspector.* Phonexia. https://www.phonexia.com/use-case/audio-forensics-software/.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hanneman, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G, & Vesely, K. (2011). *The Kaldi Speech Recognition Toolkit*. Paper presented at the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hawai'i, US.

Pruksachatkun, Y., McAteer, M., & Majumdar, S. (2023). *Practicing Trustworthy Machine Learning.* O'Reilly.

R Core Team. (2022). *R* (Version 4.3.0). [Computer Software]. Retrieved April 19, 2023 from https://www.r-project.org/.

Rouvier, M., Bouallegue, M., Matrouf, D., & Linares, G. (2011). *Factor analysis based session variability compensation for Automatic Speech Recognition*. Paper presented at the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Hawai'i, US.

Rudin, C. (2018). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, *Nature Machine Intelligence, 1,* 206-215. https://doi.org/10.48550/arxiv.1811.10154.

Safavi, S., Russell, M., & Jančovič, P. (2018). Automatic speaker, age-group and sex identification from children's speech. *Computer Speech & Language*, *50*, 141–156. https://doi.org/10.1016/j.csl.2018.01.001.

Sambur, M. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *23*(2), 176–182. https://doi.org/10.1109/TASSP.1975.1162664.

Schmid, M. & Bradley, E. D. (2019). Vocal pitch and intonation characteristics of those who are sex non-binary. *2019 International Congress of Phonetic Sciences*. Melbourne: Australia.

Shaver, C. D. & Acken, J. M. (2016). A Brief Review of Speaker Recognition Technology. *Electrical and Computer Engineering Faculty Publications and Presentations*, *350,* 1-7. https://core.ac.uk/download/pdf/37775846.pdf.

Sheena, Mary, B. B., Aswin, V. A., & Suprent, A. (2022). Variation of Harmonics to Noise Ratio from the Age Range of 9–18 Years Old in both the Sexs. *Indian Journal of Otolaryngology and Head & Neck Surgery*, *74*(3), 5518–5523. https://doi.org/10.1007/s12070-021-02858-5.

Skarnitzl, R., Asiaee, M., & Nourbakhsh, M. (2019). Tuning the performance of automatic speaker recognition in different conditions. *The International Journal of Speech, Language and the Law*, *26*(2), 209. https://doi.org/10.1558/ijsll.39778.

Son, J., Kyung, C., & Cho, H. (2019). Practical Inter-Floor Noise Sensing System with Localization and Classification. *Sensors,* 19(17), 3633-3652. https://doi.org/10.3390/s19173633.

Suess, N., Hauswald, A., Reisinger, P., Rösch, S., Keitel, A., & Weisz, N. (2022). Cortical Tracking of Formant Modulations Derived from Silently Presented Lip Movements and Its Decline with Age. *Cerebral Cortex*, *32*(21), 4818–4833. https://doi.org/10.1093/cercor/bhab518.

Tancock, C. (2018, November 26). In a nutshell: how to write a lay summary. *Elsevier.* https://www.elsevier.com/.

Teixeira, J. P. & Gonçalves, A. (2014). Accuracy of Jitter and Shimmer Measurements. *Procedia Technology, 16,* 1190-1199. https://doi.org/10.1016/j.protcy.2014.10.134.

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal Acoustic Analysis - Jitter, Shimmer and HNR Parameters. *Procedia Technology, 9,* 1112-1122. https://doi.org/10.1016/j.protcy.2013.12.124.

Tiete, J., Domíniguez, F., da Silva, B., Touhafi, A., & Steenhaut, K. (2017). MEMS microphones for wireless applications. In Uttamchandani, D. (Ed.), *Wireless MEMS Networks and Applications* (pp 177-195). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100449-4.00008-7.

Titze, I. R. (2008). Nonlinear source-filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America*, *123*(5), 2733–2749. https://doi.org/10.1121/1.2832337.

Titze, I. R., & Winholtz, W. S. (1993). Effect of Microphone Type and Placement on Voice Perturbation Measurements. *Journal of Speech and Hearing Research*, *36*(6), 1177–1190. https://doi.org/10.1044/jshr.3606.1177.

Van der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., Tully, M. P., Bozentko, K., Atwood, S., Hubbard, A., Wiper, C., Oswald, M., & Peek, N. (2021). Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. *Journal of the American Medical Informatics Association: JAMIA*, *28*(10), 2128–2138. https://doi.org/10.1093/jamia/ocab127.

Van Son, R., J.,H. (2002). *Can standard analysis tools be used on decompressed speech?* Paper presented at CoCOSDA 2002, Denver, CO.

Wang, B. and Hughes, V. (2022). Reducing uncertainty at the score-to-LR stage in likelihood ratio-based forensic voice comparison using automatic speaker recognition systems. *Proceedings of Interspeech*. Incheon: Korea.

Wang, B. X. and Hughes, V. (2021). System performance as a function of calibration methods, sample size and sampling variability in likelihood ratio-based forensic voice comparison. *Proceedings of Interspeech.* Brno, Czech Republic.

Wang, B. X., Hughes, V. and Foulkes, P. (2022). The effect of sampling variability on systems and individual speakers in likelihood ratio-based forensic voice comparison. *Speech Communication*, 138, 38–49. https://doi.org/10.1016/j.specom.2022.01.009

Yumoto, E., Sasaki, Y., & Okamura, H. (1984). Harmonics-to-Noise Ratio and Psychophysical Measurement of the Degree of Hoarseness. *Journal of Speech and Hearing Research*, *27*(1), 2–6. https://doi.org/10.1044/jshr.2701.02.

Yun, E. W.-T., Nguyen, D. D., Carding, P., Hodges, N. J., Chacon, A. M., & Madill, C. (2022). The Relationship Between Pitch Discrimination and Acoustic Voice Measures in a Cohort of Female Speakers. *Journal of Voice,* Online ahead of print. https://doi.org/10.1016/j.jvoice.2022.02.015.

Zarate, J. M., Tian, X., Woods, K. J. P., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*(1), 11475–11475. https://doi.org/10.1038/srep11475.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, *15*(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683.

Zhu, J., Sun, S., Liu, X., & Lei, B. (2009). Pitch in Speaker Recognition. *2009 Ninth International Conference on Hybrid Intelligent Systems*, *1*, 33–36. IEEE. https://doi.org/10.1109/HIS.2009.14.

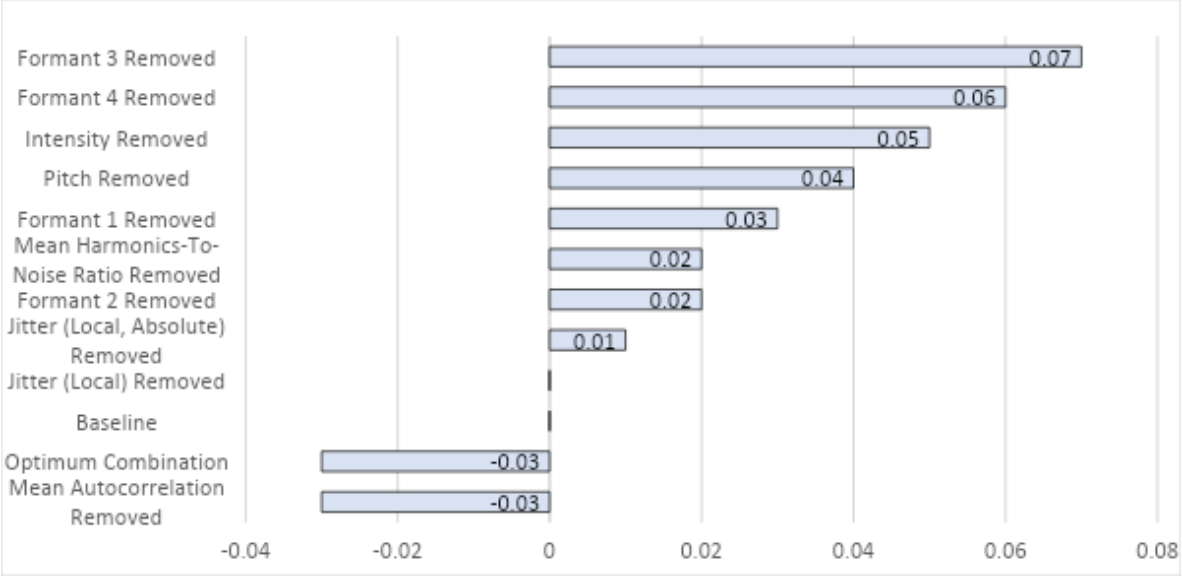
## Appendix A. Vowel Portfolios

### DyViS (TD) Portfolios

Turning first to Nolan et al.'s (2010) DyViS text-dependent sub-corpus, every segment and phoneme combined within this corpus generated a $C_{llr}$ value of 0.32. This improves upon the full combination of all data from all databases $C_{llr}$ value, as seen in the thesis above. This now serves as the baseline performance for this sociophonetic group: from here, the performance of each segment, and each feature on each segment, can be compared to this baseline to identify whether they improve performance through further tailoring of the combination of phonetic approaches.

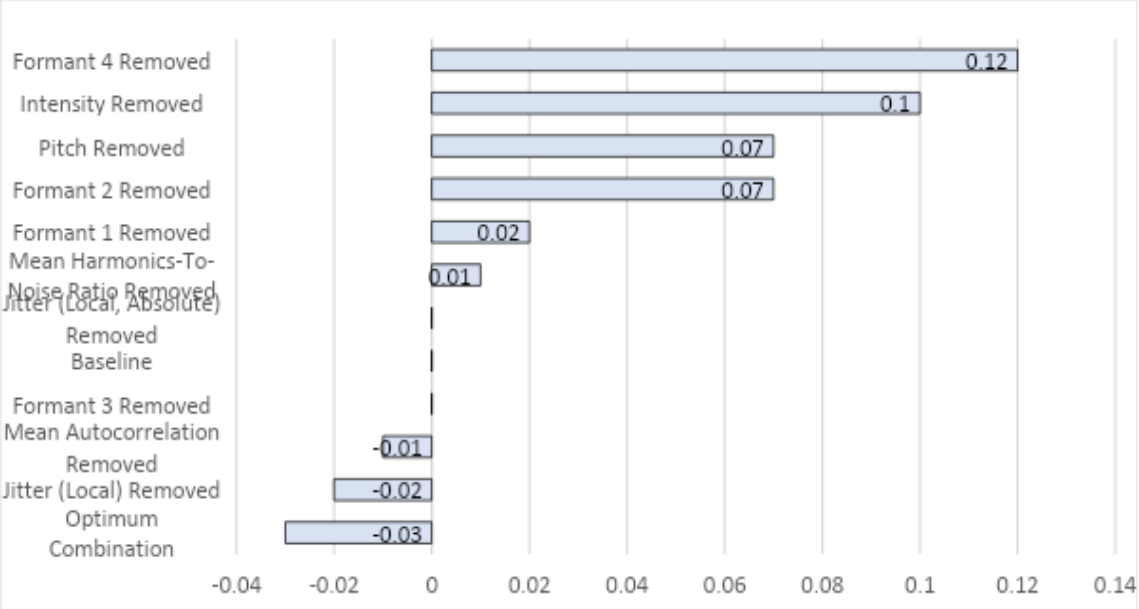
Each vowel will now be taken in turn to identify how effective each phonetic feature is when measured on that vowel before identifying which combination yields the best performance using $C_{llr}$ values. Starting with /ɑ/ (primary stress) below, the $C_{llr}$ value is 0.41 for this phoneme with all features combined together. This is worse than the baseline for this corpus. However, when mean autocorrelation is removed, this decreases this $C_{llr}$ value to 0.38. Thus, this feature should be removed to optimise performance; its inclusion makes performance worse. The other phonetic features, by contrast, increase this $C_{llr}$ value when removed. This shows that they are necessary; without them, performance worsens. These phonetic features are *f0*, intensity, formants 1-4, mean harmonics-to-noise ratio, jitter (local), and jitter (local, absolute). These are the features that will be combined in the optimised combination. The best condition for this phoneme, measured on this speaker and data group, is this optimised combination with its $C_{llr}$ value of 0.38. However, whilst this optimised phoneme combination outperforms the overall combinatory performance for this phoneme, it does not outperform the baseline $C_{llr}$ value for this specific sociophonetic group: 0.32.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɑ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



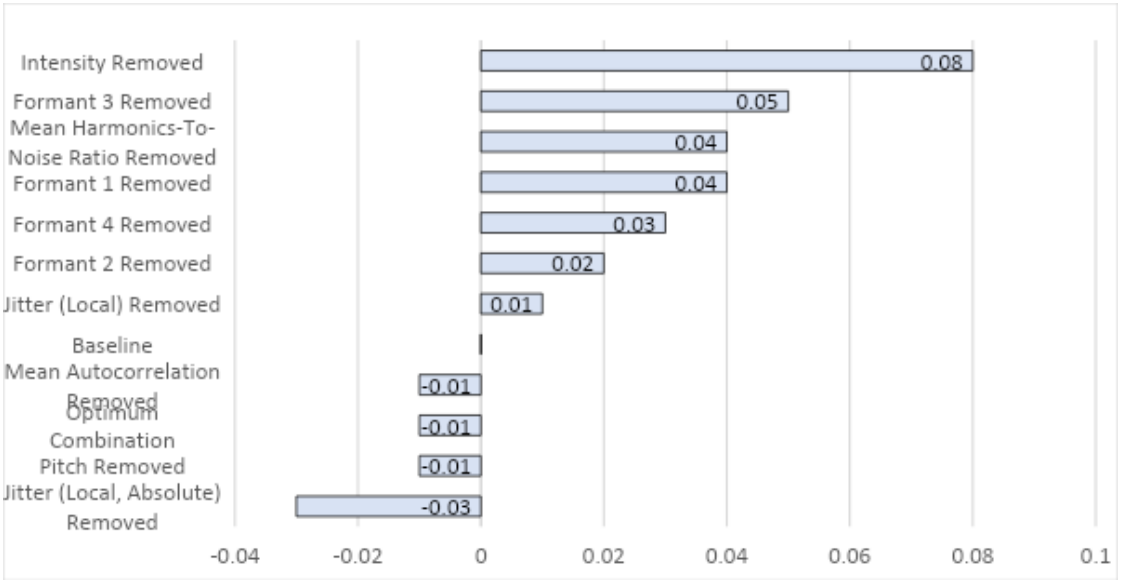

For /a/ (primary stress) below, the $C_{llr}$ value is 0.49 for this phoneme when all features are considered together. This again does not beat the baseline for this corpus overall. The phonetic features that should be removed are formant 3, mean autocorrelation, and jitter (local) and the phonetic features that should be retained are *f0*, intensity, formants 1-2 and 4, mean harmonics-to-noise ratio, and jitter (local, absolute). This optimised combination generates a $C_{llr}$ value of 0.46. The best condition for this phoneme, measured on this speaker and data group, is once again the optimised combination, but this again does not beat the overall baseline for this sociophonetic group.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /a/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



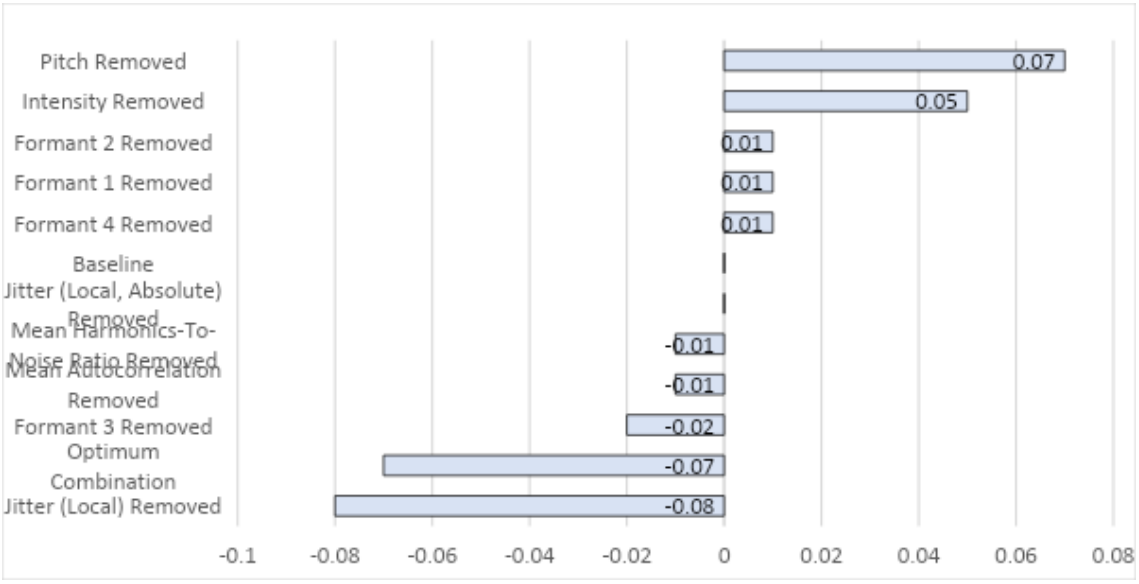

For /ə/ (no stress) below, the $C_{llr}$ value is 0.41 when all features are combined for this phoneme. This again does not beat the baseline $C_{llr}$ value for this corpus. The phonetic features that need removing are mean autocorrelation, *f0*, and jitter (local, absolute) and the phonetic features that need retaining are intensity, formants 1-4, mean harmonics-to-noise ratio, and jitter (local). This optimised combination generates a $C_{llr}$ value of 0.4. Whilst an improvement, the best profile for this phoneme, measured on this speaker and data group, is actually all features included but jitter (local, absolute) removed. This generated a $C_{llr}$ value of 0.38. Whilst this beats the optimised combination, it again does not beat the baseline $C_{llr}$ value for this sociophonetic group.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ə/ (no stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



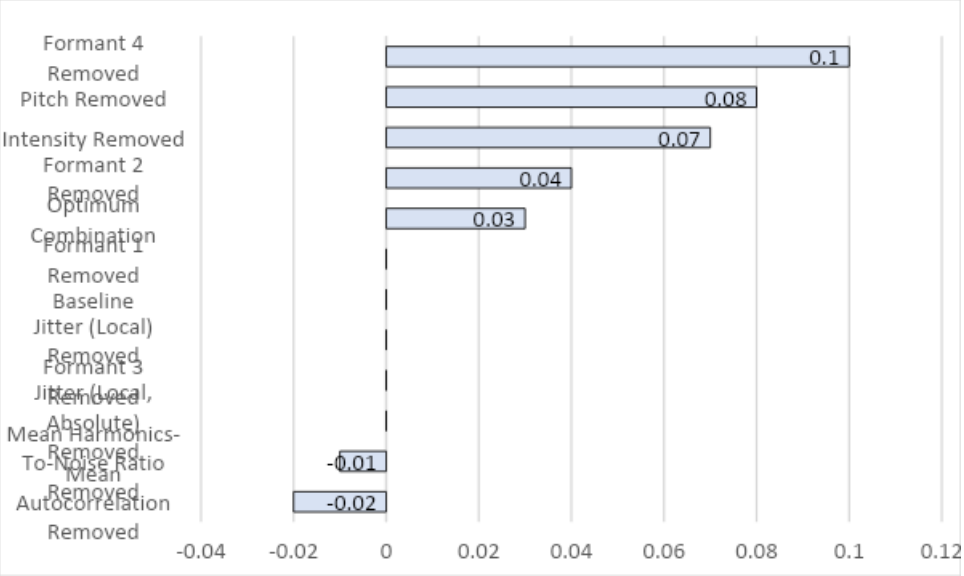

For /ʌ/ (primary stress) below, the $C_{llr}$ value is 0.59 for this phoneme when all features are combined, again not beating the baseline $C_{llr}$ value for this corpus. The phonetic features that need removing are jitter (local, absolute), mean harmonics-to-noise ratio, mean autocorrelation, formant 3, and jitter (local). The phonetic features that need retaining are *f0*, intensity, and formants 1-2 and 4. This optimised combination generates a lower $C_{llr}$ value of 0.52. The best condition for this phoneme, measured on this speaker and data group, is the sole removal of jitter (local) and the retention of all other features, however. This lowered the score to 0.51, improving upon the optimised combination, but this still does not outperform the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ʌ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



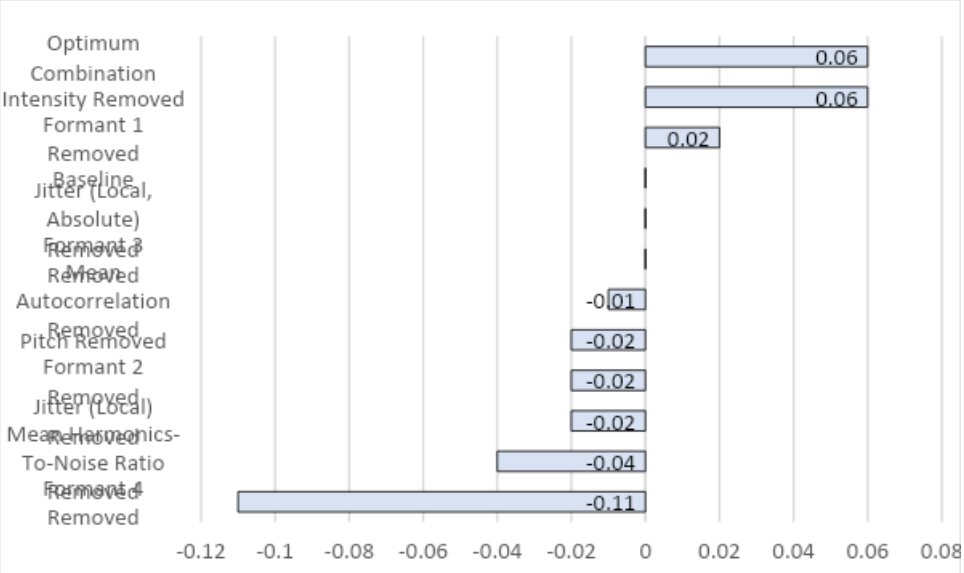

For /ɔ/ (primary stress) below, the $C_{llr}$ value is 0.55 for this phoneme when all features are combined. The phonetic features that need removing are jitter (local), formant 3, jitter (local, absolute), mean harmonics-to-noise ratio, and mean autocorrelation and the phonetic features that need retaining are *f0*, intensity, formants 1-2, and formant 4. However, this optimised combination generates a $C_{llr}$ value of 0.58, worsening performance. The best condition for this phoneme, measured on this speaker and data group, is the removal of only mean autocorrelation, however. This generates a $C_{llr}$ value of 0.53, which still does not outperform the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɔ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



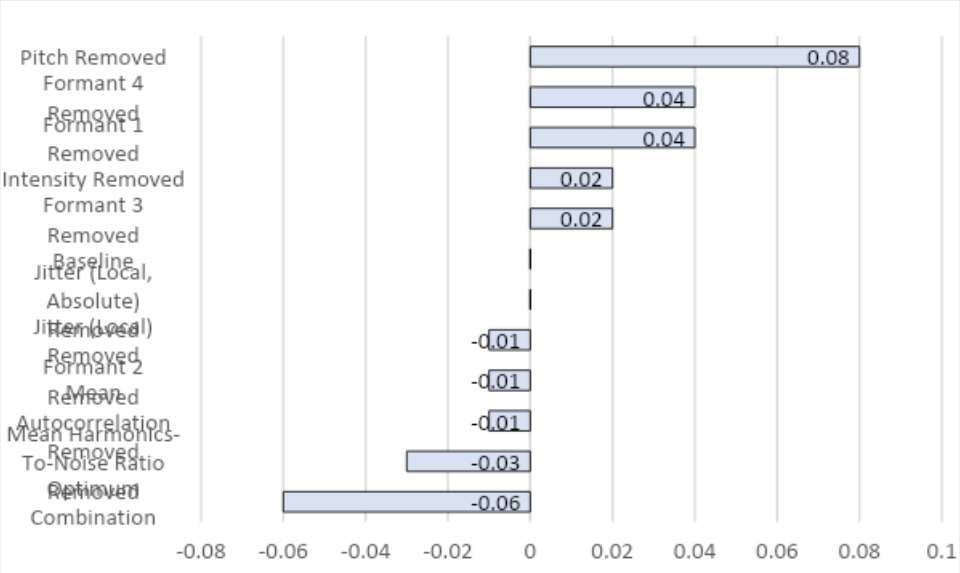

For /ɛ/ (primary stress) below, the $C_{llr}$ value is 0.7 for this phoneme when all features are combined. The phonetic features that need removing are jitter (local, absolute), formant 3, mean autocorrelation, *f0*, formant 2, jitter (local), mean harmonics-to-noise ratio, and formant 4 and the phonetic features that need retaining are intensity and formant 1. This optimised combination generates a score of 0.77, actually worsening performance. The best condition for this phoneme, measured on this speaker and data group, is the sole removal of formant 4, however. This generates a $C_{llr}$ value of 0.59, which still does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɛ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



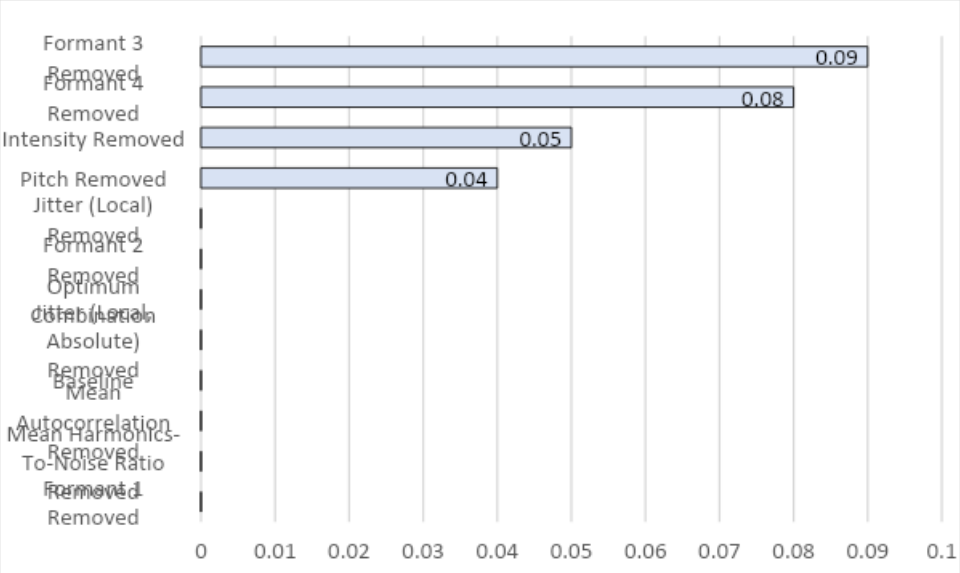

For /ɚ/ (no stress) below, the $C_{llr}$ value is 0.67 for this phoneme when all features are combined together. The phonetic features that need removing are jitter (local, absolute), jitter (local), formant 2, mean autocorrelation, and mean harmonics-to-noise ratio and the phonetic features that need retaining are formants 1 and 3-4, *f0*, and intensity. This optimised combination generates a $C_{llr}$ value of 0.61. The best condition for this phoneme, measured on this speaker and data group, is this optimised combination, but it again does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɚ/ (no stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



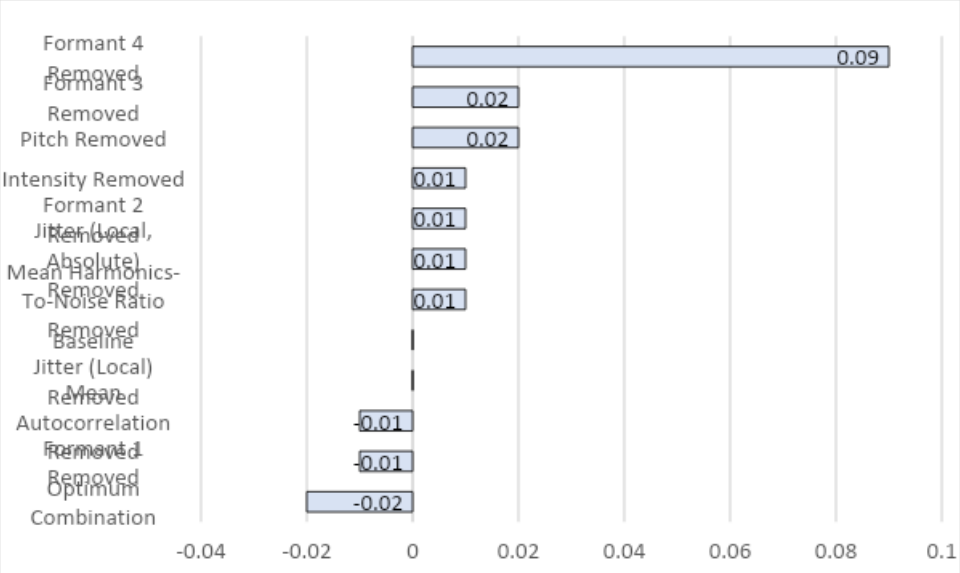

For /ɜ/ (primary stress) below, the $C_{llr}$ value is 0.49 for this phoneme when all features are combined together. The phonetic features that need removing are mean autocorrelation, mean harmonics-to-noise ratio, and formant 1 and the phonetic features that need retaining are *f0*, intensity, formants 2-4, jitter (local), and jitter (local, absolute). This optimised combination generates a $C_{llr}$ value of 0.49. The best condition for this phoneme, measured on this speaker and data group, is the sole removal of formant 1, however. This still generates a $C_{llr}$ value of 0.49, which still does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɜ˞/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



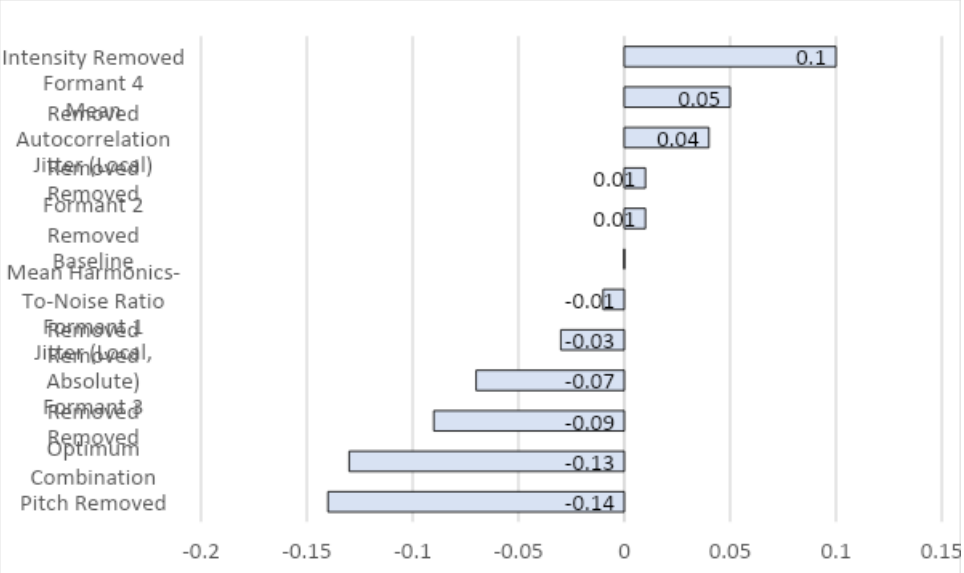

For /ɪ/ (no stress) below, the $C_{llr}$ value is 0.61 for this phoneme when all features are combined together. The phonetic features that need removing are jitter (local), mean autocorrelation, and formant 1 and the phonetic features that need retaining are *f0*, intensity, formants 2-4, mean harmonics-to-noise ratio, and jitter (local, absolute). This optimised combination generates a $C_{llr}$ value of 0.59. The best condition for this phoneme, measured on this speaker and data group, is the optimised combination, but it again does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɪ/ (no stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



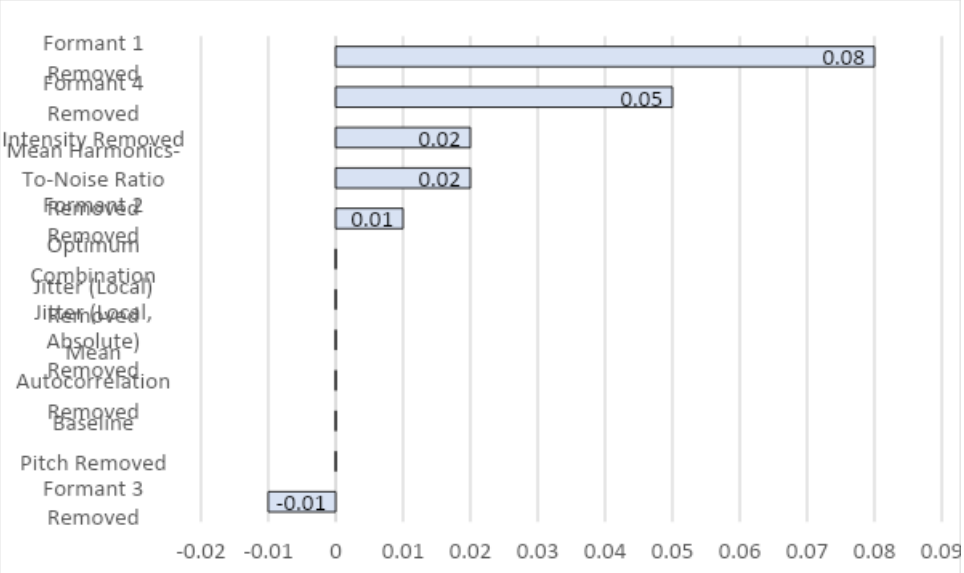

For /ɪ/ (primary stress) below, the $C_{llr}$ value is 0.75 for this phoneme when all features are combined together. The phonetic features that need removing are mean harmonics-to-noise ratio, formant 1, jitter (local, absolute), formant 3, *f0*, and the phonetic features that need retaining are formant 2, jitter (local), mean autocorrelation, formant 4, and intensity. This optimised combination generates a $C_{llr}$ value of 0.62. The best condition for this phoneme, measured on this speaker and data group, is the sole removal of *f0*, however. This generates a score of 0.61, again not beating the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɪ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



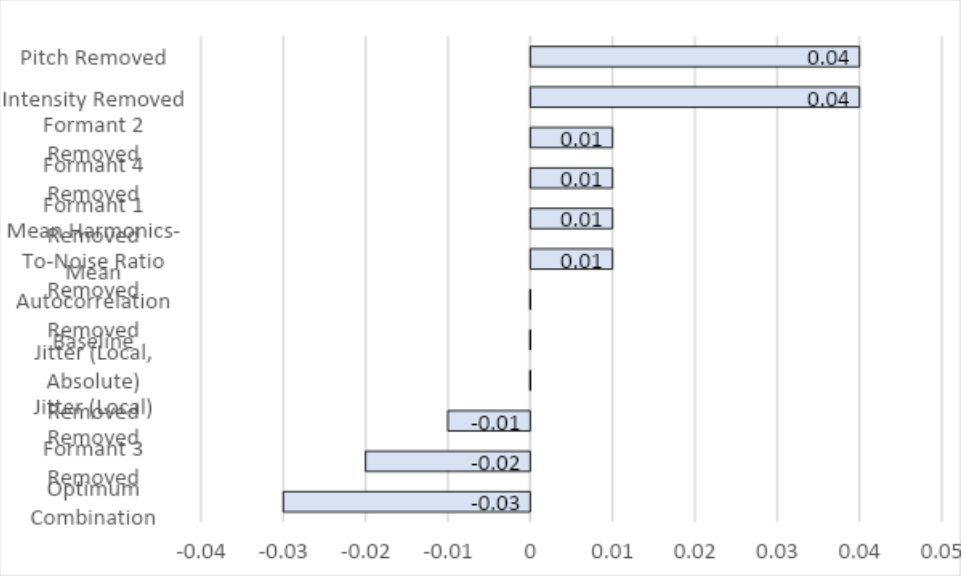

For /i/ (no stress) below, the $C_{llr}$ value is 0.63 for this phoneme when all features are combined together. The phonetic features that need removing are *f0* and formant 3 and the phonetic features that need retaining are mean autocorrelation, jitter (local, absolute), jitter (local), formants 1-2 and 4, mean harmonics-to-noise ratio, and intensity. This optimised combination generates a $C_{llr}$ value of 0.64. The best condition for this phoneme, measured on this speaker and data group, is the sole removal of formant 3, however. This generates a score of 0.62, which again does not outperform the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /i/ (no stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



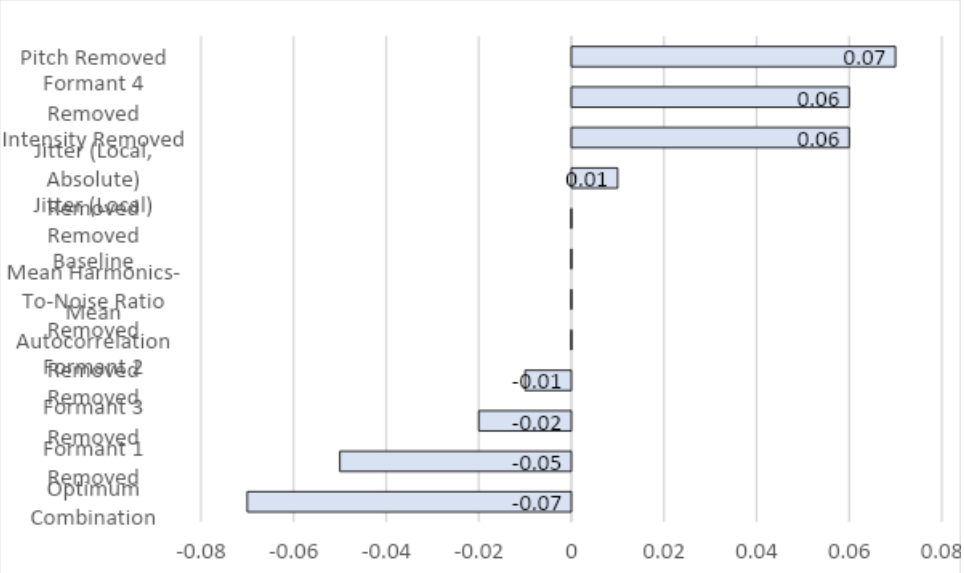

For /i/ (primary stress) below, the $C_{llr}$ value is 0.55 for this phoneme when all features are combined together. The phonetic features that need removing are jitter (local), jitter (local, absolute), and formant 3 and the phonetic features that need retaining are mean autocorrelation, mean harmonics-to-noise ratio, formants 1-2 and 4, intensity, and *f0*. This optimised combination generates a $C_{llr}$ value of 0.52. The best condition for this phoneme, measured on this speaker and data group, is the optimised combination, but it again does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /i/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



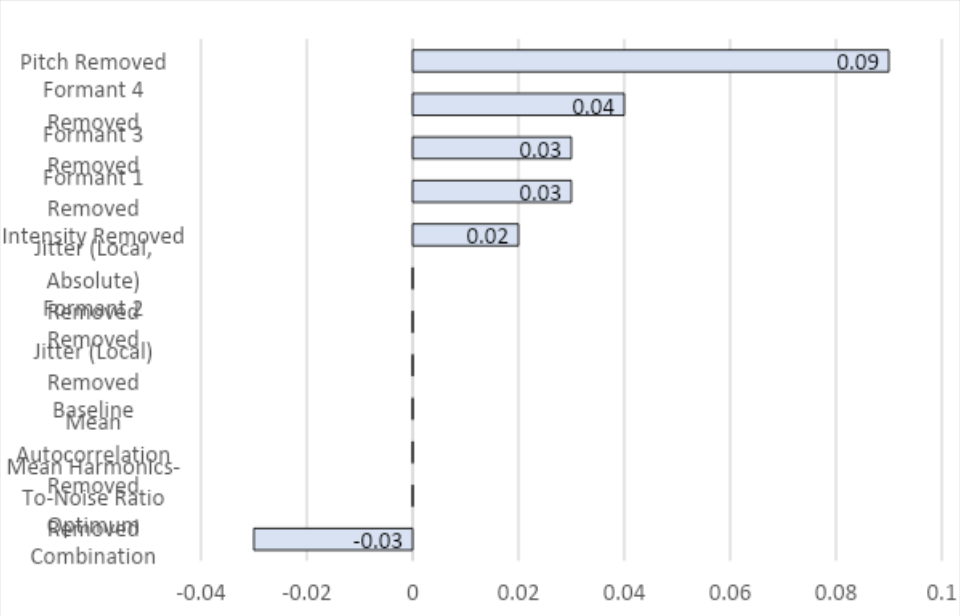

For /ʊ/ (primary stress) below, the $C_{llr}$ value is 0.81 for this phoneme when all features are combined together. The phonetic features that need removing are formants 1-3, mean autocorrelation, and mean harmonics-to-noise ratio and the phonetic features that need retaining are jitter (local), jitter (local, absolute), intensity, formant 4, and *f0*. This optimised combination generates a score of 0.74. The best condition for this phoneme, measured on this speaker and data group, is the optimised combination, but it does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ʊ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**




For /u/ (primary stress) below, the $C_{llr}$ value is 0.73 for this phoneme when all features are combined together. The phonetic features that need removing are mean autocorrelation and mean harmonics-to-noise ratio and the phonetic features that need retaining are jitter (local), jitter (local, absolute), intensity, formants 1-4, and *f0*. This optimised combination generates a score of 0.7. The best condition for this phoneme, measured on this speaker and data group, is the optimised combination, but it again does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /u/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Dependent Sub-Corpus**



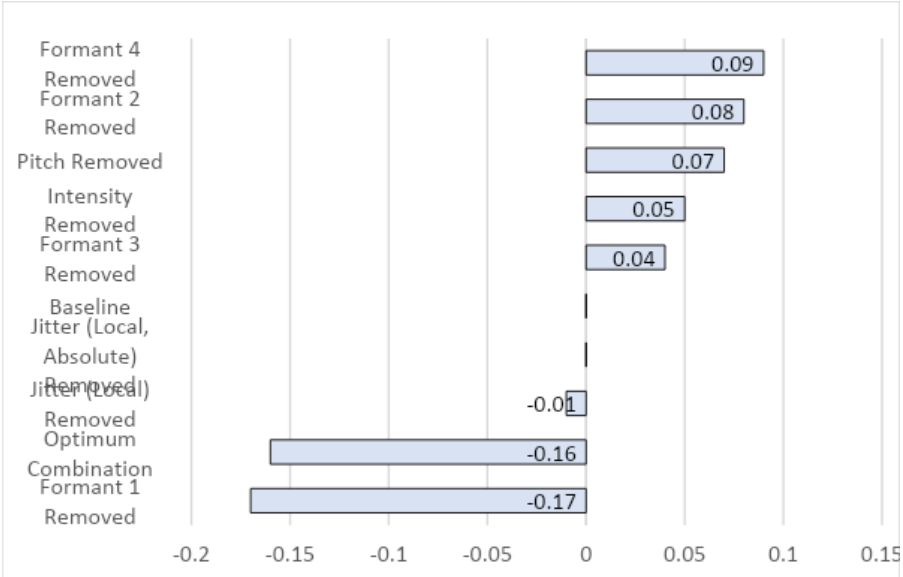

For this database, there is variation in the portfolios which will be explained in the following chapter, but there are some important trends to flag here. The sociophonetically-specific baseline combination for this corpus always outperforms any vowel-specific combination. However, on these vowels, the optimised combination regularly represents the best possible combination. Finally, the features that are most frequently trimmed are the mean autocorrelation and mean harmonics-to-noise ratio measurements and the included jitter measurements.

**DyViS (TI) Portfolios**

Turning now to Nolan et al.'s (2010) DyViS text-independent sub-corpus, the baseline $C_{llr}$ value with all features and segments considered is 0.15. This improves upon the full combination $C_{llr}$ value, as seen in the prior chapter.

Starting with phoneme /ɑ/ (primary stress) below for the phoneme-specific investigations, the $C_{llr}$ value is 0.61 for this phoneme when all features are considered on this phoneme. The phonetic features that need removing are jitter (local), jitter (local, absolute), and formant 1 and the phonetic features that need retaining are *f0*, intensity, and formants 2-4. This optimised combination generates a $C_{llr}$ value of 0.45, but the best condition for this phoneme, measured on this speaker and data group, is the sole removal of formant 1. This generates a $C_{llr}$ value of 0.44, but this does not beat the baseline $C_{llr}$ value for this database.

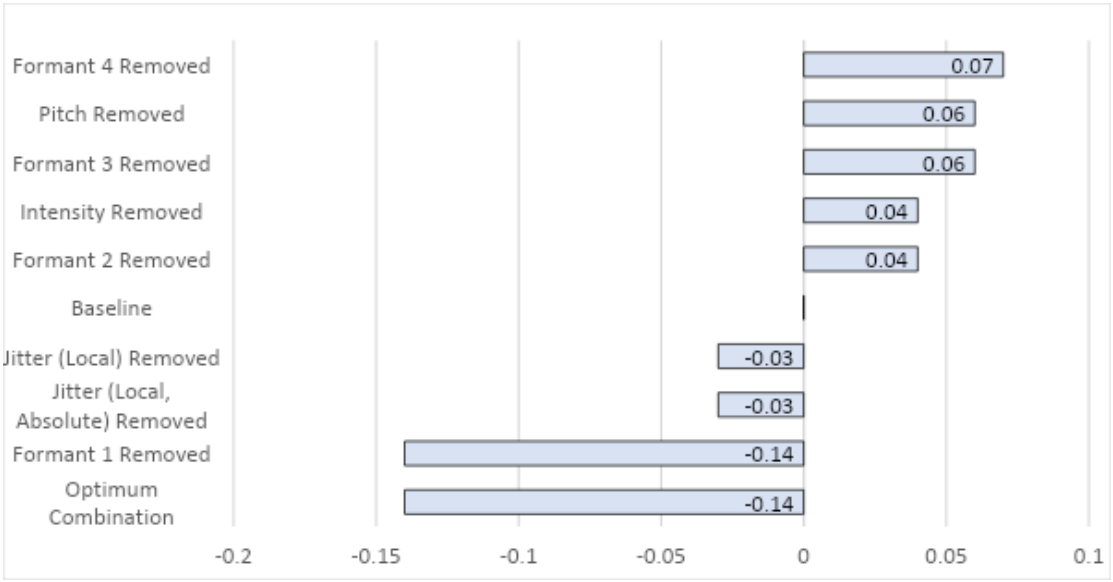
**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɑ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**




For /a/ (primary stress) below, the all-features $C_{llr}$ value is 0.48 for this phoneme. The phonetic features that need removing are jitter (local), jitter (local, absolute), and formant 1 and the phonetic features that need retaining are *f0*, intensity, and formants 2-4. This

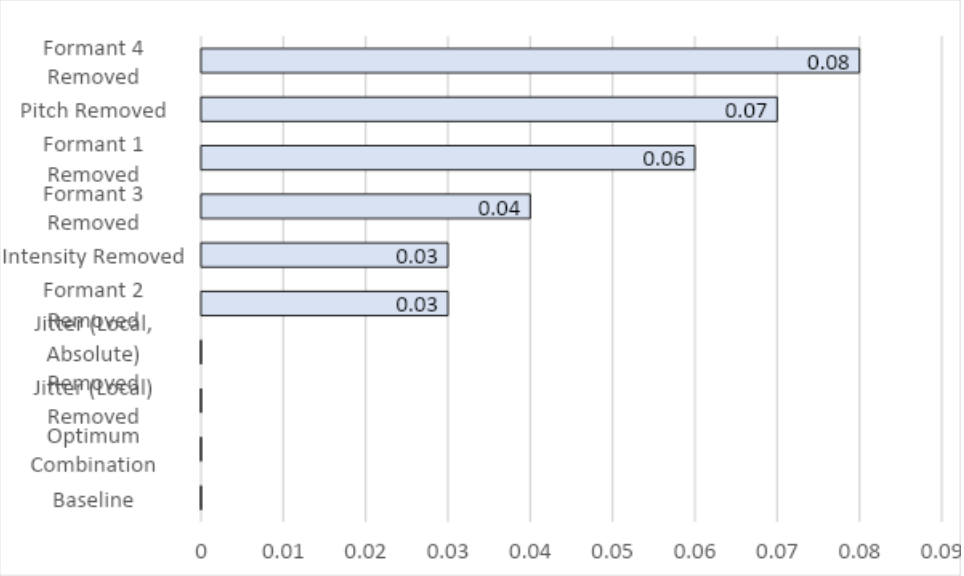
optimised combination generates a $C_{llr}$ value of 0.34 and the best condition for this phoneme, measured on this speaker and data group, is the optimised combination. This does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /a/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



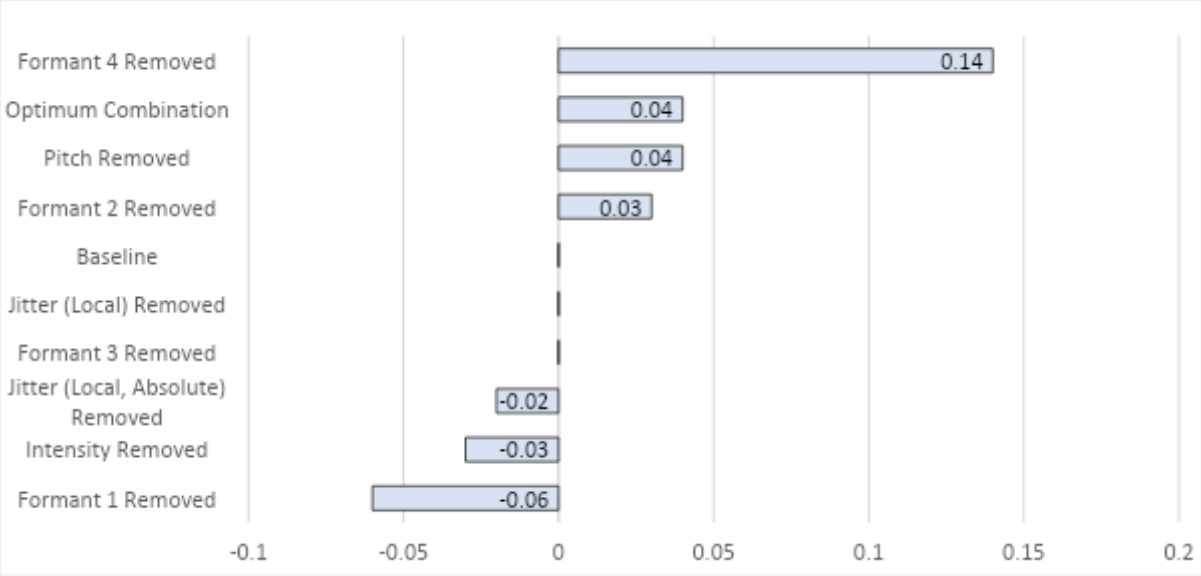

For /ə/ (no stress) below, the all-features $C_{llr}$ value is 0.23 for this phoneme. As seen below, this was the optimised combination; removing any features worsened performance, so all are necessary for best performance here. This was also the best individual vowel performance seen in any database in this thesis, but it still did not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ə/ (no stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



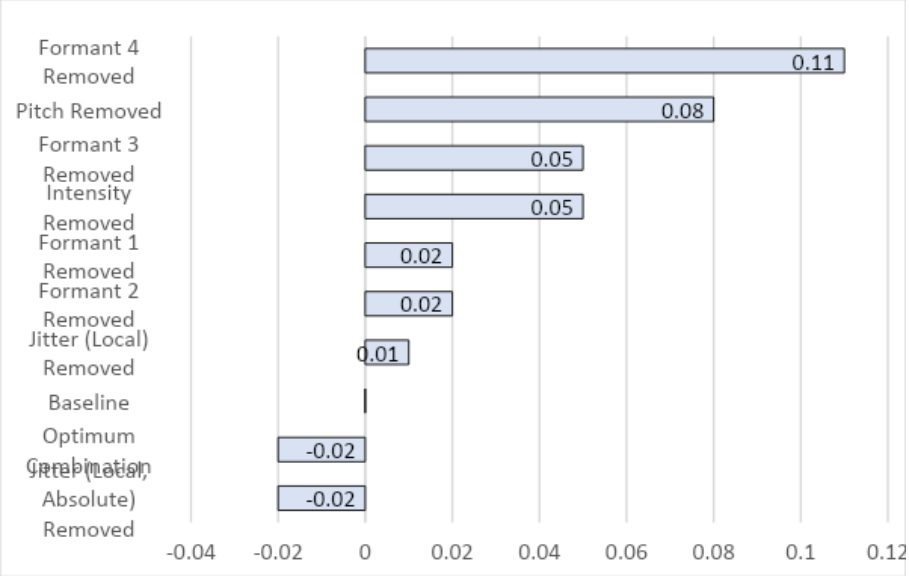

For /ʌ/ (primary stress) below, the all-features $C_{llr}$ value is 0.52 for this phoneme. The phonetic features that need removing are jitter (local), jitter (local, absolute), intensity, formant 1, and formant 3 and the phonetic features that need retaining are *f0*, formant 2, and formant 4. This optimised combination generates a $C_{llr}$ value of 0.56, worsening performance. The best condition for this phoneme, measured on this speaker and data group, is the sole removal of formant 1 and the retention of all other features. This generates a $C_{llr}$ value of 0.46, but this again does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ʌ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



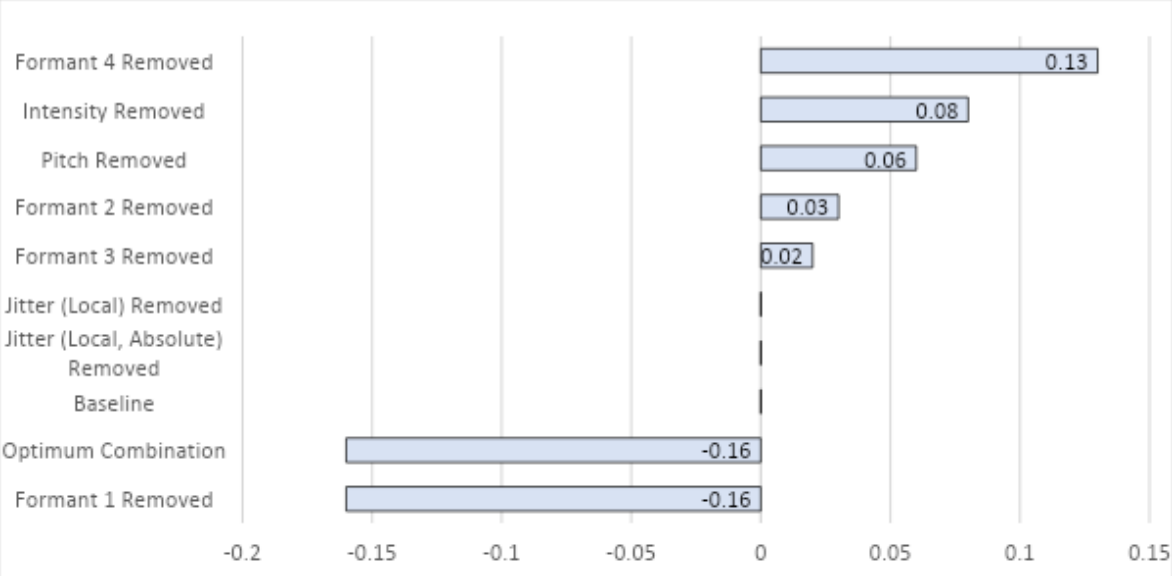

For /ɔ/ (primary stress) below, the all-features $C_{llr}$ value is 0.52 for this phoneme. The phonetic features that need removing are jitter (local, absolute) and the phonetic features that need retaining are *f0*, intensity, formants 1-4, and jitter (local). This optimised combination generates a $C_{llr}$ value of 0.5. The best condition for this phoneme, measured on this speaker and data group, is this optimised combination as only the singular feature detrimental to performance was cut, but it did not beat the baseline $C_{llr}$ value for this database.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɔ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



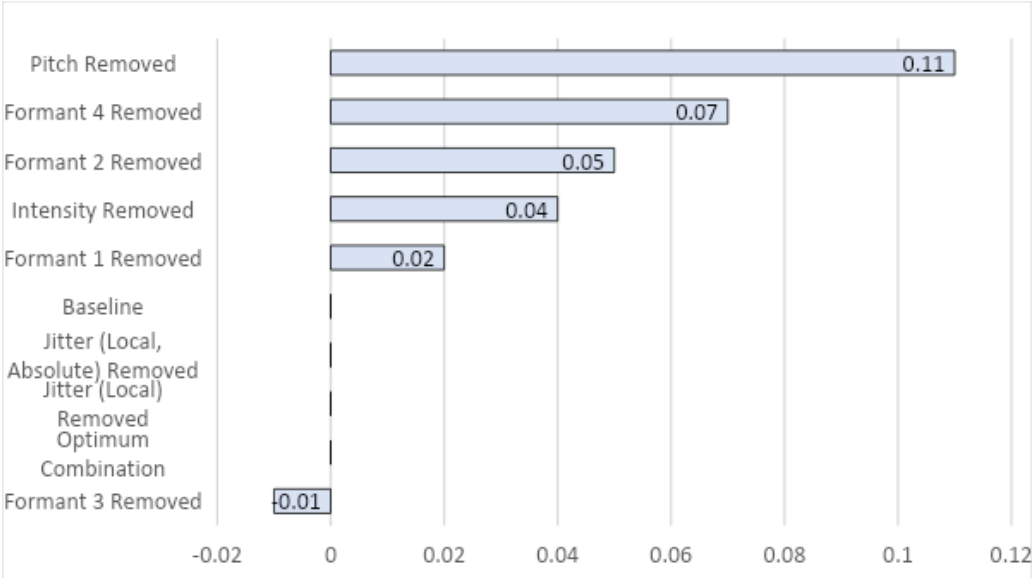

For /ɛ/ (primary stress) below, the all-features $C_{llr}$ value is 0.45 for this phoneme. The phonetic features that need removing are formant 1 and the phonetic features that need retaining are *f0*, intensity, formants 2-4, jitter (local), and jitter (local, absolute). This optimised combination generates a $C_{llr}$ value of 0.29, and the best condition for this phoneme, measured on this speaker and data group, is this optimised combination as only the singular feature detrimental to performance was cut. This still does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɛ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



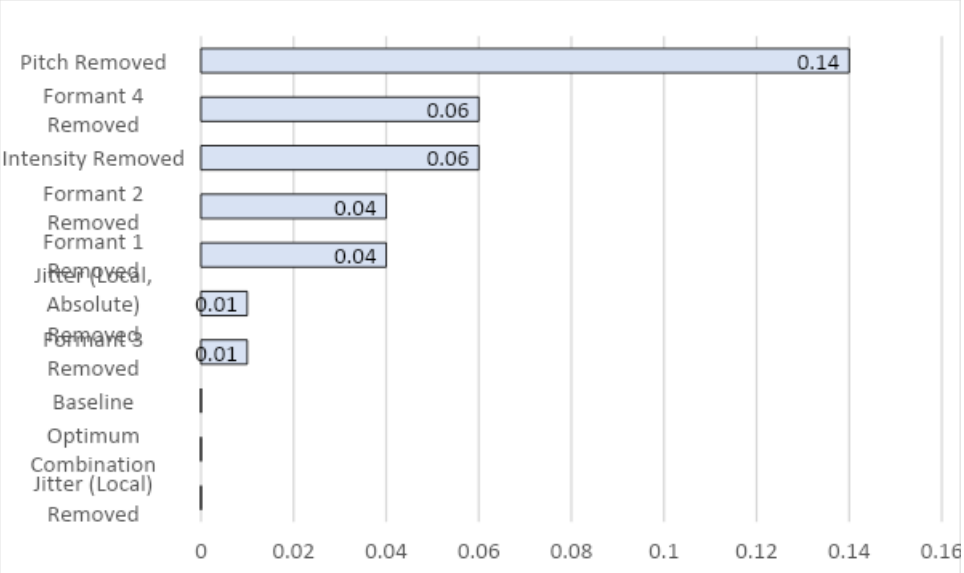

For /ɚ/ (no stress) below, the all-features $C_{llr}$ value is 0.44 for this phoneme. The phonetic features that need removing are formant 3, jitter (local), and jitter (local, absolute) and the phonetic features that need retaining are *f0*, intensity, formants 1-2, and formant 4. This optimised combination still generates a $C_{llr}$ value of 0.44, so the best condition for this phoneme, measured on this speaker and data group, is the sole removal of formant 3. This generates a score of 0.43, but this still does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɚ/ (no stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



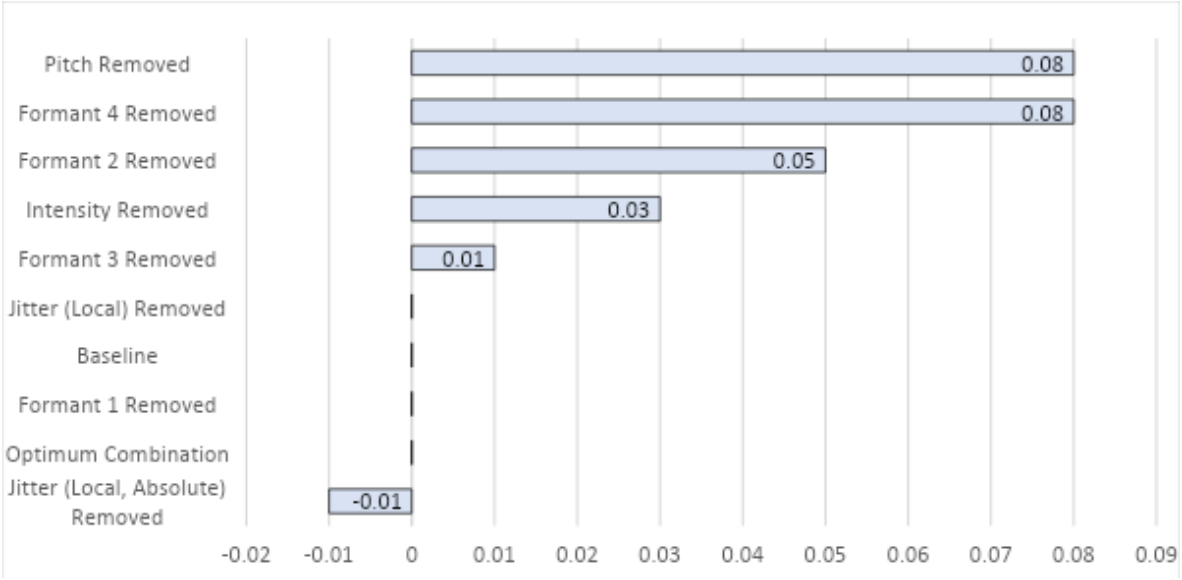

For /ɝ/ (primary stress) below, the all-features $C_{llr}$ value is 0.45 for this phoneme. The phonetic features that need removing are jitter (local) and the phonetic features that need retaining are *f0*, intensity, formants 1-4, and jitter (local, absolute). This optimised combination still generates a $C_{llr}$ value of 0.45 and the best condition for this phoneme, measured on this speaker and data group, is this optimised combination as only the singular feature detrimental to performance was cut. This still does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɜ˞/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



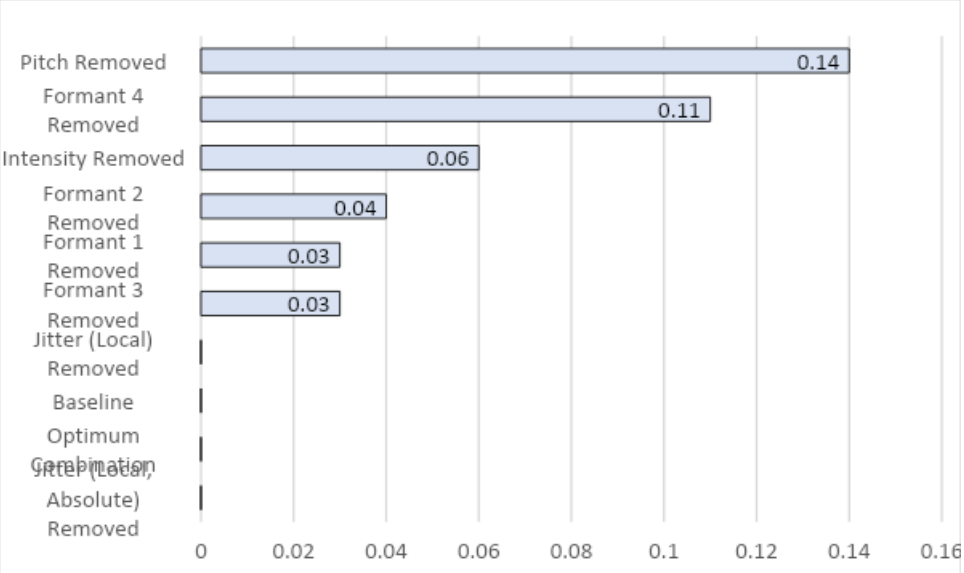

For /ɪ/ (no stress) below, the all-features $C_{llr}$ value is 0.43 for this phoneme. The phonetic features that need removing are formant 1 and jitter (local, absolute) and the phonetic features that need retaining are *f0*, intensity, formants 2-4, and jitter (local). This optimised combination still generates a $C_{llr}$ value of 0.43 but the best condition for this phoneme, measured on this speaker and data group, is the sole removal of jitter (local, absolute). This generates a score of 0.42, but this still does not beat the baseline $C_{llr}$ value for this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɪ/ (no stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



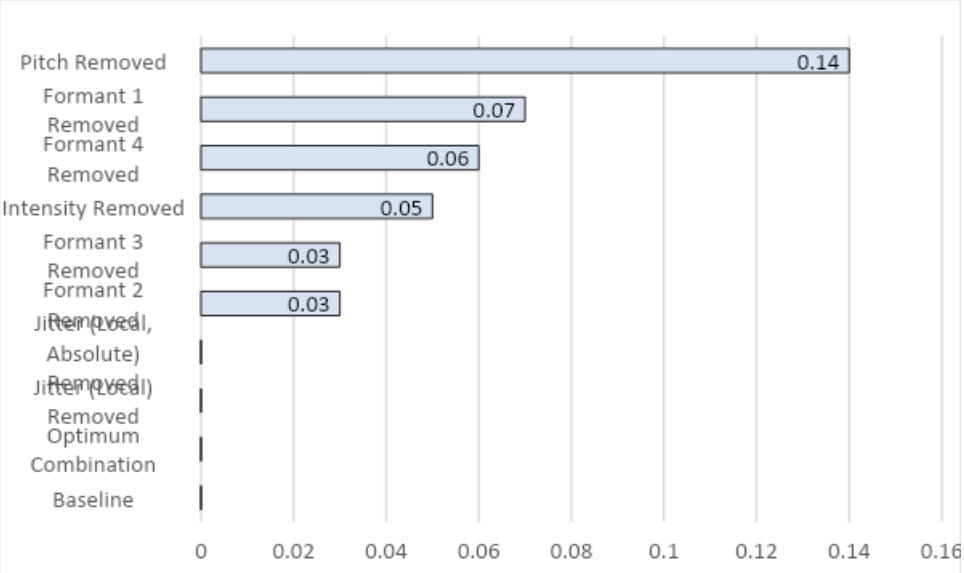

For /ɪ/ (primary stress) below, the all-features $C_{llr}$ value is 0.37 for this phoneme. The phonetic features that need removing are jitter (local, absolute) and the phonetic features that need retaining are *f0*, intensity, formants 1-4, and jitter (local). This optimised combination still generates a $C_{llr}$ value of 0.37 and the best condition for this phoneme, measured on this speaker and data group, is this optimised combination as only the singular feature detrimental to performance was cut. This still does not beat the baseline $C_{llr}$ value for this database.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɪ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



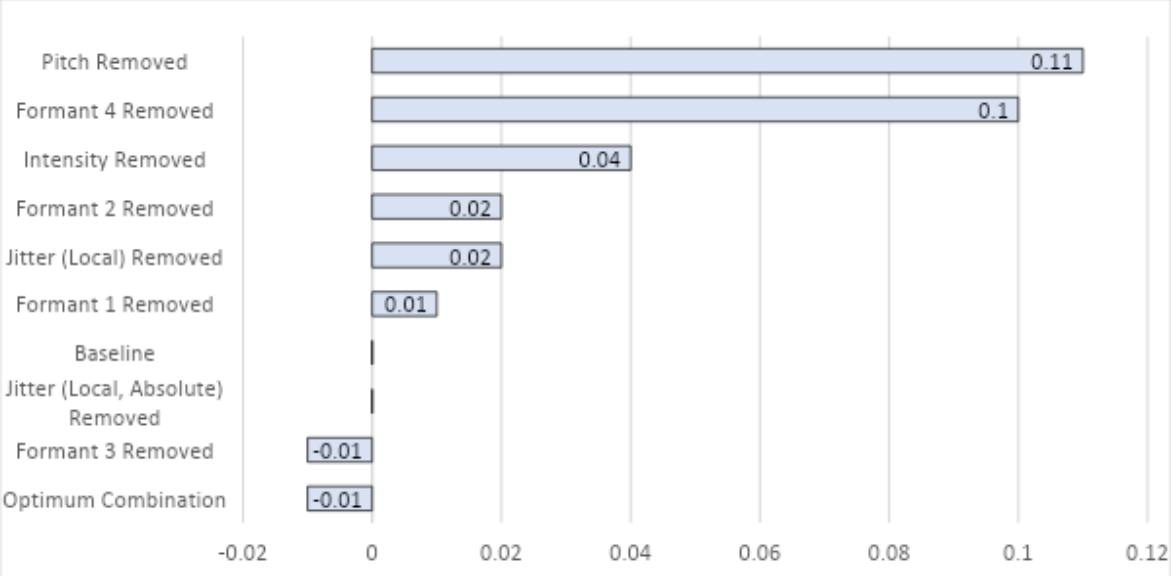

For /i/ (no stress) below, the all-features $C_{llr}$ value is 0.4 for this phoneme. As seen below, this was the optimised combination; removing any features worsened performance, so all are necessary for best performance here. This did not, however, beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /i/ (no stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



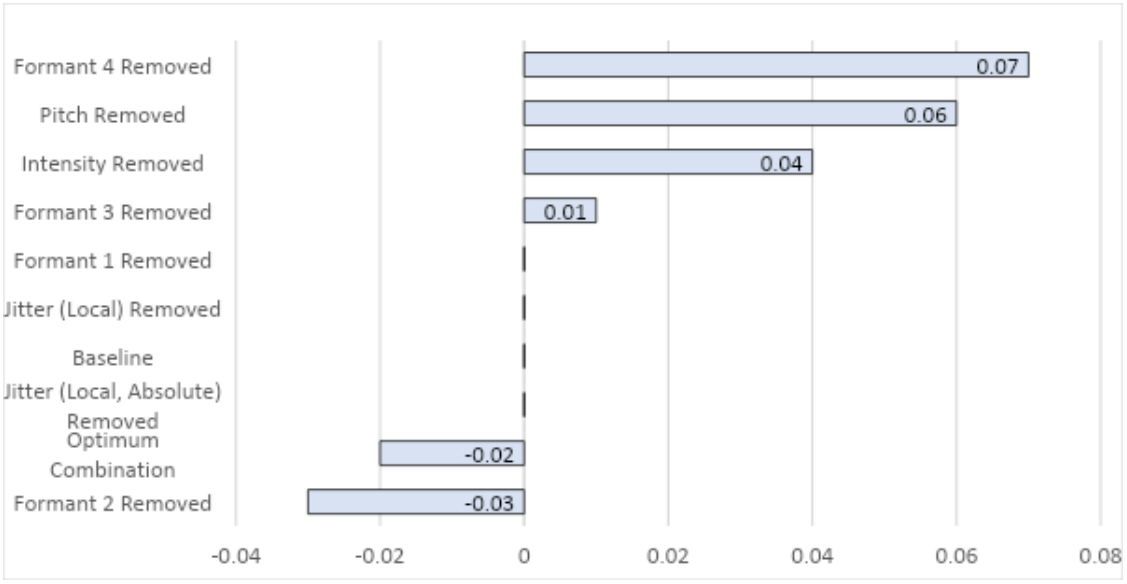

For /i/ (primary stress) below, the all-features $C_{llr}$ value is 0.43 for this phoneme. The phonetic features that need removing are jitter (local, absolute) and formant 3 and the phonetic features that need retaining are formants 1-2 and 4, intensity, *f0*, and jitter (local). This optimised combination generates a $C_{llr}$ value of 0.42 and the best condition for this phoneme, measured on this speaker and data group, is the optimised combination. This still does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /i/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**



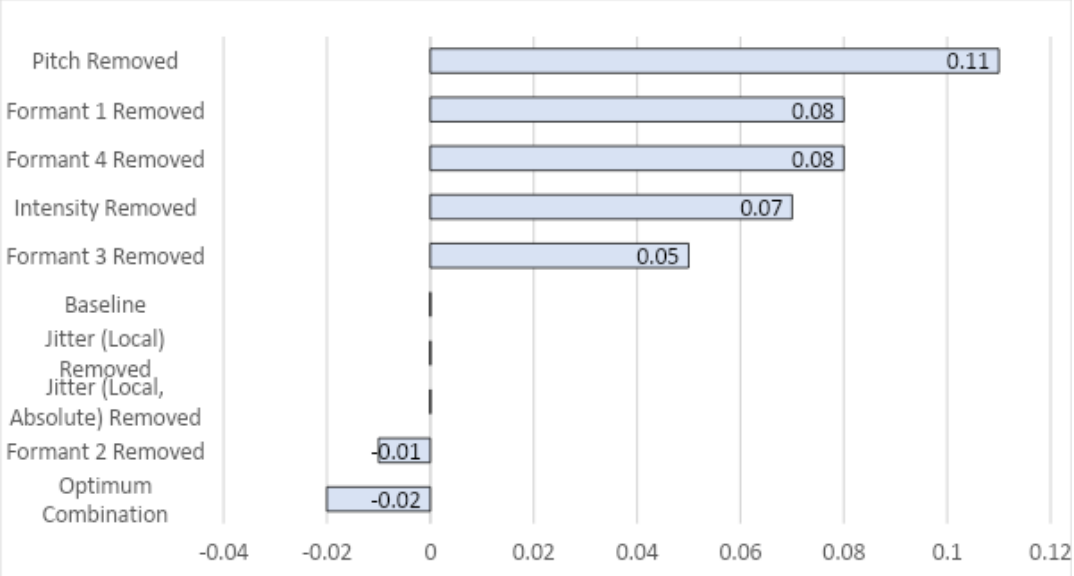

For /ʊ/ (primary stress) below, the all-features $C_{llr}$ value is 0.8 for this phoneme. The phonetic features that need removing are formant 2 and jitter (local, absolute) and the phonetic features that need retaining are *f0*, intensity, formant 1, formants 3-4, and jitter (local). This optimised combination generates a $C_{llr}$ value of 0.78 but the best condition for this phoneme, measured on this speaker and data group, is the sole removal of formant 2. This generates a score of 0.77 which still does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ʊ/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**




For /u/ (primary stress) below, the all-features $C_{llr}$ value is 0.49 for this phoneme. The phonetic features that need removing are jitter (local), jitter (local, absolute), and formant 2 and the phonetic features that need retaining are *f0*, intensity, formant 1, and formants 3-4. This optimised combination generates a score of 0.47 and the best condition for this phoneme, measured on this speaker and data group, is the optimised combination. This still does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /u/ (primary stress) in Nolan et al.'s (2010) DyViS Text-Independent Sub-Corpus**

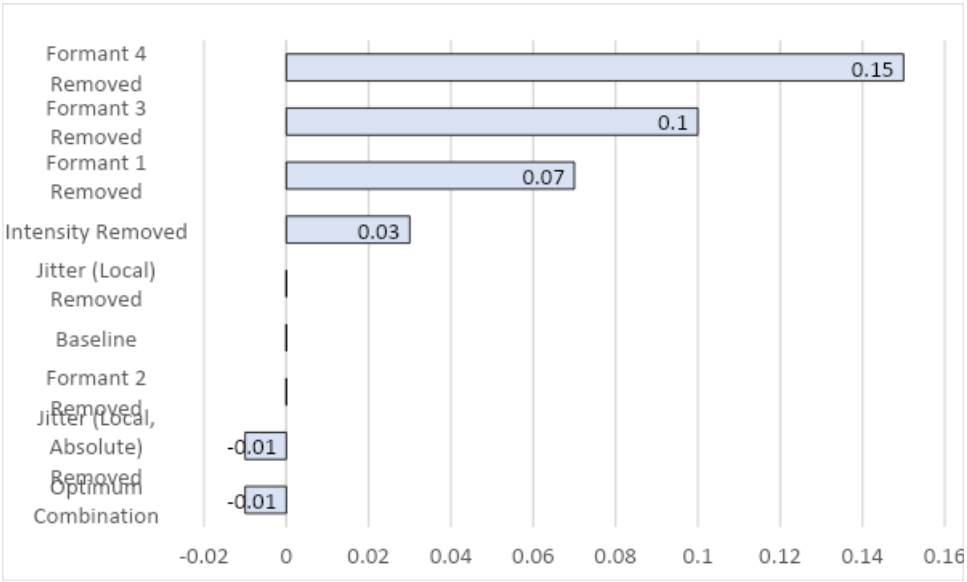



There are, again, trends to spot in this database. The baseline $C_{llr}$ value for this corpus is again the best performance seen, the optimised combination is again used frequently, and the jitter measurements are once again cut regularly. Additionally, formant 1 is now regularly trimmed.

**WYRED (TI) Portfolios**

Turning finally to Gold et al.'s (2018) WYRED text-independent sub-corpus, the baseline $C_{llr}$ value combining every feature and segment is 0.21 and improves upon the full combination $C_{llr}$ value. Starting with phoneme /ɑ/ (primary stress) below, the all-features $C_{llr}$ value is 0.52 for this phoneme. The phonetic features that need removing are jitter (local, absolute), and formant 2 and the phonetic features that need retaining are intensity, formants 1 and 3-4, and jitter (local, absolute). This optimised combination generates a $C_{llr}$ value of 0.51 and the best condition for this phoneme, measured on this speaker and data group, is the optimised combination. This still does not beat the baseline $C_{llr}$ value in this corpus.

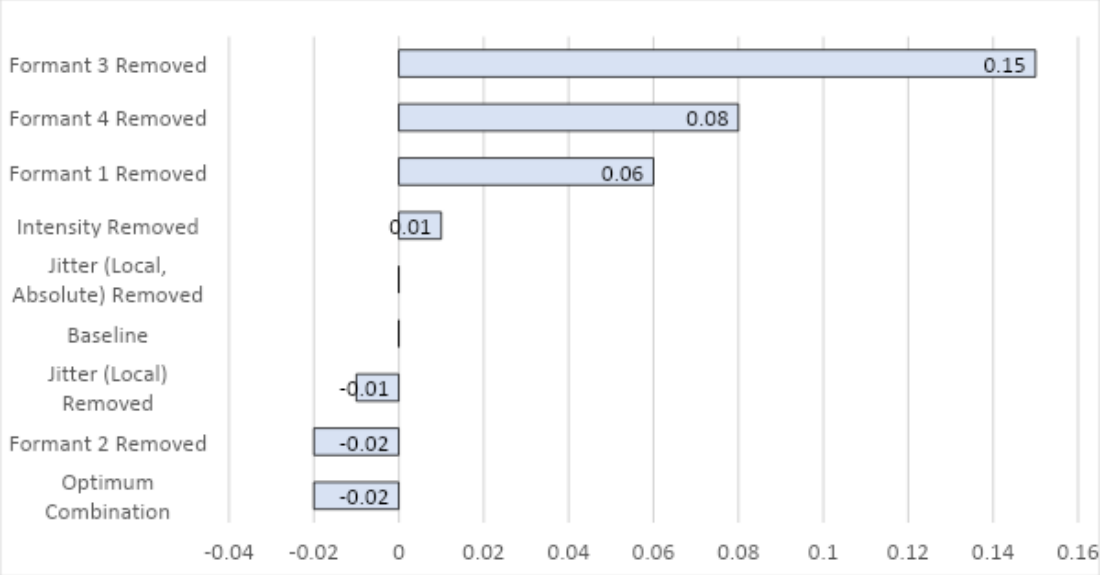
**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɑ/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**




For /a/ (primary stress) below, the all-features $C_{llr}$ value is 0.38 for this phoneme. The phonetic features that need removing are jitter (local) and formant 2 and the phonetic features that need retaining are intensity, jitter (local, absolute), and formants 1 and 3-4. This optimised combination generates a score of 0.36 and the best condition for this phoneme,

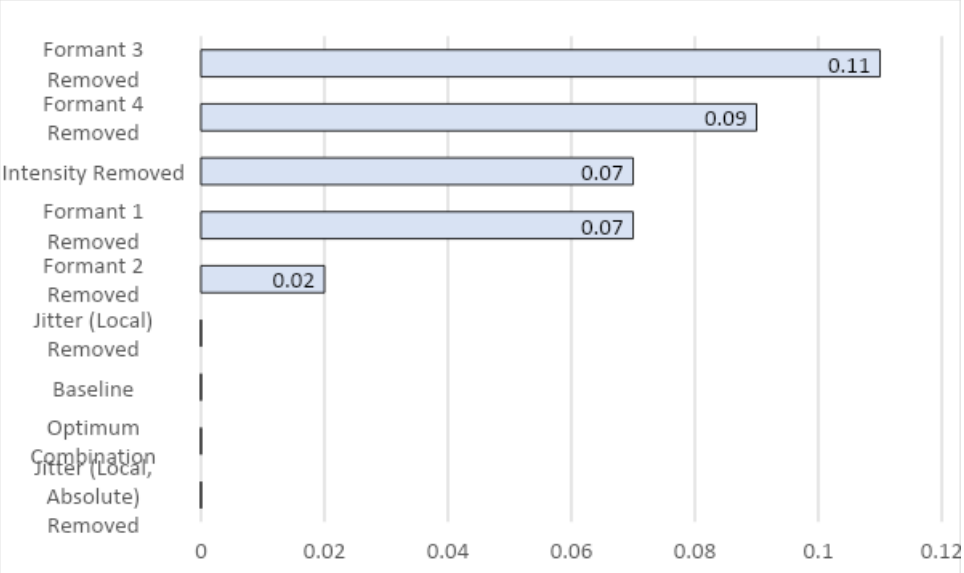
measured on this speaker and data group, is the optimised combination. This again does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /a/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



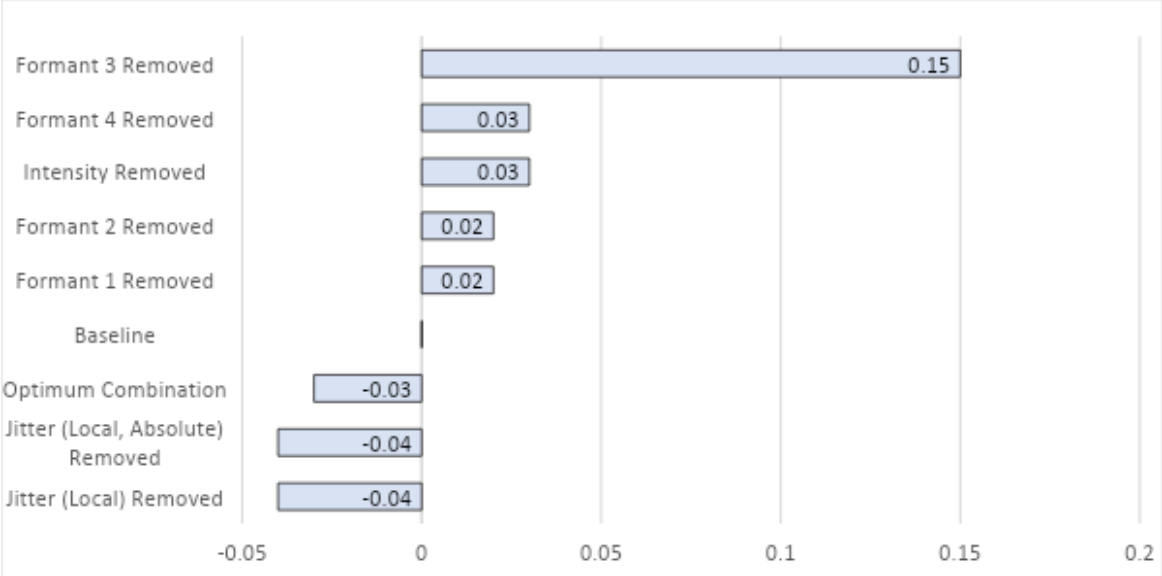

For /ə/ (no stress) below, the all-features $C_{llr}$ value is 0.39 for this phoneme. The phonetic features that need removing are jitter (local, absolute) and the phonetic features that need retaining are intensity, formants 1-4, and jitter (local). This optimised combination still generates a $C_{llr}$ value of 0.39 and the best condition for this phoneme, measured on this speaker and data group, is this optimised combination as only the singular feature detrimental to performance was cut. This again does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ə/ (no stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



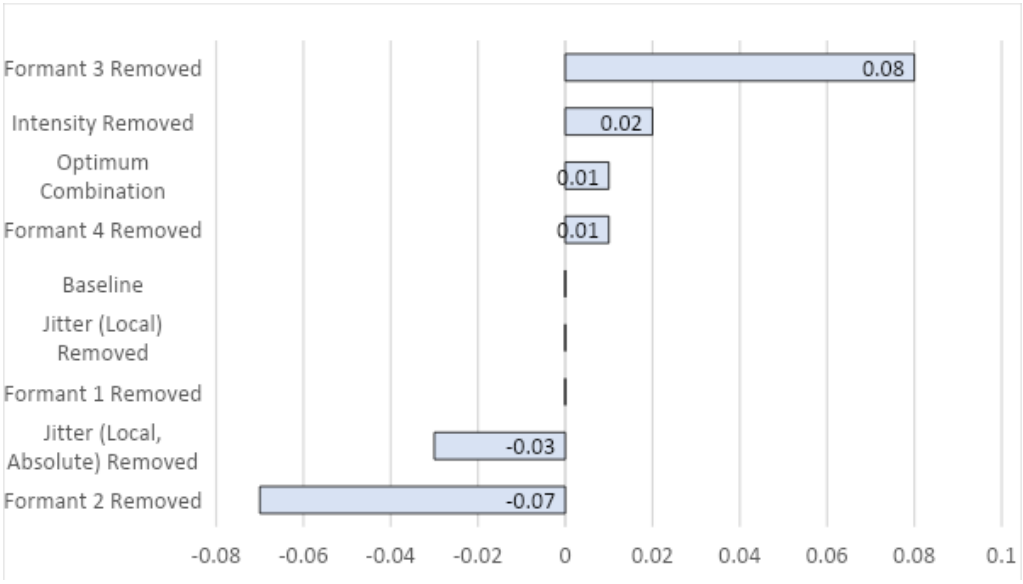

For /ʌ/ (primary stress) below, the all-features $C_{llr}$ value is 0.55 for this phoneme. The phonetic features that need removing are jitter (local) and jitter (local, absolute) and the phonetic features that need retaining are intensity and formants 1-4. This optimised combination generates a $C_{llr}$ value of 0.52 but the best condition for this phoneme, measured on this speaker and data group, is the removal of jitter (local) and the retention of all other features. This generates a score of 0.51 which still does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ʌ/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



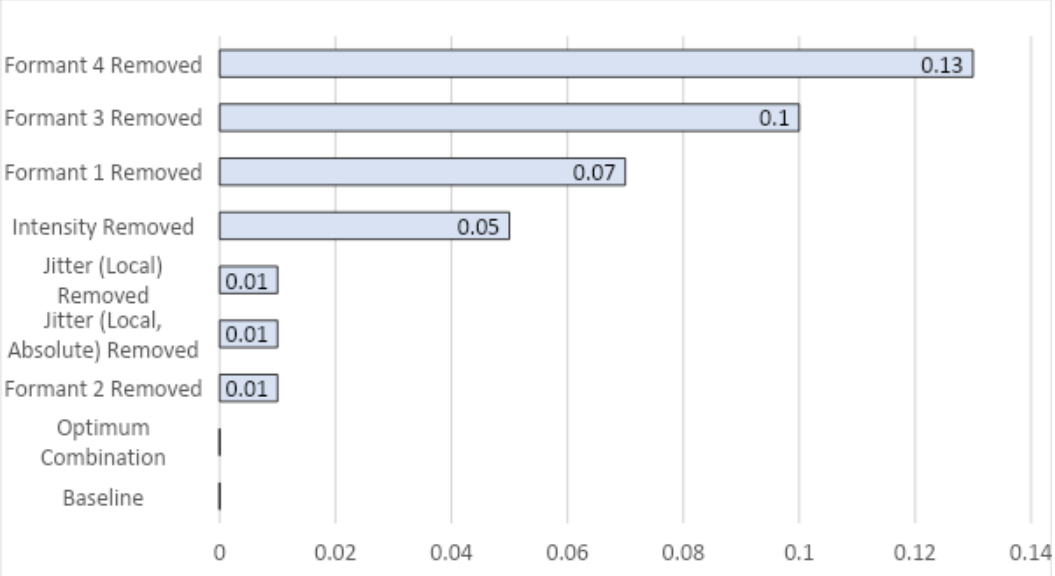

For /ɔ/ (primary stress) below, the all-features $C_{llr}$ value is 0.69 for this phoneme. The phonetic features that need removing are formants 1-2, jitter (local), and jitter (local, absolute) and the phonetic features that need retaining are intensity and formants 3-4. This optimised combination generates a $C_{llr}$ value of 0.7, worsening performance. The best condition for this phoneme, measured on this speaker and data group, is this sole removal of formant 2. This generates a score of 0.62 which again does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɔ/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



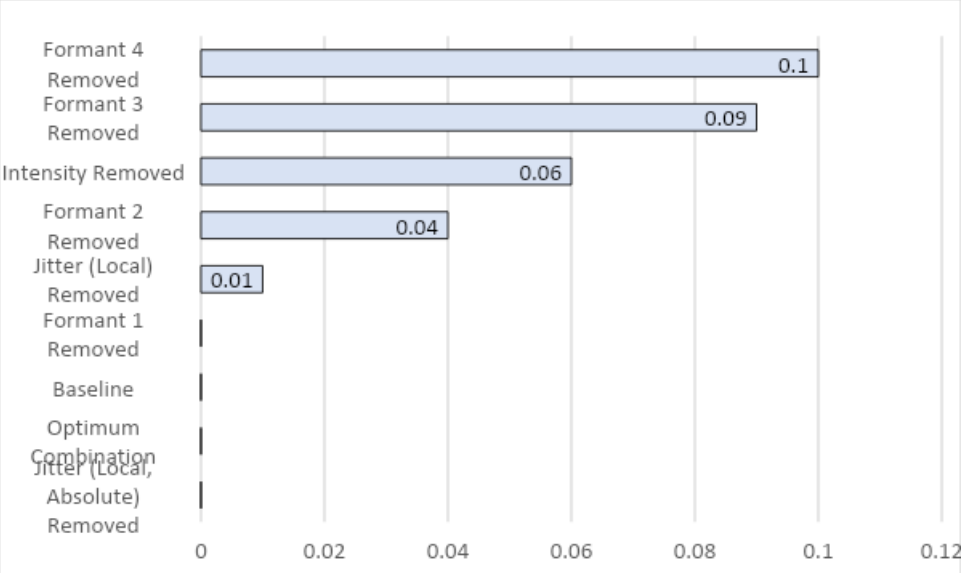

For /ɛ/ (primary stress) below, the all-features $C_{llr}$ value is 0.33 for this phoneme. As seen below, this was the optimised combination; removing any features worsened performance, so all are necessary for best performance here. This does not, however, beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɛ/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



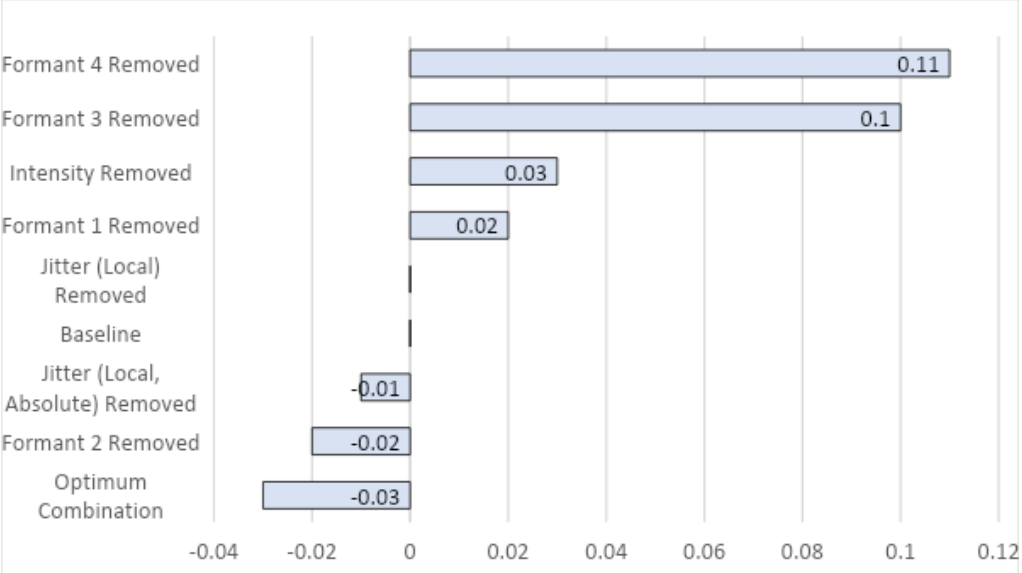

For /ɚ/ (no stress) below, the all-features $C_{llr}$ value is 0.53 for this phoneme. The phonetic features that need removing are jitter (local, absolute) and the phonetic features that need retaining are intensity, formants 1-4, and jitter (local). This optimised combination still generates a $C_{llr}$ value of 0.53 and the best condition for this phoneme, measured on this speaker and data group, is this optimised combination as only the singular feature detrimental to performance was cut. This again does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɚ/ (no stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



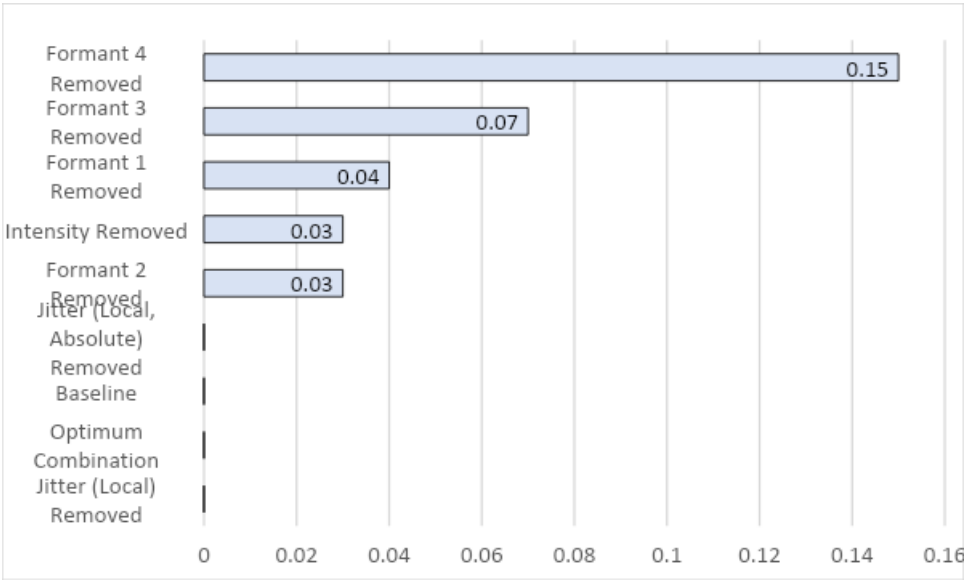

For /ɝ/ (primary stress) below, the all-features $C_{llr}$ value is 0.69 for this phoneme. The phonetic features that need removing are jitter (local, absolute) and formant 2 and the phonetic features that need retaining are intensity, formants 1 and 3-4, and jitter (local). This optimised combination generates a $C_{llr}$ value of 0.66 and the best condition for this phoneme, measured on this speaker and data group, is this optimised combination. This generates a $C_{llr}$ value of 0.66 but does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɜ˞/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



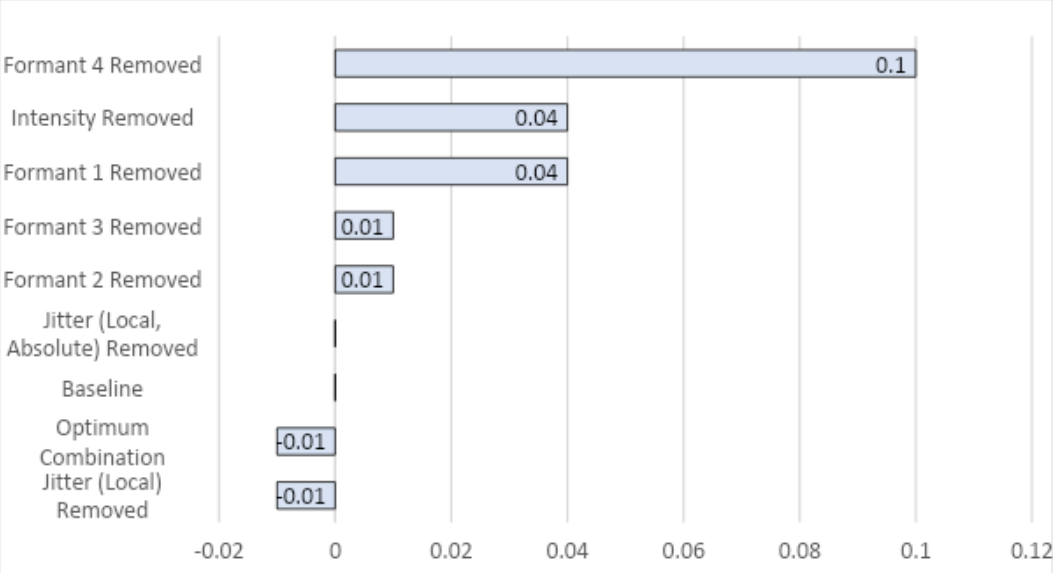

For /ɪ/ (no stress) below, the all-features $C_{llr}$ value is 0.51 for this phoneme. The phonetic features that need removing are jitter (local) and the phonetic features that need retaining are intensity, formants 1-4, and jitter (local, absolute). This optimised combination still generates a $C_{llr}$ value of 0.51 and the best condition for this phoneme, measured on this speaker and data group, is this optimised combination as only the singular feature detrimental to performance was cut. This still does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɪ/ (no stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



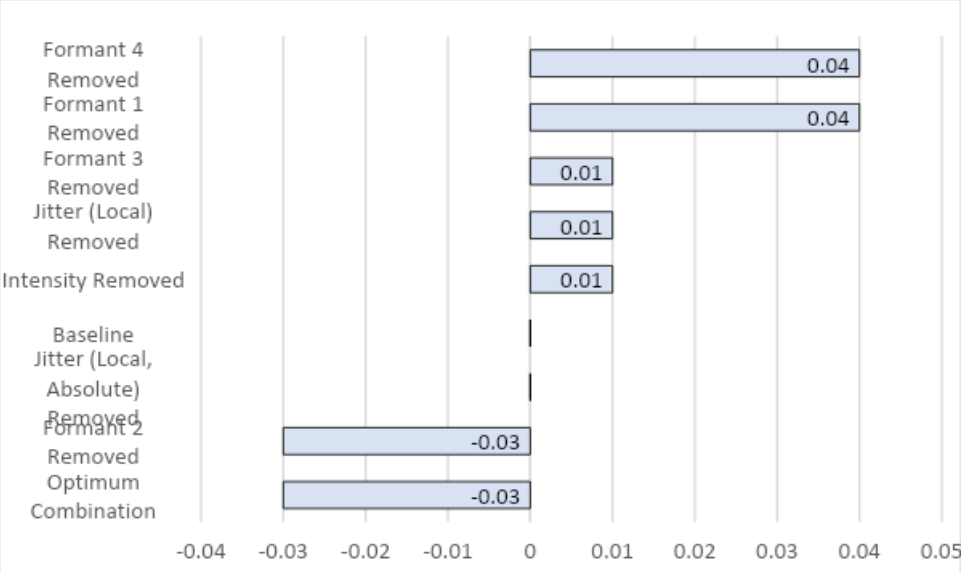

For /ɪ/ (primary stress) below, the all-features $C_{llr}$ value is 0.59 in this corpus. The phonetic features that need removing are jitter (local) and the phonetic features that need retaining are intensity, formants 1-4, and jitter (local, absolute). This optimised combination generates a $C_{llr}$ value of 0.58 and the best condition for this phoneme, measured on this speaker and data group, is this optimised combination as only the singular feature detrimental to performance was cut. This still does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ɪ/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



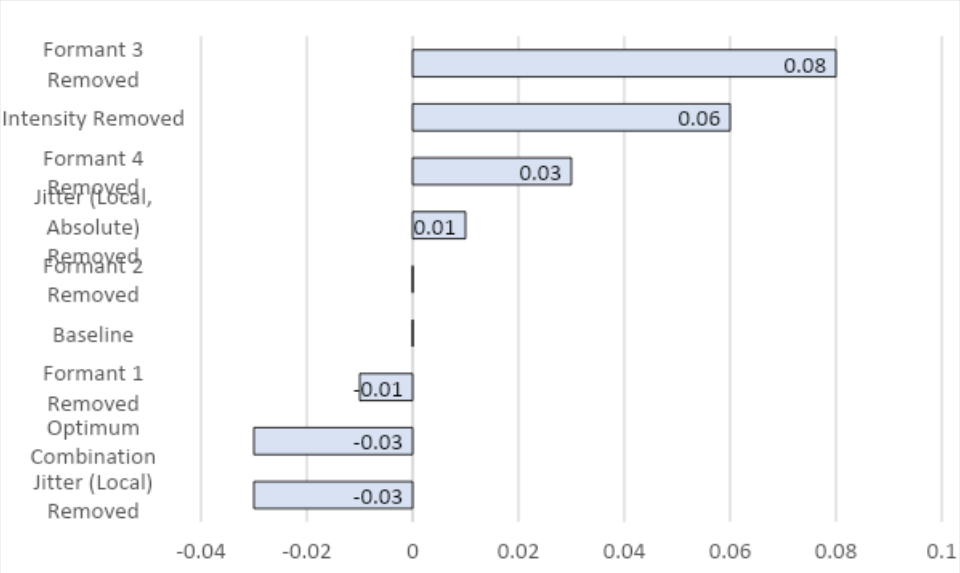

For /i/ (no stress) below, the all-features $C_{llr}$ value is 0.72 for this phoneme. The phonetic features that need removing are jitter (local, absolute) and formant 2 and the phonetic features that need retaining are formants 1 and 3-4, intensity, and jitter (local). This optimised combination generates a $C_{llr}$ value of 0.69 and the best condition for this phoneme, measured on this speaker and data group, is the optimised combination. This still does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /i/ (no stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



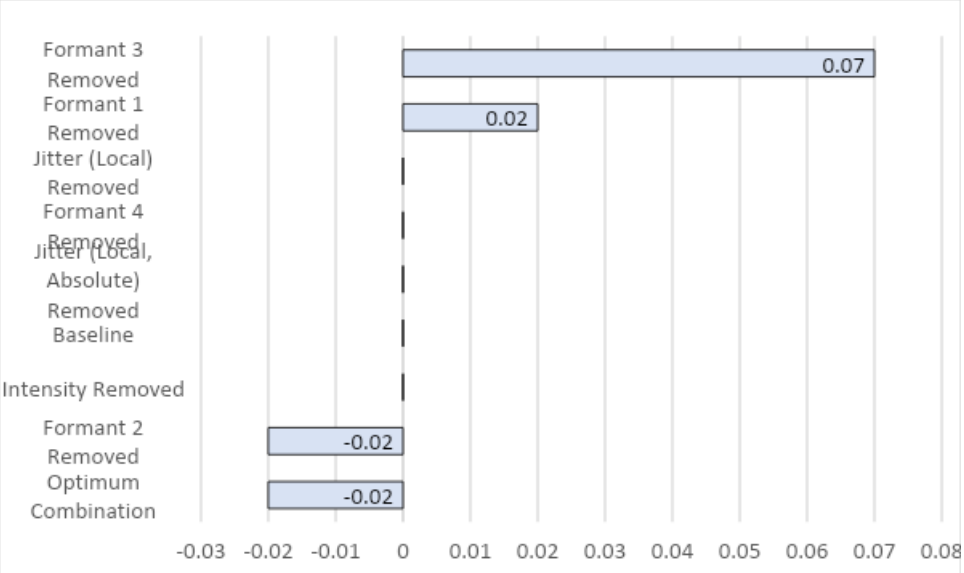

For /i/ (primary stress) below, the all-features $C_{llr}$ value is 0.71 for this phoneme. The phonetic features that need removing are jitter (local) and formant 1 and the phonetic features that need retaining are formants 2-4, intensity, and jitter (local, absolute). This optimised combination generates a $C_{llr}$ value of 0.68 but the best condition for this phoneme, measured on this speaker and data group, is the sole removal of jitter (local), however. This generates a $C_{llr}$ value of 0.68 but does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /i/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**



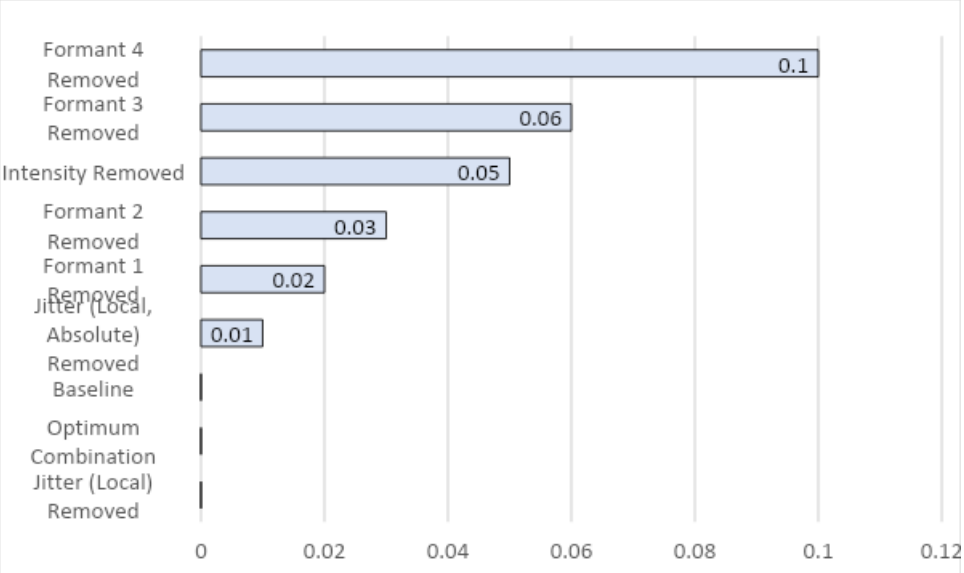

For /ʊ/ (primary stress) below, the all-features $C_{llr}$ value is 0.93 for this phoneme. The phonetic features that need removing are formant 2 and intensity and the phonetic features that need retaining are formant 1, formants 3-4, and jitter (local, absolute). This optimised combination generates a $C_{llr}$ value of 0.91 and the best condition for this phoneme, measured on this speaker and data group, is the optimised combination. This still does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /ʊ/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**




For /u/ (primary stress) below, the all-features $C_{llr}$ value is 0.57 for this phoneme. The phonetic features that need removing are jitter (local) and the phonetic features that need retaining are intensity, formants 1-4, and jitter (local, absolute). This optimised combination still generates a $C_{llr}$ value of 0.57 and the best condition for this phoneme, measured on this speaker and data group, is this optimised combination as only the singular feature detrimental to performance was cut. This still does not beat the baseline $C_{llr}$ value in this corpus.

**Change in $C_{llr}$ values in relation to the Baseline Measurement for /u/ (primary stress) in Gold et al.'s (2018) WYRED Text-Independent Sub-Corpus**




There are some trends to spot here too. The baseline $C_{llr}$ value is again the best, jitter measurements are regularly trimmed again, and the optimised combination is, again, regularly used at the segment-specific level. However, this time formant 2 is also regularly trimmed.